Published in final edited form as:
Smart Health (Amst). 2020 November ; 18: . doi:10.1016/j.smhl.2020.100137.

Automatic Depression Prediction Using Internet Traffic Characteristics on Smartphones

Chaoqun Yue[a], **Shweta Ware**[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]Department of Computer Science & Engineering, University of Connecticut, 371 Fairfield Way, Unit 4155, Storrs, 06269, CT, USA

[b]University of Connecticut Health Center, 263 Farmington Avenue, Farmington, CT 06030, USA

[c]Seldera LLC, USA

Abstract

Depression is a serious mental health problem. Recently, researchers have proposed novel approaches that use sensing data collected passively on smartphones for automatic depression screening. While these studies have explored several types of sensing data (e.g., location, activity, conversation), none of them has leveraged Internet traffic of smartphones, which can be collected with little energy consumption and the data is insensitive to phone hardware. In this paper, we explore using coarse-grained meta-data of Internet traffic on smartphones for depression screening. We develop techniques to identify Internet usage sessions (i.e., time periods when a user is online) and extract a novel set of features based on usage sessions from the Internet traffic meta-data. Our results demonstrate that Internet usage features can reflect the different behavioral characteristics between depressed and non-depressed participants, confirming findings in psychological sciences, which have relied on surveys or questionnaires instead of real Internet traffic as in our study. Furthermore, we develop machine learning based prediction models that use these features to predict depression. Our evaluation shows that Internet usage features can be used for effective depression prediction, leading to F_1 score as high as 0.80.

Keywords

Depression Prediction; Smartphone Sensing; Machine Learning; Data Analytics; Internet Traffic Characteristics

CRediT author statement:
Chaoqun Yue: Methodology, Investigation, Formal analysis, Validation, Writing - Original Draft. **Shweta Ware:** Software, Methodology, Data Curation. **Reynaldo Morillo:** Software, Data Curation. **Jin Lu:** Software, Methodology. **Chao Shang:** Software, Methodology. **Jinbo Bi:** Conceptualization, Methodology, Funding acquisition. **Jayesh Kamath:** Conceptualization, Funding acquisition. **Alexander Russell:** Conceptualization, Methodology, Funding acquisition. **Athanasios Bamis:** Conceptualization. **Bing Wang:** Conceptualization, Methodology, Funding acquisition, Resources, Project administration, Writing - Review & Editing.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Declaration of interests
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.

1. Introduction

Depression is a serious and widespread mental illness that affects 350 million people worldwide [46]. Current diagnosis has been based on clinical interviews or patient self-reports [42]. Both are limited by recall bias. In addition, clinical interviews require direct attention of a skilled clinician, which is problematic due to the lack of trained professionals [5]. Furthermore, the interviews typically take place in clinics or treatment centers (instead of natural environment), leading to limited ecological validity [45]. Patient self-reports require that a patient fills in the reports consistently over time to monitor depression conditions, which is burdensome and hence difficult to execute on a continuous basis.

Recent studies have proposed novel approaches that use data passively collected on smartphones for automatic depression screening (e.g., [17, 36, 11, 41], see Section 2). These studies have found that sensing data collected on smartphones can be used to infer a user's behavioral characteristics, which can then be fed to machine learning algorithms (with pre-trained models) to detect depression. Specifically, they have explored using several types of sensing data, related to location, activity, phone usage, conversation, and sleep. In this paper, we leverage a new type of data, Internet traffic that destines to or originates from smartphones, for depression screening. Such data can be easily collected, incurring significantly lower energy consumption than collecting location and activity data [14, 56], and is not sensitive to phone hardware. Our approach is motivated by the studies in psychological sciences (e.g., [47, 10, 34, 29, 55]), which have demonstrated the relationships between a user's Internet usage and mental health. These studies, however, rely on self-report surveys or questionnaires to characterize a user's Internet usage, which are subjective, and may suffer from recall and desirability biases. Our study, in contrast, uses real Internet traffic that is passively collected, which is objective, and the data can be mined to extract a wide range of user behavioral characteristics. To preserve user privacy, we only collect coarse-grained meta-data (the timing, source and destination IP addresses, and size) of the Internet packets; the payloads of the packets are discarded. In addition, such meta-data is accessible under most encryption techniques used in the Internet, and hence our approach is applicable to most Internet traffic.

Our goal is to explore the feasibility of using coarse-grained Internet usage meta-data for depression screening. Similar as other types of sensing data collected from phones, Internet usage data is subject to missing data and measurement noises (e.g., automatic traffic not generated by human activities). In addition, the Internet traffic meta-data in itself does not directly provide insights into human behaviors—meaningful features need to be extracted from the data that can differentiate the behavioral characteristics of depressed and non-depressed populations. In this paper, we first develop data preprocessing techniques to identify Internet usage sessions and then explore three types of features: two based on usage sessions and the third based on traffic volume. After that, we develop machine learning based prediction models that use different types of features to predict Patient Health Questionnaire (PHQ-9) [28] scores and depression status (i.e., whether one is depressed or not). Our study makes the following three main contributions.

- We develop techniques to identify Internet *usage sessions* (i.e., time periods when a user is online) from the coarse-grained Internet traffic meta-data. Based on the identified usage sessions, we extract a novel set of features, including (i) *aggregate usage features* that represent the overall Internet usage characteristics such as the amount of time a user is online and the number of usage sessions in a day, and over different periods of the day, and (ii) *category-based usage features* that represent the usage characteristics for a set of application categories (e.g., related to mail, social activities, watching video, playing game, shopping) such as the amount of time spent on each application and the number of sessions of each application. These features are insensitive to the volume of network traffic and robust to measurement noises.

- We investigate using three types of features extracted from Internet traffic meta-data for predicting depression. The first two are usage based features, i.e., aggregate usage features and category-based usage features, as described above, and the third type is volume-based features as used in an existing study [27]. We find that (i) usage based features are more correlated with PHQ-9 scores than volume-based features, (ii) using usage based features generally leads to better prediction than using volume-based features, (iii) combining aggregate usage and category-based usage features leads to better prediction than using these two types of features in isolation, and leads to better prediction than combining aggregate usage and volume-based features, and (iv) adding volume-based features to aggregate usage and category-based usage features does not lead to further improvement in prediction accuracy. Overall, the usage based features that we propose are more effective in predicting depression than volume-based features. When combining the two types of usage based features, the resulting F_1 score can be as high as 0.80, comparable to that when using other types of sensing data (e.g., when using location data [17, 57]). Our results demonstrate that Internet usage data is a valuable source of data for depression screening.

- We explore the impact of the amount of historical Internet usage data that is being used on the accuracy of the prediction. Our results indicate that the prediction accuracy generally improves as more data is incorporated, consistent with the fact that depression is a chronic disease, and longer-term data collection provides more insights into a user’s behavior, and hence more accurate prediction.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 outlines our high-level approach. Section 4 describes data collection. Section 5 presents our methodology for identifying usage sessions. Section 6 describes feature extraction. Section 7 describes the prediction results and the impact of historical data. Last, Section 8 concludes the paper and presents future work.

2. Related Work

Existing studies that are related to our work are broadly in three categories: (i) studies that use subjectively collected Internet usage characteristics, (ii) studies that use objectively collected Internet data from a campus network, and (ii) studies that use other types of data (location, phone usage, conversation, social interactions) on smartphones for depression screening. Our work differs from them in that we use coarse-grained Internet usage characteristics collected on smartphones for depression screening. We next describe the related work in more details.

Studies using subjective Internet usage characteristics.

In psychological sciences, the relationship between Internet usage and depression has been explored extensively. For instance, the studies in [47, 10, 34, 29, 55] found that people with depressive symptoms used the Internet much more than those without symptoms. The studies in [52] showed that excessive online video viewing was associated with symptoms of depression. Existing studies [37, 40] also showed that activities on Facebook can reveal the depressive states of users. In all the above studies, the Internet usage characteristics were gathered through self-report surveys or questionnaires, which are subjective, and have recall and desirability biases. Our study, in contrast, collects Internet meta-data continuously and passively on smartphones that requires no interaction from the users, and uses the objective data for depression prediction. While our approach differs from the above studies in psychological sciences, our study is inspired by the findings in those studies. Indeed, the features that we derive, including aggregate usage features (in terms of how often a user is online, the timing when they are online) and category-based usage features (that are related to the types of applications they prefer to run on their smartphones), are motivated from the findings in those studies.

Studies using objective data from a campus network.

To the best of our knowledge, the only study that uses objective Internet traffic data for mental health application is [27], which collected Internet traffic from a campus network, with the goal of associating Internet usage with depressive behaviors among college students. The data was collected using Cisco NetFlow, which provides flow-level information. The authors then extracted aggregate traffic features (e.g., total number of flows, bytes, packets, and durations of the flows), application-level features based on port numbers (e.g., peer-to-peer, HTTP, video/audio streaming, chat, email), and entropy based features (e.g., entropy of source or destination IP addresses or port numbers). Our study differs from [27] in several important aspects. Firstly, we develop a novel set of features that are based primarily on usage sessions, which, unlike the features used in [27], do not rely on destination port numbers and are not sensitive to the traffic volume. Our features are advantageous over the features in [27] for two reasons: (i) destination port numbers used in [27] do not provide reliable identification of application categories (e.g., due to limited set of well-known reserved port numbers, mis-use or abuse of such port numbers) [26], which is particularly true for smartphone traffic since it is predominantly HTTP-based (i.e., using the same HTTP or HTTPS port number), and hence not differentiable based on port numbers [13, 19], and (ii) features based on traffic volume as used in [27] are more vulnerable to measurement noises (e.g., automatically generated traffic not due to human activities). Our evaluation results in Section 7 confirm that our proposed features are indeed more effective in predicting depression than volume-based features. Secondly, we explore using different types of features, including both the usage based features that we propose as well as volume-based features used in [27]; our results in Section 7 demonstrate that our proposed features are more effective than volume-based features in predicting depression. Thirdly, our study develops machine learning based models (both regression and classification models) for predicting depression, while the study in [27] only investigated the features that were correlated with depressive symptoms and did not consider depression prediction.

Studies using other types of sensing data from smartphones.

Other studies related to our work include the studies that used data collected from smartphones or wearable devices for detecting depression or depressive mood (e.g., [18, 20, 21, 8, 32, 58, 48, 36, 43, 15, 53]). These studies differ from ours in that they extracted various features from location, activity, phone usage, conversation, physiological information, or sleep data, instead of Internet usage characteristics on smartphone. As an example, Wang et al. [49] studied the impact of workload on stress and day-to-day activities of students. They found significant correlation between a number of behavioral traits (in terms of conversation duration, number of locations visited, sleep) and depressive mood. Saeb et al. [41] found significant correlation between the phone usage and mobility patterns with respect to the self-report PHQ-9 scores. Canzian and Musolesi [11] studied the relationship between the mobility patterns and depression, and developed individualized machine learning models for predicting depression. The study in [25] collected physiological signals, location, smartphone logs, and survey response, and applied machine learning models to predict happiness, with the ultimate goal of understanding the factors that contribute to resistance to depression. Farhan et al. [17] found that the features extracted from the location data are complementary to PHQ-9 scores and can predict depression with good accuracy. Yue et al. [57] investigated fusing two types of location data (GPS and WiFi association data) collected from smartphones for depression detection. Lu et al. [31] developed a heterogeneous multi-task learning approach for analyzing location data collected over multiple smartphone platforms. Ware et al. [50] explored predicting eight major categories of depressive symptoms (including both behavioral symptoms such as appetite, energy level, sleep and cognitive symptoms such as interests, mood, concentration) using smartphone data and found that a wide range of depressive symptoms can be predicted accurately. Xu et al. [53] proposed a new approach to capturing the co-occurrence relationships across multiple sensor channels by generating contextually filtered features. Another recent study [51] developed a novel approach that uses meta-data collected from an institution’s WiFi infrastructure for large-scale depression screening. These studies have not explored using Internet usage data to predict depression, which is the focus of this study. To the best of our knowledge, our study is the first that leverages Internet traffic on smartphones for depression screening.

