2016 IEEE First Conference on Connected Health: Applications, Systems and Engineering Technologies

Multi-view Bi-Clustering to Identify Smartphone Sensing Features Indicative of Depression

Asma Ahmad Farhan[1], Jin Lu[1], Jinbo Bi[1], Alexander Russell[1], Bing Wang[1], Athanasios Bamis[2]
[1]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA
[2] Seldera LLC
{asma.farhan, jin.lu, jinbo, acr, bing}@engr.uconn.edu, athanasios.bamis@gmail.com

***Abstract*—Depression is a major public health issue with direct and significant effects on both physical and mental health. In this study, we analyze smartphone sensing data to find differential behavioral features that are correlated with depression measures such as patient health questionnaire (PHQ-9). Our approach uses an innovative multi-view bi-clustering algorithm. It takes multiple views of sensing data as input to identify homogeneous behavioral groups and simultaneously the key sensing features that characterize the different groups. Using a publicly available dataset, we discover that these behavioral groups with differential sensing features are highly discriminative of PHQ-9 scores that are self reported by the study subjects. For instance, the group comprising less active users in the sensed activities corresponds to overall higher PHQ-9 scores. We then employ the key sensing features that distinguish the different groups to create predictive models to predict the group assignment of individuals. We verify the generalizability of these models using the support vector machine classifier. Cross validation studies show that our classifiers can classify individuals into the correct subgroups with an overall accuracy of 87%.**

I. Introduction

Depression is a severe public health problem. It is estimated that depression affects 350 million people worldwide, and is ranked the 2nd among all the major illnesses in Years Lived with Disability (YLDs), accounting for 9.6% of all YLDs from all major illnesses [27]. It is a significant contributor to death by suicide [27]. In the United States, reports in 2010 show that suicide is the 10th leading cause of death, and 70% of these suicide victims are reported to have a mood disorder such as depression [1]. Depression is found to be associated with life threatening diseases like diabetes and heart related issues [4]. Currently, depression is diagnosed based on physician-administered or patient self-administered survey instruments that require significant effort and cost, rely on accurate introspection and reporting, and are inappropriate for continuous monitoring of depression or its onset.

The ubiquitous adoption of smartphones has created new opportunities for uncovering the relationship between behavior and depression. Smartphones are highly portable to their users as a small device, making them effective "human sensors" appropriate for cataloging and analyzing broad aspects of human behavior. In addition, sensing data from smartphones can be more objective than self-reports to reflect a user's behavior in the diagnosis of depression. Therefore, depression screening using smartphones could be a significant initiative to identify depression timely. On the other hand, human behavior is extremely stochastic in nature and is affected by many external factors. In addition, the correlations between behavior and depression are complex. Until now the exact causes of depression are unknown as they involve a complex interaction of genes and environment. How to use smartphone sensing data to effectively understand, monitor, detect and predict depression remains a challenging task.

Several recent studies [7], [23], [28] have demonstrated that sensing data collected from smartphones can be used to extract features related to depressive mood. Various features have been identified, ranging from activity, conversation, to locations visited, that each provide a different view into a user's behavior. It is thus important to ask whether there exist subgroups of user behavior measured by different sensing views that can be predictive or indicative of depression. Existing studies use standard correlation analysis that identifies individual features correlated with a depression measure (e.g., through PHQ-9 examination with or without clinical oversight) [28]. However, depression can be diagnosed from heterogeneous behavior, such as some patients with insomnia but others with oversleeping. There can exist substantial difference in sensing features indicative of depression. Differentiating subgroups of behavior based on sensing data may shed light on different characteristics of depression or mood swings and help us understand if sensing data can play an essential role in the screening of depression.

In this paper, we propose an innovative machine learning approach to identify subgroups of user behavior and the key features of these behavior that are indicative of depression measures, in particular, the patient health questionnaire (PHQ-9) scores. The central component of our approach is a novel multi-view bi-clustering method we recently published [26], that identifies the clusters (i.e., homogeneous groups) of individuals and the key features that characterize the behavior of these clusters simultaneously. Specifically, we extract features from different views of smartphone data, including activities (e.g., walking, running), conversation statistics, phone light features and locations that a user visited, and then represent the users with their associated features using several data matrices, one for each view. Our multi-view bi-clustering method aims to identify consistent row clusters across the views and column clusters in each view as shown in Fig. 1. The column clusters will specify the features from each view for a row cluster. Different from merging all views into a single

978-1-5090-0943-5/16 $25.00 © 2016 IEEE
DOI 10.1109/CHASE.2016.27

data matrix and performing a standard cluster analysis, which may identify clusters of users that only differ in a subset of features from a single view, our approach can guarantee the confirmatory evidence of a behavior subgroup from all sensing feature inputs. Support vector machine (SVM) classifiers are then constructed to distinguish a subgroup of subjects from others based on the identified features. We demonstrate the capability and effectiveness of our approach using a publicly available dataset collected from the StudentLife project [28] at the Dartmouth College.

The rest of the paper is organized as follows. Section II discusses the related work, Section III is dedicated to the proposed approach. Evaluation results and discussion are included in Sections IV. Section V concludes the paper and describes future work.

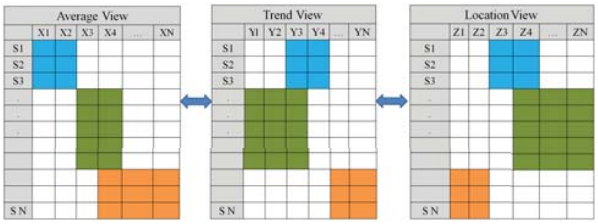


Fig. 1. Multi-view bi-clustering: rows are grouped in the same way across the three matrices. The users in each row cluster are homogeneous over a subset of features from each of the views (shown as column clusters).

