An Integrated Crowdsourcing Application for Embedded Smartphone Sensor Data Acquisition and Mobility Analysis

Article · January 2022

DOI: 10.12720/jait.13.5.503-511

6 authors, including:

Moontaha Nishat Chowdhury
Ahsanullah University of Science & Tech
2 PUBLICATIONS **1** CITATION

SEE PROFILE

An Integrated Crowdsourcing Application for Embedded Smartphone Sensor Data Acquisition and Mobility Analysis

Kazi Taqi Tahmid, Khandaker Rezwan Ahmed, Moontaha Nishat Chowdhury, Koushik Mallik, Umme Habiba, and H. M. Zabir Haque
Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh
Email: {170104015, 170104044, 170104099, 170104024, 170104004, zabir.haque.cse}@aust.edu

***Abstract*—The proliferation of smartphones has become a ubiquitous platform for acquiring and analyzing data. Smartphones' embedded sensors have become an effective source for human spatial and activity-based analysis. Machine Learning (ML) has made significant progress in learning features from these raw sensor data with high accuracy. However, domain experts, knowing ML, can apply machine learning techniques for various aspects. In this research, we have introduced—a smartphone sensor data collection and analysis platform for people in general who have little or no knowledge of machine learning but can avail the services of machine learning for their purpose. We have built an Android application for collecting sensor data and developed an Automated Machine Learning (AutoML) based web platform for data pre-processing, visualization, and analysis. Spatial analysis has been conducted on our AutoML based web application on GPS sensor data. We evaluated the most visited places of our app users using clustering techniques. The experiment shows that the DBSCAN clustering algorithm gives superior performance over K-means clustering for our spatial analysis on GPS sensor data.**

***Index Terms*—smartphone sensor, AutoML, Android application, sensor data, GPS, spatial analysis, DBSCAN, K-means clustering**

I. Introduction

The proliferation of smartphones has made it a ubiquitous platform to capture and analyze data [1]. These smartphones comprise various sensors, such as accelerometer, magnetometer, gyroscope, GPS, battery sensor, proximity sensor, light sensor, temperature sensor, humidity sensor, etc. Smartphones can collect data passively and incessantly with their strong sensing ability, the capability of computation, and serve real-time information [2], [3]. These sensors detect and respond to any change in physical input and steer a process or offer information as the input of a system. Moreover, sensor data helps capture information related to a specific task so that the data can optimize operations and increase efficiency. Former studies have recognized the importance of embedded smartphone sensors [4] as an innovative source for analyzing human activity, personality, spatial behavior, etc. Passively captured sensor data measure human spatial behavioral patterns without any participation or action from the user. As a result, the captured data forms unobtrusive observational records, which are more naturalistic [5]. Hence, embedded smartphone sensor data could provide valuable insight into human behavior patterns that surveys cannot attain.

Machine Learning (ML) has become a pivotal instrument to analyze smartphone sensors' large-scale and complex data. Using ML pipeline, these large-scale data are cleaned, analyzed, and fit the best model. ML gives opportunities to systems to learn automatically, improve their learning by experience, and identify relevant insights in data [6]. Analyzing smartphone sensor data with Machine Learning (ML) techniques has far-reaching significance in identifying human spatial patterns [2], [7], surveillance of human activities [8], intelligent homes [9], context-aware computing [10], etc. Moreover, machine learning is a fast-growing trend in the healthcare business because sensor data can be used to analyze a patient's health in real-time. By analyzing smartphone sensor data, doctors can monitor their patients remotely. Through accelerometer and gyroscope sensor data, doctors can monitor Parkinson patients' motion behavior, such as falls, walking, and so on, which aids them with medication with individual needs. Retailers rely on GPS and humidity sensor data analysis with ML techniques to identify target locations to arrange a marketing campaign or supply planning [11].

Much Prior research has been conducted on analyzing smartphones' sensor data using machine learning. To analyze smartphone sensor data using ML, prior knowledge is required in the machine learning field. Though machine learning provides better support for many industries, some enterprises are still struggling to implement ML models for deployment. They depend on data scientists for implementing ML models, but hiring data scientists is costly. Moreover, collecting a large amount of data is a complex and tiring task. For example, restaurant companies depend on data scientists to analyze GPS sensor data of target customers to identify suitable locations for opening a new branch. Hence, for people in general, the automation process of applying machine

Manuscript received February 1, 2022; revised June 1, 2022.

doi: 10.12720/jait.13.5.503-511

learning techniques is in high demand, which does not require prior knowledge of machine learning, but can avail machine learning services. The method of automating tasks or repetitive tasks to apply machine learning in real-world applications is referred to as Automated Machine Learning (AutoML). AutoML helps to automate the entire machine learning workflow [12]. The principal goal of AutoML is to alleviate users' jobs only by data set selection. AutoML tries to automate the rest of the Machine Learning pipeline, which includes data cleaning, data transformation, feature engineering, automation of model selection, finding the exemplary model architecture, and identifying which hyperparameters are the best fit for the corresponding model. But there exists a lack of concern on sensor data analysis using automated process machine learning.

