Crowdsensing: Assessment of Cognitive Fitness Using Machine Learning

Samin Ahsan Tausif *, Aysha Gazi Mouri, Ishfaq Rahman, Nilufar Hossain, and H. M. Zabir Haque

Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh;
Email: ayshagazi2477@gmail.com (A.G.M), ishfaq.rahman9@gmail.com (I.R), hossainnilufar12@gmail.com (N.H), zabir.haque.cse@aust.edu (H.M.Z.H)
*Correspondence: ahsantausif30@gmail.com (S.A.T.)

***Abstract*—The expanded use of smartphones and the Internet of Things have enabled the usage of mobile crowdsensing technologies to improve public health care in clinical sciences. Mobile crowdsensing enlightens a new sensing pattern that can reliably differentiate individuals based on their cognitive fitness. In previous studies on this domain, the visual correlation has not been illustrated between physiological functions and the mental fitness of human beings. Therefore, there exists potential gaps in providing mathematical evidence of correlation between physical activities & cognitive health. Moreover, empirical analysis of autonomous smartphone sensing to assess mental health is yet to be researched on a large scale, showing the correspondence between ubiquitous mobile sensors data and Patient Health Questionnaire-9 (PHQ-9) depression scales. This research systematically collects mobile sensors' data along with standard PHQ-9 questionnaire data and utilizes traditional machine learning techniques (Supervised and Unsupervised) for performing necessary analysis. Moreover, we have conducted statistical t-tests to find similarities or to differentiate between people of distinct cognitive fitness levels. This research has successfully demonstrated the numerical evidence of correlations between physiological activities and the cognitive fitness of human beings. The Fine-tuned regression models built for the purpose of predicting users' cognitive fitness score, perform accurately to a certain extent. In this analysis, crowdsensing is perceived to differentiate several people's cognitive fitness levels comprehensively. Furthermore, our study has addressed a significant insights to assessing people's mental fitness by relying upon their smartphone usage.**

***Keywords*—mental fitness, machine learning, crowdsensing, physiological functions**

I. Introduction

Depression is a mood disorder that causes a persistent feeling of sadness and loss of interest [1, 2]. It affects a person's emotions, thoughts, and behavior and can cause various emotional and physical problems, which may further create trouble in our daily activities. Collective efforts of researchers across diverse fields are involved in depression detection. Most of the data collection has extensively relied upon interview-based and mobile crowdsensing based techniques. Crowdsensing affects diverging people based on their mental fitness score [3].

Capponi and Fiandrino *et al.* [4] can be foreseen as a technical term consisting of a large cluster of people with mobile devices qualified for sensing, collectively sharing data, and extracting information to analyze, measure, and estimate any info of shared interest. Crowdsensing can also assess the mental fitness of people [5] besides the interpretation of depressive symptoms within patients via the standard PHQ-9 mental fitness measurement scale.

The Patient Health Questionnaire-9 (PHQ-9) is an instrument utilized for making a criteria-based diagnosis of depressive symptoms encountered by patients [1, 2]. Researchers that concentrate on identifying and evaluating the mental fitness of people have chronologically approached the PHQ-9 depression measurement scale for this purpose. However, the usage of mobile sensors in this domain has not yet been recognized or emphasized. A qualitative study by Mohr and Zhang *et al.* [3] demonstrated a framework for applying personal sensing to mental health where potentially large amounts of raw sensor data are converted into meaningful information related to behaviors, thoughts, and emotions can be performed. However, empirical analysis of autonomous and continuous smartphone sensing to assess mental health is yet to be researched on a large scale.

This research enables us to demonstrate the possible correlations between physiological activities and the mental fitness of human beings. The primary objective of this research study is to distinguish precisely the cognitive fitness levels of several participants via crowdsensing. Additionally, based on our outcomes from this analysis, we further presented a strong correlation between crowdsensing and the PHQ-9 cognitive fitness scale. This study collects mobile sensors' data and standard PHQ-9 questionnaire data from more than 200 participants. An android application "Sensor Data Collector" is used from Taqi and Rezwan *et al.* [6] of geofencing domain and we have further developed and optimized this application to collect valuable raw sensor data in a battery-optimized efficient manner [7] from which data has been received for approximately two weeks [8]. Mainly, students of various

Manuscript received September 29, 2022; revised November 16, 2022, accepted November 30, 2022; published June 16, 2023.

doi: 10.12720/jait.14.3.559-570

universities in Bangladesh have been focused on our research of evaluating mental fitness. Throughout the two weeks, we have collected precise, fine-grained data that would enable us to examine the mental fitness and depressive traits of different users on a large scale with the help of machine learning techniques.

The remaining article is organized as follows. Section II contains a brief representation of topics that are incorporated into our proposed system. Moreover, some existing works regarding crowdsensing and mental fitness are discussed in the same section. Section III elaborates a thorough explanation of the two data sets along with our proposed method. In Section IV, detailed scrutiny of the results of our proposed model is conducted and graphical correlations between physiological functions and mental fitness are illustrated. Section V shows a precise breakdown of our contribution and limitations. Finally, Section VI outlines the conclusion and future work of our research.

II. Background Study

In this section, we have exhibited the integral concepts of our study on mental fitness evaluation tools and techniques. In addition, this section also depicts some previous studies on this domain.

A. *Machine Learning*

Machine Learning (ML) stands as the scientific analysis of statistical models and algorithms that computer systems utilize to fulfill a specific task without being explicitly programmed [9]. Also, it can learn from data, recognize patterns and make decisions with minimal human intervention. Furthermore, machine Learning can be supervised or unsupervised. Supervised learning [9] is the machine learning task of learning a function that maps an input to an output based on standard input-output pairs. In contrast, unsupervised learning [10] can be regarded as discovering patterns in the data above and beyond what would be viewed as pure unstructured noise. Moreover, machine learning is a fast-growing trend [11] in the healthcare business because sensor data can be used to analyze a patient's health in real-time.

B. *Mental Fitness*

Mental fitness [12] exhibits an individual's emotional, psychological, and social well-being. Many factors contribute to mental health problems, including stress, social anxiety, depression, and personality disorders. The essence of machine learning algorithms and Artificial Intelligence (AI) can be availed [13] to predict the onset of mental illness. When implemented in real-time, such applications will help the community by performing as a monitoring tool for individuals with anomalous behavior.

C. *Related Work*

Recently, there has been a growing interest in using crowdsensing to infer human dynamics and mental fitness [3, 4]. Collecting data and generating precise models based on daily human activities are pretty challenging as it requires perceiving information based on unlabeled dataset. Even though there has been exploration on crowdsensing [14], there is little work on proposing the correlations between smart-phones' persistent and automatic sensing data and mental health outcomes such as PHQ-9. In this section, we have briefly introduced former studies on mental fitness assessment using smartphone sensor data.

Mohr and Zhang *et al.* [3] have proposed a framework and comprehensive practical analysis of how autonomous crowdsensing can hold great promise as a mechanism for conducting mental health research. Even though they have overviewed an architecture on how crowdsensing can be utilized to manifest people's mental health outcomes, they have not practically implemented any ground truth solution.

