Using Passive Smartphone Sensing for improved Risk Stratification of patients with Depression and Diabetes

Archana Sarda¹, MBBS, MD; Suresh Munuswamy²,³, MBBS, MPH, PhD; Shubhankar Sarda⁴, BSc (Physics); Vinod Subramanian⁵, BE, MBA;

¹Director, Sarda Centre for Diabetes and Self-care, Aurangabad, Maharashtra, India.
²Assistant Professor, Indian Institute of Public Health, Hyderabad, India
³Coordinator, DST Hi Rapid Design Lab and IIPH, Hyderabad, India
⁴Engineering Lead, Touchkin, Bangalore, Karnataka, India
⁵Independent Researcher, Bangalore, Karnataka, India

Corresponding Author:
Shubhankar Sarda,
1st Floor, Manjusha, No 532
16th Cross, 2nd Main Road, 2nd Stage
Indiranagar, Bangalore, 560038
INDIA
Phone: +91 9762665119
Email: shubhankar [at] touchkin.com

Abstract

Background:
Research studies are establishing the use of smartphone-sensing to measure mental well-being. Smartphone sensor information captures behavioral patterns and its analysis helps reveal well-being changes. Depression in diabetes goes highly under-diagnosed and under-reported. The comorbidity has been associated with increased mortality and worse clinical outcomes; including poor glycemic control and poor self-management. Clinical only intervention has been found to have very modest effect on diabetes management among people with depression. Smartphone technologies could play a significant role in complementing co-morbid care.

Objective:
To present an approach to risk-stratify people with diabetes based on symptoms of depression detected using smartphone-sensing information.

Methods:
A cross-sectional observational study (Project SHADO- Analyzing Social and Health Attributes through Daily Digital Observation) was conducted on 47 participants with
diabetes. The study smartphone-sensing app passively collected data regarding activity, mobility, sleep and communication from each participant. Self-reported symptoms of depression (using validated Patient Health Questionnaire-9) was collected once every 2 weeks from all participants. A descriptive analysis was performed to understand the representation of the participants. A univariate analysis was performed on each derived sensing variable to compare behavioral changes between depression states—those with self-reported major depression (PHQ-9 > 9) and those with none (PHQ-9 <= 9). A classification predictive modeling, using supervised machine-learning methods, was explored using derived sensing variables as input to construct and compare classifiers that could risk-stratify people with diabetes based on symptoms of depression.

**Results:**
A noticeably high prevalence of self-reported depression (30 out of 47 participants, ~65%) was found among the participants. Low correlation was found between self-reported depression state and each of the 53 derived sensing variables. Between depression states, a significant difference was found for average activity rates (day time) among participant-day instances with symptoms of major depression (M=16.06, SD=14.90) and those with none (M=18.79, SD=16.72); P = .005. For average number of people called, a significant difference was found between participant-day instances with symptoms of major depression (M=5.08, SD=3.83) and those with none (M=8.59, SD=7.05); P < .001. These results suggest that participants with diabetes and symptoms of major depression exhibited lower activity through the day and maintained contact with fewer people. Using all the derived sensing variables, the XGBoost machine-learning classifier provided the best performance with an average cross-validation accuracy of 79.07% (95% CI: 74%, 84%) and test accuracy of 81.05% to classify symptoms of depression.

**Conclusions:**
Participants with diabetes and self-reported symptoms of major depression were observed to show lower levels of social contact and lower activity levels during the day. While findings must be reproduced in a broader RCT, the study shows promise in use of predictive modeling for early detection of symptoms of depression in people with diabetes using smartphone-sensing information.

**Keywords:**
depression; diabetes; smartphone sensing; classification; machine learning; predictive model; digital health; risk stratification

**Introduction**
There exists growing evidence regarding the bidirectional adverse interaction between diabetes and depression [1-3]. Depression appears two to three times more common in people with diabetes than in those without. An estimated 8-35% of people with diabetes mellitus also suffer from depression [4]. Depression increases the risk of non-adherence
to medical treatment by 27-30% [6], which is a significant problem in diabetes self-care. Further, people with diabetes who also have depression are 2 to 4.5 times more expensive to treat than those with diabetes alone [5, 7]. Depression is known to be associated with abnormalities in metabolism of biologics (for e.g., increased counter regulatory hormone release and action, changes in glucose transport function and increased immuno-inflammatory activation) [8]. Depression may also increase the risk of developing type 2 diabetes with increase in insulin resistance and reduction of glucose uptake [9]. A significant association exists between depression and diabetes complications, cognitive impairment, poor self-management, quality of life and mortality [10-17].