II. Related Work

Researchers have used the rich set of built-in sensors in smartphones, for example, accelerometer, GPS, bluetooth and gyroscope, to infer location [9], [10], co-location [11], activity [12], and social relationships [18]. In terms of health related applications, recent studies have shown that human behavioral patterns identified through smartphone sensors have potential to provide useful insights into the mental and physical health of the smartphone users. For instance, the authors of [19] analyze smartphone sensor data to find the relationship between sleep, mood and sociability. They observe that sleep and sociability have significant relationship as good night sleep leads to better sociability. In another work [17], smartphone is used as a tool for understanding the effect of social interactions on weight. The authors find that social interactions can influence the weight changes and dietary habits. The authors in [16] analyze smartphones' co-location and communication sensing data to find behavioral changes when a person suffers from physical or mental health issues like common cold or stress. They find that the data can give interesting insights regarding health status of the smartphone owners. In some other studies, smartphones have been used to monitor sleep [8], stress [3], [5], [15], [24], mood [13], and general wellbeing [14], [22].

Several recent studies use smartphone sensing data for depression detection, prediction and intervention. In studentLife [28], the authors find that conversation frequency and duration, sleep duration and the number of co-locations are correlated with depression symptoms. While in [7], the authors analyze mobility patterns, extracted from GPS traces of smartphones, to understand whether mobility patterns are correlated with PHQ-9 scores. They use Support Vector Machines (SVM) to build both individual and general prediction models, and find a significant correlation between mobility patterns and depressive mood. Similarly, authors in [23] use several features extracted from smartphone GPS information and phone usage patterns, and find that they are strongly related to depressive symptoms. Another study [6] proposes Mobilyze! that provides momentary intervention to a patient that suffers from depression based on prediction of the patients mode, emotions, cognitive state, and activities using smartphone sensors.

Our work differs from the existing studies in that we use multi-view clustering [26], an unsupervised clustering algorithm, to find homogeneous behavior groups that are discriminative of depression (in particular, PHQ-9 scores). In addition, we identify a set of key features from an array of smartphone sensing data that can be good indicators of depressive mood disorder.

III. Our Approach

The proposed approach consists of three steps as shown in Fig. 2: feature extraction, clustering and feature selection, and then classification of subjects into the identified user clusters. Since our goal is to identify key behavioral features that are related to depression, we first extract features from the smartphone sensing data, and organize these features into three views according to the kind of information described by the features. The first view examines the average or cumulative behaviors of users by averaging the daily activity features, which we call the *average view*. The second view examines the variation of day-to-day dynamics in the physical activities, which we call the *trend view*. The third view extracts features from GPS location data, measuring the variability of a user's transition among locations, which we call the *location view*. We next apply the multi-view bi-clustering algorithm to the three data matrices (each corresponding to one view) to identify homogeneous behavior subgroups as well as the key features that distinguish the subgroups. The concurrent validity of the resultant subgroups is assessed not only by comparing the sensing features used in the cluster analysis but also other metrics, such as PHQ-9 scores (as we shall see, the subgroups are discriminative of PHQ-9 scores). In the last step, SVM classifiers are constructed to separate one cluster of users from another cluster of users based on the identified features, forming three classifiers. Cross validation is used to evaluate the generalizability of these classifiers. To assess the validity of the identified features, the classifiers built from using the selected features are compared with those built from using all the sensing features.

A. Dataset

We use the dataset from the Dartmouth Studentlife website [28]. The dataset contains survey self-reports, academic performance data, and passive and automatic sensing data collected from smartphone built-in sensors of 60 college students for

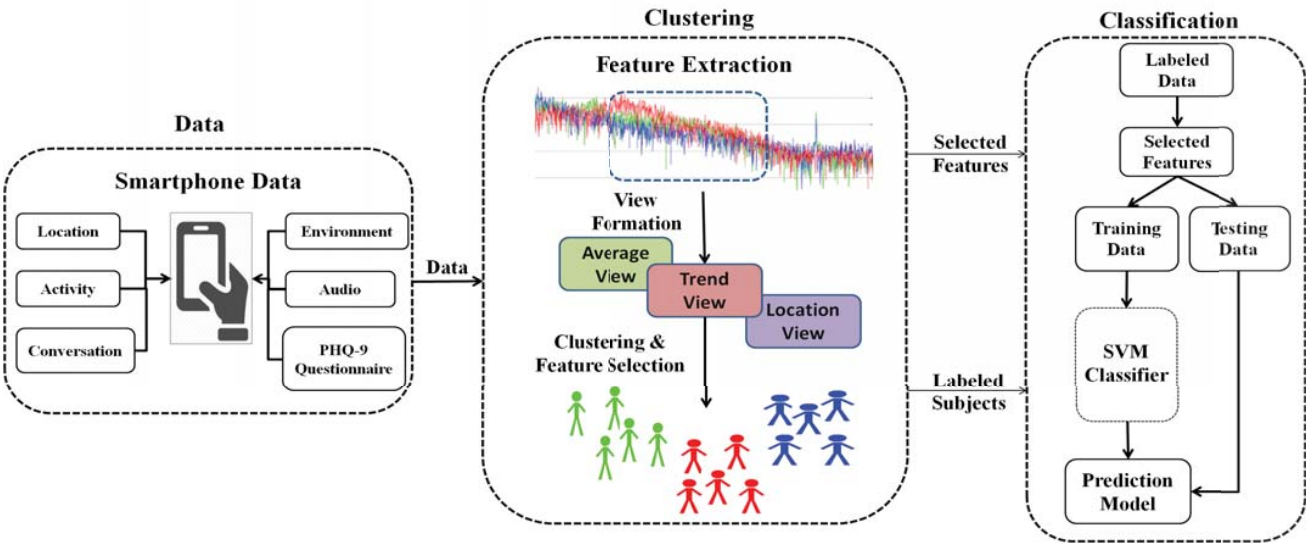


Fig. 2. High-level overview of our approach.

over 10 weeks. The specific data that we use in this study are as follows.

- Physical Activity. Physical activity is characterized in four categories: *stationary*, *walking*, *running* and *unknown*, as specified in [28]. To infer physical activity, data is collected $24/7$ with duty cycling. Specifically, the classifier runs for one minute, generating an activity inference every 2 to 3 seconds, and then pauses for 3 minutes before the next run.
- Light Information. Light sensor is used to collect the starting and ending times when a phone is in a dark environment. Data is only recorded when the phone is in a dark environment for more than an hour.
- Phone Lock. It records the starting and ending time when a phone is locked for a significant amount of time, i.e., more than one hour.
- Conversation. It records the starting and ending time for each conversation period.
- Audio. The data collection frequency is similar to that of physical activity. The audio information is classified into four categories, *silence*, *voice*, *noise* and *unknown*, using a classifier in [28].
- GPS Location. GPS coordinates are collected after every 10 minutes. Each sample contains the longitude, latitude and timestamp when the coordinate is sensed.
- PHQ-9. PHQ-9 is a standard self-report questionnaire that is used as a screening and diagnostic tool for mental health disorders of depression. Studentlife dataset contains up to two PHQ-9 reports for each participant, one at the beginning of the study and the other at the end of the study.