In contrast, Wang and Chen *et al.* [15] of have defined the evidence-based impact of workload on sleep, daily activity, stress, sociability, mental well-being, and academic performance of 48 students at Dartmouth College, through utilizing crowdsensing. Their study has strongly correlated with automatic sensing data and well-known mental fitness measures, specifically, PHQ-9 depression, Perceived Stress (PSS), and loneliness scales. Moreover, they have explored automatic and continuous smartphone sensing to investigate a student's mental fitness and behavioral patterns.

Capponi and Fiandrino *et al.* [4], Chen and Lin *et al.* [16] have attempted to draw fine distinctions between smartphone usage and sleep cycle. They have demonstrated a sensor-based inference algorithm on smartphones to optimally predict sleep duration by analyzing a collection of hints that tie sleep duration to smartphone usage patterns and environmental observations.

In a comprehensive study of stress monitoring, Ciman and Wac *et al.* [17] offered an empirical analysis that revolved around smartphone apps monitoring different physiological states correlated with stress without leveraging the sensitive private information of users. Collectively, these studies provide important insights into how crowdsensing can play a significant role in evaluating people's cognitive fitness or stress detection. However, such studies remain narrow in focus dealing with participants' privacy and illustrating correlation with standard cognitive fitness measurement scales (e.g., PHQ-9 depression, perceived stress).

Apparently, if we explore our attention to PHQ-9 mental fitness scales, there includes numerous crucial research. Kroenke and Spitzer *et al.* [1], Kroenke and Spitzer [2], and several others have proposed an efficient, reliable methodology, to measure depression severity, which is named as PHQ-9. Moreover, the questionnaire PHQ-9 has authentic reliability and its adequate validity has been tested [18]. Besides making criteria-based diagnoses, PHQ-9 can help substantially decrease depressive symptoms in people.

Considering all of this, our research has unfolded the apparent usage of crowdsensing platforms in assessing mental fitness and emphasized the interconnection of continuous and autonomous sensing data from smartphones and mental health measures such as PHQ-9. We have focused on implementing both supervised and unsupervised machine learning techniques to blend in a precise optimal outcome in our study of assessing mental fitness.

Furthermore, we have compared the findings, objectives, methodologies, and results of this study with the baseline studies and presented gaps in previous research in Table I. Additionally, we have illustrated the gaps fulfilled through this research.

TABLE I. COMPARATIVE ANALYSIS OF THE FINDINGS OF BASELINES STUDIES WITH THIS STUDY

Study	Mohr and Zhang *et al.* [3]	Wang and Chen *et al.* [15]	Capponi and Fiandrino *et al.* [4]	Thakur and Roy *et al.* [19]	Proposed study
Objectives	To utilize crowdsensing to evaluate people's cognitive fitness outcomes	To explore automatic and continuous smartphone sensing to investigate a student's mental fitness and behavioral patterns	To draw fine distinctions between smartphone usage and sleep cycle	To predict whether anybody has a mental health issue or not	To precisely distinguish the cognitive fitness levels of several participants via crowdsensing.
Methodology	Proposed a framework and empirical analysis of how cognitive fitness & behavioral markers assessment can be conducted based on data collection from smartphone sensors	Developed the "StudentLife" smartphone app to automatically infer human behavior and further defined the evidence-based impact of workload on sleep, daily activity, stress, sociability, mental well-being, and academic performance of students	Demonstrated a sensor-based inference algorithm on smartphones to optimally predict sleep duration by analyzing a collection of hints that tie sleep duration to smartphone usage patterns and environmental observations.	Derived physical activity and sociability features using smartphone usage and sensor data and developed a binary classification-based prediction model to predict mental health	Developed an android application for smartphone sensors data collection along with PHQ-9 Questionnaires data & physical activities data. Traditional ML techniques were used for performing necessary analysis. The fine-tuned regression models are built for predicting users' cognitive
Number of Users Consisting in Dataset	The authors proposed a framework but have not practically implemented it in their research	48 students at Dartmouth College participated in the dataset collection phase	8 participants	48 students at Dartmouth College participated in the dataset collection phase	Voluntary participation of 205 Users (The first of its kind large-scale research on this domain)
Smartphone Sensors Used	GPS, gyroscope, accelerometer & ambient light sensors were proposed in the framework	GPS, accelerometer & ambient light sensors were used	Accelerometer & ambient light sensors were used along with phone application features	GPS, accelerometer & ambient light sensors were used	GPS, accelerometer, gyroscope, light sensors, magnetometer, gravity etc.
Numerical Correlation between Physiological Function & Mental Health Shown or Not	No correlation was shown.	No correlation was shown.	No correlation was shown.	No correlation was shown	illustrated the numerical evidence of correlations between physiological activities and the cognitive fitness of human beings
Strong Correlation between Automatic Sensing Data and Mental Fitness Measures (PHQ-9) Shown or Not	No correlation was shown	PHQ-9 (post) & stress (value of *r*) 0.412 (*p*-value) 0.010	No correlation was shown.	Correlation was provided (Not visualized)	Demonstrated a strong correlation within this research
Results	Proposed Framework, No Practical Implementation	No Classification or regression task was performed	No results were provided	AUC in the prediction of depression: 74%	R2 Score Support Vector Regressor: 0.947 Gradient Boosting Regressor: 0.958
Key Findings	The feasibility of crowdsensing for mental health has been demonstrated, along with its challenges	Number of insights into behavioral trends, and importantly, correlations between objective sensor data from smartphones and mental well-being and academic performance	Presented a sensor-based computational model that provides automated sleep duration monitoring using smartphones	Developed a framework for automated prediction of mental health conditions but the models failed to predict the cognitive fitness score of users. They could only detect if depression existed or not.	Addressed significant insights into assessing people's mental fitness by relying on smartphone usage. Demonstrated correlations between ubiquitous sensor data from smartphones and the cognitive fitness of users. Regression models are able to predict Users' cognitive fitness

III. Proposed Methodology

In this section, a comprehensive explanation of the two dataset is elaborated. Moreover, detailed description of the techniques we utilized for working with both unlabeled and labeled data in our research study is precisely presented.

A. Dataset

In this subsection, the methodology of perceiving the dataset is elaborately described. Different approaches have been taken to collect both necessary labeled and unlabeled dataset for the intrinsic purpose of evaluating mental fitness of several participants.

1) Dataset-I

Dataset I have been collected via a robust android application, "Sensor Data Collector" developed to accumulate smartphone sensor data. Data were collected from accelerometer, gravity, gyroscope, magnetometer, orientation and GPS sensors. Gyroscope, gravity, accelerometer, and magnetometer data are recorded with their tri-axial values. GPS comprehends the device's increments in altitude, latitude and longitude. For the reliable and authenticated purpose of detailed documentation and cloud-based storage, we have managed Firebase. Autonomous sensor data was fetched from mobile phones and pushed to the cloud storage after a fixed interval of time concerning the cost. A graphical representation of our research's data collection is presented below in Fig. 1.

