Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning

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

Depression is a common mood disorder that causes severe medical problems and interferes negatively with daily life. Identifying human behavior patterns that are predictive or indicative of depressive disorder is important. Clinical diagnosis of depression relies on costly clinician assessment using survey instruments which may not objectively reflect the fluctuation of daily behavior. Self-administered surveys, such as the Quick Inventory of Depressive Symptomatology (QIDS) commonly used to monitor depression, may show disparities from clinical decision. Smartphones provide easy access to many behavioral parameters, and Fitbit wrist bands are becoming another important tool to assess variables such as heart rates and sleep efficiency that are complementary to smartphone sensors. However, data used to identify depression indicators have been limited to a single platform either iPhone, or Android, or Fitbit alone due to the variation in their methods of data collection. The present work represents a large-scale effort to collect and integrate data from mobile phones, wearable devices, and self reports in depression analysis by designing a new machine learning approach. This approach constructs sparse mappings from sensing variables collected by various tools to two separate targets: self-reported QIDS scores and clinical assessment of depression severity. We propose a so-called heterogeneous multi-task feature learning method that jointly builds inference models for related tasks but of different types including classification and regression tasks. The proposed method was evaluated using data collected from 103 college students and could predict the QIDS score with an R^2 reaching 0.44 and depression severity with an F1-score as high as 0.77. By imposing appropriate regularizers, our approach identified strong depression indicators such as time staying at home and total time asleep.

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

Additional Key Words and Phrases: Depression assessment, Multi-task learning, Prediction, Sensor data analysis

*Correspondence should be addressed to Jinbo Bi (jinbo.bi@uconn.edu).

This work was partially supported by the National Science Foundation (NSF) grant IIS-1407205. Jin Lu was also support by the NSF grants IIS-1320586 and DBI-1356655. Jinbo Bi was supported by the National Institutes of Health (NIH) grants R01DA037349 and K02DA043063, and NSF grants CCF-1514357 and IIS-1447711. Athanasios Bamis performed this work before joining Amazon.
Authors' addresses: Jin Lu; Chao Shang; Chaoqun Yue; Reynaldo Morillo; Shweta Ware, University of Connecticut, Department of Computer Science and Engineering, Storrs, CT, 06269, USA, firstname.lastname@uconn.edu; Jayesh Kamath, University of Connecticut Health Center, Department of Psychiatry, Farmington, CT, 06030, USA, jkamath@uchc.edu; Athanasios Bamis, Amazon, Inc. 101 Main St. Cambridge, MA, 02142, USA, athanasios.bamis@gmail.com; Alexander Russell; Bing Wang; Jinbo Bi, University of Connecticut, Department of Computer Science and Engineering, Storrs, CT, 06269, USA, alexander.russell@uconn.edu, bing@uconn.edu, jinbo.bi@uconn.edu.

2474-9567/2018/3-ART21 $15.00
https://doi.org/10.1145/3191753

ACM Reference Format:
Jin Lu, Chao Shang, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jayesh Kamath, Athanasios Bamis, Alexander Russell, Bing Wang, and Jinbo Bi. 2018. Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 21 (March 2018), 21 pages. https://doi.org/10.1145/3191753

1 INTRODUCTION

Depression is estimated to affect 350 million people worldwide; it is ranked 2nd among all the major illnesses in Years Lived with Disabilities (YLDs) and accounts for 9.6% of all YLDs from all major illnesses [42]. It is also a significant contributor to death by suicide [42]: 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].

Currently, diagnosis of depression is based on physician-administered survey instruments that require significant effort and cost, and rely on accurate introspection and reporting. The ubiquitous adoption of smartphones around the world creates new opportunities for depression screening. Fitbit wrist bands provide another tool for assessing behavioral patterns. The small physical form of smartphones and wrist-worn devices allow them to be constantly carried by their owners, making them effective "human sensors" appropriate for cataloging and analyzing broad spectrum of their users' behavior.

Several recent studies [6, 13, 14, 17, 18, 38, 44] have explored using smartphone sensing data for depression screening and have identified several sensor-based features as depression indicators. The present paper represents our ongoing effort in building an automatic depression diagnosis system. This system collects sensing data from smartphones and Fitbit wrist bands, extracts features from sensing data, and predicts depression via a machine learning model of the extracted features. While smartphone data can capture behavioral variables such as location changes and communication frequencies, the wearable sensors can monitor complementary variables such as heart rates and sleep quality. The premise of our study lies in the prior evidence showing that sensing variables play roles in understanding and diagnosing depression.

However, because of the different operating systems and the specific sensors used by the different sensing devices, the methods that these devices collect data vary substantially. Consequently, the behavioral parameters derived from the different sources of sensing data exhibit signficant differences. For instance, the variance in location changes calculated separately from Android and iPhone for the same individual shows clearly distinct values. There is thus great heterogeneity in the features derived from the two predominant smartphone platforms – Android and iPhone, which renders that existing sensor-based depression studies to date only utilize a single platform.

Given that our study collects data from both Android and iPhone users, a separate analysis of the data will lead to reduced sample size, diminishing power of any analysis. Using multi-task learning methods, we can jointly model the data collected from the two platforms as separate but related tasks to improve depression prediction accuracy. It also provides a way to integrate information from the two platforms by knowledge transfer during the joint model training process. Multi-task learning is a machine learning methodology that captures and exploits the relationship among multiple related tasks [19, 20, 25, 48–50, 53]. Typical multi-task learning methods assume a certain relatedness among tasks, but the definition of relatedness varies in different methods. From a Bayesian viewpoint, multi-task learning essentially seeks to learn a good prior over all tasks to capture task dependencies. It has been shown, both empirically and theoretically, that multi-task learning is more effective than learning individual tasks independently.

The proposed multi-task learning (MTL) method assumes task relatedness in feature sharing. Unlike existing MTL methods that handle homogeneous related inference tasks (e.g., all tasks are classification problems), our method extends beyond them to model both regression and classification tasks in a joint optimization framework. Specifically, we consider two distinct prediction problems of: (1) predicting the numeric score computed from the

Quick Inventory of Depressive Symptomatology (QIDS) [37] survey, that self evaluates a participant's depression status, and (2) predicting the depression severity level, a categorical value, assessed by a clinician. We show that using multi-task learning for the two tasks separately (multiple homogeneous tasks) can already gain prediction accuracy in comparison with single-task learning (STL). We further explore merging all tasks together in a heterogeneous MTL framework, including two regression tasks for predicting QIDS scores of respective Android and iPhone users and two classification tasks for predicting depression severity. This framework also allows us to combine self-reports and clinical decision on depression assessment. It thus creates an opportunity to compare and transfer knowledge between clinician's assessment and patient's self evaluation, potentially leading to more accuract prediction models. Our work makes the following two main contributions:

- We design a new MTL method that can not only model jointly the data collected from different smartphone platforms but also integrate different methods for depression assessment. To the best of our knowledge, this formulation represents the first effort to integrate different smartphone platforms in depression analysis. This MTL method allows us to transfer knowledge between clinician's assessment and individual's self recognition. Our empirical evaluation demonstrated the benefits of the proposed method. For predicting QIDS scores, the MTL approach improved the regression accuracy by 34.3% over single-task learning. When predicting depression severity for depressed participants, the classification performance was improved by 48.1%. This MTL approach, as validated in the present work, will provide a more powerful alternative to existing analytics to benefit future analyses of many other human behaviors.
- Equipped with this newly developed MTL method, in this study, we are now able to use much more diverse data sources than existing works to identify depression indicators. For the first time, Android and iPhone features are merged in a joint analysis, and complementary Fitbit features are used in conjunction with smartphone variables. In addition, unlike many other works, including our own prior works [12, 13, 52], that use the 9-item Patient Health Questionniare (PHQ-9) [23], we use a more comprehensive questionniare, a 16-item QIDS-SR_{16} survey. Furthermore, beyond self-reports, we examine the depression prediction using clinical ground truth. Although our prior works [13, 52] already used clinical diagnosis of whether an individual is depressed, we now get to use the fine grade of four-level severity assessment.

The rest of the paper is organized as follows. Section 2 briefly reviews related work. Section 3 describes our procedure of data collection and feature extraction. Section 4 motivates the need for multi-task learning. Section 5 introduces the proposed multi-task learning method. Empirical evaluation results are included in Section 6. We then discuss the results and future work, and conclude the paper in Section 7.

2 RELATED WORK

The sensing data collected from smartphones can naturally reflect user behavior, which has led to a variety of innovative applications that detect interesting patterns in sensor data [7–10, 26, 36], and infer human behavioral characteristics [9, 27, 32, 44]. There is a rich literature on analyzing smartphone sensing data for smart health applications, focusing on physical, behavioral, or mental health [5, 24, 30–32, 34, 39]. For instance, BeWell [24] is a personal health monitoring app that analyzes physical activities, sleep, and social interactions in order to provide feedback on user lifestyle. An approach was developed in [5] to automatically recognize stress from smartphone's social interaction data, weather data, and self-reported personality information. The study in [31] examined the effect of illness and stress on human behavior by analyzing the communication and co-location data collected from smartphones, and demonstrated the change in behavior with the onset of a disease.