Primary care physicians fail to diagnose as much as 50%-70% of patients who present with current depressive disorder [18]. In diabetes, depression gets identified only half of the time (51%) of which only 31% receive adequate dose of anti-depressants [19]. Clinical only interventions seem to have very modest effects in diabetes management of patients with depression [20-22]. Compared to patients with no depressive symptoms, patients with symptoms of depression and who have diabetes are more likely to be non-adherent to medication regimens [23] and exhibit worsening diabetes management [24]. The ADA (American Diabetes Association) recommends that patients with Diabetes be screened for psychosocial and psychological problems or disorders, such as depression [4, 25]. However, this appears rarely to happen [26].

Global smartphone users are expected to surpass 2.3 billion by 2017 [27]. Smartphones carry sensors such as accelerometer, GPS- Global Positioning System and ambient light sensors that captures data and that could provide information on someone's behavior. In this context, the smartphone could be the most ubiquitous data collection device today. It also presents with huge privacy and security concerns. Developed nations have higher smartphone penetration and the ownership rates in emerging and developing nations has been rising at an extraordinary rate [28]. Passive data from smartphone sensors has been known to detect patterns of behavior in people with depression [11, 29-32]. Research has been establishing the link between smartphone-sensing data and its application in overall well-being [33-36], application in depression [37-39] and its application in comorbid conditions [11]. Smartphones for social sensing [40, 41], monitoring and possibly as an intervention in mental health [42, 43], has the advantage of ubiquity, discretion and low cost.

Diabetes and depression have been associated with increased healthcare costs, thereby posing a large economic burden [44]. To use limited resources efficiently, risk stratification is important to prioritize people with diabetes. Many studies have explored how smartphone-sensing data can be used as a predictor for depression and mental health [32, 43, 45-48], but very few studies have applied passive sensing to predict symptoms of depression among people with diabetes thereby enabling improved risk stratification. There are recent studies that attempt to predict depression among
patients with diabetes using longitudinal patient records or data from clinical trials or from surveys but not with sensing data as an indicator [49-51].

The pilot study named Project SHADO (acronym for "Analyzing Social and Health Attributes through Daily Digital Observation") was conducted in 2016 on a cross-section of participants with diabetes located in peri-urban India and owning low-cost smartphones. The pilot study used a smartphone-sensing app (app) developed by Touchkin. The app assisted family members to care for their loved ones remotely and non-intrusively and to check on their well-being. The app’s machine learning (ML) platform helped detect probable well-being changes by using activity rates, communication levels, sleep patterns and mobility information collected from the user’s smartphone sensors. The aim of the study was to analyse smartphone-sensing parameters in participants with diabetes and to evaluate its association with depression using various methods including predictive modeling. To the best of knowledge, this study was the first such implementation of using automatically captured smartphone-sensing data such as activity, communication, mobility and sleep to screen for symptoms of depression among primary care patients with diabetes.

Methods

Design

The study design was completed in association with a diabetes clinic situated in Aurangabad, a city in the state of Maharashtra, India. The institutional Ethics Committee at PHFI’s (Public Health Foundation of India) Indian Institute of Public Health at Hyderabad, India approved the study (Approval # IIPHH/TRCIEC/073/2016).

A cross-sectional observational study was conducted on a sample of patients undergoing diabetes treatment at the clinic. The study did not require any intervention or change in treatment or lifestyle for the participants. It did not involve a control group. The study was designed to have no financial burden on the participant, nor any drug or device hazard. The care providers at the clinic administered the study. Formal training and onboarding was provided to the clinic administrators. Effective monitoring and support was established to manage any issues that could occur during the study.

A list of 100 peri-urban patients undergoing treatment for diabetes and who satisfied the study inclusion criteria (Textbox 1) were contacted for their interest and participation in the study.

1. 1+ year on diabetes treatment, with recent six months consulting a diabetologist
2. Had 1-2 clinic visits per month
3. 18 years and above
4. Own a smartphone with android operating system
Textbox 1: Study Inclusion Criteria

5. Had normal mobility, no known debilitating co-morbidities at time of study
6. Willing to participate for the duration of the study
7. Willing to carry smartphone at all times
8. Willing to download the smartphone-sensing app

The period of the study was extended to 20 weeks from the original 14 weeks to collect sufficient smartphone-sensing data. The study app passively and anonymously collected data regarding activity, mobility, sleep and communication from each participant. The actual conversation from the call was never collected. Socio-demographic data of participants was collected once at the time of enrolment. For identifying symptoms of depression among enrolled participants, a globally validated screening tool, PHQ-9 (Patient Health Questionnaire-9) was used [52, 53]. The first self-report depression score was assessed in person at the clinic followed by collection over telephone once every 2 weeks during the study period.