Sensor Data analysis using Automated Machine Learning (AutoML) based platforms can serve people of various domains as it will facilitate everyone to use the machine learning technology instead of a small group of people. For example, by using AutoML based platforms, retailers can observe the product purchase behavior of customers and make recommendations without hiring a data scientist. It could analyze users' smartphone motion, location sensor data to the tracker and identify store behavior. It can record which sections they have visited first and which next and apply the next basket recommendation techniques [11]. Non-technical users can analyze the GPS sensor data [1], [13] to organize big events to manage a multitude of people at various occasions, festivals, or events, like fairs, [14]concerts, etc. Furthermore, security concerns, such as analyzing the accelerometer and gyroscope sensor data anomaly can be detected by applying k-means partitioning [15] without having any knowledge of clustering algorithms.

Qualitative research has been carried out on smartphones' embedded sensors, but no single study exists that studies raw sensor data with automated machine learning (AutoML) techniques. In this research, we have developed an AutoML based web application as an analysis tool for any kind of survey analysis on smartphone sensor data. For data collection purposes, we have built an android application named "Sensor Data Collector" that captures embedded smartphone sensor data. Motion sensor data, such as accelerometer, gyroscope, magnetometer, gravity sensor, GPS, etc., are collected to track and analyze users' motion, location [2], [3], [16], [17]. The captured sensor data is then analyzed on the AutoML based web platform. The steps of a machine learning analysis like pre-processing, filtering training, etc., are done on our web server. As a result, the user does not need an analyst or prior knowledge of data analysis but will get an instant solution from our server. Besides collecting sensor data through our android application, users are also asked a different question every day based on some personality-based questions for figuring out the users' personalities and behavior. These questionnaire-based surveys will facilitate further personality-based analysis. Moreover, to make the application more interesting, there is also a Tic-Tac-Toe game for users to play for their pastimes or whenever they want.

Throughout this research, we have implemented the described framework and conducted a survey of 90+ users for several months, collected six smartphone sensor data through an android app, and ran automatic analysis via our web platform. For our study, we have extracted the GPS sensor data from the captured sensor data list and performed spatial analysis on this GPS data on our web platform. Our study has focused on data visualization for human spatial analysis through smartphone sensors, which will provide a significant opportunity to advance the understanding of unraveling human spatial patterns. GPS sensor data helps researchers to collect information on human spatial behavior or pattern [2]. For example, the GPS sensor data helps to identify the physical address by analyzing the user's most visited place at night. We have applied the DBSCAN algorithm to find the top 10 most visited places.

The remaining paper is organized as follows. Section II represents a portrayal of topics incorporated with our proposed system. Moreover, some existing works devoted to human mobility and spatial analysis are discussed in the same section. In Section III, the approaches or methodology we have followed to implement our proposed AutoML based system is explained. In Section IV, the results of our AutoML based system have been illustrated. Finally, Section V outlines the concluding points and some future works of our study.

II. Background Study

As mentioned above, in our proposed end-to-end system, we acquire sensor data from our android application and analyze this data through our web applications' AutoML platform using different clustering mechanisms. In this section, we have demonstrated the integral concepts of our study on sensor data analysis tools and techniques. This section also depicts some former studies in this field. *Sensor Data Analysis Tools.*

A. Smartphone Sensors

The smartphone embedded sensors have opened a golden opportunity for researchers to collect sensor data and conduct much-complicated research indisputably and efficiently. A couple of sensors could be integrated into smartphones connected to the SoC, using a sensor hub, for example, Accelerometer, Gyroscope, Magnetometer, GPS, Ambient light sensors, etc. These embedded sensors have a wide range of applications. Based on the smartphones having adequate sensors, any digital mobile application can add this sensor data collection and analysis as a feature to their application and apply this data for their user activity, mobility, etc., analysis. Some basic functionalities of these sensors for analyzing human mobility or activity patterns are discussed below.

Accelerometer sensors in smartphones are used to detect the orientation of that device. It is a tri-axial sensor that measures the value of three axes- x-axis, y-axis, and z-axis. The tri-axial values measure the linear acceleration

movement and can identify the direction of the earth's gravity (in the z-axis). The phone can automatically adjust the landscape and portrait view for this accelerometer sensor. Accelerometer sensor data is applied in various domains such as health care [17], fall detection, mental-health monitoring system, physical activity recognition [18], [10], transport mode detection [19], [3], etc.

To get the accurate orientation of the device, a gyroscope sensor value is supplied with an accelerometer. A gyroscope is also a tri-axial sensor like an accelerometer. It can measure the angular velocity of the device. Watching 360 videos or playing a racing game with a smartphone is possible for this embedded gyroscope sensor. The domain of gyroscopes is also vast- in 3D gaming, health monitoring [17], activity recognition, transport mode detection [19], etc. Like an accelerometer and gyroscope, the magnetometer also measures 3-dimensional values- x, y, and z-axis. A magnetometer, as the name indicates, measures the magnetic field. This sensor is used to determine the smartphone's orientation relative to the magnetic north of the earth. Many applications like digital maps or digital compass, for embedded magnetometers, can rotate with the phone's physical orientation. As the magnetic sensor knows the magnetic north of earth, with the help of this sensor smartphone can automatically turn digital maps according to the user's physical orientation.

In the case of spatial analysis-based studies, the Global Positioning System sensor known as GPS provides more accurate results than an accelerometer and gyroscope. GPS receiver gives the information of the device's location anywhere on the earth but varies inaccuracy, depending on the number of satellites in that network. Cell phones' built-in receivers communicate with multiple satellites to compute the location information [20]. GPS receivers receive the signal from satellite networks and triangulate the devices' relative position. It calculates the location by using the intersecting points of overlapping spheres of satellites and the built-in receiver. To be more specific, trilateration is determined by the distance between the GPS receiver and satellites to create overlapping spheres, which form a circle by the intersecting points. The location information is measured using the coordinate value, longitude, and latitude. In our study, we have extracted the only GPS sensor data among all raw sensor data collected by our android application to perform spatial analysis.