A number of studies that are most relevant to ours are those using smartphone sensor data to predict depressive mood or disorder [3, 6, 12–14, 17, 18, 33, 35, 38, 41, 43, 44, 52, 54]. Existing research has largely relied on self-reported surveys (e.g., PHQ-9 responses) in order to train and evaluate predictive models. The study in [44] reported a significant correlation between depressive mood and social interaction (specifically, the conversation

duration and the number of co-locations). Another study [38] found that phone usage and mobility patterns were strongly correlated with self-reported QIDS scores. The study in [6] further explored the relationship between depression and mobility patterns where the authors trained both general and individual-featured support vector machine (SVM) models, and found that individual models outperformed general models. The study in [33] demonstrated the association of depressive states with the smartphone interaction features (including phone usage patterns and overall application usage logs). A recent study developed a deep learning based approach that forecasts (instead of detects) severely depressive mood based on self-reported histories [41]. Our prior work showed that behavioral data collected from smartphones could predict the clinical diagnosis of depression with good accuracy after separately examining iPhone and Android data [13].

In particular, a recent study used multi-task and multi-kernel learning (MTMKL) to assess individual wellbeing [22]. Specifically, each of five wellbeing components (happiness, health, alertness, energy, and stress) was used to form a task in the MTL, and all tasks were homogeneously classification problems. Multiple modalities of data from a smartphone platform, surveys and a physiology sensor were used in a kernel combination. This MTMKL method was shown to outperform support vector machines and multi-kernel learning. However, because this method used all features from a modality to form a kernel, it was not designed to select important features, thus could not identify sensor-based indicators.

The present study differs from all existing studies in multiple ways as discussed in the Introduction section. Especially, the multi-task learning approach provides a new alternative for analyzing sensor data when multiple sensing platforms are employed. The proposed MTL method provides a general framework of heterogeneous multi-task feature learning by extending one of our recent MTL approaches [45, 46]. This recent approach decomposed each task's model parameters into a multiplication of two components and applied different regularizers to the components to select features important for all tasks as well as features for individual tasks. Although multi-task learning has been used in a wide range of applications (e.g., [4, 47, 51]), our study is the first that develops and uses an adapted heterogeneous MTL to model the two major smartphone platforms in a joint manner.

3 DATA COLLECTION AND FEATURE EXTRACTION

In this section, we describe the data used in our analysis and how the data were collected and processed.

3.1 Data Collection

Four types of data were collected from February 13, 2017 to May 21, 2017, including smartphone sensing data, Fitbit data, QIDS questionnaire [37] responses (via a smartphone app), and clinician assessment. To preserve privacy of the study participants, we anonymized the data by assigning a random user ID to each participant. The data collected by our apps were encrypted before being stored on the phone, and then sent to a secure server. To maximize the continuity of the data collection, if no sensing data were received from a participant for three consecutive days, the system sent an email to the participant to check the status. Similarly, when a QIDS questionnaire was not received from a participant three days after the due date, we sent an email reminder to the participant.

Participants. A total of 103 college students were recruited for the study. The participants were aged 18 - 25 and enrolled as full time students at the University of Connecticut (UConn). Of these students, 34 were Android users (including 12 depressed and 22 non-depressed), and 69 were iPhone users (27 depressed versus 42 non-depressed). Of all the participants, 76.7% were female and 23.3% were male. In terms of ethnicity, 58.3% were white, 25.2% were Asian, 3.9% were African American, 7.8% had more than one race, and 4.9% were other or unknown.

Smartphone sensing data were collected through an app called *LifeRhythm* that we developed recently. It runs in the background on a participant's phone, collecting location and physical activity data. Due to the

differences between the smartphone platforms, we had to develop separate apps for Android phones (including a variety of manufacturers) and iPhones. In this paper, we only considered location information from the phones. The activity data from the phones were not used. Instead, we used the activity data collected by Fitbit[1], which has been shown to be highly accurate [2]. Each location sample contained the longitude and latitude coordinates of a participant at a particular time point. On Androids, this information was collected periodically every 10 minutes. On iPhones, no APIs could be used to schedule periodical data collection. Hence, our app subscribed to the location services provided by the iOS operating system and used an event trigger mechanism to collect location data. The collected data were pre-processed to remove samples with large errors as reported by iOS and handle missing data. Specifically, the sample records with error larger than 165 meters were removed.

Fitbit data. We used the Fitbit Charge heart rate (HR) for this study. A Fitbit device was given to each participant during the study period. Three main types of data (related to activity, sleep and heart rate, respectively) were collected through Fitbit. The Fitbit data from a user were transmitted to and stored at the Fitbit server. We then collected the data from the Fitbit server. Specifically, we set up a dedicated server that collected notifications from the Fitbit server using the Fitbit Subscription API. When the data from a user changed, the subscription server would be notified about the change. All notifications were stored in a database. We then processed the notifications (once per day, at midnight) to obtain the latest data (corresponding to the notifications) from the Fitbit server. Through this subscription mechanism, we might receive multiple notifications about a particular type of data for a user in a day; these notifications were processed in a batch, at the end of a day, to obtain the daily summary for the user. The subscription service allowed us to obtain a user's latest data without having to implement a polling or scheduling system to retrieve the user's data. It required users' authentications, which we obtained at the time of enrollment. The Fitbit data, once retrieved from the Fitbit server, were stored anonymously on our data collection server (each data record was associated with a random ID assigned by our study instead of the actual Fitbit ID for a user).

QIDS scores. The QIDS survey can assist clinicians in diagnosing and monitoring depression. Each of the questions evaluates a person's mental health from a specific aspect of major depressive disorder. Participants in our study first responded to the QIDS questionnaire during the initial assessment, and then continued to respond on her or his phone every 7 days (one week) through a smartphone app that we developed. The QIDS is an extended version of the PHQ-9 questionnaire, and is widely used in clinical practice. In the QIDS, 16 questions are asked, concerning the behavior of a subject (including activity levels, sleeping duration or interests in activities), and the cognitive state (including feelings about self) in the past seven days. Similar to the PHQ-9, the 16-item QIDS only takes a few minutes to fill in. The QIDS provides more detailed information than the PHQ-9. For instance, it has four questions regarding sleep instead of only one question in the PHQ-9. It differentiates between decreased appetite and increased appetite. The more thorough information in the QIDS enables refined labeling to our sample records.

For each participant, we normalized her/his QIDS scores y to correct for person-level variation. For instance, a participant who used a baseline score of 7 might give very different scores from an individual who used 0 as baseline, but they may be actually similar in their daily fluctuation. We computed the average of QIDS scores collected for each individual over the study period, $\hat{y} = \left(\sum_{n=1}^{N} y^n\right)/N$, where N was the total number of intervals. Then, the QIDS score y^n at a specific time point was augmented by an average for each individual as $y^n \leftarrow \frac{1}{2}y^n + \frac{1}{2}\hat{y}$, which revised the baseline for an individual using his/her long-term (stable) status. Mathematically, the revised QIDS score has the same expectation ($\mathbb{E}$) as the original one as shown in Eq.(1) assuming that each participant's QIDS scores are independently and identically distributed (i.i.d.) during the observation; the i.i.d. assumption is

[1]We expect activity data from Fitbit to be more reliable than that from a phone because Fitbit is a wearable device and tends to be with a user more consistently than a phone.

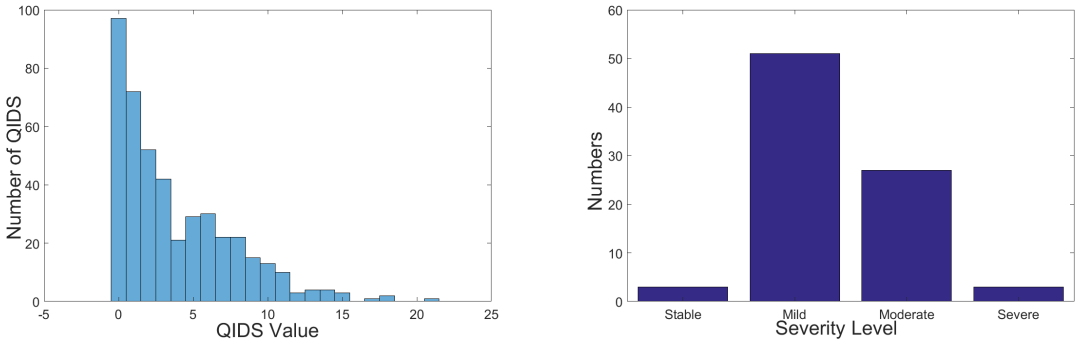


Fig. 1. QIDS histogram and Severity Level histogram

based on the rationale that depression is a recurrent and chronic disorder which can not change significantly during any treatment.

$$\mathbb{E}[y^n] = \frac{1}{2}\mathbb{E}_n[y^n] + \frac{1}{2}\mathbb{E}_n[y^n] = \mathbb{E}_n\left[\frac{1}{2}y^n\right] + \mathbb{E}_n\left[\frac{1}{2N}\sum_{m=1}^{N} y^m\right] = \mathbb{E}_n\left[\frac{1}{2}\left(y^n + \frac{1}{N}\sum_{m=1}^{N} y^m\right)\right] \quad (1)$$

The left subfigure of Fig. 1 shows the histogram of the QIDS scores that we collected. The values we observed ranged from 0 to 21. We could have multiple QIDS scores (ideally once a week) for each user.