B. Feature Extraction and Views

We extract features from the dataset described in Section III-A and arrange them into three views as follows.

1) The Average View: Intuitively, the average behavior of a low PHQ-9 scorer differs from that of a high PHQ-9 scorer. The average view contains a set of features, each being an average value (the average is over all the the days when a participant is enrolled in the study) that reflects an individual's overall behavior in one category (e.g., conversation, activity) or environment information (e.g., light level, noise). Specifically, the features include the following.

- Activity features, specifically, $Activity_s$, $Activity_w$ and $Activity_r$, which represent respectively the total duration when a participant is in stationary, walking, and running activities in a day.
- Conversation features, specifically, $Conv_d$ and $Conv_c$, which represent respectively the total duration and number of the conversations on average that a participant has in a day.
- Light features, specifically, $Dark_d$ and $Dark_c$, which represents respectively the total duration and number of times when a participant is a dark environment in a day.
- Audio features, specifically, $Audio_q$, $Audio_n$, and $Audio_v$, which represents respectively the total duration when the audio is classified as quiet, noisy and voice in a day.
- Phone lock features, specifically, $PhoneLock_d$ and $PhoneLock_c$, which represents respectively the total duration and number of times when a participant's phone is locked in a day.

2) The Trend View: Although the average behavior across multiple days can be indicative of depressive mood swings, the average of daily features may also cancel out the effects from day-to-day fluctuation or dynamics of a user's behavior. In the trend view, we calibrate the variation of several quantities

(including walking activity, noise audio duration and conversation duration) over the study period using signal processing techniques, which is another innovative aspect of our study.

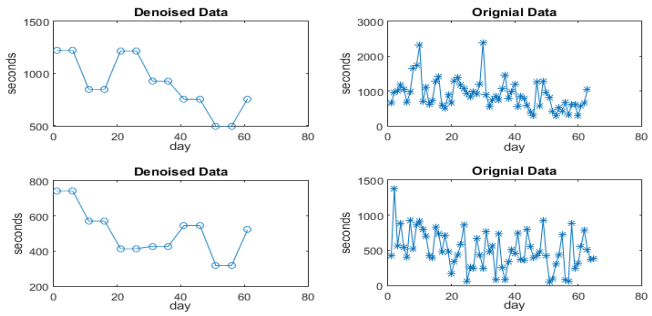
We next use the daily conversation duration $Conv_d$ as an example to illustrate how we calculate the variance of a sensor feature. First we use wavelet transformation to filter the noise from the time series, which is in a low resolution presented as one value per day across the 60 days. In particular, we used Haar wavelet in the transformation because Haar wavelet filtering can preserve the peak and trend of a time series curve. As a result of the filtering, smooth time series curves are obtained. Fig. 3 illustrates the transformation of the conversation duration time series of four users. We can see from the figure that the overall (low frequency) trend is retained in the denoised data. In particular, we compare the curves of two low PHQ-9 scorers with those of two high PHQ-9 scorers. The two users with high PHQ-9 scores (indicating depression) tend to have a declined trend in the time they spent in their daily phone conversations, while the two users with low PHQ-9 scores tend to have more periodical behavior during the study period. This was the intuition for us to derive this view of features from the variation of the sensing features. Second, after the wavelet transformation to obtain the denoised time series, we solve a least squares problem in Eq.(1) to extract four features: the amplitude (c_1), period (c_2), phase (c_3), and intercept (c_4) from each individual's denoised conversation duration time series [21].

$$\min_{\mathbf{c}} \sum_{d} \left(f(\mathbf{c}, d) - \bar{y}_d\right)^2$$
$$\text{subject to} \quad f(\mathbf{c}, x) = c_1 \sin\left(\frac{2\pi}{c_2} x + c_3\right) + c_4, \tag{1}$$

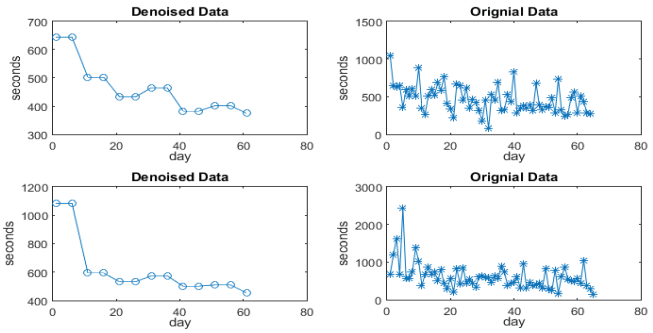
where $\bar{y}_d$ is the denoised daily-average conversation duration on the d-th day, and $\mathbf{c}$ represents the four parameters, $c_1, \ldots, c_4$, to be determined from this optimization problem. For instance, we extract the amplitude ($ConD_a$), period ($ConD_p$), phase ($ConD_{ph}$), and intercept ($ConD_i$) of the conversation duration time series of an individual user. The Levenberg-Marquardt Method [20] was employed to solve this optimization problem. These four features have been shown to provide important additional information to the average view in our experimental results.

The following features are obtained by applying the above process to the respective raw sensing features.

- Variation of daily walking activity: specifically, we use $Walk_a$, $Walk_p$, $Walk_{ph}$ and $Walk_i$ to represent the respective amplitude, period, phase and intercept of daily walking duration.
- Variation of daily noise audio: specifically, we use $Noise_a$, $Noise_p$, $Noise_{ph}$ and $Noise_i$ to represent the amplitude, period, phase and intercept of daily duration of noise audio signals.
- Variation of daily conversation duration: specifically, we use $ConD_a$, $ConD_p$, $ConD_{ph}$ and $ConD_i$ to represent the amplitude, period, phase and intercept of the daily conversation duration.