Enrolment

47 out of the 100 patients provided their consent and were enrolled at the clinic. Study administrators were identified from amongst the clinical staff. Administrator guidelines ensured effective participant enrolment and onboarding process. A web-based patient administration system, was used to manage participant details. Clinical staff were oriented about the study, familiarized about the study app and trained on the administrator system and guidelines.

All study participants followed a formal onboarding process where they were provided information about the study, educated about the privacy, security and consent process, had the study app setup, provided explicit consent and completed the initial PHQ-9 survey. The PHQ-9 survey was made available in both English and the local language (Marathi). The language modified version of PHQ-9 had been validated in other studies [54, 55]. In order for the participants to feel comfortable keeping data services enabled on their smartphones, they were provided with a 1GB data recharge per month to cover for usage costs. No other incentive was provided to the participants.

Data Collection

Privacy and security of the study data was an integral part of study design and execution. All data captured and processed for the study were after obtaining explicit consent from all participants.

Socio-Demographic Data

Participant’s socio-demographic information such as gender, marital status, occupation, age, education and family particulars were captured at enrolment. Personally identifiable information (name, email id and contact number) was securely stored in the administration system and used only for communication purposes. All other data was de-identified before use for research purposes.
**Passive Sensing Data**

Smartphone-sensing data was captured by the app automatically every 2 mins and stored on-device using a read/write SD Card. The app was designed to capture only hashed identifiers and the collected data was secured and anonymized on-device before being transferred to the storage servers for aggregate analysis. All transmissions were in encrypted form using HTTPS SSL. On the server side, these files were merged, parsed and synced by Python based post-processing infrastructure and stored in NoSQL based servers. The servers themselves and the data thereof were access restricted allowing only the engineering lead to retrieve the minimal needed data for research. The raw passive sensing data was processed and daily values derived for the sensing variables. Running of the ML models on the entire de-identified dataset was performed securely on the cloud with the research analyst getting to view only the performance results.

The social interaction data was captured from three main sensors and call logs. These were the accelerometer, the GPS and the ambient light sensor. All these sensors reflect pairwise communication and face-to-face proximity, intensity and nature of social ties, the dynamics of network and amount of light in the background. To ensure that no loss of sensing data occurred, in the event of network drops, it was stored in the participant smartphone for up to 3 days.

53 sensing variables were derived from the activity, mobility, sleep and communication data collected from smartphone sensors. *Figure 1* outlines the sensor-feature map. Activity variables were derived based on participant’s usage of the phone at various times in a day. Mobility variables were derived based on number of locations and the distance travelled. Sleep variables were based on relative gravity and screen on time. Call related variables were based on number of total calls, missed calls and call duration. The values for these variables were derived as day-wise aggregates from the raw sensing data collected for each participant.

![Figure 1: Sensor-Feature Map](image)

**Screening Instrument Data**

Screening instrument data included capturing of the participants self-reported depression score using the PHQ-9 survey. PHQ-9 consists of a validated 9-item depression screening tool, with each item having 4 options (scored 0 to 3) as responses namely; 'not at all', 'several days', 'more than half the days' and 'nearly every day'. PHQ-
9 helps screen for presence of and severity of depression with a maximum total score of 27. The PHQ-9 English and Marathi language questionnaire are exhibited in multimedia appendix 1 and 2 respectively. The PHQ-9 score was recorded at enrolment and once every 2 weeks during the study period. The values for each of the 53 derived sensing variables was aggregated day-wise whereas the screening data (PHQ-9 scores) was collected 14 days apart for each participant. Hence, it was decided to impute the average of two consecutive PHQ-9 submitted as the score for all the participant-day instances occurring between the two screening dates for each participant.

Data Quality
Of the 47 participants, only one participant did not complete the study. The final analysis was performed on data from 46 participants, of which 29 were men and 17 were women. Over the course of the study, a total of 2,694 "participant-day" records (instances) were collected. This formed the dataset for subsequent analysis.

All PHQ-9 surveys had reminders for the participants as well as administrators to follow-up. Opportunity to miss capture of sensing data could arise from many reasons and mostly these are random in nature. Some include, participants changing phone settings, forgetting to carry phones or are errors during capture, storage and manipulation. This resulted in values missing in some instances across the derived sensing variables. Sensing data quality was monitored automatically using custom scripts and alerts. Any gaps observed were followed up with respective participants towards closure.