Clinical assessment. Every participant was assessed by a clinician at the beginning of the study. Specifically, using an interview that was designed based on the Diagnostic and Statistical Manual of Mental Health (DSM-5) and the QIDS survey, the clinician classified individuals as either depressed or non-depressed during the initial screening. A participant with a diagnosis of depression must participate treatment to remain in the study.

In addition, depressed participants had follow-up meetings with the clinician periodically (once a month or less frequently, as determined by the clinician). Each meeting took 15-20 minutes and only involved interviews to assess psychiatric symptoms. The purpose of the interviews was to correlate and confirm their self-reported QIDS with their verbal report during the clinician interview. The clinician also assessed each depressed individuals their depression severity using four levels, “Stable(0)”, “Mild(1)”, “Moderate(2)” or “Severe(3)”. Note that severity levels were assessed by clinicians, whereas the QIDS scores were derived from self-reports. The right subfigure of Fig. 1 shows the histogram of all the severity levels that we collected.

3.2 Feature Extraction

3.2.1 Smartphone Features. We extracted eight features from the smartphone location data. Four features were directly from the raw location data and four features were based on a cluster analysis of locations, which were obtained by applying DBSCAN [11] to the data (we also experimented with the K-means method, and found that DBSCAN was more suitable for clustering our location data [13]). All the features were used and reported to correlate with depression previously [6, 13, 38]. Additional location-based features (e.g., circadian movement and transition time) have been used in the literature [6, 38]. They were not used by our study since most of our participants lived on campus; the locations of their major activities (dining, sleeping and studying) might be close and difficult to be differentiated by a GPS. The eight features that we used are described briefly below for completeness.

Location variance. The feature (*Loc_var*) measures the variability in a participant's location [38]. It is calculated as

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

where σ^2_{long} and σ^2_{lat} represent respectively the variance of the longitude and latitude of the GPS coordinates.

Time spent in transition. The feature, denoted as *Move*, represents the percentage of time that a participant is moving. We differentiate moving and stationary samples using the same approach as that in [38]. Specifically, the moving speed is estimated at a sensed location. If the speed is larger than 1km/h, then we classify it as moving; otherwise, we classify it as stationary.

Total distance. Given the longitude and latitude of two consecutive samples of location for a participant, we use Harversine formula [40] to calculate the distance traveled in kilometers between these two samples. The total distance traveled during a time period, denoted as *Distance*, is the total distance normalized by the length of the time period.

Average moving speed. In QIDS questionnaire, one question evaluates the mental health of a person based on whether (s)he is moving too slowly or too quickly. Inspired by this, we compute average moving speed, *AMS*, as another feature.

Number of unique locations. This feature, denoted by N_{loc}, is the number of unique clusters obtained by the DBSCAN algorithm when it is applied to an individual's location data.

Entropy. Entropy measures the variability in the durations that a participant spends at different locations. Let p_i denote the percentage of time that a participant spends in a location cluster i. The entropy is calculated as

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

Normalized entropy. Since the number of location clusters varies from person to person and entropy increases as the number of location clusters increases, we also adopt the normalized entropy [38], which is invariant to the number of clusters and depends solely on the distribution of the clusters of locations visited. It is calculated as

$$\text{Entropy}_N = \text{Entropy}/\log N_{\text{loc}} \tag{4}$$

where N_{loc} is the number of unique clusters for an individual.

Time spent at home. We use the approach described in [38] to identify "home" for a participant as the location that the participant is most frequently found between midnight to 6am. We calculate the percentage of time when a participant is at this location, denoted as *Home*.

3.2.2 Fitbit Features. We retrieved daily summaries about a user's activity, sleep and heart rate from the Fitbit server. From the summaries, we selected 28 features, including 12 features describing physical activity, 12 features for sleep, and 4 features for heart rate. Specifically, the following 7 Fitbit features were significantly correlated with depressive symptoms as shown in our experiments. We describe these features below (the names of the features are the ones used by Fitbit).

Lightly active minutes. This feature is related to activities. Fitbit classifies the extent of activeness into four categories, which are, in increasing order of activeness, sedentary, lightly active, fairly active, and very active. This feature represents the amount of time (in minutes) that a user is in a lightly active state during a day.

Sedentary minutes. This feature is related to activities. Analogous to the above feature, it represents the amount of time (in minutes) that a user is in a sedentary state during a day.

Total distance. This feature is also related to activities. It represents the total amount of distance traveled in miles.

Awake duration. This feature is related to sleep. Fitbit records sleep during a day in sessions. A user may have multiple sleep sessions during a day. One session is marked as the main session, which typically corresponds to the session with the longest sleep period. We only consider main sleep sessions in this study. This feature represent the duration (in minutes) that a user is awake during a sleep session.

Restless count. This is one of the sleep-related features. This feature records the number of times that a user is restless during a sleep session.

Minutes after wakeup. This feature measures the minutes after wakeup in a sleep session. It can correspond to the amount of time (in minutes) that it took a user to fall back asleep after waking up during a sleep session.

Minutes of heart rate in fat-burn zone. This feature is related to heart rate (i.e., heart beats per minute). Fitbit defines four zones/modes based on heart rate. In decreasing order of intensity, these four zones are peak, cardio, fat-burn, and out-of-range. This feature represents the amount of time (in minutes) that a user's heart rate is in fat-burn zone in a day.

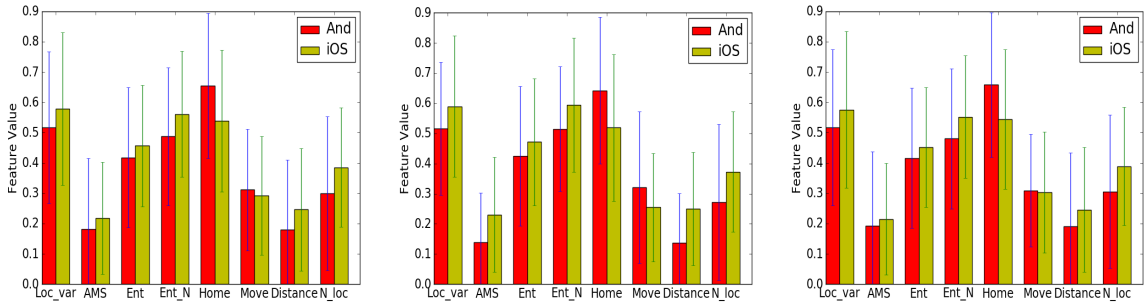
4 MOTIVATION FOR MULTI-TASK LEARNING

In this section, we analyze the collected data to motivate the need of multi-task learning. A QIDS response interval was 7 days in length, including the day when a participant responded to a QIDS questionnaire and the previous 6 days. We used a sample record that comprised the data collected in each QIDS interval. Each participant provided multiple QIDS responses during the study period. For data quality control, we applied a filter to remove the QIDS intervals that did not have sufficient amount of data. Specifically, a QIDS interval (i.e., a sample record) is used only when it had at least 3 days of smartphone sensing data, and the total amount of daily smartphone data covered more than 50% of the 7-day interval. Given the high reliability of Fitbit, Fitbit data were present in all QIDS intervals. Features were calculated for each QIDS interval. Our analysis used 145 sample records in total from Android users and 298 records from iPhone users. Of them, 29 and 65 records were, respectively, from clinically depressed participants.

Because we only used data passively collected from location-related smartphone sensors, the extent of usage on phone apps or calls should not affect our results. Given samples with data less than 50% of the interval were removed from subsequent analyses, the amount of data captured from different individuals was comparable within a range. Additionally, because we used the so-called leave-one-interval-out cross validation (described in a later section), at most one sample from each user could be selected for testing. Hence, the variation in the number of samples per user minimally affects our result.

4.1 Feature Values

Figure 2 plots the mean and the standard deviation of the various feature values obtained from the Android and iPhone platforms. Specifically, Figures 2(a), (b) and (c) plot respectively the results when using all the data, the data from the depressed users, and the data from the non-depressed users for each of these two platforms. In all three cases, data collected from the two smartphone platforms exhibit remarkably different distributions. The discrepancy between these two platforms might be partially due to the different data collection mechanisms (see Section 3). Specifically, location data were collected at a fixed sampling interval of every 10 minutes on Android phones; while they were collected using an event-triggering mechanism on iPhones (see [13, 52] for a detailed discussion). The discrepancy showed up also because different sensor types with distinct accuracy levels are adopted by the two platforms. The heterogeneity of the features across the two platforms was an important motivation for us to use MTL.



(a) All participants. (b) Depressed participants. (c) Non-depressed participants.

Fig. 2. Features comparison between the two datasets collected respectively from Android and iPhone.

4.2 Correlation between Features and QIDS Scores

We next present how individual behavioral features are correlated with QIDS scores. Tables 1 and 2 show Pearson's correlation coefficients between smartphone sensing features and QIDS scores for Android and iPhone participants, respectively. Both tables show the correlation results within three subject groups: all participants, depressed and non-depressed participants; the last column shows the difference of the correlation values between the second and third groups. One can explicitly observe that for both platforms, if only taking the depressed participants into account, their feature values were in general more correlated with their QIDS scores than those when counting in non-depressed participants, or when only considering non-depressed participants. For instance, for depressed Android users, the three features most correlated with QIDS scores were *entropy*, *time spent at home*, *the number of unique locations*, with correlation values of -0.65, 0.70, and -0.63 (and the corresponding small p-values indicated statistical significance) respectively; these same categorical features computed on all users or non-depressed users were less significantly correlated with QIDS scores.