B. Automated Machine Learning (AutoML)

Machine learning allows machines to learn from data without being explicitly programmed. With the increasing demand for hands-free machine learning solutions, the area of automated Machine Learning (AutoML) has forthwith risen [21]. Automated Machine Learning (AutoML) is a complete process from data filtering or pre-processing to the outcome of a model. It automates the tasks of applying machine learning algorithms to make the machine learning process more accessible to non-technical individuals. Moreover, AutoML significantly impacts applications in commercial settings [22]. It has been widely observed that data-driven model building and decision-making may lead to increased degrees of automation and more informed judgments are a significant motivator driving the digitalization of industry and society [23]. Applying AutoML frameworks in a system makes complex tasks easier by concerning the data and trying out every possible kind of Machine Learning model using machine learning algorithms.

AutoML aims to automatically choose, assemble, and parameterize machine learning algorithms to achieve optimal performance on a particular dataset. The hyperparameter will get automatically tuned, and it will select a parameter, and based on that, it will generate an accuracy. When the hyperparameters are adjusted, the Auto ML process runs some iterations with the selected parameters and terminates the process to achieve the highest precision from this particular model. AutoML will potentially serve as a more efficient way to complete tasks supplied with a user-friendly UI to the users who are not aware of the intricacy of machine learning techniques. The client typically uses the service by providing a dataset, and the AutoML platform does the rest of the work for the client by applying the optimum machine learning algorithm to that dataset and outputting a result [12]. Through open-source packages like Auto-WEKA (JAVA), Auto-SKlearn (Python), Auto-Keras (Python), TPOT (Python), the machine learning community has dramatically benefited such users by making a wide array of sophisticated learning algorithms and feature selection methods available.

C. Sensor Data Analysis Techniques

1) Spatial analysis

Smartphone sensor data have a direct link with user identity and user networks. Spatial analysis is performed widely using embedded smartphones or other wearable devices' GPS sensor data. The spatial analysis also referred to as spatial statistics, is a formal approach that investigates an individual or community's geographical or geometric characteristics [24], [25]. The spatial analysis technique is used to analyze structures on a human scale broadly in the studies of geospatial data. It is a subset of geographical analysis that helps to explain human mobility and behavioral pattern and represents the resultant locational analysis or spatial analysis in terms of mathematics and geometry.

DBSCAN: To extract knowledge from a large amount of spatial data collected from various applications including GPS, satellite images, and remote sensing, clustering techniques play a significant role. Different spatial data clustering approaches have been introduced to discover valuable patterns from these complex data. Among them, a well-known data clustering approach is density-based spatial clustering of applications with noise (DBSCAN) [26], [27]. DBSCAN can identify clusters of any arbitrary form and shape in the databases, even if it contains noise or outliers. Real-world data may have abnormalities such as arbitrary cluster shapes and data with noise. This algorithm clusters together points near each other by using distance measurement (typically Euclidean distance) and calculates a minimal number of points. This algorithm has two parameters, namely eps (epsilon) and min points. The eps signify those two points

are considered neighbors if their distance is less than or equal to the value of eps. The term min points refer to the smallest number of points required to establish a dense zone. Because of the importance of the two parameters, they must be anticipated correctly to obtain a decent result [26].

Silhouette Score: Determining the optimal number of clusters for a data set is critical in specific clustering algorithms. The silhouette score is one of the various strategies for determining the ideal number of clusters [21]. The silhouette plot shows the distance or closeness of each point in one cluster with the neighboring clusters' points and visually examines factors like cluster count. Silhouette's score lies between -1 to 1. If the silhouette score is near 1, the sample is referred to as clustered ideally and has already been assigned to a highly appropriate cluster. If the score is close to 0, the sample might be assigned to the cluster closest to it, and the sample would be equally far from both. That is to say, it denotes overlapping clusters. If the silhouette value is close to –1, the sample is misclassified and placed in the middle of the clusters [28].

2) Geofencing

Geofencing is a Location-Based Service (LBS) in which the GPS sensor data, Wi-Fi, or cellular data is used in an application to trigger an action when a mobile device ingress or pass a virtual boundary set up around a geographical area. This virtual boundary is called geofence. Propagation of spatial analysis-based studies using GPS sensor data helps to serve various LBS. LBS has become proactive by allowing smart alerts when a user enters or departs a specified geographic region, a feature known as Geofencing [29]. Remote surveillance from geographic areas encircled by a virtual barrier (geofence) and automated detections when mobile objects entered or departed these areas were made possible by geo-fencing [13].

Geofence applications and utilities kept track of the devices and other physical items that entered or departed the geo-fenced region and alerted administrators when their status changed. Geofencing is mainly recognized in the research community for two challenges: the influence on mobile device energy consumption and traffic load inside wireless access networks [30], [31]. Furthermore, Geofencing creates serious privacy concerns for mobile phone users if their position is constantly calculated by network operator equipment or sent to a third-party service [30]. There are countless geofencing applications, such as social networking apps, marketing, smart appliances, security systems, child position monitoring [32], etc. In our system, we create geofence using an automated Machine Learning algorithm (AutoML).