Monitoring and Support
During the study period, active support was provided to ensure minimal dropouts, priority resolution of issues and monitoring of data quality. The administrators monitored the web administration system for alerts and reached out to participants as required. Alerts included unusual smartphone usage or when 2-week surveys were due and/or other reasons (for e.g. not receiving sensing data).

Smartphone Battery and Memory Optimization
The app technicals and proprietary data collection methods ensured that the participant's smartphone battery impact was kept low. The app occupied less than 10 MB of storage space on a typical android smartphone and consumed less than 2% of total battery. This was lower than that consumed by other apps usually installed on a smartphone and as measured over a 24 hour period.

Analysis
Data analysis was performed in three parts: descriptive, univariate and classification modeling.

Descriptive analysis was performed to understand the representation of the participants based on their socio-demographics, clinical presentations and their mental well-being.
**Univariate analysis** was performed to understand whether there were observable differences in behavior between set of instances tagged with symptoms of Major Depression (Depressed-D) and those with none (Not Depressed-ND). Depressed class included instances with PHQ-9 > 9 (moderate to severe severity) and Not Depressed class included instances with PHQ-9 <= 9 (none to mild severity). The PHQ-9 cut-off for major depression was decided based on published studies [56]. An independent t-test was carried out to compare the two classes. T-test has been found to be a robust statistical test for large number of instances [57]. According to central limit theorem, the means of large sample sizes are often well-approximated by a normal distribution irrespective of the population distribution. Hence, given the large number of instances available in each class, it was safe to assume that the data in each class was normally distributed for each derived sensing variable. It was also observed that the number of instances in the two class for each derived sensing variable were significantly unbalanced. It was therefore decided that the unequal variance independent t-test be applied (with and without outliers) to compare the two classes [58]. It was observed that any imputation of missing sensing values would potentially introduce bias and hence no treatment was affected on the missing values.

**Classification modeling** was performed with the objective to explore, compare and identify the best performing classifier method to build a risk stratification model for early detection of symptoms of depression (mild to severe) in participants with diabetes. Five supervised ML methods and their ensemble was explored. Tree-based supervised ML methods were mostly considered given their robustness to multicollinearity, outliers and missing values. These include: Support Vector Machine (SVM), Single Decision Tree (DT), Random Forest (RF), Adaptive Boosting (ADA) and XGBoost (XGB). RBF kernel SVM methods are known to handle large feature sets and their non-linear interactions. The other four methods were tree-based methods that included the basic DT along with boosting trees (ADA, XGB) and bagging trees (RF). Both bagging and boosting trees combine several decision trees to reduce error and improve classification performance. Boosting trees help reduce bias while bagging trees help reduce variance. A voting ensemble was also trained, that combined each of the five methods to check for improved classification performance. Each of the five methods provide a class (depressed or not depressed) prediction (vote) for each participant-day instance while the voting ensemble counts these votes and tags the majority class voted for that instance. Open-source python software on Jupyter Notebook was used for modeling.

A lower PHQ-9 cutoff of 5 was considered for the modeling to broaden the scope and include instances with mild symptoms of depression. Mild symptoms or subclinical depression in people with diabetes has been found to be common, associated with high levels of diabetes related distress, psychological distress and lower quality of life and also a risk indicator of major depression [59-61]. Those instances with a self-reported symptoms of depression (PHQ-9 >=5- mild to severe) were grouped under “Depressed-ML (D-ML)” class and those with none (PHQ-9 < 5) under “Not Depressed-ML (ND-ML)” class. Instances that contained missing values in any of the 53 derived sensing variable
were entirely dropped. The dataset so obtained was divided into training set (90%) and test set (10%) and a K-fold Stratified Cross Validation (CV) with K equal to 10 folds was performed on the training set for each of the five classifier methods. Refer Figure 2 for the modeling approach (Train-Validate-Test) followed. The K-fold CV has been widely used to compare different classifier methods. A major advantage of using 10-fold CV approach, in that every data instance gets to be in a validation set exactly once and gets to be in a training set 9 times, leading to lower variance in the resulting estimate. Stratification ensures that the ratio of both classes (D-ML and ND-ML) gets equally represented in each fold. A K-fold CV splits the training set into K equal parts. A model gets trained on K−1 parts and gets validated on the remaining part. This process leads to development of K models for each method. The average performance of K models then gets compared for each method. A nested K-fold CV was not opted for due to computational challenges.