The observation that the behavioral features played a more lucid role within the group of the depressed participants motivated us to explore feature selection within only depressed participants and to predict their clinical severity and QIDS values separately. It was more likely to find the relationship between behavioral features and depression symptoms when analyzing the depressed users separately from the non-depressed users. We speculate that QIDS variation among non-depressed participants may be largely due to incidental variations in lifestyle rather than psychological changes associated with depression. In Section 6, we develop separate predictive models to predict QIDS scores for the depressed and non-depressed groups. Our results indeed show that the prediction for the depressed group was more accurate.

Comparing Tables 1 and 2, we also remark that the r-values and p-values for Android users were generally better than those for iPhone users, which might be due to the different data collection mechanisms for these two platforms. This data discrepancy also makes it unjustified to simply merge data from the two platforms in a STL method. It makes more sense to treat them as separate but related learning tasks and jointly model them through knowledge transfer of MTL, which is one of the major motivations of our work.

4.3 Summary

To summarize, the discrepancy of the two smartphone platforms, in both the feature values themselves and their correlation with the QIDS scores, motivated us to explore MTL approaches to jointly model the data from the two platforms. In addition, severity levels assessed by clinicians and QIDS scores as self-reports brought up another reason of exploiting the MTL framework to transfer information between clinical assessment and

Table 1. Correlation (**r**) between QIDS scores and each individual behavioral feature for Andoid users with p-values.

	All		Depressed		Non-depressed		Difference
Features	**r**	**p**	**r**	**p**	**r**	**p**	$\|\|r_{dep}\|\| - \|\|r_{nodep}\|\|$
Loc_var	-0.24	5×10^{-5}	-0.50	1×10^{-4}	-0.20	2×10^{-3}	0.30
AMS	-0.19	9×10^{-4}	-0.61	5×10^{-7}	-0.09	0.19	0.52
Entropy	-0.25	2×10^{-5}	-0.65	5×10^{-8}	-0.17	0.01	0.48
$Entropy_N$	-0.13	0.03	-0.40	3×10^{-3}	-0.11	0.09	0.29
Home	0.28	2×10^{-6}	0.70	2×10^{-9}	0.20	2×10^{-3}	0.50
Move	-0.23	1×10^{-4}	-0.51	5×10^{-5}	-0.15	0.02	0.36
Distance	-0.20	8×10^{-4}	-0.62	4×10^{-7}	-0.09	0.17	0.53
N_{loc}	-0.26	7×10^{-6}	-0.63	2×10^{-7}	-0.15	0.02	0.48

Table 2. Correlation (**r**) between QIDS scores and each individual behavioral features for iPhone users with p-values.

	All		Depressed		Non-depressed		Difference
Features	**r**	**p**	**r**	**p**	**r**	**p**	$\|\|r_{dep}\|\| - \|\|r_{nodep}\|\|$
Loc_var	-0.04	0.34	-0.12	0.22	-0.05	0.37	0.07
AMS	-0.03	0.46	-0.08	0.44	-0.05	0.29	0.03
Entropy	-0.13	3×10^{-3}	-0.38	7×10^{-5}	0.10	0.04	0.28
$Entropy_N$	-0.09	0.03	-0.28	4×10^{-3}	-0.12	0.02	0.16
Home	0.12	0.01	0.28	4×10^{-3}	0.12	0.02	0.16
Move	-0.05	0.22	-0.06	0.54	0.01	0.78	0.05
Distance	-0.08	0.09	-0.17	0.07	-0.07	0.18	0.10
N_{loc}	-0.14	2×10^{-3}	-0.34	4×10^{-4}	-0.07	0.20	0.27

self evaluation. MTL improves the generalization of the estimated models for multiple related learning tasks by capturing and exploiting the task relationships. It has been theoretically and empirically shown to be more effective than learning tasks individually. Especially when STL suffers from limited sample size, MTL reinforces a single learning process with the transferable knowledge learned from the related tasks. MTL has been widely applied in many scientific fields, such as robotics [47], computer aided diagnosis [4], and computer vision [51].

5 MULTI-TASK LEARNING FOR DEPRESSION ASSESSMENT

Our overall approach of using MTL for depression assessment is illustrated in Fig. 3. As mentioned earlier, four types of data were collected for training the models: smartphone sensing data, Fitbit data, QIDS self-reports, and clinician assessment. After feature extraction, the extracted features were used in the MTL modules. One type of MTL is to predict QIDS scores using smartphone sensing features (with or without Fitbit features), which contains two regression tasks, separately, for Android and iPhone. Another type of multi-task learning is to predict clinical severity of depression, which contains two regression or classification tasks (depending on how we use the severity labels), again one for each smartphone platform. We can further combine these two different types of learning problems, thus forming a four-task MTL problem.

We first describe a widely-used MTL formulation, which helps depict the reason why our formulation is necessary. Our formulation belongs to a family of multiplicative MTL methods [45, 46], which decomposes individual model's coefficient vector into a multiplication of two vectors where one vector is shared across multiple of the tasks and the other is specific to a task itself. These methods have shown good performance. Our formulation extends the formulation in [45, 46] to support both homogeneous and heterogeneous learning tasks, which are used for various depression assessment in Section 6.

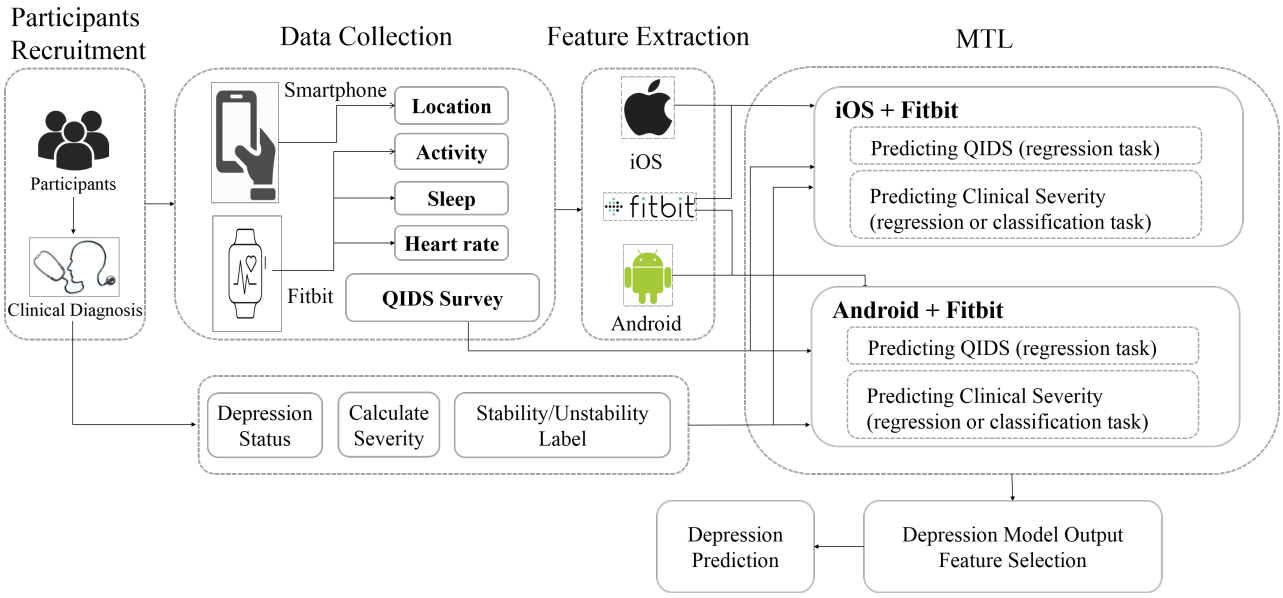


Fig. 3. An overview of using MTL for depression assessment. For smartphone sensing data, only location information is used in this paper.

Let T denote the number of learning tasks, and d denote the number of features. For each task $t \in \{1, \cdots, T\}$, we have a sample set $(\mathbf{X}_t \in \mathbb{R}^{\ell_t \times d}, \mathbf{y}_t \in \mathbb{R}^{\ell_t})$ that has ℓ_t examples, where the i-th row corresponds to the i-th example $\mathbf{x}_i^t$ of task t, $i \in \{1, \cdots, \ell_t\}$, and each example is represented by a vector of d features. The vector $\mathbf{y}_t$ contains y_i^t, the label of the i-th example for task t. We adopt functions of the linear form $y_i^t = (\mathbf{x}_i^t)\boldsymbol{\alpha}_t$ where $\boldsymbol{\alpha}_t \in \mathbb{R}^d$, which corresponds to computing $\mathbf{X}_t\boldsymbol{\alpha}_t$ on the training data. We define the parameter matrix or weight matrix $\mathbf{A} = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_T]$ where each column corresponds to a task. Denote the rows of $\mathbf{A}$ by $\boldsymbol{\alpha}^j$ where $j \in \{1, \cdots, d\}$ indexes features. To introduce the proposed multiplicative MTL method, let us start with a widely-used MTL formulation as follows:

$$\min_{\boldsymbol{\alpha}_{t:t=1,\cdots,T}} \sum_{t=1}^{T} L(\boldsymbol{\alpha}_t, \mathbf{X}_t, \mathbf{y}_t) + \Omega(\mathbf{A}) \tag{5}$$

where $L(\boldsymbol{\alpha}_t, \mathbf{X}_t, \mathbf{y}_t)$ is a loss function that computes the discrepancy between $\mathbf{X}_t\boldsymbol{\alpha}_t$ and $\mathbf{y}_t$ occured on task t, and L might be the least square loss for regression problems or the logistic regression loss for classification problems. We formulate the self-reported QIDS prediction problem as a regression task, and attempt two options for predicting clinical severity as either a regression or a classification task.