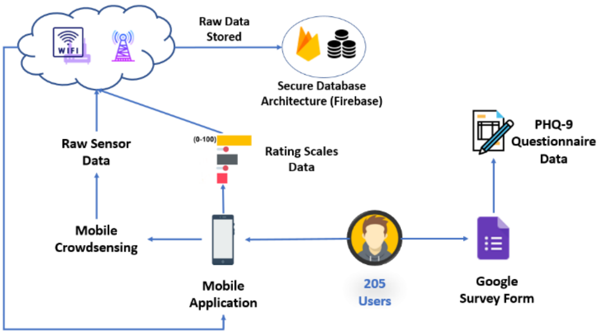
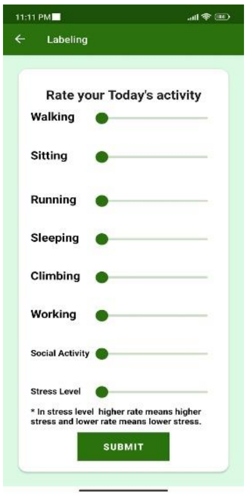


Figure 1. Architecture of data collection.

In our android application, we have also provided a questionnaire section (Fig. 2). In the questionnaire section, five physiological functions (walking, running, sleeping, climbing, sitting) were furnished along with social activity, working and stress level rating scales. In addition, the social interaction and working rating scales and all five physiological function rating scales were built to provide ratings from 0 to 100. The significant reason behind selecting range between 0 to 100 is that, it would be more preferable for users to accurately provide ratings if they assumed their whole day physical & social interaction activity on a percentage ratio rather than 1–10 scales. Users were asked to contribute their ratings on those scales based on how much physiological activities they had performed, how much they were involved in social interaction, and how much mental stress they were in during the whole day.

A stress level rating scale was developed to rate between 0–27, considering the standard PHQ-9 mental health measurement. Users were informed that a higher rate in physiological function rating scales means higher productivity in work and higher physiological activity in a day. In contrast, a higher rate of stress level means higher stress and vice-versa. The rating scales were added to perceive information regarding users' physiological and social interactive functions. Another intention was to highlight a perception of how physiological processes alleviate strong interdependence on mental fitness, which we have visually shown later by merging the PHQ-9 mental health measurement data and physiological functions rating scales data.



Rate your Today's activity

Walking

Sitting

Running

Sleeping

Climbing

Working

Social Activity

Stress Level

* In stress level higher rate means higher stress and lower rate means lower stress.

Figure 2. "Sensor data collector" android application.

We have conducted this data collection phase for around 22 days to attain a distinguished mental health fitness pattern [8] within users. After the completion of 22 days data collection phase, we obtained 21690 recorded samples of autonomous smartphone sensor data. For labeling purpose, we have collected 597 rating scales data from same 205 users.

TABLE II. Sample Data from Android Rating Scale (Physiological Functions)

User ID	Mental Health	Running	Sitting	Sleeping	Social Activity	Walking	Working
1	26.35%	24.25%	15.33%	18.71%	48.09%	29.55%	68.09%
2	3.75%	19.68%	8.82%	18.24%	65.53%	48.40%	23.06%
3	24.26%	18.48%	59.42%	0.00%	15.22%	41.30%	0.00%
4	3.42%	0.00%	14.13%	44.05%	18.23%	54.19%	57.57%

The bias in providing ratings in various rating scales was acknowledged beforehand. Therefore, users were requested to provide accurate input based on their daily physiological activities, social interaction, and stress level to acquire efficient and credible data.

In Table II, the sample data of physical activities and social interaction for random four users have been demonstrated. The hashed user ID (for example hashed ID = 5iytFliuAMh2mrcNXzqIeQJsc3B2) for each user is not shown due to privacy issues, rather each user is labelled with numbers such as 1, 2, 3, and 4.

Privacy concerns were taken into significant consideration in our research study. Focusing on protecting participants' personal information [20], we have entirely anonymized each participant's identity with a random hashed user id. We have also preserved the user id separate from all other task data so that the data cannot be traced back to individuals. Data is stored on secured servers. Users' information is uploaded using encrypted SSL connections to certify that their data cannot be intercepted by third parties. Consent forms have been collected from 205 participants which depict the fact that they have willingly participated in our research.

2) *Dataset-II*

We have accommodated google forms for our second dataset to perform surveys on 205 potential users. In the google forms, we have furnished nine standard questionnaires bearing in mind the standard PHQ-9 mental health measuring methodology beforehand. Questions related to fatigue, sleep disruption or insomnia, poor concentration, suicidal thoughts, hopelessness, etc., were prioritized within the nine questions.

TABLE III. PHQ-9 OPTIONS

Numerical Value	Option Meaning
0	Not at All
1	Several Days
2	More than Half the Days
3	Nearly Everyday

As a severity measure, the PHQ-9 score ranges from 0 to 27 because each of the nine items can be scored from 0 ("not at all") to 3 ("nearly every day") [1, 2] (Table III). For the intrinsic purpose, we have calculated the mental fitness score of every user and further, based upon that, we have added an extra feature named mental fitness score in our PHQ-9 evaluation dataset.

It provides a profound benefit as we can perceive probable information about the cognitive fitness of a participant in our research study.

Apparently, for analysis, the PHQ-9 score was divided into the following categories [1, 2] of increasing severity: 0–4, 5–9, 10–14, 15–19, 20–27 from Table IV. Category 0–4 represents minimal depression, the category 5–9 enlightens mild depression, the category 10–14 emphasizes moderate depression, the category 15–19 portrays moderately severe depression and lastly, the category 20 to upward depicts severe depression within a participant.

TABLE IV. PHQ-9 SCORES

PHQ-9 Score	Depression Severity
0–4	Minimal
5–9	Mild
10–14	Moderate
15–19	Moderately Severe
20–27	Severe

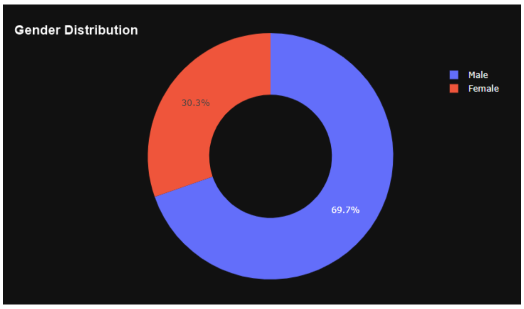


Figure 3. Percentage of male and female users.

From Fig. 3 above, we can visualize that around 30.3% of the 205 survey participants were female and the rest 69.7% participants were male.

In Table V, the sample data of PHQ-9 questionnaires for random four users have been demonstrated. The hashed user ID (for example hashed ID = 5iytFliuAMh2mrcNX) for each user is not shown due to privacy issues, rather each user is labelled with numbers such as 1, 2, 3, and 4.

TABLE V. SAMPLE DATA FROM GOOGLE FORM SURVEY (PHQ-9)

User ID	Little Interest	Feeling Hopeless	Sleep Disruption	Fatigue	Poor Appetite	Low Self Esteem	Less Concentration	Weakness	Suicidal thoughts
1	1	0	2	0	1	1	0	2	1
2	3	3	1	0	1	3	1	0	1
3	0	3	3	3	3	3	3	3	3
4	1	1	0	1	0	0	0	1	0

B. *Proposed Architecture (Unlabeled Data)*

We have collected raw sensor data from around eight sensors available in smartphones through our android application. The proposed methodology of working with unlabeled data using unsupervised learning within our research consists of eight stages that are illustrated in Fig. 4.