The final model for each method was then trained on the entire training set and tested with the unseen test set. Test performance results were then compared to identify the best ML method. Classification performance in terms of accuracy, specificity, sensitivity, precision along with the confusion matrix was used to compare the ML methods. For this study, a lower number of false negatives (wrongly classifies symptoms of depression to be absent) was important as a wrong classification would lead to patients with symptoms of depression getting missed out for priority diabetes care. Hence apart from high accuracy, a high recall with a reasonably high precision formed the basis to compare and select the appropriate method to build risk stratification models.
Results

Descriptive Analysis
The study showed a noticeably high percentage of participants (31 out of 47) to have self-reported symptoms of depression (Figure 3). Of the 31 participants, one was known to have the symptom at baseline. 6 out of 31 reported severe symptoms of depression, including suicidal tendency. These cases were referred to psychologists appropriately.

![Figure 2: Classification Modeling (Train-Validate-Test) Approach](image)

![Figure 3: Prevalence of Depression](image)
At the beginning, 30 men and 17 women participated in the study. One male participant dropped out making it 29 men and 17 women at the end of study. As seen in Table 1, participants had a mean age of 35 years (std. dev= 12). 28 out of 46 (62%) participants were in the age group of 21 to 40. 27 out of 46 (60%) participants were married. Majority of the participants were office goers at 32 out of 46 (70%) while 8 out of 46 (17%) were students. 27 out of 46 (60%) of the participants held a bachelor’s or master’s degree while 16 out of 46 (34%) had completed schooling. Some of the clinical characteristics of the participants included an almost equal mix of condition (Type 1 / Type 2) and 29 out of 46 (62%) had moderate level of control on their illness.

<table>
<thead>
<tr>
<th>Participant Characteristics</th>
<th>Category</th>
<th>Absolute Value</th>
<th>Relative Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>15-20</td>
<td>4</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>21-30</td>
<td>14</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>31-40</td>
<td>15</td>
<td>32%</td>
</tr>
<tr>
<td></td>
<td>41-50</td>
<td>5</td>
<td>11%</td>
</tr>
<tr>
<td></td>
<td>51+</td>
<td>8</td>
<td>17%</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>29</td>
<td>64%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>17</td>
<td>36%</td>
</tr>
<tr>
<td>Marital Status</td>
<td>Single</td>
<td>18</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>Married</td>
<td>28</td>
<td>60%</td>
</tr>
<tr>
<td>Education</td>
<td>Grade 10-12</td>
<td>15</td>
<td>34%</td>
</tr>
<tr>
<td></td>
<td>Bachelors</td>
<td>19</td>
<td>41%</td>
</tr>
<tr>
<td></td>
<td>Masters</td>
<td>9</td>
<td>19%</td>
</tr>
<tr>
<td></td>
<td>Vocational Education</td>
<td>3</td>
<td>6%</td>
</tr>
<tr>
<td>Occupation</td>
<td>Student</td>
<td>8</td>
<td>17%</td>
</tr>
<tr>
<td></td>
<td>Home</td>
<td>6</td>
<td>13%</td>
</tr>
<tr>
<td></td>
<td>Office</td>
<td>32</td>
<td>70%</td>
</tr>
<tr>
<td>Family</td>
<td>Living Alone</td>
<td>4</td>
<td>9%</td>
</tr>
<tr>
<td></td>
<td>Family</td>
<td>42</td>
<td>91%</td>
</tr>
<tr>
<td>Chronic Condition</td>
<td>Diabetes Type 1</td>
<td>21</td>
<td>45%</td>
</tr>
<tr>
<td></td>
<td>Diabetes Type 2</td>
<td>25</td>
<td>55%</td>
</tr>
<tr>
<td>Patient Location</td>
<td>Outstation</td>
<td>14</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>In-City</td>
<td>32</td>
<td>70%</td>
</tr>
<tr>
<td>Level of Control over condition</td>
<td>Low</td>
<td>11</td>
<td>23%</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>7</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>28</td>
<td>62%</td>
</tr>
</tbody>
</table>

Table 1: Participant Demographics

Univariate Analysis
All the 2,694 instances (participant-day records) were included for this analysis. The results with and without outliers are summarized in Table 2 and 3:
1. A significant difference was observed in the average activity rates in the morning hours (from 6 AM until 11:59 AM) among those with symptoms of major depression (M=13.70, SD=14.04) compared to those with none (M=18.48, SD=18.44); P < .001. A significant difference was also observed in average activity rates in the remaining part of the day (from Noon until 4:00 PM) among
those with symptoms of major depression (M=16.06, SD=14.90) than those with none (M=18.79, SD=16.72); P = .005. These results suggest that those with symptoms of major depression exhibited lower and irregular activity rates through the day as compared to those with none.