D. Related Works

A growing body of literature recognizes the importance of embedded smartphone sensor data for various aspects of human activity recognition, personalization, and spatial behavior analysis. Analyzing and mining these smartphone sensor data has potential applications. In this section, we have briefly introduced former studies on human spatial behavior using smartphone sensor data. Several studies have examined human spatial behavior or mobility by analyzing smartphone sensor data. The authors at [25] investigated for GPS data of 6 months on 1,00,000 anonymous smartphone users and concluded that human mobility has a highly regularized pattern, and distinctly reproducible patterns have been detected in their daily travel distance. This reproducible travel pattern of individuals can serve various applications driven by human spatial or mobility patterns, such as industrial construction or urban planning and traffic forecasting. Furthermore, a framework named predestination was proposed by Horvitz *et al*. [7], predicting users' short time trips' destination location based on smartphone GPS trajectory history. On the contrary, privacy has become a major concern as spatial information is collected. Moreover, to determine human spatial mobility there are contradictions in studies about how many days of data is required to get a complete activity space of an individual. The authors in [2] presented that at least data of 14 days is required to get a complete pattern of activity space or spatial mobility of a person using only smartphone GPS sensor data. The researchers here have made an important contribution to spatial behavior. They have used KL divergence of the spatial histogram of daily mobility through time and found that less than 14 days of data is required to get a pattern of an individual's spatial behavior.

Prior studies have shown that social ties influence human spatial behavior or mobility. Physical location is intrinsically linked with social relationships. This fact helps researchers to predict the individual or location prediction of a group of people more accurately. P. A. Grabowicz *et al*. [33] introduced a model by considering explicit feedback of human mobility while forming social ties using social media data. The integration of mobility and social interaction analysis results in several characteristics, such as the total number of connected components, the distance between individual components, user clusters based on distance. Similarly, the authors at [24], the accuracy of human mobility prediction increases where the correlation of social interaction is also considered.

Moreover, using smartphone sensor data for analyzing mobility patterns, driving behavior analysis, and road condition monitoring has become a highly popular field in academic and industrial studies. In [14], authors have used embedded smartphone sensor data for estimating smartphone orientation concerning the vehicle frame. They have discussed some methods using applied classification techniques for smartphone-based driver classification and road condition monitoring. An accelerometer-based activity classification algorithm has been described by Thiagarajan *et al*. [34] by using smartphone sensor data and route information to identify whether a user is riding on a vehicle or not. The authors also distinguished if the vehicle type is a bus or not. Reddy *et al*. [35] have used both the data from GPS receivers and an accelerometer to predict the transportation modes of users or if the users are walking, stationary, running, biking, or in motorized transport. They used a machine

learning approach, a decision tree followed by a first-order discrete Hidden Markov Model, which achieved a promising accuracy of 93.6%.

Although qualitative research has been made on human mobility or spatial analysis based on smartphone sensor data, none of the existing approaches appears suitable for people with non-IT or any domain who need data visualization of sensor data. Considering this fact, in this study, we have developed our web application as an analysis tool for any kind of survey analysis of smartphone sensor data for people with any domain.

III. Methodology

A detailed description of our approach is discussed in this section. From data collection to visualizing the output, we have divided our whole system into three parts –A mobile app for data collection, Data analysis using ML algorithm, and a Web application to automate the entire process (Fig. 1). Firstly, the system collects embedded sensor data through our android app. Moreover, it also conducts users' personality-based analysis through questionnaires for further studies. After that collected data is transmitted to the cloud for storage, and finally, the AutoML platform for data cleaning or preprocessing visualization.

Figure 1. System architecture.

A. Sensor Data Collector App

In this study of "Spatial Analysis Based on Smartphone Sensor Data", we have developed an Android app called "Sensor Data Collector" that collects data from users' embedded smartphone sensors (Fig. 2). To walk through the features of our app "Sensor Data Collector", we use the following scenario.

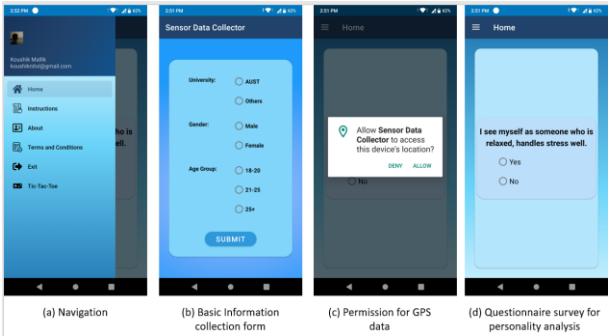

Figure 2. Data acquisition through our android application.

Mallik is a smartphone user who has participated as a volunteer in our research. At first, when Mallik opens the app after installation, a pop-up form has arrived which asks for his university name, his gender, and which age group he belongs to, for further analysis purposes. To access the GPS data "Sensor Data-Collector" asked Mallik to accept location access permission. When he was permitted to access GPS data, the app started to collect a total of six sensors- accelerometer, orientation, magnetometer, light, gravity, and GPS data along with the timestamp and battery life in the background.

After successful login, Mallik can see side navigation, where he found an option of Tic-Tac-Toe and plays in his leisure time. Whenever he opens the app, the landing page asks a question, and he discovers a new question every day. This survey questionnaire has been conducted for further psychological and behavioral research.