1) *Data pre-processing*

Prior to dataset scaling, data pre-processing and data cleaning has been performed across all datasets to get rid of redundant and null data. Users were requested to keep their location on to receive GPS longitude and latitude data. But, as some users failed to keep the location on for a certain amount of time, null values were obtained in our research for GPS sensors. As a consequence, 5663 recorded samples consisting of null values of GPS sensor longitude & latitude values, that can be assessed as invalid data, are removed from the dataset. As a result, we are left with 16,027 recorded autonomous mobile sensor data for our research purpose.

2) Outlier detection and removal

Outliers in dataset can negatively impact as they can skew and mislead the training process of machine learning algorithms, which can further result in less accurate models and poor outcomes. Therefore, we have scrutinized and detected the potential outlier percentage for GPS sensor & accelerometer triaxial data and have removed those outliers.

After removing the outlier and redundant data, we are left with 8642 recorded samples of 71 unique android users from 205 individual users and 16,027 recorded samples.

Therefore, the size of the gap between originally fetched data and pre-processed data is large due to the impact of outliers and redundant null GPS values. If less outliers and lesser number of null GPS values existed in the dataset, then better results could be expected from the number of existing samples in the research.

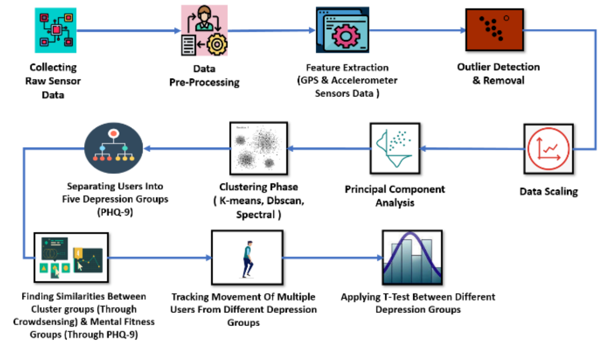

Collecting Raw Sensor Data
Data Pre-Processing
Feature Extraction (GPS & Accelerometer Sensors Data)
Outlier Detection & Removal
Data Scaling
Principal Component Analysis
Clustering Phase (K-means, Dbscan, Spectral)
Separating Users Into Five Depression Groups (PHQ-9)
Finding Similarities Between Cluster groups (Through Crowdsensing) & Mental Fitness Groups (Through PHQ-9)
Tracking Movement Of Multiple Users From Different Depression Groups
Applying T-Test Between Different Depression Groups

Figure 4. Flowchart of our proposed model.

3) Data scaling

In our research, we have performed standard scaling operations feature-wise independently to manage highly varying values and resize the distribution of values. The sensor-based dataset of this research comprises instances from embedded sensors containing information from diverse motions and positions. As a result, the ranges and magnitudes of the features differ significantly. Moreover, the outcomes would fluctuate extensively amongst units. Likewise, while estimating distance between sample points, features of high magnitudes will matter much more heavily than features with low magnitudes. Consequently, our proposed model will be inaccurately biased towards sensor aspects consisting of high magnitude. Therefore, we need to bring all features to a similar magnitude level to neutralize this impact. This is where feature scaling plays a vital role. In our case, we utilized normalization as our scaling strategy.

Normalization is performed during the preprocessing phase. However, we have performed it after feature selection to analyze the variance more appropriately. L2 norm was utilized in this research, which is the square root of the sum of the squared values. Squaring the values emphasizes large values and less influence on small ones.

4) Principal component analysis

Afterwards, we implemented Principal Component Analysis (PCA). PCA is a strategy for decreasing the dimensionality of such datasets while it improves interpretability and minimizes information loss. In our research, we have only taken tri-axial accelerometers and GPS longitude and latitude sensors for analytical consideration. However, a higher number of features could cause an overfitting issue due to increased noise. Moreover, it was challenging to visualize and interpret the data in high dimensions.

Therefore, we have utilized PCA to transform high dimensional data to low dimensional data (2D), so it can be easily visualized. PCA also helps to overcome the overfitting issue by lessening the number of features and reduction of noise.

5) Clustering phase

To grasp an understanding of the optimal number of clusters into which data may be clustered, the elbow method has been availed. From the elbow method, we have decided that the number of clusters of 5 or 6 can be appraised as ideal for clustering.

Considering the variation and dimension of our dataset, we have maneuvered three potential clustering algorithms named K-Means, Spectral and DBSCAN Clustering to better analyze and cluster our dataset according to our expected outcome. Determining the optimal number of clusters for a data set is an important problem in certain clustering algorithms. The silhouette score is one of the various strategies for determining the ideal number of clusters [21]. Determining the optimal number of clusters for a data set is an important problem in certain clustering algorithms. The silhouette score is one of the various strategies for determining the ideal number of clusters [21].

K-Means clustering has manifested lower silhouette scores, while Spectral DBSCAN clustering algorithms have emerged with remarkable silhouette scores (Table VI). In our case, the silhouette score that we got is 0.708 which portrays that the samples are differentiable by the decision boundary between two neighboring clusters.

The DBSCAN algorithm [22] is most likely used when data is robust with outliers. The reason behind the significant performance of DBSCAN clustering algorithm (Fig. 5) over K-Means is that, DBSCAN can potentially handle clusters of multiple sizes and structures and is not powerfully influenced by noise or outliers. While, K-means has difficulty with nonglobular clusters and clusters of multiple sizes. We have acknowledged the clusters of DBSCAN Clustering algorithm for our further empirical analysis of separating users based on autonomous smartphone sensor data.

TABLE VI. SILHOUETTE SCORE OF DIFFERENT CLUSTERS

Clustering Models	Silhouette Score
K-Means Clustering	0.462
Spectral Clustering	0.665
DBSCAN Clustering	0.708

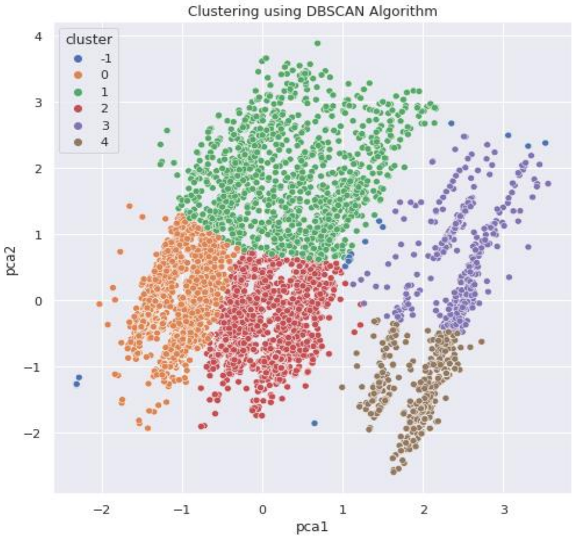


Figure 5. Clustering using DBSCAN clustering algorithm.

C. *Proposed Architecture (Labeled Data)*

For labeled data, we have two different sets of data to analyze. One is the survey form data that depicts the mental fitness of survey participants. At the same time, another one is the labeled rating scales data which represents the evaluation of physiological functions and social interaction rates of the participants.