The second term $\Omega(\mathbf{A})$ regularizes the model parameters for the T tasks. A common strategy is to impose a blockwise joint regularization [15, 21, 28] on the matrix $\mathbf{A}$ to shrink the effects of an individual feature across the tasks. Commonly the $\ell_{1,p}$ matrix norm is employed to encourages sparsity on the rows of the matrix (i.e., zeroing out an entire row). Hence, the $\ell_{1,p}$ regularizer is used to rule out irrelevant features for all tasks by shrinking the corresponding rows in $\mathbf{A}$ to zeros.

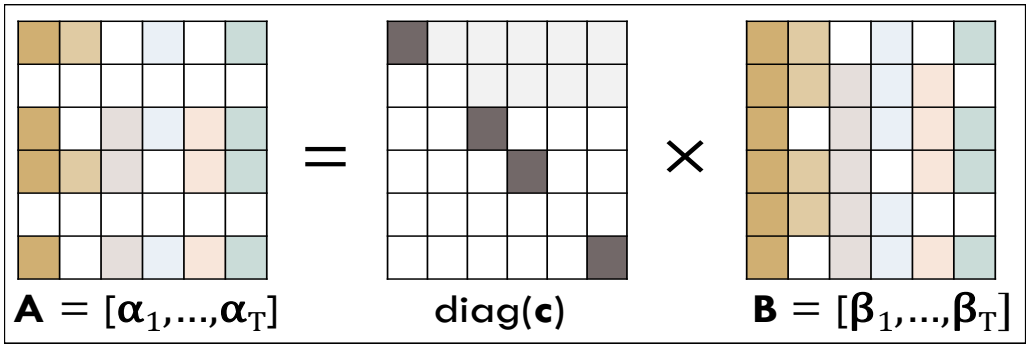


Fig. 4. Coefficient matrix **A** derived from the shared vector **c** and the task-specific vector $\boldsymbol{\beta}_t$.

5.1 The Proposed Formulation

To simultaneously construct models for heterogeneous tasks (e.g., one type as regression tasks for predicting QIDS scores and another as a classification task for classifying depression severity), we have to enable MTL methods to employ different types of loss function for different kinds of tasks. We hence revise the formulation (5) so that for each task t, we can choose its own specific loss function L_t. In our study,

$$L_t = \begin{cases} \|\mathbf{X}_t^T \boldsymbol{\alpha}_t - \mathbf{y}_t\|_2^2, & \text{if task } t \text{ is regression} \\ \log(1 + e^{-\mathbf{y}_t(\mathbf{X}_t^T \boldsymbol{\alpha}_t)}), & \text{if task } t \text{ is classification} \end{cases} \tag{6}$$

The use of Eq.(6) in a MTL setting supports both homogeneous learning tasks (i.e., when all tasks are of the same type) [45] and heterogeneous learning tasks (when some tasks are regression tasks while others are classification tasks). The resultant predictive models are of a generalized linear model form. For a test subject with an observed vector $\mathbf{x}$ of sensing features, to predict his/her QIDS score, we compute $y_t = \mathbf{x}^T \boldsymbol{\alpha}_t$ which gives a real-valued y_t; to classify his/her depression severity if the subject is depressed, we compute $y_t = 1/(1 + exp(-\mathbf{x}^T \boldsymbol{\alpha}_t))$ which gives a probability of being in the severe depression class.

Feature selection in a MTL setting can be very informative and insightful to suppress noise as well as understand the commonality and distinction between tasks. However, the widely-used MTL method in Eq.(5) has a major limitation. Because it regularizes models for different tasks in the same way, it either selects a feature as relevant to all tasks or excludes it from all models, which is very restrictive in practice because the tasks may share some features but may also have their own specific features that are not relevant to other tasks.

We propose to decompose each model's parameters $\boldsymbol{\alpha}_t$ into a multiplication of two components $\mathbf{c}$ and $\boldsymbol{\beta}_t$ with different regularization conditions imposed on each component. The model parameter vector thus becomes $\boldsymbol{\alpha}_t = \text{diag}(\mathbf{c})\boldsymbol{\beta}_t$, where diag($\mathbf{c}$) is a diagonal matrix with its diagonal elements composing $\mathbf{c}$. The vector $\mathbf{c}$ is used across all tasks, indicating if a feature is useful for any tasks. The vector $\boldsymbol{\beta}_t$ is only used for task t, containing the task-specific model parameters. The overall parameter matrix $\mathbf{A}$ can be derived as shown in Fig. 4. Let j index the entries in these vectors. We have $\alpha_t^j = c_j \beta_t^j$. Typically $\mathbf{c}$ comprises binary entries that are equal to 0 (false) or 1 (true) indicating if a feature is useful across tasks, but the integer constraint is often relaxed to require just non-negativity (i.e., $\mathbf{c} \geq 0$). The proposed method minimizes a regularized loss function as follows for the best $\mathbf{c}$ and $\boldsymbol{\beta}_t$.

$$\min_{\boldsymbol{\beta}_t, \mathbf{c} \geq 0} \sum_{t=1}^{T} L_t(\mathbf{c}, \boldsymbol{\beta}_t, \mathbf{X}_t, \mathbf{y}_t) + \gamma_1 \sum_{t=1}^{T} ||\boldsymbol{\beta}_t||_p^p + \gamma_2 ||\mathbf{c}||_k^k. \tag{7}$$

Algorithm 1: The blockwise coordinate descent algorithm for solving problem (7)

Input: $\mathbf{X}_t, \mathbf{y}_t, t = \{1, \cdots, T\}$, as well as γ_1, γ_2, p and k.
Initialize: $c_j = 1, \forall j = 1, \cdots, d$ **and** $s = 1$
repeat
  Conpute $\mathbf{X}_t\mathbf{diag}(\mathbf{c}^{s-1}) \rightarrow \hat{\mathbf{X}}_t, \forall t = 1, ..., T$;
  for $t = 1, ..., T$ **do**
    Solve $\min_{\boldsymbol{\beta}_t} L_t(\boldsymbol{\beta}_t, \hat{\mathbf{X}}_t, \mathbf{y}_t) + \gamma_1 ||\boldsymbol{\beta}_t||_p^p$ **for** β_t^s
  end
  Compute $\boldsymbol{\alpha}_t^s = \mathbf{diag}(\mathbf{c}^{(s-1)})\boldsymbol{\beta}_t^s$, **and compute** $\mathbf{c}^s$ **according to the formula (8);**
  Set $s = s + 1$ **Output:** $\boldsymbol{\alpha}_t$, **c and** $\boldsymbol{\beta}_{t:t=1,\cdots,T}$
until $\max(|\boldsymbol{\alpha}^{(s+1)} - \boldsymbol{\alpha}^{(s)}|) < \epsilon$;

Different regularizers are imposed on the two decomposed components in Eq.(7), where p and k are positive integers. The notation $||\cdot||_p$ denotes the ℓ_p-vector norm $||\mathbf{x}||_p = \sqrt[p]{\sum_i (x_i)^p}$, and $||\mathbf{x}||_p^p$ corresponds to the ℓ_p-norm of $\mathbf{x}$ to the power of p. The tuning parameters γ_1 and γ_2 are used to balance the empirical loss and regularizers. The learned model also selects features. Specifically, at optimality, if $c_j = 0$, the j-th variable is removed for all tasks, and the corresponding row vector $\boldsymbol{\alpha}^j = \mathbf{0}$; otherwise the j-th variable is selected for use in at least one of the $\boldsymbol{\alpha}$'s. Then, a specific $\boldsymbol{\beta}_t$ may rule out the j-th variable from a specific task t if $\beta_t^j = 0$.

Now, let us discuss the impact of the different choices of regularizers. Note that appropriate regularizers may be problem specific. When $p = k = 2$, which amounts to an early method discussed in [4], Eq.(7) does not impose strong sparsity on the model parameters, and consequently a majority of features may be selected and may be shared across tasks. When $p = k = 1$, similar to that in [29], Eq.(7) is suitable for learning from tasks with persistently sparse models. In other words, a large portion of features may be disgarded, and even for selected features, only few of them are shared from task to task. Two new formulations have been proposed in our recent work [45] that examines theoretical properties of the multiplicative decomposition Eq. (7). The two new formulations consist of the case of $p = 2$, but $k = 1$ and the case of $p = 1$, but $k = 2$. The first new formulation is favorable to learning problems where many features are irrelevant to any of the tasks, however, among the selected features, a lot of them might be shared by different tasks. In other words, $\mathbf{c}$ is sparse, but the features used by each task are not sparse with respect to the selected features indicated by $\mathbf{c}$. The second formulation may help those learning tasks when the union of the features that are relevant to any given task includes many or even all features, but different tasks may share only a limited number of features. In other words, $\mathbf{c}$ is not sparse, but each individual task uses a sparse $\boldsymbol{\beta}_t$ to select from those indicated by $\mathbf{c}$.