The collected sensor data from Mallik's smartphone has been stored in two steps. Firstly, the collected data is stored in the local database after a particular time. After that, the stored data in the local database is pushed to the cloud storage after a fixed interval of time concerning the cost. From smartphone devices to cloud storage, the data flow is maintained by controlling the time intervals from the cloud storage. The interval variable values are stored in the cloud storage which allows us to control the time after which we want to store the data in the cloud. (Fig. 3)

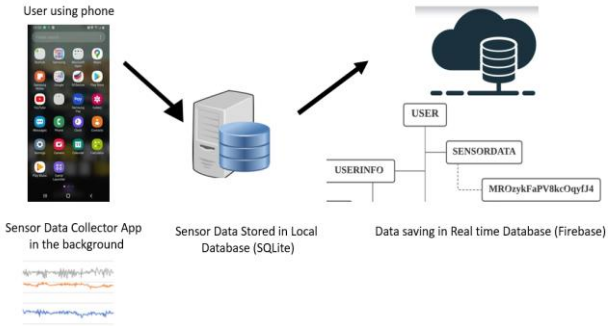

Figure 3. App architecture.

Concerning the fact that internet connection might not be always available to Mallik's smartphone, offline capabilities have been enabled in the app. So that when an internet connection is not available the data is kept in the Mallik's device's local storage and that data is automatically stored in the cloud whenever the internet connection is available and also the stored data is deleted from his device so that it doesn't take up extra storage. Another major concern is GPS sensor data [14] collected among the six sensors, which increases the battery consumption rate of Mallik's phone [10]. Hence, we programmed our app in a manner so that it doesn't collect GPS data at regular intervals if the location of the user doesn't change. For example, if the location of a certain user doesn't change then the app collects the GPS data for say after 1 hour and if the GPS data starts to change then the app collects GPS data after every 15 minutes. To reduce battery consumption, we store the data locally using SQLite. After collecting and saving the data for a while in SQLite, we are pushing the batch data into firebase while the user has an active internet connection.

1) Sensor data overview

We have collected a total of six different sensor data through our sensor data collector app. Fig. 4(a) shows the

frequency distribution of sensor data collected per month. The x-axis represents the timeline or corresponding months from January 2021— to August 2021, and the y axis shows the frequency of captured data. From this graph, it can be shown that our app collects the lowest amount of data in January, February, and June. The reason behind capturing the lowest amount of data in these three months is due to the corona pandemic situation. Our sensor data collector application collects a moderate amount of data in April and August. Hence, we choose to analyze the data of August for spatial analysis.

Based on the captured sensor data, demographic analysis shows (Fig. 4(b)), the majority of the participants are male. Among 90+ volunteers, there were 70% male participants in our study.

We have collected a total of six sensor data through our "Sensor Data Collector" app. Among them, Table I is a demonstration of GPS and Accelerometer data. Here, UID is the unique identifier of the user that each user receives when he logs into our app for the first time. The Date time represents the corresponding time and date when this data was collected. Latitude value represents the coordinate that specifies the north-south position and Longitude value represents the coordinate that specifies the east-west position. In one of the cells, the value is 0 that is because that particular user didn't permit the location sensor during that point in time. The X-axis, Y-axis, and Z-axis represent the tri-axial values of the accelerometer (see Table II-III).

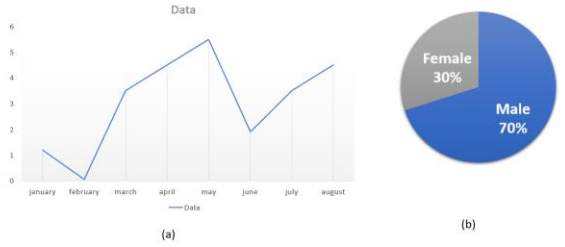

Figure 4. Frequency distribution of collected sensor data (a), and participants' demographic data (b).

TABLE I. VALUES OF 3 MOTION SENSORS AT A GIVEN TIME

Axis	Accelerometer	Gyroscope	Magnetometer
X-axis	0.21068975329399	1.7453292093705386E-4	-27.0625
Y-axis	4.73094272613525	0.0643851980566978	-18.5
Z-axis	9.53849983215332	0.117757365107	-36.5

TABLE II. SAMPLE DATASET OF GPS SENSOR

UID	Date Time	GPS	
		Latitude	Longitude
1auGDa4mMKb	18 Apr 2021 4:15:30	0	0
1auGDa4mMKb	18 Apr 2021 4:40:44	23.762	90.4181

TABLE III. SAMPLE DATASET ACCELEROMETER SENSOR

UID	Date Time	Accelerometer		
		X-Axis	Y-Axis	Z-Axis
1auGDa4mMKb	18 Apr 2021 4:15:30 am	0.271742	0.0706289	9.66659
1auGDa4mMKb	18 Apr 2021 4:40:44 pm	1.16238	5.70179	7.89009

From the collected data of six sensors, in this study, we have filtered only the GPS sensor data to perform spatial analysis using Machine Learning.

B. Data Analysis Using Machine Learning Algorithm

The entire data collection procedure was discussed in the preceding section. This section depicts our sensor data analysis approaches using machine learning algorithms. From the collected sensor data, we have extracted GPS sensor data filtered out the remainder, and performed machine learning analysis on the GPS data. The intuition behind extracting GPS sensor data was to conduct spatial analysis of users' most visited places.