3. High-level Approach

Our study is motivated by the findings in psychological sciences that Internet usage is correlated with mental health and status (see Section 2). Indeed, people spend a significant amount of time online, and the Internet traffic on smartphones represents an important aspect of their daily life, particularly because smartphones are personal devices and are being conveniently used by their owners anytime anywhere. As a result, intuitively, Internet traffic on smartphones can be analyzed to understand the behavioral features of their owners. Our high-level approach is to collect Internet traffic data on a smartphone (i.e., the data packets going to or coming out of the smartphone), extract features that represent human behaviors, and then use the features to train machine learning models to predict depression.

Methodology.

We develop the following methodology and the associated techniques in this study.

- **Preserving user privacy.** The Internet traffic (e.g., which websites were accessed and the content of the websites) contains sensitive user information. To protect user privacy, we only collect coarse-grained meta-data, including source and destination IP addresses of the packets, the size of the payload, and the timing information (when a packet arrives or departs from a phone). Application-level information (e.g., the URL of a website, the content) is never captured. The above coarse-grained meta-data is in plaintext for most encryption techniques used in the Internet, including application-level encryption, transport-layer encryption such as TLS and HTTPS, and IP-layer encryption such as IPsec. Therefore, our approach is applicable to most Internet traffic. One of the goals of our study is to investigate whether such coarse-grained information can still provide sufficient insights into one's behavior to help detect depression.

- **Extracting features that represent human behaviors.** The Internet traffic collected on smartphones does not directly provide insights into human behavior. We design techniques to extract features from the Internet traffic. Considering that some traffic on smartphones may not be associated with human activities, e.g., those due to background services/apps [14, 24, 33] or advertisement [12], we focus particularly on extracting timing-based features that are less sensitive to such non-human activity related traffic. Specifically, we develop techniques to identify *usage sessions*, i.e., when a user is online, and extract two types of features based on the identified usage sessions. One type of features is based on the *aggregate* Internet usage, related to the overall timing and extent of the Internet usage, and the other is related to *applications*, including seven commonly used applications (mail, social, video, audio, game, shopping and study), representing the timing and extent of usage of each of these applications.

- **Developing machine learning techniques.** Using the extracted features, we develop regression based techniques to predict PHQ-9 scores and Support Vector Machine (SVM) based techniques to predict the depression status (i.e., whether one is depressed or not). For both types of prediction, we explore feature selection, and quantify the prediction accuracy when using the selected features. We further compare the effectiveness of the two types of usage based features that we propose with volume-based features as used in [27], and explore whether different types of features are complementary to each other in improving prediction accuracy.

Missing data, measurement noises and inference errors.

Similar as other types of sensing data (e.g., location, activity), Internet usage data collected on smartphones is subject to missing data and measurement noises such as automatically generated Internet traffic that are not due to human activities. It is difficult to completely identify and eliminate such measurement noises solely based on the coarse-grained meta-data that is used in this paper. Furthermore, we use heuristics to infer application categories (Section 6.1.2), which may lead to inference errors (in fact, identifying application categories based on coarse-grained meta-data is a challenging problem in itself [26]). The focus of this paper is to investigate the feasibility of depression prediction using coarse-grained Internet usage data, despite missing data, measurement noises and inference errors, instead of completely eliminating measurement noises and inference errors. To handle missing data, we use data filtering to select time periods with sufficient amount of data (see Section 7.1); to reduce the impact of measurement noises and inference errors, we propose a novel set of features that are based on usage sessions (see Section 6.1), and hence are not sensitive to non-user generated traffic. Since depression is a chronic disease, providing us the opportunity of using data collected over a long time (several days or longer) for depression detection, we envision that occasional noises and errors will not affect the overall effectiveness of our approach.

Internet usage versus other sensing modalities.

Internet usage is one of the many types of sensing data that can be captured on smartphones. It is complementary to other sensing modalities such as GPS location and movement in that it represents an individual’s activities in the cyber world. Collecting Internet traffic on phones incurs significantly lower energy consumption [14, 56] than collecting location and movement data, and is not sensitive to phone hardware. In addition, as we shall show in Section 6, a rich set of features can be extracted from Internet usage data that reflect human behavior. As a result, we expect that Internet usage data can be used to complement other types of sensing data (e.g., it can be used when GPS sensors are turned off when the battery level is low), or be used together with other types of sensing data for more effective depression screening. Another advantage of using Internet usage data is that such data can be captured in an institution’s campus network, e.g., as that in [27]. Specifically, an institution may capture the meta-data of Internet traffic at its gateway router and then analyze the meta-data to detect depression. While many user privacy issues need to be carefully considered in the design and deployment of such a service, e.g., only coarse-grained packet headers should be captured and only the data from the users who select to use the service should be analyzed, such a service has the potential of achieving large-scale depression screening using passively collected data with little cost. We leave the investigation of such scenarios as future work. Last, our approach is complementary the approaches that use screen time for mental health applications [30, 39]. Specifically, these approaches keep track of the amount of time spent on individual apps to provide insights into a user’s behavior. However, a user may use a generic app such as a web browser for a wide range of applications, e.g., games, videos, music, shopping, making it difficult to account for the amount of time spent on individual application categories. In contrast, our approach is based on destination IP addresses and can categorize such applications accordingly. On the other hand, our approach only considers applications that incur network traffic.

4. Data Collection

The data was collected from 79 participants during the months of October 2015 to May 2016. All participants were full-time students of the University of Connecticut, aged 18-25. Of them, 73.9% were female and 26.1% were male; 62.3% were white, 24.6% were Asian, 5.8% were African American, 5.8% had more than one race, and 1.5% with unknown ethnicity information. Each participant met with our study clinician for informed consent and initial screening before being enrolled in the study. The clinician also assessed the depression status of the participants, and classified 19 as depressed and 60 as non-depressed. Each participant used a smartphone (either Android or iOS phone) to participate in the study. To ensure the privacy of the participants, we assigned a random ID to each participant, which was used to identify the participants.

Four types of data were collected: network traffic data, PHQ-9 questionnaire responses, clinician assessment, and screen on-off events (only for Android phones). We next describe the methodologies used to collect these four types of data in more detail.

4.1. Network Traffic

Network traffic contains the packets coming to and going out of a user’s phone, which can be directly captured at the phone, e.g., using a local man-in-the-middle proxy, or a virtual private network (VPN) service [6]. We chose to use a third-party application, OpenVPN [3], to capture network traffic from all the participants at a central OpenVPN server that we set up. This data collection approach is flexible and platform independent (it can be easily used on both Android and iOS phones). Specifically, OpenVPN is an open-source software that implements VPN techniques for creating secure point-to-point or site-to-site connections between an OpenVPN client and a server. It uses a custom security protocol based on SSL/TLS for key exchange, and is able to traverse network address translators (NATs) and firewalls. Each participant was asked to run a OpenVPN client on her smartphone in the background, which accessed the Internet via the OpenVPN server that we set up. As a result, all Internet traffic (WiFi or cellular traffic) from/to a user passed through our OpenVPN server. At the OpenVPN server, we used a data capture tool, tcpdump [4], to capture the incoming and outgoing packets from individual phones. For user privacy, only packet headers were captured; see details below. To differentiate the network traffic data from different participants, the traffic flows pertaining to a participant were identified based on the source IP address. Specifically, we created a mapping file that associated each participant with a unique IP address. After that, we created an authentication certificate for each user, which contained a specific signature based on the mapping file. When the OpenVPN client on a user’s phone connected to the server, it sent the certificate to the server, and obtained the pre-assigned unique IP address.

To preserve user privacy, we only collected up to 40 bytes of the packet header (including the IP-layer and transport-layer headers) for each packet, not including any content of the packet. From the packet header, we obtained the source and destination IP addresses, and the payload size (in bytes). In addition, we recorded the timestamp when the packet was being captured. The above four types of information were stored temporarily on the data collection server and the rest of the data was discarded. After that, we further aggregated the retained information to a coarser granularity (see Section 5) and only stored the coarser granularity data for data analysis, i.e., the packet-level meta-data were removed from the server, to further protect user privacy.

4.2. PHQ-9 Questionnaire Responses

PHQ-9 is a 9-item self-reported questionnaire that assists clinicians in diagnosing and monitoring depression. Each of the nine questions evaluates a person's mental health on one aspect of depressive disorder. Participants were asked to fill in a PHQ-9 questionnaire during the initial assessment. After that, they filled in the questionnaire on their phones every 14 days using an app that we developed. The app generated a notification when a PHQ-9 questionnaire was due. When missing a report, we sent a reminder to a participant three days after her/his PHQ-9 filling due date.

4.3. Clinician Assessment

Using a Diagnostic Statistical Manual (DSM-V) based interview and PHQ-9 evaluation, a clinician associated with our study classified a participant as either depressed or not during the initial screening. A participant with depression must be in treatment to remain in the study. In addition, depressed participants had follow-up meetings with the clinician periodically (once or twice a month determined by the clinician). Each meeting lasted for 10-20 minutes and only involved interviews to assess psychiatric symptoms. The purpose of the interviews was to correlate and confirm their self-reported PHQ-9 scores with their verbal report.

4.4. Screen On-off Events

On Android phones, it has been reported that a wide variety of applications/services can run in the background [14, 24, 33], when the screens are off and users are not actively interacting with their phones. Therefore, on Android phones, we log screen on-off events, and only consider the traffic when the screen is on. Specifically, let a *screen-on interval* be an interval that starts with a screen-on event and ends with a screen-off event. We observe that 80% of the screen-on intervals are less than 60 minutes. However, some screen-on intervals are as long as several hours, which are likely caused by missing events, e.g., a screen-off event is not captured between two screen-on events, causing an inflated screen-on interval. In the rest of the paper, we assume that the screen-on durations are no more than 60 minutes. That is, if a screen-on interval is longer than 60 minutes, we assume that the screen is on only for the first 60 minutes, and the screen is off for the remaining time in the interval.

Unlike Android, iOS prevents applications from generating background traffic at will. Specifically, an application needs to use a mechanism called *push notification* [2], in which the iOS (not the application) decides when to send or receive data, and the iOS only sends/receives data when the screen is on. As a result, we believe the amount of traffic on iOS phones when the scree is off is very low (if any). Therefore, we did not log screen on-off events on iOS phones and consider all the collected traffic in data analysis.

Last, some traffic generated during screen-on intervals may not be due to user activities (e.g., they can be generated by some apps that automatically fetch updates, sync with the cloud server, or advertisements to the phone). We expect that the amount of such traffic is lower than the amount of traffic due to user activities. In addition, since such traffic is concurrent with user traffic, we propose features that are based on usage sessions (i.e., the period during which a user is online); see Section 6.1. Such features are not sensitive to the amount of traffic, and are less impacted by automatically generated traffic.

5. Identifying Usage Sessions

On the high level, a user's Internet usage behavior can be described by an on-off process: the user accesses the Internet during the *on* period, and is disconnected from the Internet during the *off* period. In the rest of the paper, we also refer to an on-period as a *usage session*. Since we do not know the exact time periods when a user is using the Internet, we infer the on-off process using the information of the packets. In the following, we first describe a methodology for identifying keep-alive packets in long-lived TCP connections, which can lead to overestimation of on-periods. We then describe packet-level prepressing and inference of on-off periods.

Keep-alive packets.