2. A significant difference was observed in the number of screen-on times at night (hours after midnight till 6 AM), among those with symptoms of major depression (M=6.70, SD=9.33) compared to those with none (M=3.16, SD=8.91); P < .001. The results suggest that those with symptoms of major depression possibly had an impacted sleep quality due to higher screen-ons.

3. A significant difference was observed in the average total number of calls, among those with symptoms of major depression (M=12.61, SD=9.15) compared to those with none (M=22.28, SD=50.76); P < .001. A significant difference was also observed in the average number of people called, among those with symptoms of major depression (M=5.08, SD=3.83) compared to those with none (M=8.59, SD=7.05); P < .001. The results suggest that those with the symptoms of major depression maintained contact with fewer people and attended fewer calls.

4. Mobility variables showed limited to no statistical significance at 95% and 99% CI between the two depression states and hence not reported.

<table>
<thead>
<tr>
<th>Measurements</th>
<th>Activity Rate (AM)</th>
<th>Activity Rate (Day)</th>
<th># Screen-On (Night)</th>
<th>Total # Calls</th>
<th># of people called</th>
<th>Call Duration (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean values (D: PHQ-9 &gt; 9)</td>
<td>13.70</td>
<td>16.06</td>
<td>6.70</td>
<td>12.61</td>
<td>5.08</td>
<td>18.95</td>
</tr>
<tr>
<td>Mean values (ND: PHQ-9 &lt;= 9)</td>
<td>18.48</td>
<td>18.79</td>
<td>3.16</td>
<td>22.28</td>
<td>8.59</td>
<td>37.59</td>
</tr>
<tr>
<td>Std. Dev (D)</td>
<td>14.04</td>
<td>14.91</td>
<td>9.33</td>
<td>9.15</td>
<td>3.83</td>
<td>19.32</td>
</tr>
<tr>
<td>Std. Dev (ND)</td>
<td>18.44</td>
<td>16.72</td>
<td>8.91</td>
<td>50.76</td>
<td>7.05</td>
<td>174.88</td>
</tr>
<tr>
<td>P values with unequal variance</td>
<td>P &lt; .001</td>
<td>P = .005</td>
<td>P &lt; .001</td>
<td>P &lt; .001</td>
<td>P &lt; .001</td>
<td>P &lt; .001</td>
</tr>
<tr>
<td>n1 (D)</td>
<td>194 (11%)</td>
<td>228 (12%)</td>
<td>130 (9%)</td>
<td>262 (11%)</td>
<td>262 (11%)</td>
<td>262 (11%)</td>
</tr>
<tr>
<td>n2 (ND)</td>
<td>1598 (89%)</td>
<td>1761 (88%)</td>
<td>1301(91%)</td>
<td>2057 (89%)</td>
<td>2057 (89%)</td>
<td>2057 (89%)</td>
</tr>
</tbody>
</table>

Table 2: Univariate Analysis Results (with outliers)
Table 3: Univariate Analysis Results (without outliers)

Univariate trends over the weeks also show those with symptoms of major depression (D) exhibited irregular and lower day-time average activity rates (Figure 4) compared to those with none (ND).

\*n1: number of instances with values for Depressed. n2: number of instances with values for Not Depressed.
Trends over the week show those with symptoms of major depression (D) having irregular and higher average number of screen-ons at night-time (Figure 5) than those with none (ND).

![Figure 5: Trend of average Screen-ons (night)](image)

Trends over the week also show those with symptoms of major depression (D) withdrawing socially with lower average number of calls, lower average number of people contacted and lower average duration per call (Figure 6) than those with none (ND).

![Figure 6: Trends of average Calls-People-Duration](image)

Classification Analysis
All instances with missing values in any of the derived sensing variables were removed which resulted in 950 out of 2,694 instances available for analysis. A 90:10 (training: test) split resulted in 855 instances in the training set and 95 instances in the test set. A 10-fold stratified CV was performed on this training set before testing on unseen test set.
Correlation Findings

Very low correlation (r values ranging from -0.15 to 0.13) was observed between self-reported depression state (PHQ-9 cut-off 5) and each of the 53 derived sensing variable as measured by Pearson correlation.

High pair-wise correlation or collinearity (> 80%) was observed among some of the derived sensing variables (multimedia appendix 3). Mobility and Communication based variables showed higher collinearity. This was primarily due to association between the variables and their subset, for e.g. “total calls at peak” and “total calls at off-peak” variables were subsets of “total calls” variable. Activity and Sleep based variables showed lower collinearity. This was again due to the association between variables such as “total activity rates”, “screen-on times” and their “time of day” subsets. For e.g. “total activity by day/night/eve/am” were all subsets of the “total activity” variable. As each derived sensing variable, of its own, provided rich behavioral context, it was decided to retain all 53 variables and their subsets as inputs to modeling.