(a) Results of two low PHQ-9 score users



(b) Results of two high PHQ-9 score users

Fig. 3. Illustration of the denoising process for conversation duration signals of four users across the 60 days in the study period.

3) The Location View: Recent studies [7], [23] have shown significant correlation between location, mobility patterns and depressive mood disorder. In our location view, we used similar features as proposed in [7], [23]. Specifically, the features include the following:

- Location variance, $Location_{var}$, which measures the variability in a participant's location. Using the approach in [23], we calculate the location variance as

$$Location_{var} = (\sigma^2_{\text{long}} + \sigma^2_{\text{lat}}), \tag{2}$$

  where σ^2_{long} and σ^2_{lat} represent respectively the variance of the longitude and latitude of the GPS coordinates.
- Time in location clusters, $Time_{C_1}$, $Time_{C_2}$, and $Time_{C_3}$, which represent respectively the amount of time that a participant spends in the top three clusters, normalized by the number of days that the participant is enrolled in the study. We identify unique clusters from the location data using k-means clustering. To find the optimal k, we initially set k from 1 to 10, apply k-means clustering for each of these values. We then calculate the distance between a point in a cluster and the centroid of the cluster. Intuitively, as the number of clusters increases, the distance to the centroid decreases. The optimal k is the value beyond which the decrease is minimal. Using the described approach, we first find the optimal number of clusters and then calculate the amount of time that a

user spent respectively in each cluster. We find that for all the participants, the top three clusters cover 90% of the locations (sorting cluster w.r.t size).

- Entropy of a participant's locations, specifically, *Entropy*, measures how uniformly a participant spends time at different locations. Let p_i denote the percentage of time that a participant spends in location cluster i. The entropy of the participant is calculated as

$$\text{Entropy} = -\sum (p_i \log p_i) \tag{3}$$

- Normalized Entropy, specifically, *Entropy*$_N$. Since the number of location clusters varies among the participants and entropy increases as the number of location clusters increases, we also adopt normalized entropy [23], which is invariant to the number of clusters and depends solely on the distribution of the visited location clusters. The range of normalized entropy is in $[0, 1]$, where 0 implies that all location data points belong to the same cluster, while 1 means that all points are uniformly distributed among all the clusters. Let N represent the total number of location clusters for a participant. Then the normalized entropy for the participant is calculated as the participant's entropy in location normalized by $\log N$, i.e.,

$$\text{Entropy}_{\text{N}} = \text{Entropy}/\log N \tag{4}$$

- Percentage of time that a participant spends at home/dorm, specifically, *Home*$_d$. As in [23], we identify "home" for a participant as the location where the participant is found most often between 12am to 6am. For a participant, let T_d denote the total amount of time that the participant spends in all GPS locations and H_d denote the amount of time that the participant spends at home on the dth day. Then

$$\text{Home}_{\text{d}} = \sum H_d / \sum T_d \tag{5}$$

- Percentage of time that a participant is moving, specifically, *Move*$_{percent}$. Let M_d denote the total amount of time when a participant is moving on the dth day. Then

$$\text{Move}_{\text{percent}} = \sum M_d / \sum T_d \tag{6}$$

- Total distance coved by a participant while enrolled in the study, specifically, *Dist*. Given the longitude and latitude for a particular participant, we use Harversine formula [25] to calculate the distance traveled in kilometers by the participant, which is then normalized by the number of days that the participant is enrolled in the study.

C. Identifying Homogeneous Behavior Groups

We now briefly describe our recently-developed multi-view bi-clustering method [26], which is used to identify homogeneous behavior groups and the key features. Given a data matrix $\mathbf{X}$ of size n-by-d with n users and d features, a subgroup of its rows and a subgroup of its columns can be simultaneously achieved by decomposing the matrix into a pair of left and right (singular) vectors that are both sparse. Let $\mathbf{u}$ of size n and $\mathbf{v}$ of size d be the left and right vectors, respectively, resulted from the decomposition. Their outer product forms a sparse rank-one approximation of the original matrix, i.e., $\mathbf{X} \approx \mathbf{u}\mathbf{v}^T$. Eq.(7) is solved to obtain $\mathbf{u}$ and $\mathbf{v}$.

$$\begin{aligned} \min_{\mathbf{u},\mathbf{v}} \quad & \|\mathbf{X} - \mathbf{u}\mathbf{v}^T\|_F^2 \\ \text{subject to} \quad & \|\mathbf{u}\|_0 \le s_u, \|\mathbf{v}\|_0 \le s_v, \end{aligned} \tag{7}$$

where the objective function measures the approximation error with a Frobenius norm of the matrix difference, and $\|\cdot\|_0$ is the ℓ_0 vector norm (although commonly called a norm, it is not really a norm) that returns the number of non-zeros in a vector. The rows in $\mathbf{X}$ corresponding to non-zero components in $\mathbf{u}$ form a row subgroup; and the columns in $\mathbf{X}$ corresponding to non-zero components in $\mathbf{v}$ form a column subgroup. The resultant row and column clusters help to define each other. Afterwards, subsequent clusters can be obtained by solving Eq.(7) with an updated matrix $\mathbf{X}$, specifically by excluding the rows in $\mathbf{X}$ that correspond to subjects that are in the clusters already identified (other approaches are also possible, see [26]).