To determine the most visited location, machine learning clustering algorithms were applied. Primarily, we have used the K-means clustering technique. The k-means method divides an array of coordinates into k clusters by grouping rows together. However, because it reduces variance [36], k-means is not a good method for geographical data analysis. Moreover, resulting in a higher number of rows near a specific location in the data set indicates a higher likelihood of more rows being randomly selected for that area. One problem with this approach is many locations would be missing from any clusters due to the random seed, and increasing the number of clusters would still result in gaps throughout the data set. As a result, we decided to use the DBSCAN algorithm instead.

DBSCAN uses two parameters to cluster a GPS data set: one is a physical distance between each point, and the other is a minimum cluster size. For latitude-longitude

data, this technique performs substantially better. To determine great circle distances between points, we employed the haversine metric and the ball tree algorithm. Because the haversine metric employs radian units, the epsilon and coordinates are transformed to radians. Moreover, after applying the DBSCAN algorithm, we got our desired result.

C. AutoML Based Web Application

For analyzing the raw sensor data from embedded smartphone sensors, we have built an automated machine learning-based web application. As shown in Fig. 5, our web application consists of three main segments. The front-end section built with the NextJS framework fetches data from the firebase and presents them to the viewers. On the webserver section, we have deployed our machine learning analysis code. We have used three firebase data storage features for our whole process namely: real-time database, firestore, and storage.

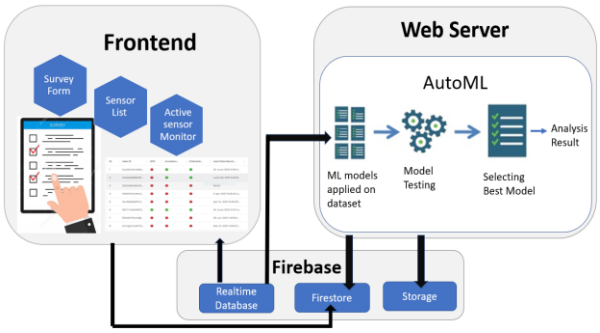

Figure 5. Web application architecture.

The main goal of our web application is to automate machine learning analysis on the dataset given by the client having no or little knowledge of machine learning. It offers various functionality to users. Users will be able to create multiple surveys from this platform. For each survey, they will have to enter the survey name, institution name, survey duration, and survey period. All of these details will be stored in the firestore. So, the information is fetched and presented for the user's viewing.

This web application enables users to visualize the survey participants' data in a tabular format and users can export the participants' data in their desired format. Users can see the list of the sensors that are currently active in the survey participants. For this, we have fetched the sensor data along with the timestamp of each data from the real-time database and then compared it with the current time to show the participant as active or not active. Moreover, a data analysis done with the machine learning algorithms will be presented for the users on the sensors they have selected. For example, if a user selects GPS sensor data, then we will provide analysis like the most visited places by different age groups and genders, etc. Furthermore, Users or survey creators can see the survey participants' statistical analysis based on age and gender.

IV. Results and Discussion

Our goal was to collect sensor data, apply some machine learning analysis, and present the result on our web app. In the previous sections, we discussed our data collection procedure and our whole working methodology. In this section, we will talk about the results that we achieved from our experiment. We have used the GPS sensor data to conduct spatial analysis. As a result of our survey through the app, from August 13 to August 19 when the system was fully used, we collected 2,834 hours of sensor data. Hence, we have analyzed the most visited places by our participants from August 13 to August 19, 2021. We had to pre-process the data for our machine learning analysis at first. Our dataset has a total of 4,336 rows. We eliminated the other sensor data columns and worked with the GPS data values because we did our analysis on GPS data. Then we filtered out any rows with '0' in the latitude and longitude data values.

Below Table IV shows the longitude and latitude, along with the number of data counts of most visited places by the app users. The latitude and longitude values were then paired and transformed to coordinates. 6371.00 was chosen as the number of kilometers per radian. The epsilon value was translated to radian and determined as 1.5 kilometers. The DBSCAN clustering algorithm was then used to reduce the 4,336 points to 16 clusters. Then we found out the most central point of all the clusters. The cluster centers of each cluster are depicted on a map in Fig. 6.

TABLE IV. Top 10 Most Visited Places

Cluster	Latitude	Longitude	Count
4	23. 8359	90.3739	1120
3	22.3307	91.8286	615
0	23.7443	90.4110	311
2	23.7127	90.4132	121
16	37.4220	-122.0840	112
1	23.7623	90.3652	83
5	24.2293	89.9599	38
6	23.7800	90.4065	22
9	23.7427	90.3927	21
11	23.875	90.2985	16

Figure 6. Cluster centers with K-means(a), Cluster centers with DBSCAN (b).

Initially, we applied K-means on the extracted GPS sensor data. The k-means method divides an array of coordinates into k clusters by grouping rows together. However, because it reduces variance, k-means is not a good method for geographical data [36]. More rows near a specific location in the data set indicate a higher likelihood of more rows being randomly selected for that area. Therefore, many locations are missing from clusters due to the random seed, and increasing the number of clusters results in gaps throughout the data set. To overcome this shortcoming, we decided to apply the DBSCAN clustering algorithm. Fig. 6 shows that, with k-means clustering on the same set of GPS sensor data, a total of 7 cluster centers are identified, whereas, with DBSCAN, the system gets 8 cluster centers. Hence, we have used DBSCAN clustering for further analysis.