Keep-alive packets are packets sent during inactivity period of a TCP connection (i.e., there is no data or acknowledgement packet) to keep the connection alive so that it is not closed by the corresponding server or an intermediate middlebox device (e.g., firewall). While the original recommendation is to send a keep-alive packet after a long period of inactivity (every two hours [35]), we observed in our dataset that the period can be much shorter (in minutes). Fig. 1 shows an example, where each circle represents one TCP packet from a source to a destination. The burst of the packets at the beginning corresponds to data transfer; the periodic packets afterwards are the keep-alive packets. We identify keep-alive packets as periodic packets from a source to a destination with payload of 0 or 1 byte. Since keep-alive packets are not related to data transmission or user behaviors, and including them can lead to overestimation of an on-period (or usage session), we removed them before data analysis.

Packet-level preprocessing.

After removing keep-alive packets, we performed packet-level preprocessing that aggregate packets to further protect user privacy. Recall that we recorded a tuple, (timestamp, source IP address, destination IP address, payload size), for each packet going to or coming from a smartphone at our OpenVPN server. The packets going to an external IP address a and coming from the external IP address a in a short time interval, from t to $t + \delta$, are likely to be in the same application session (they may represent the data packets and the corresponding acknowledgement packets). We therefore do not differentiate these packets and regard them as in the same *application session*. Specifically, we aggregate these packets together, and represent the resultant application session as (t, a, s), where t is the start time, a is the external IP address, and s is the sum of the payload sizes of all the packets (going to and coming from a) from t to $t + \delta$. We choose δ to be 1 minute empirically, since it is unlikely that the packets corresponding to one activity are separated by more than one minute. In other words, for each phone, we store per-minute traffic between the phone and an external IP address, which is of coarser granularity than per-packet information, and leads to better user privacy.

Inferring on-off periods.

After packet-level preprocessing, we infer on-off periods, i.e., the time periods when a user is online and offline. Specifically, we aggregate the traffic collected on the phone and identify the time periods with traffic as on-periods, and the interval between two on-periods as an off-period. The aggregation is first over each external IP address, and then across all the external IP addresses. Fig. 2 illustrates the process. It shows the traffic for three external IP addresses, a_1, a_2 and a_3, over a time interval of 15 minutes. Fig. 2(a) shows per-minute traffic for each of the three external IP addresses, where an oval represents that there is traffic in that minute. We see that for IP address a_1, there is traffic in the 3rd, 4th, and 6th minute; for IP address a_2, there is traffic in the 8th and 9th minute; and for IP address a_3, there is traffic in the 15th minute. Fig. 2(b) illustrates the aggregation for an individual IP address. Specifically, consider an IP address a. Suppose there is traffic for a in the i_1, i_2, … , i_n th minute. An interval with no traffic may be because the user had finished the current session and then started a new activity session at a later time, or may be because the user was “thinking” (e.g., the user was reading a webpage after it was downloaded). We differentiate the above two scenarios based on the duration of the intervals. Specifically, if $i_{k+1} - i_k \leq 1$ minute, then we assume the users was in the same session; otherwise, we assume the user started a new session in the i_{k+1} th minute. The threshold of 1 minute is chosen empirically, assuming that the “thinking” time tends to be less than 1 minute. In this way, the traffic for each IP address is aggregated into sessions. Following the above aggregation process, we identify one session associated with each of the three IP addresses, marked by the three rectangular bars in Fig. 2(b). Fig. 2(c) illustrates the aggregation across IP addresses. It considers the sessions of all the external IP addresses, and aggregate two sessions into one if the gap between the ending time of one session and the beginning time of another session is less than 1 minute apart. Applying the above process leads to two sessions in Fig. 2(c), one from the 3rd to the 9th minutes (aggregated over the sessions from IP addresses a_1 and a_2) and the other for the 15th minute (from IP address a_3), which are referred to as two *on-periods*; the rest of the intervals are *off-periods*. In the rest of the paper, for convenience, we also refer to an on-period as a *usage session* since it marks a time period when a user uses the phone.

Applying the above aggregation procedures (within an IP address and across multiple IP addresses), we identify a set of on-periods and off-periods for each user during a day. Fig. 3(a) shows an example for one Android user over 15 days, where the black bars represent the on-periods. We see that the on-periods are spread out during a day, with most of them between 7am to 10pm; there are also a small amount of on-periods during other time (e.g., early morning from 1 to 4am). For some days, we do not see any on-periods, which is likely due to issues in data collection (see more details in Section 7.1). Fig. 3(b) plots the distributions of the on-periods for the data collected on Android and iOS phones, respectively. For each of these two platforms, we plot the distributions of four time periods, morning, afternoon, night and midnight, separately. We observe that the distributions are close to each other, and most of on-periods are within 10 minutes. The similar distributions for the same time periods on the two platforms are expected since the general user behaviors should not be much affected by the platform they use.

6. Feature Extraction

For each participant, we extracted features from the data collected in a *PHQ-9 interval*, which is a two-week time period, defined as the time when a PHQ-9 questionnaire was filled in and the two weeks before that (since PHQ-9 asks about the previous two weeks). To reduce the impact of non-user generated traffic, we propose a set of features that are based on usage sessions. For comparison, we further explore volume-based features that were proposed in [27], and use the prediction results based on such features as baseline in Section 7.

6.1. Features based on Usage Sessions

We propose two types of features based on usage sessions, one being high-level aggregate usage features and the other being more detailed category based features that are related to applications, as detailed below. These features, being based on usage sessions (i.e., when a user is online), are less sensitive to non-user generated traffic (such as background apps or advertisements) since they are concurrent with user traffic.

6.1.1. Aggregate Usage Features—The aggregate usage features quantify the duration and the number of the usage sessions (i.e., on-periods, see Section 5) from a user during a PHQ-9 interval. Specifically, for a user, let s_i and e_i represent the start and end time of the ith usage session in a PHQ-9 interval, respectively. Then $\{(s_i, e_i)\}$ represents the set of usage sessions. To differentiate user behaviors during different times of the day, we further divided a day into four time periods: morning (6am-12pm), afternoon (12pm-6pm), night (6pm-12am), and midnight (12am-6am). For time period i, $i = 1, \dots, 4$, representing morning, afternoon, night and midnight, respectively, let $s_{i,j}$ and $e_{i,j}$ represent the start and end time of session j, respectively. Then $\{(s_{i,j}, e_{i,j})\}$ represent the set of usage sessions for time period i. We define the following aggregate usage features.

Total duration.: This feature represents the total duration of the usage sessions, i.e., $\sum_i(e_i - s_i)$. To take account of missing data, let D denote the number of days with data during a PHQ-9 interval. We then normalize the total duration by D, and the resultant feature is $\sum_i(e_i - s_i)/D$.

Total number of sessions.: Let N be the total number of usage sessions during a PHQ-9 interval. This feature is defined as N/D, where D is the number of days with data. Again the normalization by D is to take account of missing data.

Total off-duration.: This feature represents the total duration of off-periods. It is defined is $\sum_i(s_i - e_{i-1})/D$, where $s_i - e_{i-1}$ represent the duration of the off-period between the $(i-1)$th and ith usage sessions, and D is the number of days with data.

Internet usage in each time period.: We quantify the Internet usage of a time period (morning, afternoon, night or midnight) in terms of the duration and the number of usage sessions of a user during that time period. Specifically, we define $\sum_j (e_{i,j} - s_{i,j}) / \sum_i \sum_j (e_{i,j} - s_{i,j})$, which is the total duration of the usage sessions in time period i normalized by the total duration of the usage sessions over all the four time periods, leading to four features on duration, one for each time period. The normalization allows the quantities obtained from different PHQ-9 intervals to be comparable. Let N_i be the number of usage sessions in time period i. Then $\sum_i N_i = N$, where N is the total number of usage sessions across all individual time periods during a PHQ-9 interval. Similarly, we normalize N_i by N, i.e., define N_i/N, which lead to four features on the number of usage sessions, one for each time period.

6.1.2. Category-based Usage Features—The category-based features represent a user’s activities in multiple application categories and provide insights into a user’s interests. The categorization is based on the destination IP addresses. Specifically, we used a public online database, DBIP [1], to look up information about each destination IP address. We then used the information to determine the application category by matching the information with a set of keywords (see details below). Table 1 shows an example, where the destination IP address is 31.13.69.197. The response of the database query includes hostname, ASN (autonomous system number), ISP, organization and description. We found keyword “Facebook” in the response, and hence set the category as social. Table 2 lists the keywords for each application category. Specifically, we considered 7 common application categories, including mail, social, video, audio, game, shopping, and study, based on common Internet browsing activities for our participant population. The categories of mail and social correspond to online communication that a user has with others, motivated by the studies that show relationship between depression and social applications [37, 40]. The categories of video, audio and game correspond primarily to entertainment activities (while they may be used for other purposes, e.g., for learning, we speculated that the majority of the time they were used for entertainment), motivated by studies that relate excessive video-gaming applications and mental health [52]. The category of shopping is motivated by existing studies [23, 7] that relate shopping behavior with stress and anxiety. The last category (i.e., study related applications) is of interests since all of our participants are university students. The keywords in Table 2 were built on the keywords used in [44] by adding new keywords based on manual inspection of our dataset and avoiding keywords that may correspond to a large number of services (e.g., google, yahoo). When matching keywords, we followed the order of the keywords in Table 2; a destination IP address was associated with one category, i.e., the first match in the table.

We now describe the category-based features, which were obtained from the usage sessions in a PHQ-9 interval. Specifically, for category i, we considered all the usage sessions with the destination IP addresses belonging to that category, and merged these sessions together, i.e., if two sessions overlaped in time, then they were merged into a longer session. For category i, let N_i denote the number of sessions belonging to category i in a PHQ-9 interval, and $s_{i,j}$ and $e_{i,j}$ denote the start and end time of session j in a PHQ-9 interval. Then the set of usage sessions associated with category i is $\{(s_{i,j}, e_{i,j})\}$. For each category, we extract two types of features, one on duration and the other on the number of sessions in a PHQ-9 interval.

Duration of each application category.: This set of features represents the total duration of the usage sessions for each application category. In total, we have 7 such features, one for each category. The feature for category i is defined as $\sum_j (e_{i,j} - s_{i,j}) / \sum_i \sum_j (e_{i,j} - s_{i,j})$, i.e., the total duration of the usage sessions in this category normalized by the total duration of the usage sessions over all categories.

Number of sessions of each application category.: This set of features represents the number of sessions for each application category. Again, we have 7 such features, one for each category. The feature for category i is $N_i / \sum_i N_i$, i.e., the number of usage sessions in this category normalized by the total number of sessions over all categories.

6.2. Volume-based Features

We further explore volume-based features that were used in an existing study [27]. As mentioned earlier, such features may be biased by traffic that are not generated by user activities. On the other hand, since applications differ in the amount of traffic that they generate (e.g., video applications tend to lead to significantly more traffic than other applications), traffic volume can shed light on the applications used by a user, and hence provide insights into user activities. Specifically, we consider the following five volume-based features.

Total volume.—Let B be the total amount of Internet traffic in bytes during a PHQ-9 interval. This feature is defined as B/D, where D is the number of days with data. Again the normalization by D is to take account of missing data.

Volume in each time period.—Let B_i be the total amount of Internet traffic in bytes in time period i, where $i = 1, \ldots, 4$, corresponding to morning, afternoon, night, and midnight, respectively. Then the volume feature for period i is $B_i / \sum_i B_i$, where the normalization allows the quantities obtained from different PHQ-9 intervals to be comparable.

7. Depression Prediction

In this section, we use the features described in Section 6 to predict PHQ-9 scores and depression status. In the following, we first describe data filtering, and then describe the prediction results. We present the prediction results for the iOS and Android datasets separately because these two platforms differ in handling background traffic (see Section 4.4).