When there are multiple views of input data (and hence multiple data matrices), for instance, three different views as in our study, the objective is then to find the same row clusters from all the views. We use a binary vector $\boldsymbol{\omega}$ to connect the different $\mathbf{u}$ vectors decomposed from the data matrices as shown in Eq.(8):

$$\begin{aligned} \min_{\boldsymbol{\omega},\mathbf{u}^k,\mathbf{v}^k,k=1,2,3} \quad & h(\boldsymbol{\omega}, \mathbf{u}^k, \mathbf{v}^k) = \sum_{k=1}^{3} \|\mathbf{X}^k - \text{diag}(\boldsymbol{\omega})\mathbf{u}^k{\mathbf{v}^k}^T\|_F^2 \\ \text{subject to} \quad & \|\boldsymbol{\omega}\|_0 \le s_\omega, \ \|\mathbf{v}^k\|_0 \le s_{v^k}, \\ & k = 1, 2, 3, \\ & \boldsymbol{\omega} \in \mathcal{B}_n. \end{aligned} \tag{8}$$

where $\text{diag}(\boldsymbol{\omega})$ is a diagonal matrix with diagonal entries equal to $\boldsymbol{\omega}$, s_ω and s_{v^k}'s are hyper-parameters that are predetermined to enforce sparsity of $\boldsymbol{\omega}$ and $\mathbf{v}^k$'s, and $\mathcal{B}_n$ is the set that contains all binary vectors of length n. When $\omega_i = 0$, regardless the value of the i-th components of $\mathbf{u}^k$, the i-th row will be excluded from the subgroup in all views. Hence, row clusters that are the same across the different views can be identified directly by finding rows that correspond to non-zero entries in the optimal $\boldsymbol{\omega}$. In other words, in multi-view bi-clustering, we use $\boldsymbol{\omega}$ instead of $\mathbf{u}^k$'s to enforce consistent row clusters (and hence consistent user subgroups) across the multiple views. The non-zero values in $\mathbf{v}^k$ specifies the key features that are selected from view k. Again, after one cluster is identified, subsequent clusters can be identified by updating $\mathbf{X}^k$'s.

D. Assessing Group Separability

After the clusters have been identified, we label a subject with respect to the cluster that he/she belongs to. Specifically, we identify three clusters, and correspondingly label subjects into three subgroups. To assess the separability of the subgroups (i.e., user clusters), we construct SVM classifiers to

separate the subjects into clusters using the key features that have been selected (recall multi-view bi-clustering identifies clusters and simultaneously key features). SVM is a supervised learning algorithm that transforms training examples to a higher dimensional space. In our work, we use linear SVM with one-vs-all approach. The reason for using linear SVM is that our dataset is small with relatively more features. Our SVM classifier evaluates the separability of the clusters. In addition, as we shall see, it verifies that indeed the selected features can represent the clusters reasonably well.

IV. Evaluation

In this section, we first present the clustering results obtained from the multi-view clustering algorithm. Once the clusters are found, we then use SVM for classification. We further compare the classification results when using all the features and the key features that are selected by the multi-view clustering algorithm. At the end, we present the overall generalizability of our approach using 10-fold cross validation.

A. Multi-view Clustering

We use multi-view clustering [26] to cluster the users into distinct groups. Specifically, we use 49 participants from the dataset; the rest of the participants are ignored because of missing values. In the following, we first describe parameter tuning for the multi-view clustering algorithm and then present the clustering results.

1) Parameter Tuning: Our multi-view clustering algorithm uses four hyper-parameters, i.e., s_ω that controls the size of the cluster, and s_{v^1}, s_{v^2} and s_{v^3} that determine the number of key features for clustering in the three views, respectively.

We first tune s_{v^1}, s_{v^2}, and s_{v^3} since our recent study [26] indicates that these three parameters are more sensitive than s_ω, and hence they need to be more carefully selected. We use Principal Component Analysis (PCA), a commonly-used tool for dimension reduction, to find the optimal values for these three parameters. Specifically, through dimension reduction, PCA can find a subset of features, i.e., principal components that can represent the feature space. To find the principal components, we first calculate the covariance matrix $\mathbf{C}$ for a view's data matrix. Using PCA, we then compute all the principal components of $\mathbf{C}$. We then pick the first m principal components that represent at least 90% of the covariance matrix $\mathbf{C}$. Through simulation, we found the optimal values for s_{v^1}, s_{v^2} and s_{v^3} to be 5.

To tune s_ω, we start with a range of possible values, i.e., $5-20$, and search the range to find the optimal value. For each possible value, we apply multi-view clustering to the dataset. Once the clusters are identified, we label the data using the cluster index. After that, we run SVM for classification over the labeled data. In this paper, we use 5-fold cross validation for performance evaluation. Fig. 4 summarizes our simulation results to find the optimal s_ω for cluster 1. We observe that the prediction error is the lowest when $s_\omega = 9$, and hence we choose $s_\omega = 9$ for cluster 1. Using the same approach, we choose $s_\omega = 7$ for cluster 2.

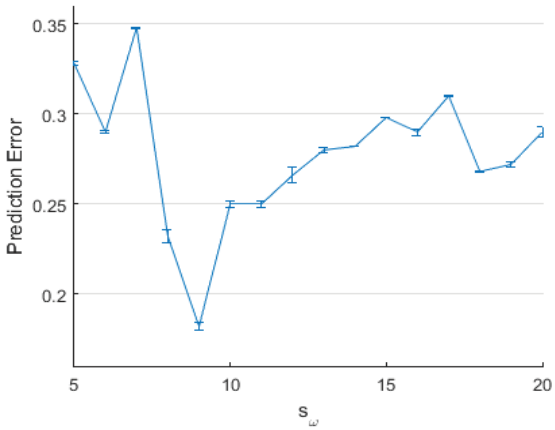


Fig. 4. Cross validation for searching the proper s_ω.

2) Clustering Results: After parameter selection, we apply multi-view clustering to find clusters in the data. The algorithm identifies two clusters with 9 and 7 participants respectively; the remaining 33 participants are in the third cluster. Recall that the clustering algorithm only takes the sensing features, and does not take PHQ-9 scores as input. An interesting question is whether the resultant clusters are discriminative of the PHQ-9 scores, and in addition, what key features are identified to be useful in each of the three views. To answer the above two questions, we categorize the clusters and the related PHQ-9 scores and feature information as follows. For each view, we draw a bar plot that presents the relative mean values of the PHQ-9 scores (the PHQ-9 score for a participant is the average of the pre- and post- PHQ-9 scores) and the key features for each of the three clusters, where the relative mean value is calculated as

$$\frac{Mean(Sample\_in\_Cluster) - Mean(Entire\_sample)}{STD(Entire\_sample)}$$

where STD represents to the standard deviation of a feature over the entire sample.