Model Development

Five ML methods (SVM, decision tree, random forest, adaptive boost and XGBoost) and their ensemble were trained and compared for performance. Accuracy, provided the fraction of correctly classified samples of both classes (D-ML and ND-ML). XGB and RF performed the best in terms of accuracy. They reported an average cross-validated accuracy of 79.1% (95% CI, 74%, 84%) and 78.3% (95% CI, 71%, 85%) respectively and a higher test accuracy of 81.1% and 80.0% respectively (Refer Table 4). Both the methods also reported a higher recall of 75.0% and 70.0% respectively and a reasonable precision of 78.9% and 80.0% respectively as compared to other methods. Recall can be interpreted as “Of the participant-day instances that were actually symptoms of depression, what proportion were classified as having symptoms of depression”. Precision can be interpreted as “Of the participant-day instances that were classified as symptoms of depression, what proportion actually had symptoms of depression”.

<table>
<thead>
<tr>
<th>Performance</th>
<th>SVM (RBF)</th>
<th>Decision Tree - Single</th>
<th>Random Forest</th>
<th>XGBoost</th>
<th>Adaptive Boost</th>
<th>Voting Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. cross-validation accuracy</td>
<td>73.8%</td>
<td>69.1%</td>
<td>78.3%</td>
<td>79.1%</td>
<td>74.3%</td>
<td>75.3%</td>
</tr>
<tr>
<td>95% CI</td>
<td>(67%, 81%)</td>
<td>(57%, 88%)</td>
<td>(71%, 85%)</td>
<td>(74%, 84%)</td>
<td>(67%, 81%)</td>
<td>(68%, 82%)</td>
</tr>
<tr>
<td>Test Accuracy</td>
<td>80.0%</td>
<td>66.3%</td>
<td>80.0%</td>
<td>81.1%</td>
<td>73.7%</td>
<td>77.9%</td>
</tr>
<tr>
<td>Precision (Test)</td>
<td>86.2%</td>
<td>60.5%</td>
<td>80.0%</td>
<td>78.9%</td>
<td>75.9%</td>
<td>80.6%</td>
</tr>
<tr>
<td>Sensitivity / Recall (Test)</td>
<td>62.5%</td>
<td>57.5%</td>
<td>70.0%</td>
<td>75.0%</td>
<td>55.0%</td>
<td>62.5%</td>
</tr>
<tr>
<td>Specificity (Test)</td>
<td>92.7%</td>
<td>72.7%</td>
<td>87.3%</td>
<td>85.5%</td>
<td>87.3%</td>
<td>89.1%</td>
</tr>
<tr>
<td>Confusion Matrix</td>
<td>Training counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positive</td>
<td>228</td>
<td>304</td>
<td>324</td>
<td>330</td>
<td>241</td>
<td>321</td>
</tr>
<tr>
<td>True Negative</td>
<td>487</td>
<td>503</td>
<td>521</td>
<td>520</td>
<td>485</td>
<td>519</td>
</tr>
</tbody>
</table>
Discussion

Principal Findings
A noticeably high prevalence of self-reported symptom of depression (~65%) was observed in this study as compared to the 8-35% normally reported in other studies [4]. This could be attributable to the single study site and the characteristics of the recruited participants. A detailed analysis was not in scope of the study.

Low correlation was observed between self-reported symptoms of depression and each of the derived sensing variable. This contrasts from the results observed in the Dartmouth Student Life study [41] where results indicated strong correlation between automatic sensing data (derived for sleep, conversation and location) and PHQ-9 scores. The difference can be attributed to the use of different set of derived sensing variables and possibly due to missing values in the sensing variables. The study did show, at 95% and 99% significance, lower levels of social contact (total calls made/received), higher phone access at night (number of screen-ons) and lower day time activity (activity rates during day) among those with self-reported symptoms of depression (PHQ-9 > 9).