The proposed methodology of working with labeled data using supervised machine learning techniques within our research consists of six stages that are illustrated in Fig. 6.

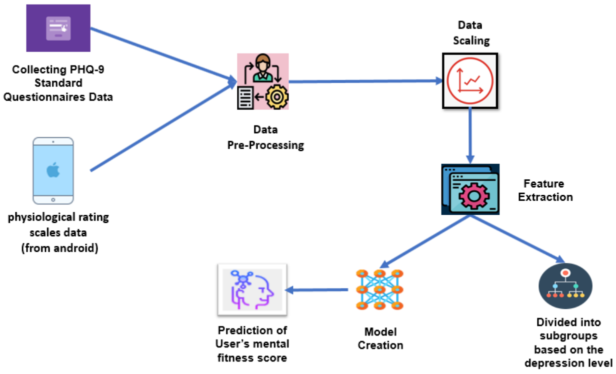

Collecting PHQ-9 Standard Questionnaires Data
physiological rating scales data (from android)
Data Pre-Processing
Data Scaling
Feature Extraction
Divided into subgroups based on the depression level
Model Creation
Prediction of User's mental fitness score

Figure 6. Flowchart of proposed method for working with labeled data.

D. *Data Scaling, Feature Extraction and Implementation*

Before building and evaluating models, we independently carried out standard scaling operations and split the data into two sets for training testing purposes. We have used all the physiological functions' ratings, social interaction rating, and PHQ-9 standard nine questions as features for our models and have employed the Mental fitness score as our target variable or label to make predictions. Furthermore, we have utilized six different regression models in our research to acknowledge a clear understanding of which models are suitable for our dataset.

IV. Experimental Results and Evaluation

A more detailed description of the outcomes of proposed methodology is given in the following section. Besides, the evaluation of regressor models built upon multiple features will be demonstrated in this section.

A. *Separating Survey Users into Multiple Depression Groups*

In addition, we have distinguished survey participants based on their mental fitness scores. Fig. 7 visually represents the number of people in different mental health depression groups. The plot amplifies the fact that there are around 72 people who exist within the range of 5–9 cognitive fitness score groups that resemble mild depression. 31 participants lie within the scope of 0–4 fitness score group (minimal depression), 62 participants dwell within the 10–14 score group (Moderate Depression), and 23 participants exist in the range of 15–19 score, which exaggerates moderately severe depression.

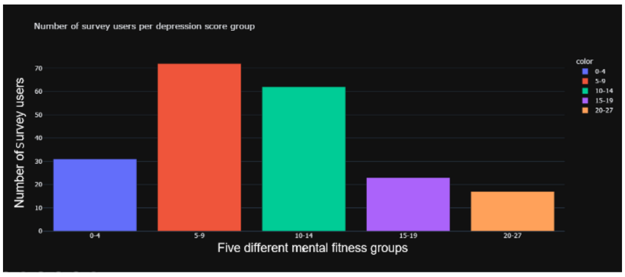


Figure 7. Number of people per depression group according to fitness score.

The matter of great concern is that the bar plot vividly represents that around 18 participants are in the severe depression region (mental fitness score of 20–27), and henceforth they are in need of medical treatment according to their potential depressive traits.

B. *Correlation between Physiological Functions and PHQ-9 Measurements*

We have merged both sets of datasets into one to blend in a visual representation of how physiological functions possess a profound correlation with the mental fitness of participants. Prior to merging both datasets, data pre-processing and data cleaning were performed across all datasets to get rid of redundant and null data.

From the heatmap below, we manifest a significant conception about how physiological functions and social interactions in the daily living of humans unfold a radical correlation with the cognitive fitness of people.

Fig. 8 visually shows that physiological functions and social interaction have a negative and concrete relation with depressive mental health traits such as little interest, hopelessness, low self-belief, insomnia, less concentration, and suicidal thoughts.

In Table VII we have illustrated the interpretation [23] of the range of values of correlation coefficient of a heatmap. Additionally, in Table VIII we have demonstrated the potential numerical correlation between physiological functions & PHQ-9 questionnaires data found from the significant insights of the heatmap (Fig. 8).

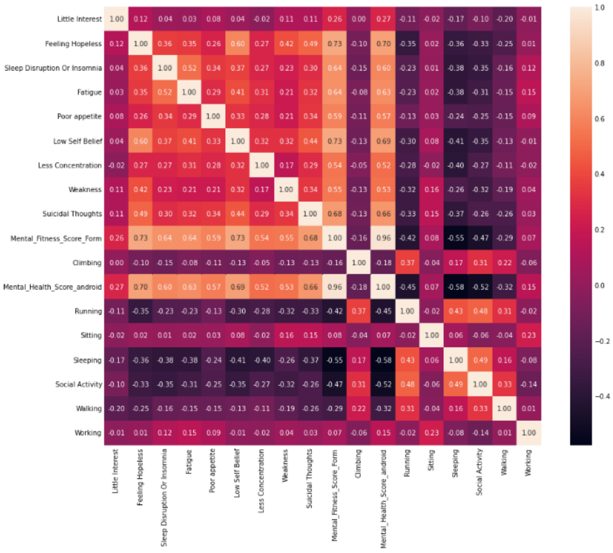


Figure 8. Visual correlation between physiological functions and PHQ-9 measurements.

TABLE VII. CORRELATIONS

Range	Correlation
−1<value<0	Negative Correlation
0–0.25	No or Poor Correlation
0.25–0.5	Below Average Correlation
0.5	Medium Correlation
0.5–0.75	Fair Correlation
0.75–1	Strong/Positive Correlation

A possible explanation for this outcome might be that more physiological functions in daily life tend to bring an enormous sense of well-being [24]. In addition, they relieve stress, improve memory, help people sleep better, and boost their overall mood. Social interaction also plays a pivotal role [25] in the reduction of depressive symptoms. The subsequent urge of humans to connect with others and gain acceptance into social groups is fulfilled through social interaction that further decreases the possibility of inheriting depressive traits [25] in life.

TABLE VIII. CORRELATIONS BETWEEN PHYSIOLOGICAL FUNCTIONS AND PHQ-9 SCORES

Physiological Function PHQ-9 Questionnaire	Sleeping	Running	Working	Walking	Social Activity	Mental Fitness Score (0–27) (Provided by Users in Android)
Little Interest	−0.17	−0.11	−0.01	−0.20	−0.10	0.27
Feeling Hopeless	−0.36	−0.35	0.01	−0.25	−0.33	0.70
Sleep Disruption	−0.38	−0.23	0.12	−0.16	−0.35	0.60
Fatigue	−0.38	−0.23	0.15	−0.15	−0.31	0.63
Poor Appetite	−0.24	−0.13	0.09	−0.15	−0.25	0.57
Low Self Esteem	−0.41	−0.30	−0.01	−0.13	−0.35	0.69
Less Concentration	−0.40	−0.28	−0.02	−0.11	−0.27	0.52
Weakness	−0.26	−0.32	0.04	−0.19	−0.32	0.53
Suicidal Thoughts	−0.37	−0.33	0.03	−0.26	−0.26	0.66
Mental Fitness Score (0–27) (Based upon Questionnaires Data)	−0.55	−0.42	0.07	−0.29	−0.47	0.96

C. *Correlation between Different Depressive Symptoms and Mental Fitness*

Fig. 9 illustrates an outline that increase in suicidal thoughts, based symptoms leads to increase in mental fitness score which resembles higher depression [26] within participants.