The results for the three views are summarized in Figures 5 to 7. In each figure, the first bar for a cluster represents the relative mean value of the PHQ-9 scores for this particular cluster. An interesting observation is that the participants in cluster 1 tend to have low PHQ-9 scores, participants in cluster 2 tend to have high PHQ-9 scores, and the remaining participants, i.e., those in cluster 3, have medium PHQ-9 scores. The above observation indicates that the subgroups identified by the multi-view clustering algorithm are indeed discriminative of the PHQ-9 scores.

We next describe the key features identified by the multi-view clustering algorithm for each of the three views. Fig. 5 plots the relative mean values of all the features in the first view (i.e., the average view). The results for all the three clusters are shown in the figure. The key features that are identified by the multi-view clustering algorithm are $Conv_d$, $Dark_d$, $Dark_c$, $Audio_v$, $Audio_q$ and $PhoneLock_d$. Psycholog-

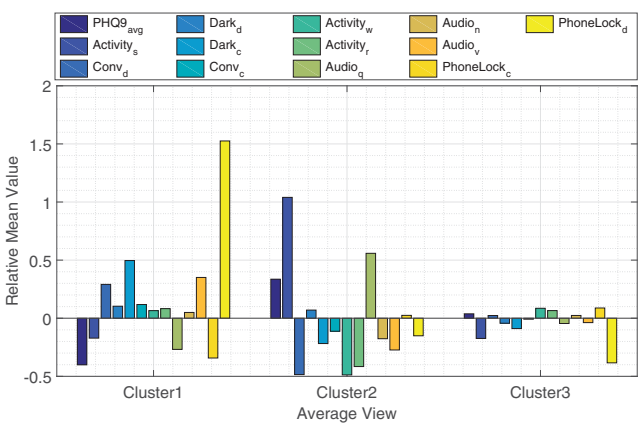


Fig. 5. The characteristics of the three clusters in the average view. The bars represent the relative mean values of the PHQ-9 scores and various features.

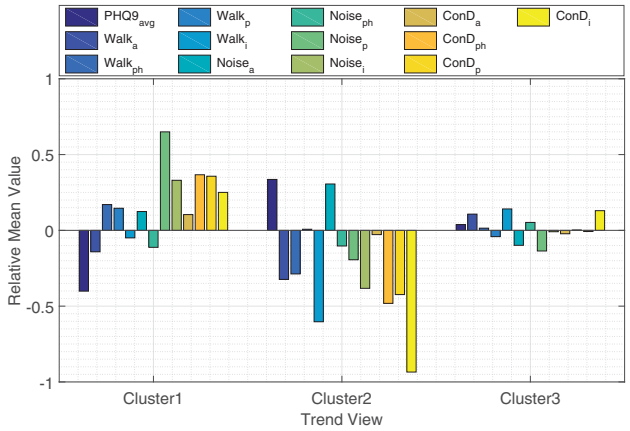


Fig. 6. The characteristics of the three clusters in the trend view. The bars represent the relative mean values of the PHQ-9 scores and various features.

ical studies such as [2] found that conversation and sleep duration can indicate depressive mood disorder. Our results on finding that one feature related to conversation ($Conv_d$, i.e., average duration of conversation in a day) and two features related to darkness ($Dark_d$ and $Dark_c$, i.e., average duration and number of times in a dark environment in a day) as two key features are consistent with these results. Specifically, cluster 1 subgroup (which turns out to contain participants with lower PHQ-9 scores) tends to have longer duration of conversations than cluster 2 subgroup (which turns out to contain participants with higher PHQ-9 scores). Similarly, if we infer sleep from $Dark_c$, then cluster 1 subgroup shows a normal sleep pattern compared to cluster 2 subgroup. A similar pattern is observed by [28]. The two key features related to audio, $Audio_v$ and $Audio_q$ (representing the average duration when audio is classified as voice and quiet, respectively) are also informative behavioral features, as cluster 2 subgroup (i.e., the high PHQ-9 subgroup) spent more time in quiet environment. Last, our results on identifying $PhoneLock_d$ (i.e., the average duration of phone being locked in a day) as another key feature are consistent with the observations in [23]: the participants in cluster 1 subgroup (the low PHQ-9 subgroup) tend to use their phones less compared to participants in cluster 2 subgroup (i.e., the high PHQ-9 score subgroup).

The key features identified in the trend view (see Fig. 6) includes $Walk_i$, $Walk_p$, $Noise_i$, $ConD_a$, $ConD_p$ and $ConD_i$. A larger $Walk_p$ indicates that the corresponding curve's frequency is low, representing a more stable pattern. From Fig. 6, we observe that $Walk_p$ for cluster 1 is higher than that for cluster 2, indicating a more stable walking routine. A similar trend is observed in $ConD_p$, i.e., the low PHQ-9 subgroup (cluster 1) has a relatively more stable conversation routine. The two features, $ConD_i$ and $Walk_i$ (corresponding to respectively the intercept of conversation and walking duration) are also interesting since intercept in wavelet transformation represents the approximate mean value. Therefore, these two intercept features being identified as key features confirms our

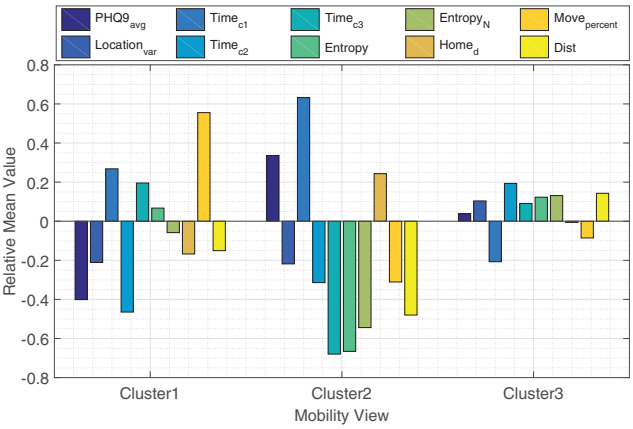
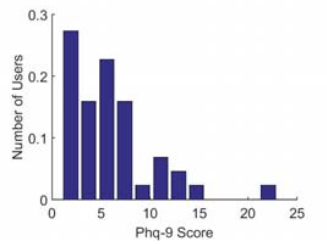


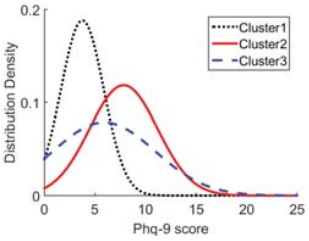
Fig. 7. The characteristics of the three clusters in the location view. The bars represent the relative mean values of the PHQ-9 scores and various features.

results from the average view that the low PHQ-9 subgroup tends to spend more time in walking and conversation.