5.2 The Optimization Algorithm

We have developed a blockwise coordinate descent algorithm to solve problem (7) in [46]. As described in Algorithm 1, the algorithm alternates between optimizing two sub-problems until convergence. The first sub-problem solves only for $\boldsymbol{\beta}$'s with a fixed $\mathbf{c}$ whereas the second sub-problem solves for $\mathbf{c}$ with fixed $\boldsymbol{\beta}$'s. In particular, we have derived a closed-form solution for the second sub-problem to directly compute $\mathbf{c}$ [46]. The trick is that at iteration s, first compute the iterate $\boldsymbol{\alpha}^s$ from the current iterates of $\boldsymbol{\beta}^s$ and $\mathbf{c}^s$, and then use the following formula to compute the next iterate $\mathbf{c}$ (readers can consult with [46] for the detailed derivation.)

$$c_j = \sqrt[p+k]{\frac{\gamma_1}{\gamma_2} \sum_{t=1}^{T} (\alpha_t^j)^p}. \tag{8}$$

The loss functions we used in Eq.(7) comprise the least squares loss for regression and the logistic regression loss for classification. Any other loss function that is convex and differentiable in terms of $\boldsymbol{\alpha}$ will also be appropriate. The first sub-problem of optimizing $\boldsymbol{\beta}$'s amounts to solving a single task learning problem, and can be solved separately for individual tasks using efficient single task learning tools already available. The second sub-problem is analytically solved as discussed above. We choose to monitor the maximum norm of the parameter matrix $\mathbf{A}$ in order to terminate the process, but it can be replaced by any other suitable termination criterion. Initialization can be important for this algorithm, and we suggest starting with $\mathbf{c} = \mathbf{1}$, which means to consider all features initially equally in the learning process.

6 MULTI-TASK LEARNING RESULTS

We present the evaluation results, and discuss the comparison between MTL and STL in this section. To validate the proposed MTL method and the usefulness of Fitbit data, we carried out two sets of experiments by: (1) only using the 8 smartphone sensing features, and (2) using the 8 smartphone sensing features plus the 28 Fitbit features in constructing the prediction models. The QIDS scores were modeled for the depressed and non-depressed groups, separately, which was motivated by our observation that the sensing features showed stronger predictive power for the depressed users (see Section 4). For depressed participants, a model was also constructed to predict their level of severity. We first describe the evaluation settings in the following subsection.

6.1 Evaluation Settings

We evaluated four variants of the proposed MTL approach by differing the regularizers used in Eq.(7) as discussed in Section 5.1, and contrasted them with the STL approaches that were adopted in [38]. Specifically, the various methods used in our comparison include

- STL L1 (lasso): Learning each task independently with the $L1$ norm as the regularizer.
- STL L2 (ridge): Learning each task independently with the $L2$ norm as the regularizer.
- MTL L11: Multiplicative multi-task feature learning using two $L1$ norms as regularizers: $||\boldsymbol{\beta}||_1$ and $||\mathbf{c}||_1$.
- MTL L12: Multiplicative multi-task feature learning using the $L1$ and $L2$ regularizers: $||\boldsymbol{\beta}||_1$ and $||\mathbf{c}||_2^2$ $(p = 1, k = 2)$.
- MTL L21: Multiplicative multi-task feature learning using the $L2$ and $L1$ regularizers: $||\boldsymbol{\beta}||_2^2$ and $||\mathbf{c}||_1$ $(p = 2, k = 1)$.
- MTL L22: Multiplicative multi-task feature learning using two $L2$ norms as regularizers: $||\boldsymbol{\beta}||_2^2$ and $||\mathbf{c}||_2^2$.

Our prediction goal was focused on the longitudinal generalization of the constructed models. In other words, we tested whether a prediction model constructed using a period of data can predict outcome of another period. We partitioned the data in the following way in a cross validation (CV) process. Each fold of the CV corresponded to a specific week, i.e., a QIDS survey interval, because we collected QIDS responses every week, and hence each training example amounted to a week's data. We left out that specific week of data from each individual for test and the remaining data were used in training. We called this process leave-one-week-out or leave-one-interval-out. The regularization parameters of individual methods were tuned within training using another round of leave-one-week-out by selecting values from pre-chosen choices of $10^{-5}, \cdots, 1, \cdots, 10^5$. The same tuning process was used to tune the hyper-parameters of every method for fair comparison.

We used the coefficient of determination (R^2) to measure the regression performance and the $F1$ score to measure the classification performance. The R^2 value ranges from 0 to 1, measuring how much percentage of the variance of the dependent variable can be accounted for by the model. A higher R^2 value indicates better regression performance. We reported the R^2 values averaged over all regression tasks in each MTL setting. For classification tasks, we reported the averaged $F1$ scores over all classification tasks as well. The $F1$ score also ranges from 0 to 1 with higher values representing better classification performance.

Table 3. Performance comparison of different methods using Fitbit+Smartphone data for predicting QIDS scores and clinical severity respectively.

	QIDS		Clinical Severity
	Depressed	**Non-depressed**	**Depressed**
Method	R^2	R^2	R^2
$L11$	0.44	0.23	0.41
$L12$	0.41	0.23	0.39
$L21$	0.33	0.21	0.41
$L22$	0.30	0.16	0.42
$STL + L1$	0.34	0.19	0.35
$STL + L2$	0.30	0.14	0.34

Table 4. Performance comparison of different methods using smartphone data for predicting QIDS scores and clinical severity respectively.

	QIDS		Clinical Severity
	Depressed	**Non-depressed**	**Depressed**
Method	R^2	R^2	R^2
$L11$	0.28	0.11	0.37
$L12$	0.23	0.10	0.37
$L21$	0.25	0.16	0.39
$L22$	0.24	0.16	0.40
$STL + L1$	0.21	0.15	0.35
$STL + L2$	0.19	0.16	0.34

6.2 MTL with Two Tasks

We compare the results when using MTL for two homeogeneous prediction tasks respectively for Android and iPhone users. We carried out three separate MTL experiments: the first setting aimed to predict QIDS scores for depressed users only; the second setting aimed to predict QIDS scores for non-depressed users; and the third setting aimed to predict the four-level severity for depressed users. Note that here we treated the severity prediction as a regression problem. Hence, all of the three experiments were concerned with regression problems. In each of the three experiments, we used two input configurations: Android and iPhone smartphone datasets (8 features as input variables for each task), Android+Fitbit and iPhone+Fitbit datasets (8+28 features as input variables for each task).

The left half of Tables 3 and 4 provides results of the QIDS prediction. We observed that the R^2 values for the MTL approach, especially the L11 method, were substantially greater than those of STL approaches (for each approach, the reported R^2 was the average of the two tasks). By cross referencing the two tables, we observed that the inclusion of Fitbit data enhanced the MTL performance by at much as 57% in the QIDS prediction for depressed users, 44% for non-depressed users. Of the four MTL approaches, the MTL L11 model performed the best in both data configurations. Precisely, the R^2 value of the L11 model was 29.4% greater than that of the best STL approach in the smartphone+Fitbit case. These results confirmed that the joint modeling was beneficial.

For non-depressed users, the MTL approaches did not improve the performance as significantly as that for the depressed users. Across all methods, the R^2 values for non-depressed users were much lower than those for the

depressed users, which was consistent with the weaker correlation observed between features and QIDS scores for non-depressed users in Section 4. Again, including variables derived from Fitbit data, the prediction accuracy was substantially improved for all comparison methods.

We next illustrate the features that were selected by the best MTL approach. Specifically, we used the MTL L11 model for depressed users, the best prediction model in Table 3, as an example. Fig. 5 shows the weights of each feature used in this model as bar plots for each of the two tasks, where Task 1 was for iPhone users and Task 2 for Android users. The L11 model selected 12 features, including 'Location variance', 'AMS', 'Home', 'Move', 'Total distance', 'Lightly active minutes', 'Sedentary minutes', 'Minutes after wakeup', 'Awake duration', 'Restless count', 'Total minutes asleep' and 'Minutes of heart rate in fat-burn zone'.

These models bring out several interpretable insights. For example, positive weights of 'Home', 'Sedentary minutes', 'Minutes of heart rate in fat-burn zone' and 'Minutes after wakeup' indicated that longer time spent at home, longer sitting time, longer duration in fat-burn zone (which is of low heart-rate intensity) and longer amount of time to fall back sleep corresponded to higher QIDS values; negative weights of 'Location variance' and 'Lightly active minutes' indicated that higher variability in locations and longer duration in outdoor environments would lead to lower QIDS scores. The selection of 'Home', 'Location variance' and 'Total distance' as important features was consistent with the observation that they had the strongest correlation with QIDS scores (see Tables 1 and 2). The 'Move' variable was selected although it had low correlation with QIDS scores, but it could be complementary to 'Location variance'. In contrast, the STL L1 method selected 'Home' only for iPhone users and 'Move' for Android users. We believe that the MTL method captured a certain level of relatedness between the two tasks (e.g., thus selecting the two features for both tasks).

The right half of Tables 3 and 4 shows the results for predicting clinical severity for depressed users. The MTL L22 method achieved the best performance. It improved the R^2 value of the best STL method by 0.07 in Table 3, by 0.05 in Table 4, showing improvement of nearly 20%. The MTL L11 and L12 models had similar performance as that of the STL methods, which could be partially because the ℓ_1 norm penalty imposed on $\boldsymbol{\beta}$'s in these two methods might be too aggressive in inducing sparsity for this task. This observation showed that careful selection of regularizers in MTL methods could be important and task-specific.