7.1. Data Filtering

Our prediction below uses the Internet traffic meta-data collected over individual PHQ-9 intervals (i.e., a two-week time period, including the day when a PHQ-9 questionnaire was filled in and the previous two weeks, see Section 6). We observed missing data in data collection, which may be due to multiple reasons, e.g., failed data capture at our OpenVPN server, mis-configuration of a phone, or lack of Internet activity on the phone. To deal with such missing data, we removed the PHQ-9 intervals that do not have sufficient amount of data. Specifically, we removed a PHQ-9 interval if less than half of the days in the interval had data, or the average duration of Internet usage was less than 20 minutes per day for the days that had data. We further removed a PHQ-9 interval if there was Internet usage only for one time period (e.g., morning) throughout the interval or none of the traffic in the interval can be categorized, since such cases were unlikely under normal circumstances.

After the above data filtering, we identified 150 valid PHQ-9 intervals from 40 iOS users (of them 10 depressed and 30 non-depressed), with 44 PHQ-9 intervals from depressed users and 106 intervals from non-depressed users. For the Android dataset, we identified 38 valid PHQ-9 intervals from 13 Android users (of them 4 depressed and 9 non-depressed), with 14 PHQ-9 intervals from depressed users and 24 intervals from non-depressed users.

7.2. Correlation Analysis

In the following, we first present the correlation between the aggregate usage features and PHQ-9 scores, and then present the correlation between category-based usage features and PHQ-9 scores. At the end, we present the results for volume-based features.

Correlation results for aggregate usage features.—Table 3 presents Pearson's correlation coefficients between various aggregate Internet usage features and PHQ-9 scores along with p-values (obtained using significance level $\alpha = 0.05$) for the Android and iOS datasets. The results are in three cases: the first is for all participants, the second is for depressed participants, and the third is for non-depressed participants. The top half of Table 3 shows the results for the iOS dataset. We see that the total amount of time online is positively correlated with PHQ-9 scores. This is consistent with studies in psychological sciences [47, 10, 34, 29, 55], which show that people with depressive symptoms spend more time online than non-depressed people. We also find that the amount of time online in individual time period (morning, afternoon, night and midnight) is correlated with PHQ-9 scores. The durations online in morning and afternoon both have significant negative correlation with PHQ-9 scores, indicating that the participants with higher PHQ-9 scores tend to spend less time online during the day; instead, they tend to spend more time online during night and midnight, as shown in the significant positive correlation between the corresponding features (i.e., the durations online during night and midnight) and PHQ-9 scores. In addition, the number of sessions online in each time periods (especially during night and midnight) has significant positive correlation with PHQ-9 scores, indicating that the participants with higher PHQ-9 scores tend to use phone more frequently. The lower half of Table 3 shows the results for the Android dataset. We see that the p-values for all the features are high, indicating no significant correlation between the features and the PHQ-9 scores. The significantly different p-values for the iOS and the Android datasets might be because the number of samples in the Android dataset is much lower than that in the iOS dataset (44 versus 150 samples). On the other hand, as we shall see later, even for the Android dataset, when combining multiple features, we can predict PHQ-9 scores and depression status accurately.

To provide further insights, we divided the samples of PHQ-9 intervals (each associated with a PHQ-9 score and the features) into four groups, with PHQ-9 scores in the ranges of [0, 5), [5, 10), [10, 15), and [15, 20), respectively (for the Android dataset, the maximum PHQ-9 score is 15 and hence the PHQ-9 scores are in the first three ranges). For each group, we obtained the duration of the usage sessions in each time period (morning, afternoon, night and midnight). Figures 4(a) and (b) plot the results for iOS and Android users separately (the duration for each time period was normalized so that the sum of the four time periods is 1, see Section 6.1.1). We observe that, for the iOS dataset, compared to the two groups with lower PHQ-9 scores, the two groups with higher PHQ-9 scores are indeed associated with a larger amount of time online during night and midnight, and a lower amount of time online during morning and afternoon, consistent with the correlation results described earlier. For the Android dataset, we also observe that the group with PHQ-9 scores above 10 spent more time online during midnight and less time online during morning than the other two groups with lower PHQ-9 scores.

Correlation results for category-based features.—Table 4 presents Pearson's correlation coefficients between category-based features and PHQ-9 scores. The top half of Table 4 shows the results for the iOS dataset. We see that the amount of time spent on social apps shows significant positive correlation, indicating that the participants with higher PHQ-9 scores tend to spend more time on social apps. The amount of time spent on mail related apps, on the other hand, has significant negative correlation with PHQ-9 scores for depressed participants. We further see that the numbers of sessions that are on social, game and shopping categories have significant positive correlation with PHQ-9 scores, indicating that the participants with higher PHQ-9 scores tend to access those types of contents more frequently, which is consistent with studies in psychological sciences [52, 23, 7]. The lower half of Table 4 shows the correlation results for the Android dataset. We again see that the p-values for most of the features are high, which might be due to limited number of samples in the Android dataset.

To gain more insights, Figures 5(a) and (b) plot the amount of time spent on each of the seven application categories versus PHQ-9 score range for iOS and Android datasets, respectively. Again, we divided the samples of PHQ-9 intervals into four groups, based on their associated PHQ-9 scores, and the amount of time spent on each application category was normalized as described in Section 6.1.2. For the iOS dataset, we see that the users with PHQ-9 scores above 10 spent more time on game and social apps, while spent less time on mail than those with lower PHQ-9 scores, consistent with the correlation results in Table 4. For the Android dataset, we see the users with PHQ-9 scores above 10 spent more time on game and video apps, while spent less time on shopping apps than those with lower PHQ-9 scores.

Correlation results for volume-based features.—Table 5 presents the correlation results for volume-based features. We see that even for the iOS dataset, the p-values for most of the features are high, indicating no significant correlation between the features and PHQ-9 scores; the only exception is the total volume, which shows positive correlation with PHQ-9 scores. The weaker correlation between volume-based features with PHQ-9 scores compared to that achieved by usage based features (i.e., aggregate usage and category-based usage features) is consistent with our intuition that volume-based features may be noisier (they are more affected by non-user generated traffic) and less descriptive than features based on usage sessions.

7.3. Multi-feature Regression

We developed both linear and non-linear multi-feature regression models to predict PHQ-9 scores. Specifically, we applied ℓ_2-regularized ϵ-SV (support vector) multivariate regression [16] and radial basis function (RBF) ϵ-SV multivariate regression. For both models, we used leave-one-user-out (i.e., we used the data from all the users except one user for training, and used the data for the user that is not used in training for testing) cross validation to optimize model parameters. The reason for using leave-one-user-out cross validation is to ensure that the data of one user was either used for training or testing, but never for both, so as to avoid overfitting the models since the data of a user over different PHQ-9 intervals may be correlated. For ℓ_2-regularized ϵ-SV regression, we optimized the cost parameter C and the margin ϵ by choosing $\log(C)$ from −13 to 13 with the step size of 0.1, and selecting ϵ from [0, 1] with the step size of 0.01. For RBF ϵ-SV regression, we optimized the cost parameter C, the margin ϵ, and the parameter γ of the radial basis functions by selecting $\log(C)$ from −13 to 13 with the step size of 0.1, selecting ϵ from [0, 5] with the step size of 0.01, and selecting $\log(\gamma)$ from −15 to 15 with the step size of 1. To assess the performance of a model, we calculated Pearson's correlation of the predicted values from the model with the PHQ-9 scores.

Our results below are for six settings. The first three settings each uses a single type of features, i.e., volume-based features (5 features), aggregate usage features (11 features), and category-based features (14 features), where the results using volume-based features serve as the baseline. The last three settings each uses multiple types of features, including volume-based combined with aggregate usage features, aggregate usage features combined with category-based features, and all three types of features. For each of the six settings, we used Joint Mutual Information (JMI) [54] for feature selection, which has been shown to provide the best tradeoff in terms of accuracy, stability, and flexibility with small data samples among all information theoretic feature selection criteria [9]. Let k be the number of features selected. We varied k from 1 to the number of features in each setting. For a given k, the features were selected using JMI and the parameters of the regression models were chosen as described above. Fig. 6(a) shows an example. It plots the Pearson's correlation of the predicted values with the PHQ-9 scores when increasing the number of features, k, under the linear model for the iOS dataset. The results for three settings, i.e., only using the aggregate usage features, only using the category-based usage features and using both types of features, are plotted in the figure; the results for the other three settings show similar trend and are omitted for clarity. We observe that indeed the number of features being selected affected the correlation results significantly, confirming the importance of feature selection. The same observation holds in Fig. 6(b), which shows the correlation results when increasing the number of features under the non-linear model.

The top and bottom halves of Table 6 list the best correlation results (i.e., by choosing the best set of features) for the iOS and Android datasets, respectively. For both datasets, the results for all the six settings with linear and non-linear models are shown in the table. When using a single type of features, we see that aggregate usage features in general lead to better prediction than the other two types of features. In addition, while category-based usage features and volume-based features both reflect application related patterns, the former in general leads to better prediction than the latter, which is perhaps not surprising since the former is less affected by measurement noises (non-user generated traffic) and provides substantially more detailed information on application usage than the latter. We also see the benefits of combining aggregate usage features and category-based usage features, which in general leads to better prediction than using them in isolation, confirming that these two types of features are complementary to each other. Combining aggregate usage features and volume-based features can also lead to benefits (e.g., for the non-linear model, Android dataset), but the benefits are not as significant as those when combining aggregate usage features and category-based usage features. In addition, adding volume-based features to the combination of aggregate usage features and category-based features does not lead to further improvement in prediction results. The best correlation results are 0.54 and 0.39 for the iOS and Android datasets, respectively, which are comparable to the results when using location data collected on the phones [17, 57].

Table 7 lists the selected features corresponding to the best non-linear model (listed in Table 6) when combining both aggregate and category-based usage features for the iOS and Android datasets; in the interest of space, the selected features for the best models for other settings are not presented. For the iOS dataset, 12 features were selected, including the aggregate usage features over a day (total duration), aggregate usage features for individual time periods (morning, afternoon, night and midnight), and features related to game, video, audio, and social applications. For the Android dataset, 6 features were selected, including the aggregate usage features over a day (total number of sessions), aggregate usage features for individual time periods (night and midnight), and the features related to video and shopping applications.

7.4. Classification Results

We next present the results on predicting clinical depression (i.e., classify whether one is depressed or not). The machine learning techniques we used were Support Vector Machine (SVM) models with an RBF kernel. The assessment from the study clinician was used as the ground truth. The SVM models have two hyper-parameters, the cost parameter C and the parameter γ of the radial basis functions. We again used a leave-one-user-out cross validation procedure to choose the optimal values of the hyper-parameters. Specifically, we varied both C and γ as $2^{-15}, 2^{-14}, \dots, 2^{15}$, and chose the values that gave the best cross validation F_1 score. The F_1 score, defined as 2(precision × recall)/(precision + recall), can be interpreted as a weighted average of the precision and recall, ranges from 0 to 1. The higher the F_1 score, the better the result is.

We repeated the above SVM training and testing procedures in three settings: 1) only using aggregate features, 2) only using category-based features, 3) using both aggregate and category-based features. For each setting, we first used SVM recursive feature elimination (SVM-RFE) [22] to select a subset of k features, where k is varied from 1 to the number of features, and then trained and tested the models. We chose to use SVM-RFE since it is a wrapper-based feature selection algorithm that was designed for SVM, and has been developed for both linear and non-linear kernels [22, 38]. Specifically, we used SVM-RFE to select features as follows. For each C and γ pair, SVM-RFE provided a ranking of the features, from the most important to the least important. We considered all the combinations of C and γ values, and obtained the average ranking for each feature across all the combinations. For n features, let $f^{(1)}, \ldots, f^{(n)}$ represent the complete order of the features in descending order of importance. That is, on average, $f^{(1)}$ is the most important feature, $f^{(2)}$ is the second most important feature, and $f^{(n)}$ is the least important feature. For a given k, the features $f^{(1)}, \ldots, f^{(k)}$ were used to choose the parameters (i.e., C and γ) to maximize F_1 score based on the above leave-one-user-out cross validation procedure.