The key features identified in the location view (Fig. 7) are $\text{Location}_{\text{Var}}$, $Time_{c_1}$, Entropy, $\text{Entropy}_{\text{N}}$, $\text{Move}_{\text{percent}}$ and Dist. We observe that cluster 1 subgroup (i.e., the low PHQ-9 subgroup) has lower $Time_{c_1}$ than cluster 2 subgroup (i.e., the high PHQ-9 subgroup), indicating that low PHQ-9 subgroup tends to spend less time at a single location. Similarly, $\text{Move}_{\text{percent}}$ is high for this subgroup, meaning that participants with low PHQ-9 scores are comparatively more active than high PHQ-9 participants. These results are intuitive as they suggest that participants with low PHQ-9 scores are in general more active than participants with high PHQ-9 scores. Entropy and normalized entropy are used to measure the uniformity in mobility patterns. From the bar plot (Fig. 7), we find that low PHQ-9 score participants have lower entropy, indicating that they have a more uniform mobility pattern, while high PHQ-9 score participants spend time less

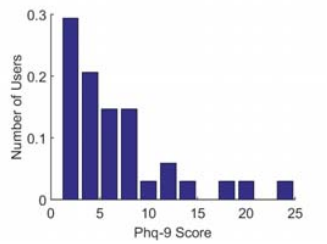


(a) Histogram of pre-PHQ-9 scores

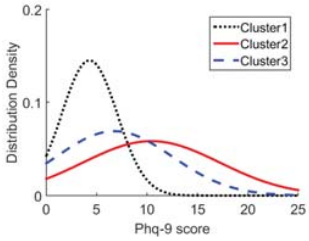


(b) Gaussian fitting of each cluster

Fig. 8. Pre-PHQ-9 score distribution and Gaussian approximation for each of the three clusters.



(a) Histogram of Post-PHQ-9



(b) Gaussian Fitting of Each Cluster

Fig. 9. Post-PHQ-9 score distribution and Gaussian approximation for each of the clusters.

uniformly, i.e., they tend to spend time in a small number of locations.

We also estimate the PHQ-9 score distribution of a cluster assuming PHQ-9 scores follow a Gaussian distribution. The results are shown in Figures 8 and 9, where the top two figures show the histograms of the pre- and post- PHQ-9 scores, respectively, over the entire sample, and the bottom two figures show the density curves of the scores for individual clusters. The modes of clusters 1 and 2 are clearly separable. Cluster 3's density curve spans across those of the other two clusters, which is expected as it contains the remaining subjects (not identified by the first two clusters). The above results further demonstrate that the subgroups identified by the multi-view clustering approach are discriminative of PHQ-9 scores.

B. Group Separability Results

After clustering, we label the subjects with their corresponding cluster indices determined by the multi-view clustering approach, and use SVM for classification to validate the separability of the clusters. Furthermore, we compare the classification accuracy when using the features selected from our multi-view clustering algorithm [26] with that when using all the features that are extracted. Specifically, now the dataset has class labels as cluster 1, cluster 2 and cluster 3. Since the dataset has only a small number of instances, before training the classifier, we first up-sample by repeating the samples for clusters 1 and 2, i.e., duplicate each row three more times for cluster 1 and four times for cluster 2. As a result, cluster 1 now has 36 rows, and cluster 2 has 35 rows. We then randomly choose 70% of the instances as the training set and the remaining 30% as the testing set. Since we are dealing with multi-class classification problem, we use one-vs-all scheme for training the SVM. One-vs-all scheme builds K different classifiers, where K represents the total number of classes. For each classifier i, the positive example will be the one that belongs to class i, while all others are considered as negative instances. Let C_i be the ith classifier's confidence score. The classifier output $C(x)$, for unseen instance x, reflects the highest confidence score and can be represented as

$$C(x) = \arg\max_i(C_i(x))$$

We now present the results of SVM classification when using features selected by multi-view clustering. To check the generalizability of the SVM model, 30% of the data is used as the test data. Table I lists the confusion matrix, where cluster 1, cluster 2 and cluster 3 represent the three classes respectively. The diagonal of the matrix represent the correct classification results. The overall accuracy of the classifier over the three classes is approximately 87.1%. We observe that SVM misclassifies instances of class 3, while performs well for the other two classes. The relatively high accuracy indicates that the clusters are indeed separable using the key features.

Actual Label \ Predicted Label	Cluster 1	Cluster 2	Cluster 3
1	11	0	0
2	0	11	0
3	3	1	5

TABLE I
CONFUSION MATRIX FOR THE SVM MODEL, USING THE KEY FEATURES IDENTIFIED FROM THE MULTI-VIEW CLUSTERING ALGORITHM.

We also compare the performance of the above SVM model with the model when using the complete feature set. Table II lists the confusion matrix when using the complete feature

set. Again, the diagonal of the matrix represents the correct classification results. The overall accuracy of the classifier over the three classes is approximately 80.6%. The comparable accuracy when using the selected features and the complete feature set (i.e., 87.1% versus 80.6%) indicate that the selected features are sufficient (they capture sufficient information of all the features).

Actual Label \ Predicted Label	Cluster 1	Cluster 2	Cluster 3
1	10	0	0
2	0	11	0
3	3	3	4

TABLE II
CONFUSION MATRIX FOR THE SVM MODEL, USING ALL THE FEATURES.

Using confusion matrices I and II, we measure recall (sensitivity) and precision. To calculate recall we use Recall $=$ TP/AP. Here TP stands for true positives, i.e., the correct true class classification, and AP represents actual positives. When using the selected features, recalls for the three classes are 1.0, 1.0 and 0.56, respectively, while when using all the features, recalls are 1.0, 1.0 and 0.4, respectively. To calculate precision, we use Precision $=$ TP/PP, PP stands for predicted positives. When using the selected features, precisions for the three classes are 0.79, 0.92 and 1.0 respectively, when when using all the features, precisions are 0.77, 0.79 and 1.0, respectively. The comparable recall and precision when using the selected features and all the features again confirm that the selected features are sufficient.