6.3 MTL with Four Tasks

We further carried out four-task MTL experiments on depressed users by merging the QIDS prediction and severity prediction, respectively for Android and iPhone users, in a joint framework. We compared the cases with and without Fitbit data as well. We noticed that the severity prediction problem was very challenging due to medical subtleness between the different severity levels and a significantly reduced sample size. We hence performed two separate settings where: (1) depression severity was treated as a regression problem with four levels, leading to a homogeneous four-task MTL problem; (2) depression severity was treated as a binary classification problem of classifying subjects into the stable and unstable classes, leading to a heterogeneous MTL problem. The unstable class includes two severe levels (namely "Moderate" and "Severe"). We treated it as classification because the number of samples in the unstable severity levels was very small. Therefore we aggregated them together to provide a reasonable sample size.

The left half of Tables 5 and 6 shows the results of the homogeneous setting, where for each approach, the R^2 value was averaged over the four regression tasks. Specifically, Table 5 contains results using both smartphone and Fitbit variables, while Table 6 contains results based only on smartphone variables. The R^2 values of the L11 and L12 models in Table 5, are 34.3% more than that of the best STL model. By cross referencing the tables with Tables 3 and 4, we see that the various MTL models in the four-task setting achieve better or comparable performance than that in the two-task setting. This observation might be an evidence that self-reported depression scores and clinical severity can be related so to improve each other's prediction performance.

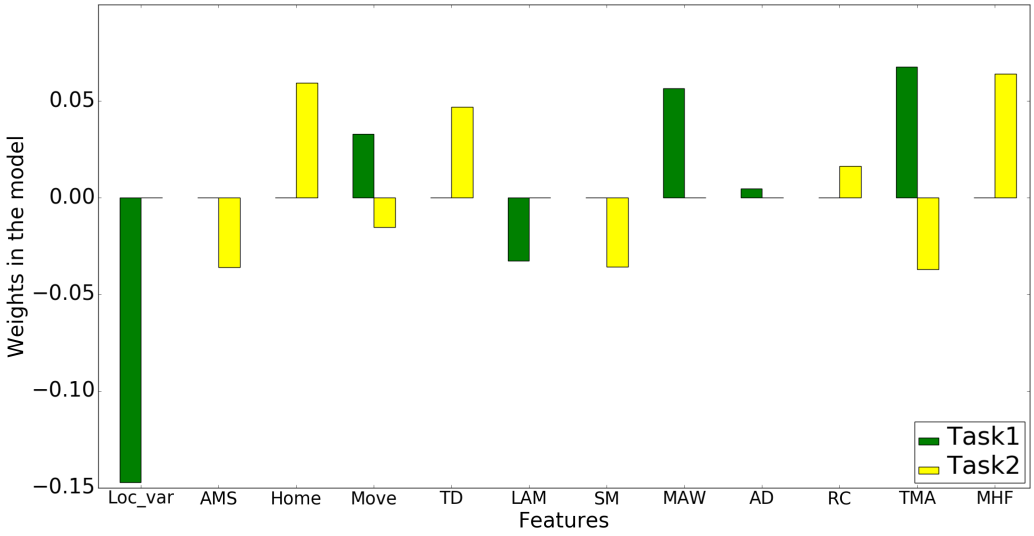


Fig. 5. Features selected by the MTL L11 model to predict QIDS scores for depressed users (the two tasks are for iPhone and Android users, respectively). Features are 'Loc_var', 'AMS', 'Home', 'Move', 'Total distance', 'Lightly active minutes', 'Sedentary minutes', 'Minutes after wakeup', 'Awake duration', 'Restless count', 'Total minutes asleep', 'Minutes of heart rate in fat-burn zone'.

Table 5. Performance comparison of different methods for predicting QIDS scores (two regression tasks) together with depression severity (two regression or classification tasks) for depressed users using Fitbit+Smartphone data.

	QIDS and Depression Severity (Homogeneous)	**QIDS and Depression Severity (Heterogeneous)**	
Method	R^2	R^2	**F1 score**
$L11$	0.43	0.35	0.77
$L12$	0.43	0.36	0.77
$L21$	0.41	0.35	0.67
$L22$	0.41	0.40	0.67
$STL + L1$	0.32	0.34	0.52
$STL + L2$	0.29	0.30	0.52

The results from the heterogeneous setting are given in the right half of Tables 5 and 6. For the QIDS regression problems, we observed that the R^2 values of the MTL models were remarkably better than those of the STL models: the R^2 values of the best MTL model were 17.6% and 19.0%, respectively, better than that of the best STL model in the smartphone+Fitbit and smartphone cases. The best R^2 value was worse than those in the homogeneous four-task setting, implying that using the four-level severity labels (in the homogeneous setting) helped more with the knowledge transfer to improve the QIDS prediction than the two-way severity labels (in the heterogeneous setting). For the classification task, the best MTL model was clearly better (by 48.1%) than the

Table 6. Performance comparison of different methods for predicting QIDS scores (two regression tasks) together with depression severity (two regression or classification tasks) for depressed users using Smartphone data.

	QIDS and Depression Severity (Homogeneous)	**QIDS and Depression Severity (Heterogeneous)**	
Method	R^2	R^2	**F1 score**
$L11$	0.35	0.21	0.67
$L12$	0.34	0.25	0.67
$L21$	0.33	0.20	0.57
$L22$	0.33	0.23	0.55
$STL + L1$	0.30	0.21	0.50
$STL + L2$	0.30	0.19	0.50

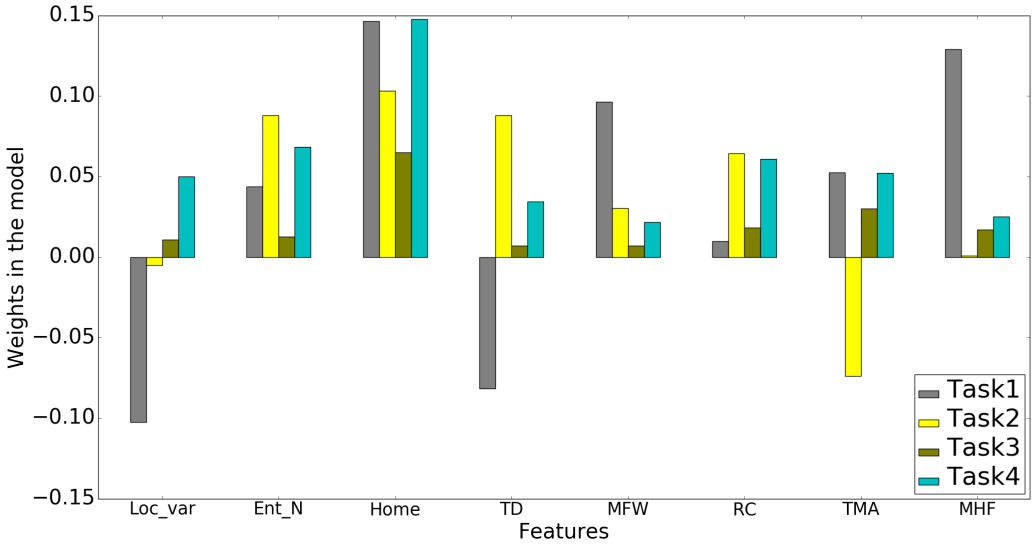


Fig. 6. Feature selected by the MTL L22 model in the heterogeneous setting with four tasks. Selected features are 'Loc_var', 'Entropy_N', 'Home', 'Total distance', 'Minutes after wakeup', 'Restless count', 'Total minutes asleep', 'Minutes of heart rate in fat-burn zone'.

best STL model in terms of the F1 score when using Fitbit and smartphone data altogether. It may be valid that the difficult severity prediction tasks can benefit from the general QIDS prediction, but not vice versa, because we observed that the R^2 values in this heterogeneous four-task setting for QIDS prediction were lower than those in Tables 3 and 4.

When both smartphone and Fitbit data were used, the best QIDS prediction model in the heterogeneous setting (considering both predicting QIDS score and classifying severity) was the MTL L22 models. The weights in this model are plotted in Fig. 6. Eight important features were selected, including 'Location variance', 'Entropy_N', 'Home', 'Total distance', 'Minutes after wakeup', 'Restless count','Total minutes asleep', and 'Minutes of heart rate

in fat-burn zone'. Besides the similar features that were also selected in Section 6.2., 'Entropy$_N$' was selected, which was consistent with the correlation analysis in Tables 1 and 2 that 'Entropy$_N$' was highly correlated with the QIDS scores. The correlation between the other two selected features ('Restless count' and 'Total minutes asleep') and the QIDS scores was not very strong as individual features. The result shows that the MTL methods learn the interplay of the individual features for use in the prediction models.

7 DISCUSSIONS AND CONCLUSIONS

In this paper, we have formulated a problem that uses MTL to model the data collected from two smartphone platforms together with Fitbit data. The statistical analysis justified the correlation between the extracted features (both from smartphone and Fitbit sensing data) and the QIDS scores. We further proposed a novel heterogeneous multi-task learning method with feature selection to predict QIDS scores and depression severity levels.