We again consider six settings: three settings use a single type of features and the other three settings use multiple types of features. For each setting, we observe significantly varying F_1 score when varying the number of features as described above (figures omitted), confirming the importance of feature selection. Table 8 presents the best F_1 scores (i.e., when using the best set of features) for the six settings. For each setting, the optimal values of the parameters are also shown in the table. When using a single type of features, consistent with the regression results in Section 7.3, we see that usage based features (i.e., aggregate usage features or category-based usage features) are more effective in classifying depression than the volume-based features. Furthermore, combining these two types of usage based features leads to better classification results than using each in isolation, and the results are better than those when combining the aggregate usage features and volume-based features. Last, adding volume-based features does not further improve the classification results obtained by using aggregate and category-based usage features, which, as explained earlier, might be because volume-based features provide insights in application usage, which are already covered by category-based usage features. The highest F_1 scores are 0.71 and 0.80 for the iOS and Android datasets, respectively, comparable to the classification results when using location data collected on the phones [17, 57]. The above results need to be further verified using larger datasets (particularly for the Android platform). Nonetheless, they demonstrate that Internet usage data can provide important insights into one's behavior that can be used for effective depression screening.

Table 9 lists the selected features that lead to the best F_1 score for the setting of combining aggregate and category-based usage features; the selected features for the other five settings are omitted in the interest of space. For the iOS dataset, 10 features were selected, including the aggregate usage features over a day (e.g., total off-duration), aggregate usage features for particular time periods, and category-based features (related to mail, shopping, and game). For the Android dataset, 10 features were selected, again including the aggregate usage features over a day (e.g., total duration, total off-duration), aggregate usage features for particular time periods, and category-based usage features (related to social, mail, shopping, and game).

7.5. Impact of Historical Data

So far, we have considered Internet traffic data collected in PHQ-9 intervals of 14 days, including the day when a PHQ-9 questionnaire was filled in and the previous two weeks. Intuitively, even though PHQ-9 asked about the depression symptoms in the past two weeks, the experience of a user during the days right before the PHQ-9 fill-in date may have a higher impact on how they fill in the questionnaire (and hence the corresponding PHQ-9 scores). In the following, we consider the Internet traffic data collected in m days, i.e., including the day when a PHQ-9 questionnaire was filled in and the previous m days, where m is varied from 1 to 14. Our goal is to investigate how the prediction accuracy changes with m so as to select the optimal m for prediction accuracy. For each interval of m days, we used the same two conditions as described in Section 7.1 to determine whether the corresponding Internet traffic data is valid or not (the only exception is when $m = 1$, where we only used the second condition, i.e., it is regarded as invalid if it only has data for one time interval). Since our goal is to evaluate the impact of m, we only consider a PHQ-9 score if it has valid Internet traffic data for each possible value of m (i.e., $m = 1, \dots , 14$). A subset of iOS and Android datasets satisfied the above filtering criteria and were used in the data analysis below. Specifically, for the data analysis, the iOS dataset contained 58 samples (each with a PHQ-9 score and the corresponding Internet traffic data for m days, $m = 1, \dots , 14$) from 23 iOS users (7 depressed and 16 non-depressed users), including 19 samples from depressed users and 39 sample from non-depressed users. For the Android dataset, the intersection contained 21 samples from 11 Android users (3 depressed and 8 non-depressed users), including 6 samples from depressed users and 15 samples from non-depressed users.

We next present the impact of m on both the multi-feature regression results and the classification results. For both tasks, the prediction results were obtained by combining the aggregate and category-based usage features, which, as shown earlier (Section 7.3 and 7.4), led to better prediction than other types of features. Figures 7(a) and (b) show the impact of m on the multi-feature regression results for the iOS and Android datasets, respectively, where m is varied from 1 to 14. For each m, we used the same procedure as described in Section 7.3 to select a subset of features that provided the best prediction result. Both the results from the linear and non-linear regression models are shown in the figures. Figures 8(a) and (b) show the impact of m on the classification results for the iOS and Android datasets, respectively. For each m, we used SVM-RFE to select a subset of features that provided the best cross validation F_1 score. We observe that both the regression and classification results tend to improve with m, indicating that a PHQ-9 score is not only affected by the several days right before the PHQ-9 fill-in date, but also affected by the days that are further away in the past. Using the past 14 days, which corresponds to the time interval that PHQ-9 questionnaire asks about, leads to good regression and classification results.

8. Conclusions, Limitations, and Future Work

In this study, we have investigated using coarse-grained meta-data of Internet traffic on smartphones for depression screening. We have developed techniques to identify usage sessions and defined a novel set of Internet usage features based on usage sessions, including both aggregate and category-based usage features. We further developed machine learning algorithms that used these features for depression screening. Using a dataset of Internet traffic meta-data collected from 79 college students over several months, our results demonstrate that users' Internet usage characteristics are correlated with their PHQ-9 scores, and using these features provides a promising direction in predicting PHQ-9 scores and depression status. Our evaluation also demonstrates that the usage-based features that we proposed are more effective for depression prediction than volume-based features used in [27]. We further explored the impact of the amount of historical data on depression prediction.

Limitations and future work.

Our study has several limitations, which will be addressed in future work. Firstly, the datasets used in this study are relatively small, particularly for the Android platform. The numbers of users for both platforms need to be expanded in future work to further validate the results in this study. In addition, a significantly higher percentage of the participants in our study are female students; recruitment of more balanced female and male participants is left as future work. Secondly, the participants in our study are all college students. While our results show that the usage features that we proposed are effective for this population, different types of features may need to be developed for other populations. For example, for senior citizens, features related to certain applications (e.g., social, gaming, study) may not be effective and features on other types of applications may be needed. Investigating the feasibility of using Internet traffic for depression screening for other populations is left as future work. Another future direction is handling missing data, which significantly reduced our sample size after data preprocessing and filtering. We will develop effective techniques to handle missing human activity data, which is a challenging task [57]. Last, we will investigate other machine learning models for more accurate prediction in future work.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful and constructive comments, and the editor for handling the paper. This work was supported by the National Science Foundation (NSF) grant IIS-1407205 and NIMH grant R01MH119678-01.

References

[1]. DBIP. https://www.npmjs.com/package/dbip.

[2]. iOS Push notification, https://developer.apple.com/notifications/.

[3]. OpenVPN. https://openvpn.net/.

[4]. tcpdump. http://www.tcpdump.org/.

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

[6]. AT&T Application Resource Optimizer (ARO): User Guide. https://developer.att.com/static-assets/documents/aro/release/att-aro-user-guide-5.0.pdf, 2016.

[7]. Atalay AS and Meloy MG. Retail therapy: A strategic effort to improve mood. Psychology & Marketing, 5 2011.

[8]. Ben-Zeev D, Scherer EA, Wang R, Xie H, and Campbell AT. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. Psychiatric Rehabilitation Journal, 38(3):218–226, 2015. [PubMed: 25844912]

[9]. Brown G, Pocock A, Zhao M-J, and Lujdn M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. Journal of machine learning research, 13(Jan):27–66, 2012.

[10]. Campbell AJ, Cumming SR, and Hughes I. Internet use by the socially fearful: Addiction or therapy? Cyber Psychology & Behavior, 9(1):69–81, 2006.

[11]. Canzian L and Musolesi M. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In Proc. of ACM UbiComp, pages 1293–1304, 2015.

[12]. Chen G, Cox JH, Uluagac AS, and Copeland JA. In-depth survey of digital advertising technologies. IEEE Communications Surveys & Tutorials, 18(3):2124–2148, 2016.

[13]. Chen X, Jin R, Suh K, Wang B, and Wei W. Network performance of smart mobile handhelds in a university campus WiFi network. In Proc. of ACM 1MC, November 2012.

[14]. Chen X, Jindal A, Ding N, Hu YC, Gupta M, and Vannithamby R. Smartphone background activities in the wild: Origin, energy drain, and optimization. In Proc. of MobiCom, pages 40–52. ACM, 2015.

[15]. Chow IP, Fua K, Huang Y, Bonelli W, Xiong H, Barnes EL, and Teachman AB. Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. J Med Internet Res, 19(3):e62, 3 2017. [PubMed: 28258049]

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

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

[18]. Frost M, Doryab A, Faurholt-Jepsen M, Kessing LV, and Bardram JE. Supporting disease insight through data analysis: refinements of the monarca self-assessment system. In Proc. of ACM International Joint Conference on Pervasive and Ubiquitous Computing, pages 133–142, 2013.

[19]. Gember A, Anand A, and Akella A. A comparative study of handheld and non-handheld traffic in campus WiFi networks. In Proc. of PAM, 2011.

[20]. Gruenerbl A, Osmani V, Bahle G, Carrasco JC, Oehler S, Mayora O, Haring C, and Lukowicz P. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolarpatients. In Proceedings of the 5th Augmented Human International Conference, page 38 ACM, 2014.

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

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

[23]. Hama Y. Shopping as a coping behavior for stress. Japanese Psychological Research, 43(4), 11 2001.

[24]. Huang J, Qian F, Mao ZM, Sen S, and Spatscheck O. Screen-off traffic characterization and optimization in 3G/4G networks. In Proc. of Internet Measurement Conference, pages 357–364. ACM, 2012.

[25]. Jaques N, Taylor S, Azaria A, Ghandeharioun A, Sano A, and Picard R. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In International Conference on Affective Computing and Intelligent Interaction (ACII), 2015.

[26]. Jin Y,Duffield N, Haffner P, Sen S, and Zhang Z-L. Inferring applications at the network layer using collective traffic statistics. ACM SIGMETRICS Performance Evaluation Review, 38(1), 6 2010.

[27]. Katikalapudi R, Chellappan S, Montgomery F, Wunsch D, and Lutzen K. Associating internet usage with depressive behavior among college students. IEEE Technology and Society Magazine, 31(4):73–80, 2012.

[28]. Kroenke K, Spitzer RL, and Williams JB. The PHQ-9. Journal of General Internal Medicine, 16(9):606–613, 2001. [PubMed: 11556941]

[29]. Lam LT and Peng Z-W. Effect of pathological use of the internet on adolescent mental health: a prospective study. Archives of pediatrics & adolescent medicine, 164(10):901–906, 2010. [PubMed: 20679157]

[30]. LiKamWa R, Liu Y, Lane ND, and Zhong L. Moodscope: Building amood sensor from smartphone usage patterns. In Proceeding of the 11th annual international conference on Mobile systems, applications, and services, pages 389–402, 2013.

[31]. Lu J, Shang C, Yue C, Morillo R, Ware S, Kamath J, Bamis A, Russell A, Wang B, and Bi J. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2017 accepted.

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

[33]. Meng L, Liu S, and Striegel AD. Characterizing the utility of smartphone background traffic. In Proc. of International Conference on Computer Communication and Networks (ICCCN), 2014.

[34]. Morrison CM and Gore H. The relationship between excessive internet use and depression: a questionnaire-based study of 1,319 young people and adults. Psychopathology, 43(2):121–126, 2010. [PubMed: 20110764]

[35]. Network Working Group, Internet Engineering Task Force. Requirements for internet hosts – communication layers. https://tools.ietf.org/html/rfc1122#page-101.

[36]. Palmius N, Tsanas A, Saunders KEA, Bilderbeck AC, Geddes JR, Goodwin GM, and Vos MD. Detecting bipolar depression from geographic location data. IEEE Transactions on Biomedical Engineering, PP(99):1–1, 2016.

[37]. Park S, Lee SW, Kwak J, Cha M, and Jeong B. Activities on Facebook reveal the depressive state of users. Journal of medical Internet research, 15(10), 2013.