In DBSCAN clustering, we have used the silhouette score to determine the optimal number of clusters. The silhouette score that we got is 0.707, indicating that the samples are distinguishable by the decision boundary between two neighboring clusters.

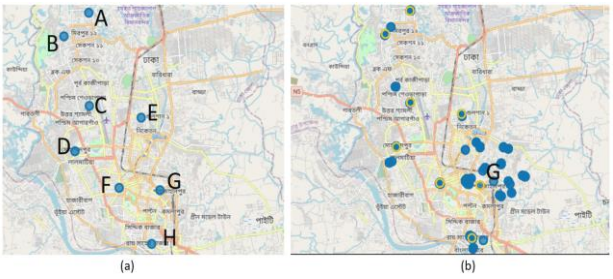

Figure 7. Cluster centers plotted on the map (a), all the Locations and cluster centers are plotted (b).

After that, the cluster with the largest number of points or coordinates is picked. Then we get the cluster's most central point. The cluster centers of each cluster are depicted on a map in Fig. 7(a). Here, A, B, C, D, E, F, G, H are cluster centers. The most visited area is identified among these center points by plotting all the points with the corresponding cluster center. As a result, the most visited coordinates are found as shown in Fig. 7(b). The yellow rings represent the center of that cluster. All the coordinates are plotted with blue circles to indicate their location. Fig. 7(b) shows that among the most visited central point (A–H), the locations around central point G is the most visited area by app users.

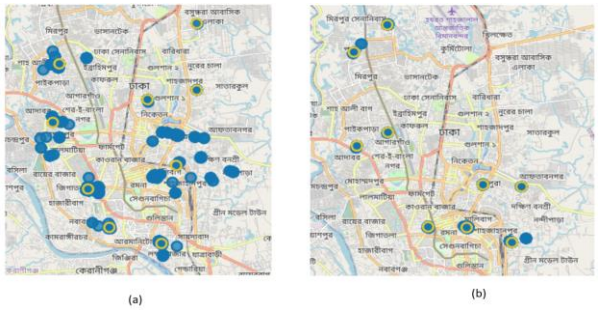

Figure 8. All of the locations and cluster center coordinates of male users are plotted(a), all of the locations and cluster center coordinates of female users are plotted (b).

From Fig. 8, it is clearly shown that there are more male users (a) than female (b) participating in our study. In both cases (a, b), the yellow marked circle represents the cluster center in the map and blue points are the most visited coordinates.

V. Conclusion

In this research, we have developed a mobile application that collects sensor data and a web application that processes and visualizes that data using machine learning algorithms. We build our web application as an analysis tool for survey analysis for any smartphone sensor data. In this study, we have collected eight smartphone sensor data through our android app. Then we extracted the GPS data for spatial analysis, applied the DBSCAN algorithm to find the most visited places, and presented this analysis on our web app. The steps of a machine learning analysis like pre-processing, filtering, training, etc., are done on our web server, which hosts our machine learning codes. Hence, we can easily view valuable insights of sensor data captured through our android application, such as demographic distribution or gender-wise spatial behavior.

The study's primary goal was to make machine learning analysis of smartphone sensor data accessible to those with no or limited machine learning experience. So that any user or company can have a machine learning analysis on smartphone data and get a result without developing and hiring any developer and analyst. Using our mobile application and the AutoML based web platform, they can quickly get their desired outcome.

The main weakness of this study was the paucity of self-optioned volunteers. Due to COVID-19, we cannot reach people to collect data. Currently, we have 90+ users who have been giving data continuously for more than a month. More research using controlled trials is needed to analyze the data more meaningfully. Further research includes the development of the auto ML function in our web platform and developing it so that it can cover most of the analysis that can be done with smartphone sensor data.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Each author has contributed equally in conducting the research, analyzing the data, and writing the paper. All authors had approved the final version.

References

[1] Sensors overview. [Online]. Available: https://developer.android.com/guide/topics/sensors/sensors_overview

[2] K. E. A. Stanley, "How many days are enough? Capturing routine human mobility," *International Journal of Geographical Information Science*, vol. 32, no. 7, pp. 1485-1504, 2018.

[3] S. Wang, C. Chen, and J. Ma, "Accelerometer based transportation mode recognition on mobile phones," in *Proc. Asia-Pacific Conference on Wearable Computing Systems*, 2010.