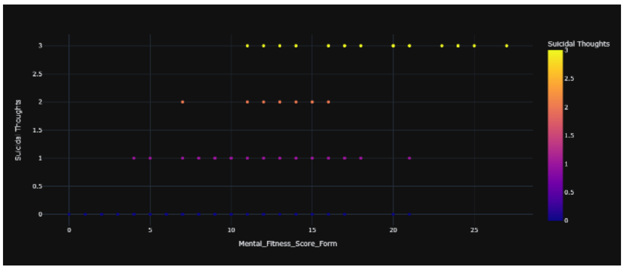


Figure 9. Correlation between suicidal thoughts and PHQ-9 scores.

If we observe Fig. 10, where the correlation between sleeping with mental fitness has been summarized. A higher rate of sleeping radicates proper sleeping within people that obtain a healthy mindset for the rest of the day and eliminates insomnia or sleep disruption [16]. Proper sleeping improves overall mood and relives stress. Therefore, the risk of fatigue and depression gradually decreases.

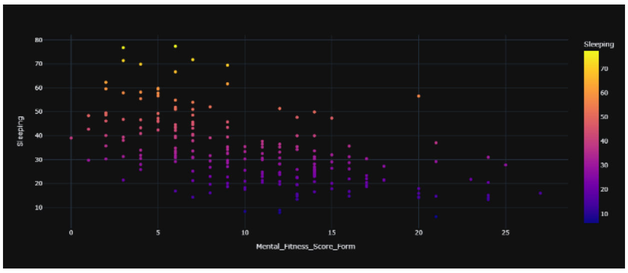


Figure 10. Correlation between physiological function ‘sleeping’ and mental fitness.

D. *Evaluation of Regression Models*

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are two of the most standard metrics utilized to measure accuracy for continuous variables. The MAE calculates the average volume of the errors in a set of forecasts and it is robust to outliers. In the case of RMSE, the errors are squared before they are averaged. Therefore, this metric gives a moderately high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable. For MAE and RMSE, it is recommended that the lower the value, the better the model's performance.

Meanwhile, the excellence of fit of regression models can be analyzed based on the R-squared process. It is advised that, the more the value of R-square near 1, the better is the model. We have implemented six different regression models in our research to acknowledge a clear understanding of which models are suitable for our dataset.

Furthermore, if we observe the remaining three models (shown in Table IX) such as, Support Vector Regressor, XGB Regressor and Gradient Boosting Regressor, we can perceive that MAE and RMSE errors have been greatly reduced than the previous three models. MAE is in the region of 0.68 to 0.97, while RMSE is in the region of 1.13 to 1.24 which can be considered as remarkably better. In addition, the R squared score results of these three models are even better. SVR, XGB and GB Regressor have provided R squared score of 0.94, 0.951 and 0.958 respectively.

TABLE IX. EVALUATION METRICS OF DIFFERENT REGRESSION MODEL

Model Name	Mean Absolute Error	Root Mean Squared Error	R Squared Score
Random Forest Regressor	1.494146	1.852094	0.852601
AdaBoost Regressor	1.391137	1.716705	0.867831
KNN Regressor	1.445886	1.785572	0.892838
Support Vector Regressor	0.683656	1.169604	0.946866
XGB Regressor	0.972755	1.242055	0.951142
Gradient Boosting Regressor	0.887458	1.135959	0.958008

Hyperparameter tuning has played a significantly vital role in the major improvement of these models. For example, using sigmoid as kernel for SVR improved its potential outcome and reduced MAE and RMSE. Adjusting the loss function to huber for Gradient Boosting Regressor has improved its prediction accuracy.

Overfitting could be a major concern for our model creation as we have lesser amount of labeled data to train and test our regression models. To counteract that, our models have been tuned in a way to be robust with overfitting and these results (illustrated in Table III) are representing the exact facts.

E. Finding Similarities between Cluster Groups and Depression Group (Permutation)

To clarify which sample belongs to which cluster, we have computed and labeled samples according to the cluster they belong to. Furthermore, we have grouped users according to the five different clusters (through crowdsensing) they are associated with and also in accordance with the five mental fitness score groups (through PHQ-9). Subsequently, we have searched for similarities and differences between users in cluster groups and five different mental health depression groups.

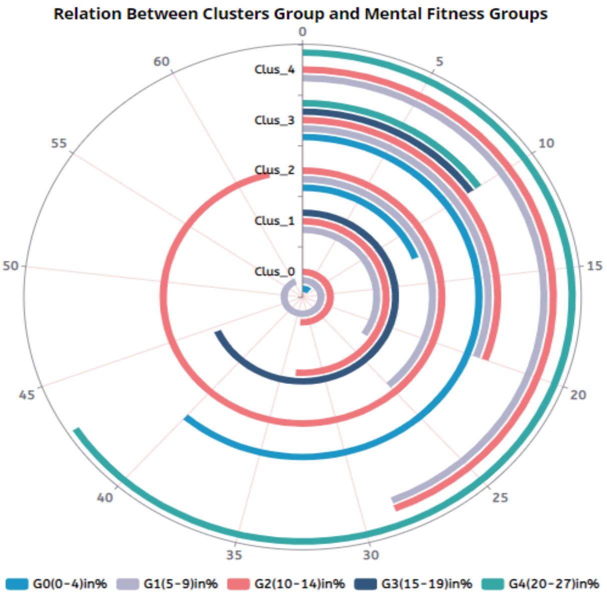


Figure 11. Stacked radial chart.

Fig. 11 illustrates the resemblance and discrepancy and portrays a comprehensible insight of which cluster group is mostly identical to which mental health depression group. If we visualize the stacked radial chart, we can observe that approximately 61% similarity exists between cluster 0 and mental fitness group 5–9 (mild depression). Furthermore, approximately 62.5% closeness can be anticipated between cluster 2 and mental fitness group 10–14 (moderate depression). 40% similarity can be observed between cluster 3 and mental fitness group 0–4 (minimal depression). And lastly from a thorough empirical observation, correlation of almost 43% between cluster 4 and mental fitness group 20–27 (severe depression) can be visualized. In the illustrated stacked radial chart, G0 G1 G2 G3 G4 actually represents minimal, mild, moderate, moderately severe, severe depression group respectively.

From the stacked radial chart, we can proclaim an analogy that cluster 0 has the most equivalence with the mental fitness score group of range within 5–9(mild depression), while cluster 2 has the most similarities with the score group of 10–14(moderate depression). Likewise, cluster 3 blends most closely to the group of range 0–4 (minimal depression). Lastly, cluster 4 unfolds most resemblance to the mental fitness score group of 20–27 that interprets severe depression within survey participants.

F. Applying T-Test between Different Depression Groups

We have performed T-Test between multiple depression groups to highlight similarity and variance among five mental fitness clusters acknowledged after data pre-processing and applying multiple clustering algorithms (see Table X).