[38]. Rakotomamonjy A. Variable selection using svm-based criteria. Journal of machine learning research, 3(Mar):1357–1370, 2003.

[39]. Razavi R, Gharipour A, and Gharipour M. Depression screening using mobile phone usage metadata: a machine learning approach. Journal of the American Medical Informatics Association, 27(4):522–530, 2020. [PubMed: 31977041]

[40]. Rosen LD, Whaling K, Rab S, Carrier LM, and Cheever NA. Is Facebook creating iDisorders? The link between clinical symptoms of psychiatric disorders and technology use, attitudes and anxiety. Computers in Human Behavior, 29(3):1243–1254, 2013.

[41]. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, and Mohr DC. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. Journal of Medical Internet Research, 17(7), 2015.

[42]. Smith KM, Renshaw PF, and Bilello J. The diagnosis of depression: current and emerging methods. Comprehensive Psychiatry, 54(1):1–6, 1 2013. [PubMed: 22901834]

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

[44]. Trestian I, Ranjan S, Kuzmanovic A, and Nucci A. Measuring serendipity: connecting people, locations and interests in a mobile 3G network. In Proc. of ACMIMC, 2009.

[45]. Trull TJ and Ebner-Priemer U. Ambulatory assessment. Annual review of clinical psychology, 9:151–176, 2013.

[46]. Vos T, Flaxman AD, Naghavi M, Lozano R, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the global burden of disease study 2010. The Lancet, 380(9859):2163–2196, December 2012.

[47]. Wahbeh H, Svalina MN, and Oken BS. Group, one-on-one, or Internet? Preferences for mindfulness meditation delivery format and their predictors. Open medicine journal, 1:66, 2014. [PubMed: 27057260]

[48]. Wang R, Aung MSH, Abdullah S, Brian R, Campbell AT, Choudhuryy T, Hauserz M, Kanez J, Merrilly M, Scherer EA, Tsengy VWS, and Ben-Zeev D. Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In Proc. of UbiComp, 2016.

[49]. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, Zhou X, Ben-Zeev D, and Campbell AT. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In Proc. of ACM Ubicomp, pages 3–14, 2014.

[50]. Ware S, Yue C, Morillo R, Lu J, Shang C, Bi J, Russell A, Bamis A, and Wang B. Predicting depressive symptoms using smartphone data. In Proc. ACM/IEEE CHASE, October 2019.

[51]. Ware S, Yue C, Morillo R, Lu J, Shang C, Kamath J, Bamis A, Bi J, Russell A, and Wang B. 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, 2018.

[52]. Weaver J III, Mays D, Weaver SS, Kannenberg W, Hopkins G, Eroglu D, and Bernhardt J. Health-risk correlates of video-game playing among adults. America Journal of Preventive Medicine, 37(4):299–305, 2009.

[53]. Xu X, Chikersal P, Doryab A, Villalba DK, Dutcher JM, Tumminia MJ, Althoff T, Cohen S, Creswell KG, Creswell JD, Mankoff J, and Dey AK. Leveraging routine behavior and contextually-filtered features for depression detection among college students. In Proc. of UbiComp, 2019.

[54]. Yang HH and Moody J. Data visualization and feature selection: New algorithms for nongaussian data. In Advances in Neural Information Processing Systems, pages 687–693, 2000.

[55]. Young KS and Rogers RC. The relationship between depression and internet addiction. Cyberpsychology & behavior, 1(1):25–28, 1998.

[56]. Yue C, Sen S, Wang B, Qin Y, and Qian F. Energy considerations for ABR video streaming to smartphones: Measurements, models and insights. In Proc. of ACM MMSys, June 2020.

[57]. Yue C, Ware S, Morillo R, Lu J, Shang C, Bi J, Russell A, Bamis A, and Wang B. Fusing location data for depression prediction. IEEE Transactions on Big Data, 10 2018.

[58]. Zhou D, Luo J, Silenzio VM, Zhou Y, Hu J, Currier G, and Kautz H. Tackling mental health by integrating unobtrusive multimodal sensing. In AAAI Conference on Artificial Intelligence, 2015.

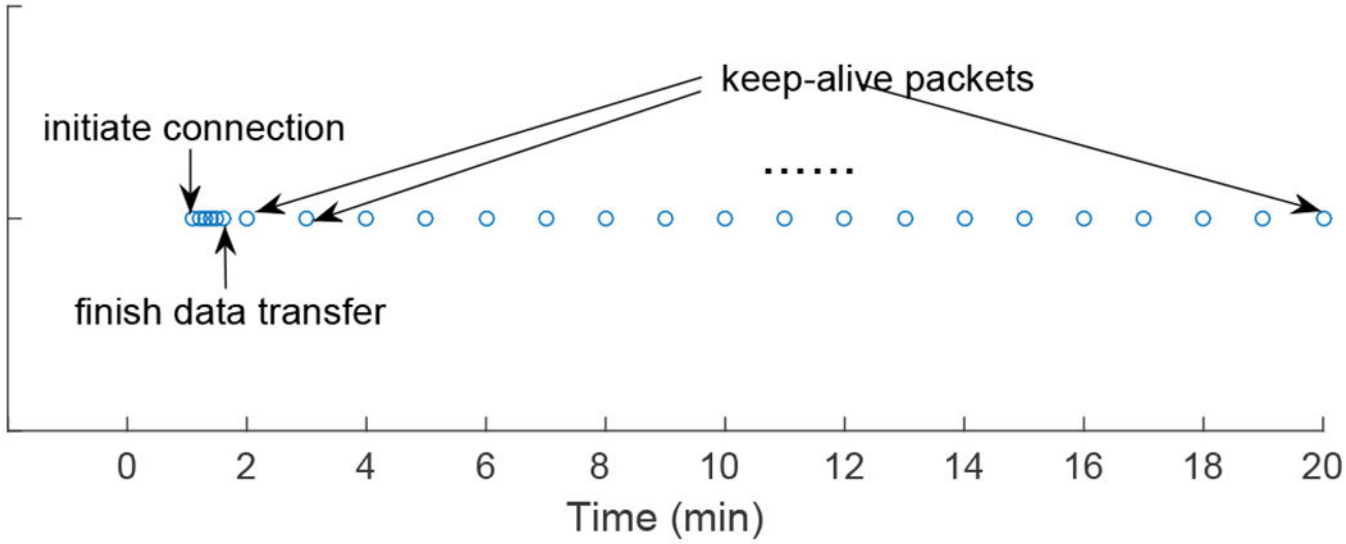


Fig. 1:

An example that illustrates keep-alive packets.

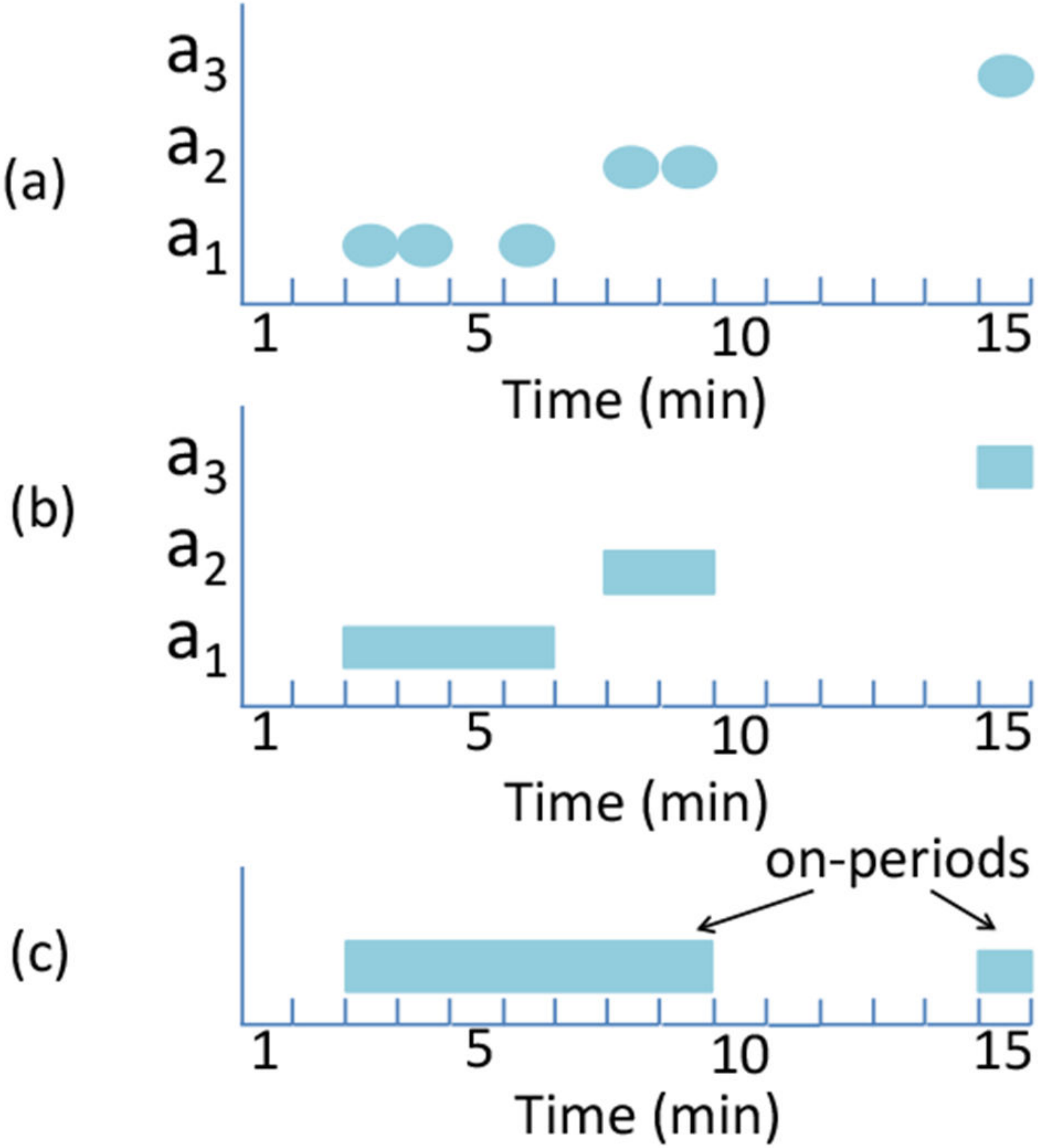


Fig. 2:

Illustration of identifying on-off periods. (a) Per-minute traffic for each of the three external IP addresses (an oval marks the presence of traffic in that minute). (b) Aggregation for each IP address. (c) Aggregation over the IP addresses.