[4] N. F. Sugie, "Utilizing smartphones to study disadvantaged and hard-to-reach groups," *Sociological Methods & Research*, vol. 47, pp. 458-491, 2018.
[5] G. M. Harari, S. R. Müller, M. S. H. Aung, and P. J. Rentfrow, "Smartphone sensing methods for studying behavior in everyday life," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 83-90, 2017.
[6] D. V. Kuppevelt, J. Heywood, M. Hamer, *et al*., "Segmenting accelerometer data from daily life with unsupervised machine learning," *PloS One*, vol. 14, 2019.
[7] J. Krumm and E. Horvitz, "Predestination: Inferring destinations from partial trajectories," in *Proc. International Conference on Ubiquitous Computing*, 2006, pp. 243-260.
[8] L. Shao, L. Ji, Y. Liu, and J. Zhang, "Human action segmentation and recognition via motion and shape analysis," *Pattern Recognition Letters*, vol. 33, pp. 438-445, 2012.
[9] S. R. Ramamurthy and N. Roy, "Recent trends in machine learning for human activity recognition—A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, 2018.
[10] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometer," *SIGKDD Explorations*, vol. 12, pp. 74-82, 2011.
[11] P. Wang, Y. Zhang, S. Niu, and J. Guo, "Modeling temporal dynamics of users' purchase behaviors for next basket prediction," *Journal of Computer Science and Technology*, vol. 34, pp. 1230-1240, 2019.
[12] A. Crisan and B. Fiore-Gartland, "Fits and starts: Enterprise use of AutoML and the role of humans in the loop," in *Proc. CHI Conference on Human Factors in Computing Systems*, 2021.
[13] G. Cardone, A. Cirri, A. Corradi, *et al*., "Crowdsensing in urban areas for city-scale mass gathering management: Geofencing and activity recognition," *IEEE Sensors Journal*, vol. 14, pp. 4185-4195, 2014.
[14] J. Wahlström, I. Skog, and P. Händel, "Smartphone-Based vehicle telematics: A ten-year anniversary," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, pp. 2802-2825, 2017.
[15] Y. Mirsky, A. Shabtai, B. Shapira, *et al*, "Anomaly detection for smartphone data streams," *Pervasive and Mobile Computing*, vol. 35, pp. 83-107, 2017.
[16] S. Hemminki, P. Nurmi, and S. Tarkoma, "Accelerometer-Based transportation mode detection on smartphones," in *Proc. 11th ACM Conference on Embedded Networked Sensor Systems*, 2013.
[17] S. Majumder and M. J. Deen, "Smartphone sensors for health monitoring and diagnosis," *Sensors*, vol. 19, p. 2164, 2019.
[18] A. Bayat, M. Pomplun, and D. A. Tran, "A study on human activity recognition using accelerometer data from smartphones," *Procedia Computer Science*, vol. 34, pp. 450-457, 2014.
[19] A. Jahangiri and H. A. Rakha, "Applying machine learning techniques to transport mode recognition using mobile phone sensor data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 2406-2417, 2015.
[20] V. Y. S. Bharadwaj and V. N. Sastry, "Analysis on sensors in a smart phone-survey," *Inter. J. Innov. Res. Adv. Eng.*, vol. 1, pp. 96-108, 2014.
[21] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *Proc. IEEE 7th International Conference on Data Science and Advanced Analytics*, 2020.
[22] K. Chauhan, S. Jani, D. Thakkar, *et al*., "Automated machine learning: The new wave of machine learning," in *Proc. 2nd International Conference on Innovative Mechanisms for Industry Applications*, 2020.
[23] L. Tuggener, M. Amirian, K. Rombach, *et al*., "Automated machine learning in practice: State of the art and recent results," in *Proc. 6th Swiss Conference on Data Science*, 2019.
[24] M. D. Domenico, A. Lima, and M. Musolesi, "Interdependence and predictability of human mobility and social interactions," *Pervasive and Mobile Computing*, vol. 9, pp. 798-807, 2013.
[25] M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779-782, 2008.
[26] M. Ester, H. P. Kriegel, J. Sander, *et al*., "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD*, 1996.
[27] K. Khan, S. U. Rehman, K. Aziz, *et al*., "DBSCAN: Past, present, and future," in *Proc. Fifth International Conference on the Applications of Digital Information and Web Technologies*, 2014.
[28] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.
[29] S. R. Garzon and B. Deva, "Geofencing 2.0: Taking location-based notifications to the next level," in *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014.
[30] D. Namiot and M. Sneps-Sneppe, "Geofence and network proximity," in *Internet of Things, Smart Spaces, and Next Generation Networking*, Springer, 2013, pp. 117-127.
[31] A. Küpper, U. Bareth, and B. Freese, "Geofencing and background tracking--The next features in LBSs," in *Proc. 41th Annual Conference of the Gesellschaft für Informatik eV*, 2011.
[32] I. N. E. Indrayana, P. Sutawinaya, N. Pratiwi, *et al*., "Android-Based child monitoring application using a smartwatch and geofence service," *Journal of Physics: Conference Series*, 2021.
[33] P. A. Grabowicz, J. J. Ramasco, B. Gonçalves, *et al*., "Entangling mobility and interactions in social media," *PloS One*, vol. 9, 2014.
[34] A. Thiagarajan, J. Biagioni, T. Gerlich, *et al*., "Cooperative transit tracking using smart-phones," in *Proc. 8th ACM Conference on Embedded Networked Sensor Systems*, 2010, pp. 85-98.
[35] S. Reddy, M. Mun, J. Burke, *et al*., "Using mobile phones to determine transportation modes," *ACM Transactions on Sensor Networks*, vol. 6, pp. 1-27, 2010.
[36] G. Dong, Y. Jin, S. Wang, *et al*., "Db-kmeans: An intrusion detection algorithm based on dbscan and k-means," in *Proc. 20th Asia-Pacific Network Operations and Management Symposium*, 2019, pp. 1-4.

H. M. Zabir Haque is an Assistant professor at Ahsanullah University of Science and Technology, Dhaka, Bangladesh. He received his Master of Science in Bioinformatics from the University of Saskatchewan, Canada, and a Bachelor's degree in Computer Science and Engineering from the Ahsanullah University of Science and Technology, Dhaka, Bangladesh. His research interests include Bioinformatics, Computational Biology, and machine learning.