TABLE X. T-TEST SCORES

Independent Two Samples (T-Test)	Difference of Mean	Degree of Freedom	T	Two Test *P* value	Difference of <0*P* Value	Difference of >0*P* Value	Pearson's Relation
Cluster 0 and 1	−5.3899	25	−4.1887	0.0003	0.0002	0.9998	0.6422
Cluster 0 and 2	−1.2639	32	−1.1596	0.2548	0.1274	0.8726	0.2008
Cluster 0 and 3	0.0111	26	0.0063	0.995	0.5025	0.4975	0.0012
Cluster 0 and 4	−6.6746	23	−3.6497	0.0013	0.0007	0.9993	0.6056
Cluster 1 and 2	4.125	23	2.8286	0.0095	0.9952	0.0048	0.508
Cluster 1 and 3	5.4	17	2.2155	0.0407	0.9797	0.0203	0.4733
Cluster 1 and 4	−1.2587	14	−0.5121	0.6165	0.3083	0.6917	0.1356

We can visualize that the T-score between cluster 0 (mild depression — score range 5–9) and cluster 3 (minimal depression — score range 0–4) is only 0.0063. Lower T-score validates that the two sampling sets have quite similarities between them. The two mental depression groups, minimal and mild depression, are pretty similar due to the closer range between their mental fitness scores.

Meanwhile, if we observe the T-score of cluster 0 (mild depression — score range 5–9) and cluster 4 (severe depression — score range 20–27), we can examine that the t-score is 3.6797. If we take the absolute value, it can be regarded as 3.6797, which portrays that the mental fitness score and depressive symptoms of users with mild depression differ a lot from the cognitive fitness score and cognitive depressive traits of users with severe depression. The higher the t-score value, the more dissimilarity between the two sample sets.

Advanced degrees of freedom generally mean larger sample sizes. A higher degree of freedom means further power to reject a false null hypothesis and find a significant result. We can examine that the degree of freedom varies from 14 to 32 for different two-sample independent t-tests. Henceforth, it can be acknowledged as a positive thing as it means additional power is there to reject a false null hypothesis.

G. *Tracking User Based on Movement Intensity of Multiple Depression Group*

We have apparently decided to adopt a novel approach to blend in perception or to prove the methodology of finding accurate similarities between five cluster groups and five mental fitness score groups. We have calculated the average movement (in kilometers) for three random users from each of the five depression cluster groups.

Through this approach, two different philosophies can be proposed. One of them is that the users of the cluster group representing minimal depression will potentially have more average movement (in kilometers) than the other four cluster groups. Meanwhile, the users of the cluster group emphasizing severe depression will alarmingly have the fewest average movement [27] among the five different depression cluster groups.

In addition, the perception is that people who tend to indulge in more physiological functions possess lesser depressive symptoms and lower cognitive fitness score. While, people who are involved in fewer physiological activities, inherit more depressive traits and encounter moderate to severe depression in their daily life which hampers their lifestyle.

Fig. 12 illustrates the average movement (in km) of three random users selected from each of the five different depression clusters.

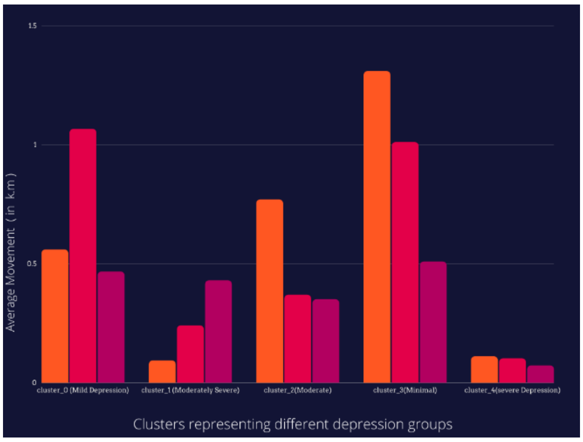


Figure 12. Average movement of participants from multiple depression groups.

If we explore the bar chart, we can visualize that the three random users of cluster 0 which portrays mild depression, augment average movement in the range of 0.6 km to 1.1 km within the three weeks data collection phase.

The random users of cluster 1 representing moderately severe depression tend to have fewer average movement (in the range of 0.1–0.4 km), while the three random users of the moderate depression group (cluster 2) have slightly higher average movement (0.4–0.8 km) than users of cluster 1.

But a significant increase in average movement can be profoundly noticed in the bar chart for cluster 3 portraying minimal depression. If we interpret numerically, the three random users of cluster 3 (minimal depression) have an average movement of 0.5 km to 1.3 km within the three weeks data collection phase, which is the maximum ratio among all five different mental fitness groups. This establishes the notion that people with more physiological activities in their daily life, tend to inherit lesser cognitive depression symptoms [24] in their life period.

V. DISCUSSION AND LIMITATIONS

Based on both the quantitative and qualitative results from our study, our observations & findings show that our research of evaluating the cognitive health of people formulated on crowdsensing can be assessed as a promising and novel approach. We have unfolded several contributions in this region of research, where fewer studies have been conducted with an intrinsic correlation shown between standard PHQ-9 mental health measures and crowdsensing.

Wearable sensors place a relatively huge burden on the users to ensure mental fitness-related data is collected correctly [16, 28]. In comparison, our approach to evaluating mental fitness via autonomous smartphone sensing and PHQ-9 standard measures possesses a lower user burden. Smartphone sensing requires no additional hardware and therefore monitoring of mental fitness can be nearly free of cost [14].

To the best of our knowledge, we are the first to work on the mental fitness of users on a large scale using ubiquitous smartphone sensors. Our research represents the cognitive fitness of 205 participants, enabling us to examine the depressive symptoms within the participants on a large scale. In comparison, prior studies regarding crowdsensing have been unable to demonstrate the cognitive fitness of a large batch of participants utilizing PHQ-9 or perceived stress scale.

Our research has shown the apparent correlation between smartphones’ continuous and autonomous sensing data and mental health measures such as PHQ-9. We have achieved the primary purpose of this research study to visualize the potential interconnection between physiological activities and the mental fitness of human beings. Additionally, from our analysis, we can identify users suffering from moderately severe and severe depression, which might further help them regarding the early treatment of depression utilizing medical science.

In our proposed research, we are not invading the privacy [20] of the participants as we are not collecting application usage information & phone call records for evaluating the social interaction of users, unlike other research studies.

However, several limitations in our study must be overcome before our research study to evaluate whether mental fitness can be ready for widespread usage. Even though we started our research with 205 participants, as an outcome of outlier detection and removal and reduction of repetitious and null GPS sensor data, we finally had to continue our research with 71 participants’ data which is relatively fewer in number. If we had more participants’ data to analyze our study, our results would generalize more appropriately to other user groups.

In our research, we further needed to handle the missing value problem. The explanation behind this problem is human labeling errors and sensitive sensors. This led to an issue where the sensor values for one or two sensors were very low in some cases. Because of the amount of variability coming from contrasts in hardware, device-usage patterns, and the environment, crowdsensing platforms will presumably require a large user base for widespread applicability.

VI. Conclusion and Future Work

Since unhealthy cognitive health conditions are rising, cognitive fitness monitoring has become an intriguing research field. The scope for research into evaluating mental fitness utilizing smartphone sensors is enormous. Potential challenges and limitations remain when using crowdsensing in this domain in broader populations.