C. Cross Validation

To further validate the accuracy of the classification, we use 10-fold cross validation. Specifically, we divide the dataset randomly into 10 equal-size disjoint subsets, and construct ten training and testing sets, each training set uses 90% of the data and the corresponding testing set uses the remaining 10% of the data. SVM classifier is trained 10 times, and each time a different set is held out as a test set. Using 10-fold cross validation, the overall accuracy is 89.4% (see Table III).

Actual Label \ Predicted Label	Cluster 1	Cluster 2	Cluster 3
1	36	0	0
2	0	35	0
3	7	4	22

TABLE III
CONFUSION MATRIX FOR SVM, 10-FOLD CROSS VALIDATION.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel approach to identify homogeneous behavioral groups using smartphone sensing data. Our approach, centered on multi-view bi-clustering, identifies three clusters that are discriminative of PHQ-9 scores (they contain participants with low, high and medium PHQ-9 scores, respectively). In addition, it simultaneously finds the key sensing features that characterize the different groups. We further used SVM for classifying the clusters. The overall prediction results when using the key features are promising, with an overall accuracy of 87.1%.

As future work, our plan is to analyze data extracted from other smartphone sensors. In addition, we are in the process of collecting data from a larger population, and will analyze the dataset to understand further the relationship of human behavior reflected by smartphone sensing data and mental health.

ACKNOWLEDGEMENTS

This work was partially supported by National Science Foundation grants IIS-1407205 and IIS-1320586. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We would also like to thank the anonymous reviewers for their insightful comments.

REFERENCES

[1] Centers for disease control and prevention. national center for injury prevention and control. 2010.

[2] P. Aselton. Sources of stress and coping in american college students who have been diagnosed with depression. *Journal of Child and Adolescent Psychiatric Nursing*, 25(3):119–123, 2012.

[3] G. Bauer and P. Lukowicz. Can smartphones detect stress-related changes in the behaviour of individuals? In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 423–426. IEEE, 2012.

[4] A. Beck, A. L. Crain, L. I. Solberg, J. Unützer, R. E. Glasgow, M. V. Maciosek, and R. Whitebird. Severity of depression and magnitude of productivity loss. *The Annals of Family Medicine*, 9(4):305–311, 2011.

[5] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland. Pervasive stress recognition for sustainable living. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 345–350. IEEE, 2014.

[6] M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr. Harnessing context sensing to develop a mobile intervention for depression. *J Med Internet Res.*, 13(3), Jul-Sep 2011.

[7] L. Canzian and M. Musolesi. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1293–1304, New York, NY, USA, 2015. ACM.

[8] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell. Unobtrusive sleep monitoring using smartphones. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 145–152. IEEE, 2013.

[9] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *Proceedings of ACM Conference on Ubiquitous Computing*, pages 481–490. ACM, 2012.

[10] Y. Chon, E. Talipov, H. Shin, and H. Cha. Mobility prediction-based smartphone energy optimization for everyday location monitoring. In *Proceedings of ACM conference on Embedded Networked Sensor Systems*, pages 82–95. ACM, 2011.

[11] T. M. T. Do and D. Gatica-Perez. Groupus: Smartphone proximity data and human interaction type mining. In *Annual International Symposium on Wearable Computers (ISWC)*, pages 21–28, June 2011.

[12] Y.-S. Lee and S.-B. Cho. Activity recognition using hierarchical hidden markov models on a smartphone with 3d accelerometer. In *Hybrid Artificial Intelligent Systems*, pages 460–467. Springer, 2011.

[13] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. Can your smartphone infer your mood. In *PhoneSense workshop*, pages 1–5, 2011.

[14] M. Lin, N. D. Lane, M. Mohammod, X. Yang, H. Lu, G. Cardone, S. Ali, A. Doryab, E. Berke, A. T. Campbell, et al. Bewell+: multi-dimensional wellbeing monitoring with community-guided user feedback and energy optimization. In *Proceedings of the Conference on Wireless Health*. ACM, 2012.

[15] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 351–360. ACM, 2012.

[16] A. Madan, M. Cebrian, D. Lazer, and A. Pentland. Social sensing for epidemiological behavior change. In *Proceedings of ACM international conference on Ubiquitous computing*, pages 291–300. ACM, 2010.

[17] A. Madan, S. T. Moturu, D. Lazer, and A. S. Pentland. Social sensing: obesity, unhealthy eating and exercise in face-to-face networks. In *Wireless Health*, pages 104–110. ACM, 2010.

[18] J.-K. Min, J. Wiese, J. I. Hong, and J. Zimmerman. Mining smartphone data to classify life-facets of social relationships. In *Proceedings of Conference on Computer Supported Cooperative Work*, pages 285–294. ACM, 2013.

[19] S. T. Moturu, I. Khayal, N. Aharony, W. Pan, and A. Pentland. Using social sensing to understand the links between sleep, mood, and sociability. In *IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 208–214. IEEE, 2011.

[20] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[21] K. Polat and S. Güneş. Classification of epileptiform eeg using a hybrid system based on decision tree classifier and fast fourier transform. *Applied Mathematics and Computation*, 187(2):1017–1026, 2007.

[22] M. Rabbi, S. Ali, T. Choudhury, and E. Berke. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 385–394. ACM, 2011.

[23] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, 17(7), 2015.

[24] A. Sano and R. W. Picard. Stress recognition using wearable sensors and mobile phones. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 671–676. IEEE, 2013.

[25] B. Shumaker and R. Sinnott. Astronomical computing: 1. computing under the open sky. 2. virtues of the haversine. *Sky and Telescope*, 68:158–159, 1984.

[26] J. Sun, J. Lu, T. Xu, and J. Bi. Multi-view sparse co-clustering via proximal alternating linearized minimization. In *Proceedings of International Conference on Machine Learning*, pages 757–766, 2015.

[27] T. Vos, A. D. Flaxman, M. Naghavi, R. Lozano, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, V. Aboyans, et al. Years lived with disability (ylds) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859):2163–2196, 2013.

[28] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14. ACM, 2014.