The XGB and RF methods were able to classify each participant-day instance with a test accuracy of 81.1% and 80.0% and with a sensitivity/recall of 75.0% and 70.0% respectively. From amongst the recently published passive sensing studies, only one study [38] was found comparable to the approach followed in our study. In that they used a smartphone app to collect sensing data, used a PHQ-9 self-report scale but with a cut-off of 11 to separate participants into two classes and also built binary classifiers with leave-one-out cross validation approach to predict symptoms of depression. That study used two classification methods, namely RF and SVM which resulted in an accuracy of 61% and 59% and a sensitivity/recall of 62% and 72% respectively. While the results obtained from Project SHADO study did show better performance, it would not be appropriate to make a direct comparison given the different study design adopted by both papers. However, both studies did show a performance superior to a random classification.
Limitations
Key limitations of the study included a single study site, small participant size and the non-random and non-control group based approach. There is a need to investigate a better approach to estimate daily symptoms of depression for each participant instead of the imputation approach taken for this study. PHQ-9 although a validated scale to screen for symptoms of depression is not a tool to firmly diagnose depression. Hence participants with high PHQ-9 scores need not necessarily have depression and vice versa. Proactive follow-up with participants to fill missing sensing values would assist in increasing the availability of day-wise instance for analysis. It would also assist in obtaining a fairer picture into the association between derived sensing variables and symptoms of depression. The observations from this pilot study are at best preliminary and a larger, randomized control based study would help in validation of the findings.

Benefits
With increasing sophistication of ML methods and availability of large continuous streams of passive smartphone-sensing data, the study presents an approach to build risk stratification models that proactively classify symptoms of depression in patients with comorbid conditions such as diabetes and hence could complement care providers in their clinical care.

Conclusions
Smartphone sensor enabled daily digital observation of health and social attributes is a promising new approach with significant potential for management of comorbid conditions. Although the findings need to be replicated with a larger multi-site randomized control study, this observational study has opened up the possibility of understanding the real world everyday mental well-being and social attributes of people with diabetes in a clinical setting. Supplementing the smartphone-sensing data with clinical records from each visit along with daily behavioral information aggregated from a smartphone-based conversational agent would help further the risk stratification objectives. It is equally important to be sensitive and treat passive sensing data as sensitive health information and ensure adequate privacy and security controls prior to wider use.

Acknowledgements
Touchkin funded the pilot study and the publication fees for the paper. Touchkin provided access to their smartphone app product along with technical support during the study period. Participant recruitment and study administration was managed by clinical staff and administrators at Sarda Centre for Diabetes and Self-care.

The paper acknowledges the following for their support during the study. Dr. Banshi Saboo, MD, FIACM, FICN, FACE, MNAMS; a diabetologist associated with Dia Care & Hormone Clinic, Diabetology, Ahmedabad, India. Dr. Girija Chetty, Associate Professor, PhD associated with University of Canberra, Computer Science, Canberra, Australia. Stephen Raj and Priyadarsini Suresh, associated with PHFI-Indian Institute of Public Health, Health Informatics, Hyderabad, India.
The paper acknowledges Soundarya Murthy, content editor at Touchkin for her paper reviews; Himanshu Panwar, AI engineer at Touchkin along with Bindhya Rajendran and Aayush Agarwal, analytics interns at Touchkin, for their support in building and validating the ML methods.

**Authors Contribution**

Study concept and design was provided by AS with support from SM. Ethics committee approval for study was provided by SM. Study administration was provided by AS, SS and staff at diabetes clinic. Data collection, secure storage and data aggregation was managed by SS. Research and data analysis was done by VS, SM and SS. VS helped in drafting of the manuscript. Data interpretation was done by all authors. The paper was reviewed by all authors.

**Conflicts of Interest**

AS (first author) is also an advisor to Touchkin and without any fiduciary relation. SM (second author) is a program coordinator with PHFI and holds no conflict of interest. SS (third author) is an Engineering Lead and an employee of Touchkin. VS (fourth author) is an independent researcher with no fiduciary relationship with Touchkin and holds no conflict of interest.

**Multimedia Appendix**

1. Patient Health Questionnaire- 9 English language version
2. Patient Health Questionnaire- 9 Marathi language version
3. Pearson Correlation of derived sensing variables and depression state

**References**


Archived at: http://www.webcitation.org/6wTzEPULT

Archived at: http://www.webcitation.org/6wU1u4wUn


30 Doryab A, Min J, Wiese J, Zimmerman J, Hong J. Detection of behavior change in people with depression. AAAI; 2014 Presented at: AAAI Workshop on Modern Artificial Intelligence for Health Analytics; 2014; Québec City, Québec, Canada. ISBN: 9781577356691


Abbreviations
CI: Confidence Interval
DST: Department of Science and Technology
HTTPS: Hyper Text Transfer Protocol Secure (HTTPS)
NoSQL: Not Only SQL is an alternative to traditional relational database
PHQ-9: 9-item Patient Health Questionnaire. An assessment tool specific to depression
RBF: Radial Basis Function Kernel. Commonly used in SVM Classification
RCT: Randomized Controlled Trial
SSL: Secure Sockets Layer
XGBoost: Extreme Gradient Boosting. A recent machine learning algorithm