When predicting QIDS scores (two regression tasks, one for each smartphone platform), the MTL approach outperforms the STL approach by 29.4%; when predicting clinical severity (treated as two regression tasks, one for each smartphone platform), the performance improvement reaches 20.0%. The improvement becomes even clearer when combining the four regression tasks of predicting QIDS scores and clinical severity into a single framework. In this case, the MTL with four homogeneous tasks improves the overall regression performance by 34.3% over STL. In addition, we have validated that the proposed heterogeneous multi-task method improves the classification accuracy (for classifying depression severity) by 48.1% and improves the regression accuracy (for predicting QIDS scores) by 17.6%, compared to the classical STL methods. In summary, our results demonstrate that MTL is a promising technique for jointly modeling the data collected from different platforms as well as the data collected for different tasks.

Our evaluation shows that Fitbit data can improve the prediction accuracy of all models. As shown in Figures 5 and 6, heart rate and sleep related features from Fitbit were selected and weighted heavily in the prediction models. Our results are consistent with the findings in [16], which shows that heart rate and activity are biologically correlated with depression. Therefore, further exploration of using sensing data from wearable devices for depression screening might be an interesting direction.

Our results are consistent with earlier studies (e.g., [6, 12, 13, 38]) that location based features, such as time spent at home, entropy and number of locations, are correlated with self-reported (QIDS/PHQ-9) depression scores. The proposed heterogeneous MTL method is compatible with the heterogeneous data from multiple platforms and heterogeneous tasks (including both classification and regression tasks).

Because depression symptoms can last for a long period of time and can be recurring, automatic prediction of QIDS scores can help monitor the depression status over time, which will be a valuable tool in practice. Similarly, for non-depressed users, automatic prediction of QIDS scores can keep a tap on one's mental health, and potentially be used to automatically detect the onset of depression. Our results provide further evidence that sensing data from smartphones and wearable devices can be used for automatic depression screening.

There are two directions of future work. First, we are in the process of collecting a larger dataset to enhance the statistical power of our analysis. Secondly, since we would keep observing participants in a long term, we will explore using both baseline and longitudinal features to improve the performance of the multi-task learning methods.

REFERENCES

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

[2] A. K. Battenberg, S. Donohoe, N. Robertson, and T. P. Schmalzried. The accuracy of personal activity monitoring devices. In *Seminars in Arthroplasty*, volume 28, pages 71–75. Elsevier, 2017.

[3] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, and A. T. Campbell. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal*, 38(3):218–226, 2015.
[4] J. Bi, T. Xiong, S. Yu, M. Dundar, and R. B. Rao. An improved multi-task learning approach with applications in medical diagnosis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 117–132. Springer, 2008.
[5] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the ACM International Conference on Multimedia*, pages 477–486. ACM Press, 2014.
[6] L. Canzian and M. Musolesi. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proc. of ACM UbiComp*, pages 1293–1304, 2015.
[7] 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.
[8] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *Proc. of ACM UbiComp*, pages 481–490. ACM, 2012.
[9] Y. Chon, E. Talipov, H. Shin, and H. Cha. Mobility prediction-based smartphone energy optimization for everyday location monitoring. In *Proc. of ACM conference on embedded networked sensor systems*, pages 82–95. ACM, 2011.
[10] 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.
[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM KDD*, volume 96, pages 226–231, 1996.
[12] A. A. Farhan, J. Lu, J. Bi, A. Russell, B. Wang, and A. Bamis. Multi-view bi-clustering to identify smartphone sensing features indicative of depression. In *Proc. IEEE CHASE*, June 2016.
[13] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang. Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data. In *Proc. of IEEE Wireless Health Conference*, October 2016.
[14] M. Frost, A. Doryab, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram. Supporting disease insight through data analysis: refinements of the monarca self-assessment system. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 133–142. ACM, 2013.
[15] P. Gong, J. Ye, and C. Zhang. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 895–903. ACM, 2012.
[16] J. M. Gorman and R. P. Sloan. Heart rate variability in depressive and anxiety disorders. *American heart journal*, 140(4):S77–S83, 2000.
[17] A. Gruenerbl, V. Osmani, G. Bahle, J. C. Carrasco, S. Oehler, O. Mayora, C. Haring, and P. Lukowicz. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *Proceedings of the 5th Augmented Human International Conference*, page 38. ACM, 2014.
[18] A. Grünerbl, P. Oleksy, G. Bahle, C. Haring, J. Weppner, and P. Lukowicz. Towards smart phone based monitoring of bipolar disorder. In *Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare*, page 3. ACM, 2012.
[19] S. Guo, O. Zoeter, and C. Archambeau. Sparse bayesian multi-task learning. In *Advances in Neural Information Processing Systems*, pages 1755–1763, 2011.
[20] L. Jacob, J.-p. Vert, and F. R. Bach. Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems*, pages 745–752, 2009.
[21] A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 964–972, 2010.
[22] N. Jaques, S. Taylor, A. Sano, and R. Picard. Multi-task, multi-kernel learning for estimating individual wellbeing. In *Proc. NIPS Workshop on Multimodal Machine Learning*, December 2015.
[23] K. Kroenke, R. L. Spitzer, and J. B. Williams. The PHQ-9. *Journal of General Internal Medicine*, 16(9):606–613, 2001.
[24] N. D. Lane, M. Lin, M. Mohammod, X. Yang, H. Lu, G. Cardone, S. Ali, A. Doryab, E. Berke, A. T. Campbell, et al. BeWell: Sensing sleep, physical activities and social interactions to promote wellbeing. *Mobile Networks and Applications*, 19(3):345–359, 2014.
[25] S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller. Learning a meta-level prior for feature relevance from multiple related tasks. In *Proceedings of the 24th international conference on Machine learning*, pages 489–496. ACM, 2007.
[26] 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.
[27] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. Can your smartphone infer your mood? In *PhoneSense workshop*, pages 1–5, 2011.
[28] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 339–348. AUAI Press, 2009.
[29] A. Lozano and G. Swirszcz. Multi-level lasso for sparse multi-task regression. In *Proceedings of International Conference on Machine Learning*, pages 361–368, 2012.
[30] 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.
[31] A. Madan, M. Cebrian, D. Lazer, and A. Pentland. Social sensing for epidemiological behavior change. In *Proc. of ACM UbiComp*, pages 291–300. ACM, 2010.
[32] 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, 2010.
[33] A. Mehrotra, R. Hendley, and M. Musolesi. Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction. In *Proc. of UbiComp*, 2016.
[34] 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 *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 208–214. IEEE, 2011.
[35] N. Palmius, A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, J. R. Geddes, G. M. Goodwin, and M. D. Vos. Detecting bipolar depression from geographic location data. *IEEE Transactions on Biomedical Engineering*, PP(99):1–1, 2016.
[36] L. Pei, R. Guinness, R. Chen, J. Liu, H. Kuusniemi, Y. Chen, L. Chen, and J. Kaistinen. Human behavior cognition using smartphone sensors. *Sensors*, 13(2):1402–1424, 2013.
[37] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber, et al. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological psychiatry*, 54(5):573–583, 2003.
[38] 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.
[39] 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.
[40] 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.
[41] Y. Suhara, Y. Xu, and A. Pentland. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proc. of WWW*, 2017.
[42] T. Vos, A. D. Flaxman, M. Naghavi, R. Lozano, 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.
[43] R. Wang, M. S. H. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhuryy, M. Hauserz, J. Kanez, M. Merrilly, E. A. Scherer, V. W. S. Tsengy, and D. Ben-Zeev. Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proc. of UbiComp*, 2016.
[44] 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 the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14. ACM, 2014.
[45] X. Wang, J. Bi, S. Yu, and J. Sun. On multiplicative multitask feature learning. In *Advances in Neural Information Processing Systems*, pages 2411–2419, 2014.
[46] X. Wang, J. Bi, S. Yu, J. Sun, and M. Song. Multiplicative multitask feature learning. *Journal of Machine Learning Research*, 17(80):1–33, 2016.
[47] C. Williams, S. Klanke, S. Vijayakumar, and K. M. Chai. Multi-task Gaussian process learning of robot inverse dynamics. In *Advances in Neural Information Processing Systems*, pages 265–272, 2009.
[48] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8(Jan):35–63, 2007.
[49] M. Yang, Y. Li, and Z. Zhang. Multi-task learning with Gaussian matrix generalized inverse Gaussian model. In *ICML (3)*, pages 423–431, 2013.
[50] K. Yu, V. Tresp, and A. Schwaighofer. Learning Gaussian processes from multiple tasks. In *Proceedings of the 22nd international conference on Machine learning*, pages 1012–1019. ACM, 2005.
[51] X.-T. Yuan, X. Liu, and S. Yan. Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing*, 21(10):4349–4360, 2012.
[52] C. Yue, S. Ware, R. Morillo, J. Lu, C. Shang, J. Bi, A. Russell, A. Bamis, and B. Wang. Fusing location data for depression prediction. In *Proc. IEEE Ubiquitous Intelligence and Computing*, August 2017.
[53] Y. Zhang, D.-Y. Yeung, and Q. Xu. Probabilistic multi-task feature selection. In *Advances in neural information processing systems*, pages 2559–2567, 2010.
[54] D. Zhou, J. Luo, V. M. B. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. A. Kautz. Tackling mental health by integrating unobtrusive multimodal sensing. In *Proc. of AAAI*, 2015.

Received May 2017; revised November 2017; accepted January 2018