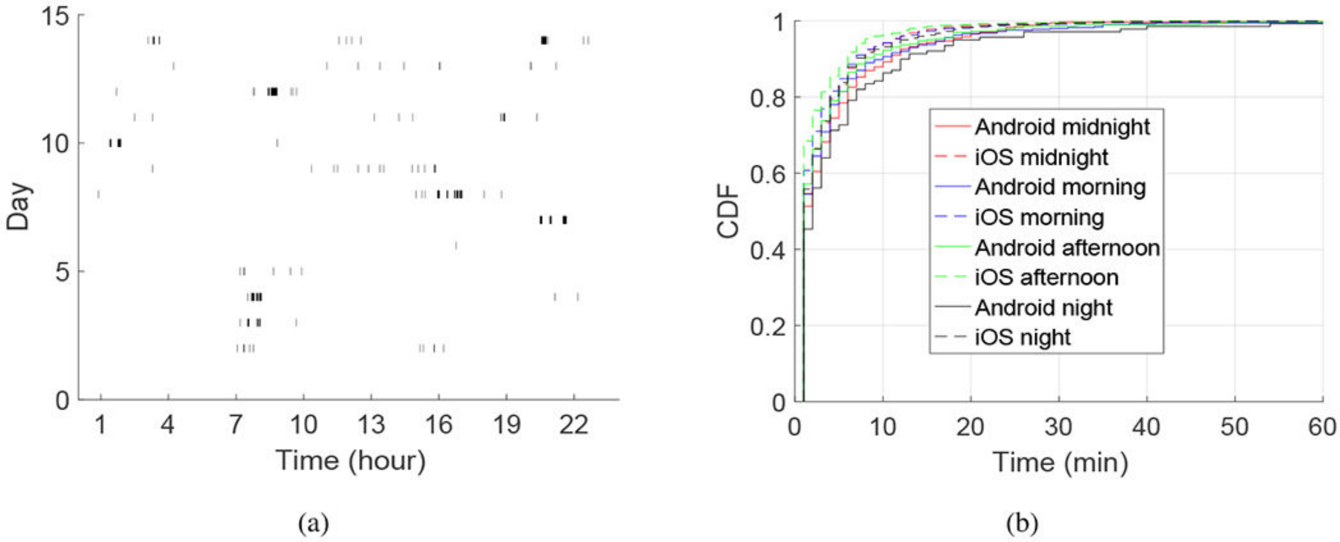
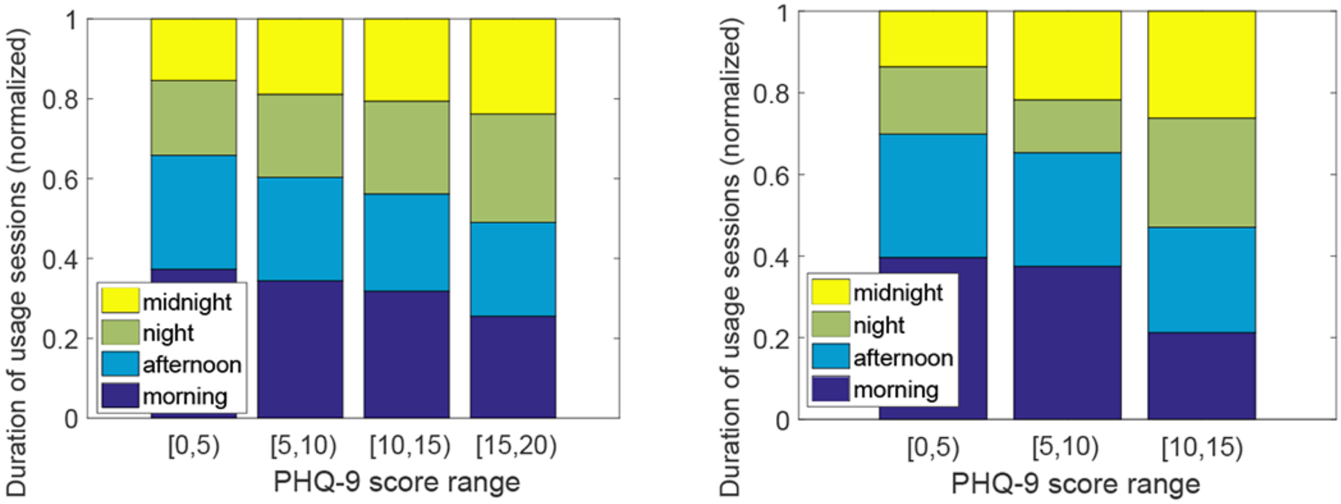


Fig. 3:

On-periods after data pre-processing. (a) Illustration of the on-off periods for one Android user in a PHQ-9 interval; the black bars represent on-periods. (b) The distributions of the on-periods during four time periods, morning (6am-12pm), afternoon (12pm-6pm), night (6pm-12am) and midnight (12am-6am), for the Android and iOS datasets.

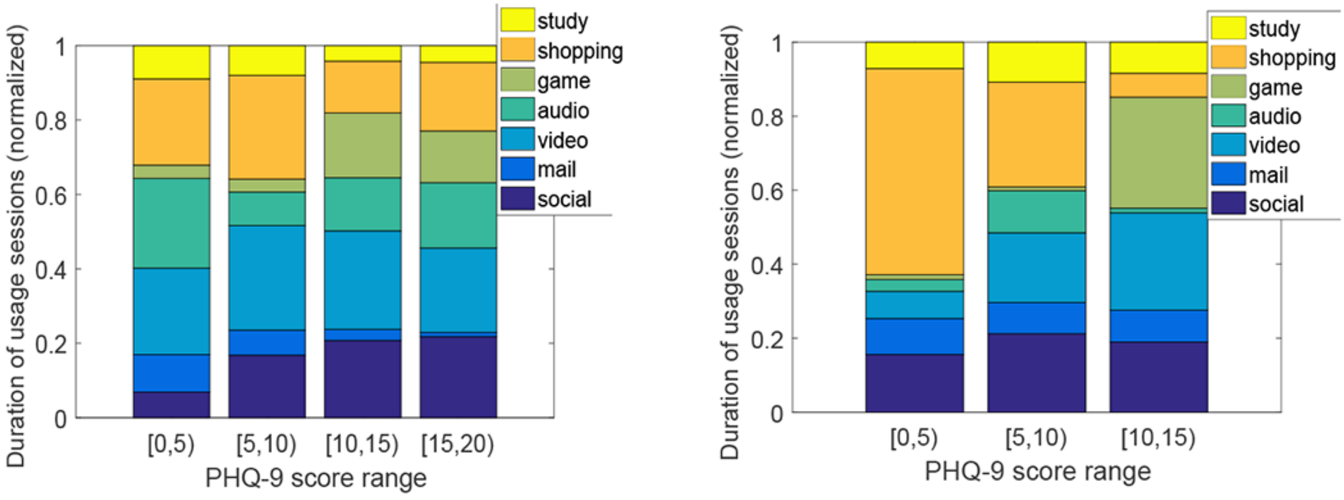


(a) iOS dataset.

(b) Android dataset.

Fig. 4:

The amount of time online in each of the four time periods (normalized value) versus PHQ-9 score range.

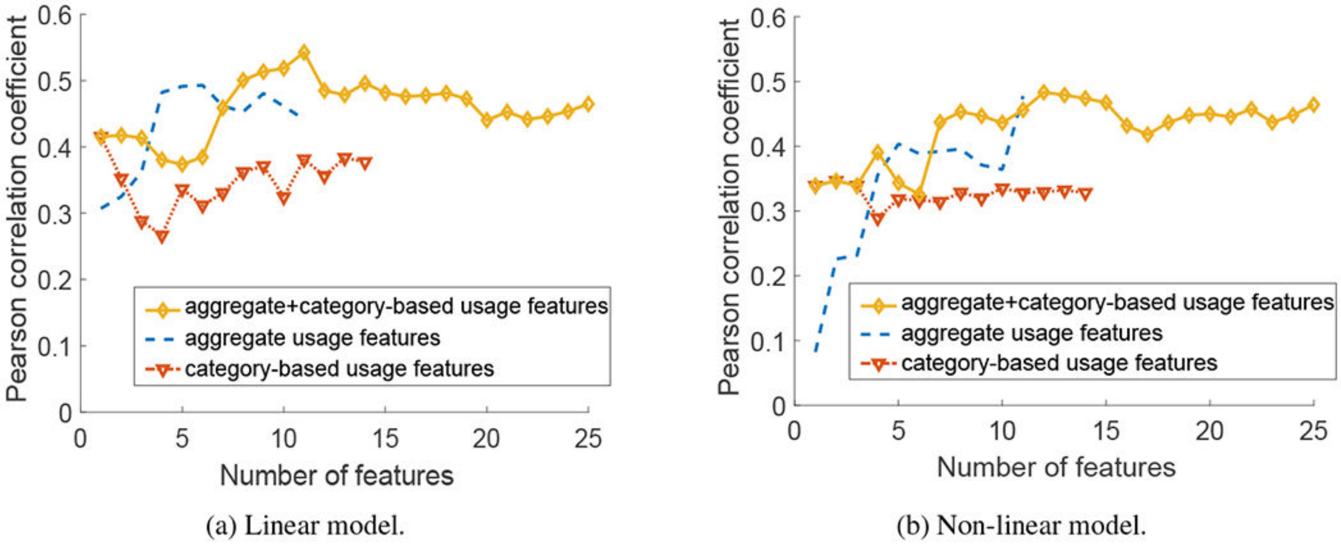


(a) iOS dataset.

(b) Android dataset.

Fig. 5:

The amount of time spent on each application category (normalized value) versus PHQ-9 score range.

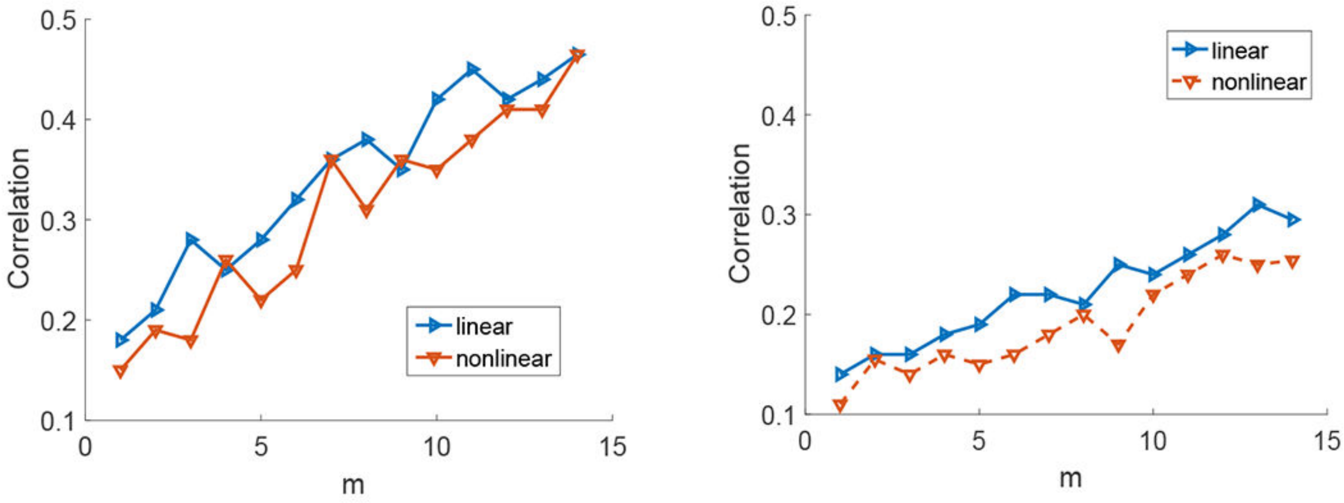


(a) Linear model.

(b) Non-linear model.

Fig. 6:

Correlation of the values predicted by the models with the PHQ-9 scores when increasing the number of the selected features for the iOS dataset.