This study developed an android application for autonomous smartphone sensor data collection from a substantial number of participants. Through illustrating the numerical evidence of correlation, we have enlightened a significant insight into how physical activities and social interactions in the daily living of humans unfold a fundamental connection with the cognitive fitness of people. Besides, this is the first-of-its-kind research study, where crowdsensing is applied in a broader population to analyze in detail and differentiate people’s cognitive fitness levels numerically. Furthermore, we have developed and fine-tuned regression models which predict cognitive fitness scores accurately to a certain extent based on multiple necessary features.

There is still abundant room for progress as limitations exist in this research in evaluating mental fitness relying upon smartphone sensors. There remains much scope for refinement in the regression models of our research, as additional valid data are required for training to achieve precise accuracy and more satisfactory performance for widespread applicability. For future studies, we tend to use even more smartphone sensors than accelerometers and GPS [15] for our experimentation. Moreover, we would also like to explore the sleep-wake cycle for sleep monitoring [16] and categorize applications usage phone calls [15] for an enhanced cross-assessment of social interaction.

To conclude, we can presume that the ultimate success of crowdsensing in assessing mental fitness will likely depend on the continued engagement of users who provide both smartphone sensor data passively and some measure of accurate labeling of rating scales [3].

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

All of the authors contributed towards conducting the research paper. Samin and Aysha proposed the research idea. Samin & Ishfaq preprocessed, analyzed the data and performed data visualization. Samin, Aysha & Nilufar developed the android and contributed in data collection phase. Samin implemented the methodology and wrote the research paper. Zabir Haque supervised the methodology, implementation and writing of the research paper. All authors had approved the final version.

References

[1] K. Kroenke, R. L. Spitzer, and J. B. Williams, “The PHQ-9: Validity of a brief depression severity measure,” *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[2] K. Kroenke and R. L. Spitzer, “The PHQ-9: A new depression diagnostic and severity measure,” *Sychiatric Annals*, vol. 32, no. 9, pp. 509–515, 2002.

[3] D. C. Mohr, M. Zhang, and S. M. Schueller, “Personal sensing: Understanding mental health using ubiquitous sensors and machine learning,” *Annual Review of Clinical Psychology*, vol. 13, p. 23, 2017.

[4] A. Capponi, C. Fiandrino, B. Kantarci, L. Foschini, D. Kliazovich, and P. Bouvry, "A survey on mobile crowdsensing systems: Challenges, solutions, and opportunities," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2419–2465, 2019.

[5] S. Majumder and M. J. Deen, "Smartphone sensors for health monitoring and diagnosis," *Sensors*, vol. 19, no. 9, 2164, 2019.

[6] K. T. Taqi, K. A. Rezwan, M. N. Chowdhury, K. Mallik, U. Habiba, and H. M. Z. Haque, "An integrated crowdsourcing application for embedded smartphone sensor data acquisition and mobility analysis," *Journal of Advances in Information Technology*, vol. 13, no. 5, pp. 503–511, 2022.

[7] K. Athukorala, E. Lagerspetz, M. Von Kügelgen, A. Jylhä, A. J. Oliner, S. Tarkoma, and G. Jacucci, "How carat affects user behavior: implications for mobile battery awareness applications," in *Proc. SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 1029–1038.

[8] K. Stanley, E.-H. Yoo, T. Paul, and S. Bell, "How many days are enough? Capturing routine human mobility," *International Journal of Geographical Information Science*, vol. 32, no. 7, pp. 1485–1504, 2018.

[9] B. Mahesh, "Machine learning algorithms — A review," *International. Journal of Science and Research (IJSR)*, vol. 9, pp. 381–386, 2020.

[10] Z. Ghahramani, "Unsupervised learning," in *Summer School on Machine Learning*, Springer, 2003, pp. 72–112.

[11] T. M. Ghazal, M. K. Hasan, M. T. Alshurideh, H. M. Al-zoubi, M. Ahmad, S. S. Akbar, B. Al-Kurdi, and I. A. Ak-Our, "IOT for smart cities: Machine learning approaches in smart healthcare — A review," *Future Internet*, vol. 13, no. 8, p. 218, 2021.

[12] Åvitsland, E. Leibinger, T. Haugen, Ø. Lerum, R. B. Solberg, E. Kolle, and S. M. Dyrstad, "The association between physical fitness and mental health in Norwegian adolescents," *BMC Public Health*, vol. 20, no. 1, pp. 1–10, 2020.

[13] M. Srividya, S. Mohanavalli, and N. Bhalaji, "Behavioral modeling for mental health using machine learning algorithms," *Journal of Medical Systems*, vol. 42, no. 5, pp. 1–12, 2018.

[14] H. Ma, D. Zhao, and P. Yuan, "Opportunities in mobile crowd sensing," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 29–35, 2014.

[15] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Student life: Assessing mental health, academic performance, and behavioral trends of college students using smartphones," in *Proc. 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014, pp. 3–14.

[16] 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 *Proc. 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, IEEE, 2013, pp. 145–152.

[17] M. Ciman, K. Wac, and O. Gaggi, "Isensestress: Assessing stress through human-smartphone interaction analysis," in *Proc. 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (Pervasive- Health)*, IEEE, 2015, pp. 84–91.

[18] Y. Sun, Z. Fu, Q. Bo, Z. Mao, X. Ma, and C. Wang, "The reliability and validity of PHQ-9 in patients with major depressive disorder in psychiatric hospital," *BMC Psychiatry*, vol. 20, no. 1, pp. 1–7, 2020.

[19] S. S. Thakur and R. B. Roy, "Predicting mental health using smart-phone usage and sensor data," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 10, pp. 9145–9161, 20.

[20] S. Gisdakis, T. Giannetsos, and P. Papadimitratos, "Security privacy, and incentive provision for mobile crowdsensing systems," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 839–853, 2016.

[21] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *Proc. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2020, pp. 747–748.

[22] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "DBSCAN: Past, present, and future," in *Proc. the Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, IEEE, 2014, pp. 232–238.

[23] A. Kumar. Correlation concepts, matrix & heatmap using Seaborn. [Online]. Available: https://vitalflux.com/correlation-heatmap-with-seaborn-pandas

[24] S. Saxena, M. V. Ommeren, K. Tang, and T. Armstrong, "Mental health benefits of physical activity," *Journal of Mental Health*, vol. 14, no. 5, pp. 445–451, 2005.

[25] I. Kawachi and L. F. Berkman, "Social ties and mental health," *Journal of Urban Health*, vol. 78, no. 3, pp. 458–467, 2001.

[26] S. Hu, D. Mo, P. Guo, H. Zheng, X. Jiang, and H. Zhong, "Correlation between suicidal ideation and emotional memory in adolescents with depressive disorder," *Scientific Reports*, vol. 12, no. 1, pp. 1–10, 2022.

[27] M. A. Harris, "The relationship between physical inactivity and mental wellbeing: Findings from a gamification-based community-wide physical activity intervention," *Health Psychology Open*, vol. 5, no. 1, 2018, doi: 10.1177/2055102917753853

[28] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.