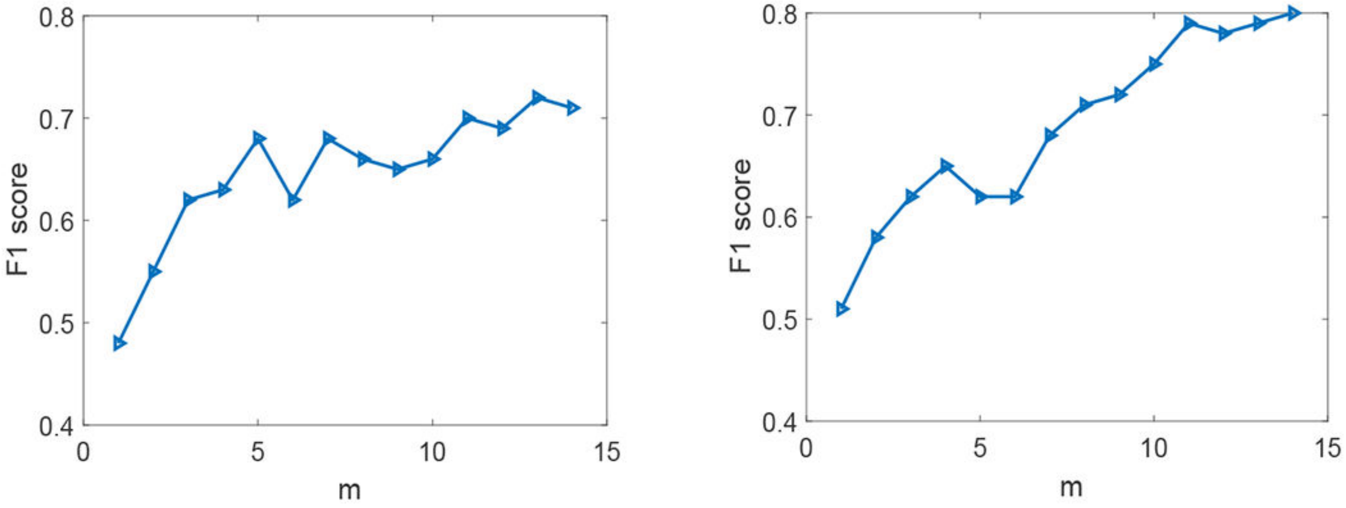


(a) iOS dataset.

(b) Android dataset.

Fig. 7:

Impact of *m* on multi-feature regression results: correlation of the predicted PHQ-9 score and the ground-truth PHQ-9 score versus *m*, i.e., when considering *m* days of Internet traffic data.



(a) iOS dataset.

(b) Android dataset.

Fig. 8:

Impact of m on classification results: predicted F_1 score versus m, i.e., when considering m days of Internet traffic data.

Table 1:

An example output when looking up an IP address from DBIP.

Address type	**IPv4**
Hostname	edge-star-shv-01-iad3.facebook.com
ASN	32934 - FACEBOOK - Facebook, Inc.
ISP	Facebook
Description	IPv4 address owned by Facebook and located in Washington D.C., District of Columbia, United States

Table 2:

Keywords used to classify Internet visits into application categories.

Category	Keywords
mail	mail, aol, zumbox
social	twitter, linkedin, snapchat, instagram, wechat, whatsapp, vlingo, snapch, tencent, myspace, blog, LivePerson
video	video, youtube, brightcov, freeWheel, cedato, youtube, conviva
audio	audio, music, pandora, spotify, rhapsody, radio, song, mp3, tone
game	game, midasplayer, neogaf, machine zone, poker, blackjack, casino
shop	shop, paypal, trade, tapa, gwallet, taobao, ebay, nobis, l brand, road runner, netbank, aciworldwide, fujitsu, digital insight, mart, gigy, hasbro, market, craigslist
study	cerfnet, intuit inc, shuyuan, blackboard

Table 3:

Correlation between aggregate usage features and PHQ-9 scores.

		All		Depressed		Non-depressed	
	Feature	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
iOS	Total duration	0.25	10^{-3}	0.26	0.006	0.28	0.05
	Total number of sessions	0.27	10^{-4}	0.28	0.003	0.31	0.03
	Total off-duration	−0.11	0.16	−0.06	0.49	−0.19	0.17
	Duration (morning)	−0.40	10^{-8}	−0.27	0.004	−0.49	10^{-4}
	Duration (afternoon)	−0.27	10^{-4}	−0.35	10^{-4}	−0.16	0.27
	Duration (night)	0.39	10^{-7}	0.34	10^{-4}	0.48	10^{-4}
	Duration (midnight)	0.28	10^{-3}	0.16	0.09	0.33	0.02
	Number of sessions (morning)	0.11	0.16	0.17	0.07	0.14	0.31
	Number of sessions (afternoon)	0.17	0.03	0.18	0.05	0.25	0.08
	Number of sessions (night)	0.36	10^{-6}	0.36	10^{-5}	0.37	0.007
	Number of sessions (midnight)	0.36	10^{-6}	0.35	10^{-4}	0.35	0.01
Android	Total duration	0.28	0.12	−0.47	0.03	0.62	0.06
	Total number of sessions	−0.04	0.84	−0.32	0.14	0.32	0.37
	Total off-duration	0.10	0.60	0.34	0.13	−0.21	0.56
	Duration (morning)	−0.11	0.56	0.01	0.95	−0.36	0.31
	Duration (afternoon)	−0.08	0.68	0.01	0.95	−0.38	0.28
	Duration (night)	−0.06	0.73	−0.19	0.40	0.31	0.39
	Duration (midnight)	0.32	0.08	0.23	0.31	0.56	0.09
	Number of sessions (morning)	−0.19	0.31	−0.32	0.14	−0.19	0.61
	Number of sessions (afternoon)	−0.21	0.26	−0.27	0.22	−0.28	0.44
	Number of sessions (night)	0.02	0.93	−0.33	0.14	0.43	0.21
	Number of sessions (midnight)	0.28	0.12	−0.13	0.56	0.50	0.14

Table 4:

Correlation between category-based features and PHQ-9 scores.

		All		Depressed		Non-depressed	
	Feature	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
iOS	Duration (social)	0.28	10^{-4}	0.21	0.02	0.26	0.07
	Duration (mail)	−0.24	10^{-3}	−0.29	10^{-3}	−0.02	0.90
	Duration (video)	0.01	0.85	0.02	0.79	10^{-3}	0.97
	Duration (audio)	−0.17	0.04	−0.12	0.23	−0.24	0.09
	Duration (gaming)	0.27	10^{-4}	0.10	0.25	0.37	10^{-3}
	Duration (shopping)	0.06	0.41	0.11	0.21	−0.03	0.85
	Duration (study)	−0.13	0.11	−0.08	0.42	−0.05	0.76
	Number of sessions (social)	0.25	10^{-3}	0.34	10^{-4}	0.12	0.37
	Number of sessions (mail)	−0.06	0.49	−0.11	0.28	0.13	0.33
	Number of sessions (video)	0.12	0.12	0.04	0.60	0.16	0.24
	Number of sessions (audio)	−0.06	0.49	0.04	0.65	−0.16	0.28
	Number of sessions (game)	0.33	10^{-6}	0.37	10^{-5}	0.27	0.04
	Number of sessions (shopping)	0.30	10^{-5}	0.42	10^{-6}	0.25	0.06
	Number of sessions (study)	0.10	0.19	0.17	0.05	0.10	0.47
Android	Duration (social)	0.20	0.27	0.13	0.56	0.26	0.47
	Duration (mail)	−0.09	0.62	−0.07	0.77	−0.2	0.58
	Duration (video)	0.44	0.01	0.28	0.20	0.74	0.01
	Duration (audio)	0.17	0.34	0.38	0.08	−0.31	0.38
	Duration (gaming)	0.52	10^{-3}	0.17	0.44	0.65	0.04
	Duration (shopping)	−0.47	0.01	−0.49	0.02	−0.51	0.14
	Duration (study)	−0.27	0.12	−0.30	0.18	−0.16	0.67
	Number of sessions (social)	0.40	0.02	0.04	0.86	0.50	0.14
	Number of sessions (mail)	0.12	0.50	−0.22	0.33	0.42	0.22
	Number of sessions (video)	0.46	0.01	−0.49	0.02	0.64	0.05
	Number of sessions (audio)	−0.04	0.84	0.05	0.82	0.01	0.90
	Number of sessions (game)	0.65	10^{-5}	0.02	0.91	0.86	10^{-3}
	Number of sessions (shopping)	−0.24	0.18	−0.40	0.06	0.43	0.22
	Number of sessions (study)	−0.18	0.31	−0.40	0.06	0.42	0.22

Table 5:

Correlation between volume-based features and PHQ-9 scores.

		All		Depressed		Non-depressed	
	Feature	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
iOS	Total volume	0.16	0.02	0.22	0.01	0.18	0.08
	Volume (morning)	−0.09	0.13	−0.09	0.24	−0.14	0.13
	Volume (afternoon)	−0.05	0.37	0.01	0.99	−0.10	0.28
	Volume (night)	0.06	0.34	0.04	0.67	0.17	0.09
	Volume (midnight)	0.17	0.01	0.12	0.15	0.22	0.03
Android	Total volume	0.16	0.35	0.35	0.23	−0.27	0.20
	Volume (morning)	0.07	0.71	0.11	0.73	−0.08	0.68
	Volume (afternoon)	−0.09	0.56	−0.01	0.97	−0.10	0.62
	Volume (night)	−0.10	0.55	−0.17	0.53	−0.02	0.91
	Volume (midnight)	0.24	0.15	0.10	0.75	0.50	0.01

Table 6:

Prediction results using multi-feature regression.

		Linear model		Non-linear model	
	Features	**r-value**	**p-value**	**r-value**	**p-value**
iOS	Volume-based features	0.29	10^{-4}	0.33	10^{-5}
	Aggregate usage features	0.49	10^{-14}	0.48	10^{-10}
	Category-based usage features	0.42	10^{-8}	0.35	10^{-6}
	Aggregate usage + volume-based features	0.46	10^{-9}	0.48	10^{-10}
	Aggregate usage + category-based usage features	0.54	10^{-14}	0.48	10^{-11}
	Aggregate usage + category-based usage + volume-based features	0.48	10^{-11}	0.48	10^{-11}
Android	Volume-based features	0.33	0.05	0.29	0.08
	Aggregate usage features	0.33	0.05	0.35	0.04
	Category-based usage features	0.29	0.08	0.31	0.07
	Aggregate usage + volume-based features	0.33	0.05	0.37	0.03
	Aggregate usage + category-based usage features	0.33	0.05	0.38	0.02
	Aggregate usage + category-based usage + volume-based features	0.33	0.05	0.39	0.02

Table 7:

Selected features for multi-feature regression for the non-linear model when combining aggregate usage and category-based usage features.

	Selected features
iOS	Number of sessions (gaming), Number of sessions (morning), Number of sessions (social), Total duration, Duration (game), Number of sessions (video), Number of sessions (midnight), Duration (midnight), Number of sessions (night), Duration (afternoon), Duration (social), Number of sessions (audio)
Android	Duration (video), Number of sessions (video), Number of sessions (night), Total number of sessions, Duration (midnight), Duration (shopping)

Table 8:

Classification results when using various features.

		F_1 Score	Precision	Recall	Specificity	log(C)	log(γ)
iOS	Volume-based features	0.60	0.61	0.59	0.62	−2	4
	Aggregate usage features	0.63	0.67	0.60	0.59	8	5
	Category-based usage features	0.65	0.62	0.68	0.61	−1	5
	Aggregate usage + volume-based features	0.69	0.62	0.78	0.62	−3	4
	Aggregate usage + category-based usage features	0.71	0.71	0.71	0.63	−2	4
	Aggregate usage + category-based usage + volume-based features	0.71	0.67	0.76	0.63	−2	4
Android	Volume-based features	0.56	0.50	0.64	0.45	3	1
	Aggregate usage features	0.62	0.7	0.57	0.51	0	1
	Category-based usage features	0.66	0.77	0.59	0.44	2	5
	Aggregate usage + volume-based features	0.71	0.71	0.71	0.69	5	2
	Aggregate usage + category-based usge features	0.80	0.75	0.86	0.77	8	−5
	Aggregate usage + category-based usage + volume-based features	0.80	0.75	0.86	0.77	8	−5

Table 9:

The features that were selected to achieve the highest F_1 score when combining both aggregate and category-based features.

	Selected features
iOS	Total off-duration, Duration (morning), Duration (night), Number of sessions (night), Number of sessions (afternoon), Number of sessions (midnight), Duration (mail), Number of sessions (mail), Number of sessions (game), Number of sessions (shopping)
Android	Total duration, Total off-duration, Duration (midnight), Number of sessions (morning), Number of sessions (night), Number of sessions (midnight), Duration (social), Duration (mail), Duration (shopping), Number of sessions (game)