Wireless Networks (2022) 28:1–28
https://doi.org/10.1007/s11276-021-02819-4

ORIGINAL PAPER

A framework for monitoring movements of pandemic disease patients based on GPS trajectory datasets

Paulinus O. Ugwoke[1,2] · Francis S. Bakpo[1] · Collins N. Udanor[1] · Matthew C. Okoronkwo[1]

Accepted: 11 October 2021 / Published online: 26 October 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

The rapid spread of contagious diseases poses a colossal threat to human existence. Presently, the emergence of coronavirus COVID-19 which has rightly been declared a global pandemic resulting in so many deaths, confusion as well as huge economic losses is a challenge. It has been suggested by the World Health Organization (WHO) in conjunction with different Government authorities of the world and non-governmental organizations, that efforts to curtail the COVID-19 pandemic should rely principally on measures such as social distancing, identification of infected persons, tracing of possible contacts as well as effective isolation of such person(s) for subsequent medical treatment. The aim of this study is to provide a framework for monitoring Movements of Pandemic Disease Patients and predicting their next geographical locations given the recent trend of infected COVID-19 patients absconding from isolation centres as evidenced in the Nigerian case. The methodology for this study, proposes a system architecture incorporating GPS (Global Positioning System) and Assisted-GPS technologies for monitoring the geographical movements of COVID-19 patients and recording of their movement Trajectory Datasets on the assumption that they are assigned with GPS-enabled devices such as smartphones. Accordingly, fifteen (15) participants (patients) were selected for this study based on the criteria of residency and business activity location. The ensuing participants movements generated 157, 218 Trajectory datasets during a period of 3 weeks. With this dataset, mining of the movement trace, Stay Points (hot spots), relationships, and the prediction of the next probable geographical location of a COVID-19 patient was realized by the application of Artificial Intelligence (AI) and Data Mining techniques such as supervised Machine Learning (ML) algorithms (i.e., Multiple Linear Regression (MLR), k-Nearest Neighbor (kNN), Decision Tree Regression (DTR), Random Forest Regression (RFR), Gradient Boosting Regression (GBR), and eXtreme Gradient Boosting regression(XGBR) as well as density-based clustering methods (i.e., DBSCAN) for the computation of Stay Points (hot spots) of COVID-19 patient. The result of this study showed clearly that it is possible to determine the Stay Points (hot spots) of a COVID-19 patient. In addition, this study demonstrated the possibility of predicting the next probable geographical location of a COVID-19 patient. Correspondingly, Six Machine Learning models (i.e., MLR, kNN, DTR, RFR, GBR, and XGBR) were compared for efficiency, in determining the next probable location of a COVID-19 patient. The result showed that the DTR model performed better compared to other models (i.e., MLR, kNN, RFR, GBR, XGBR) based on four evaluation matrices (i.e., ACCURACY, MAE, MSE, and R^2) used. It is recommended that less developed Countries consider adopting this framework as a policy initiative for implementation at this burgeoning phase of COVID-19 infection and beyond. The same applies to the developed Countries. There is indication that GPS Trajectory dataset and Machine Learning algorithms as applied in this paper, appear to possess the potential of performing optimally in a real-life situation of monitoring a COVID-19 patient. This paper is unique given its ability to predict the next probable location of a COVID-19 patient. In the review of extant literature, prediction of the next probable location of a COVID-19 patient was not in evidence using the same Machine Learning algorithms.

Keywords Trajectory dataset · Covid-19 · Data mining · Machine learning · Movement · Pandemic

Extended author information available on the last page of the article

1 Introduction

Pandemic has always been a major challenge faced by humanity throughout history. It is virulent, non-indigenous, eruptive and cuts across physical expanse. It usually spreads rapidly in a country or in one/more continents at the same time. This implies that physical distance plays an important role in the spread of pandemic which leverages on human mobility [8–13]. Human beings are by nature social and as such interact through movements. It is therefore natural for people (small or large group of persons) to move from one location to another either by foot, ship, car, airplane, truck, or motorcycle for social or economic reasons which support their means of livelihoods [1–7]. This shows that man is a mobile being and as such is closely associated with the concept of *movement*. Historically, human movement has resulted in the eruption and aided the spread of previous pandemics which include but not limited to AIDS, Influenza, tuberculosis, Spanish Flu, Swine Flu, SARS, H7N9, Hong Kong Flu, Ebola, Zika [33] [34] and the current pandemic called coronavirus popularly referred to as COVID-19**.** Each pandemic has affected human life and economic development either positively or negatively.

The Coronavirus Disease 2019 (i.e., COVID-19) is an ongoing pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [14, 35]. The disease was first discovered in Wuhan [36], China, in early December 2019 [37, 38]. According to WHO, the virus which has spread to many countries globally has common symptoms of fever, dry cough, tiredness, and so on [39]. The disease spreads primarily from person to person through small droplets from the nose or mouth, which are expelled when a person with COVID-19 coughs, sneezes, or speaks. This means that close contact with someone with COVID-19 should be restricted so as not to be infected. An individual who contacts the virus in a particular geographical location, could easily spread it, through movement from one location to another and/or social interaction with another individual in another geographical location in a matter of hours [14]. Movement and social interaction are the two potent vehicles used to spread COVID-19. In view of this, WHO in partnership with several countries enforced restriction of movement of all sorts within the cities (except for medical emergencies, movement of essentials and services), as well as keeping social distance to curtail its spread. According to WHO, as at 6:57 pm CEST, 7 October 2021, there are 236,132,082 confirmed cases of COVID-19, including 4,822,472 deaths, and a total of 6,262,445,422 vaccine doses already administered globally [40].

Despite the mitigating measures adopted globally, there are several reported cases of fleeing or 'absconding' COVID-19 patients that have been tested and admitted in Isolation and Treatment Centres to curtail further spread [41]. This ugly situation has posed a serious challenge to Scientists which prompted the need to research on scientific and systematic ways of monitoring COVID-19 patients and to understand their movement patterns and how they socially interact with other innocent people to proactively curtail the spread [41, 42]. It is obvious that to curtail the spread of COVID-19 pandemic, it will involve breaking the chain of virus transmission by identifying and managing any individual or group of individuals that have been exposed to COVID-19 patient[s] to avoid further spread [29, 30, 44].

In view of the above, this study adopted GPS and Assisted-GPS technologies [45] as part of the proposed framework for monitoring COVID-19 patients assigned with smartphones and recording of their geographical location datasets also known as Trajectory datasets. By applying AI clustering algorithm such as DBSCAN and ML algorithms such as MLR, kNN, DTR, RFR, GBR, and XGBR on the recorded datasets determine the COVID-19 patient's Stay Points and any associated contact as well as the next probable location of a COVID-19 patient respectively. The ML is an algorithm that improves its performance automatically through experience while solving problems. The algorithms can learn on their own from the given data without being explicitly programmed thereby making them a good choice by Scientists for modeling safety–critical systems, just as applied in this study.

The rest of the paper is organized thus: Sect. 2 presents a concise and critical review of related works on efforts being made to trace the movement of covid-19 carriers. The materials used in this study are presented in different subsections under Sect. 3. Section 4 presents the methodology used to obtain the results presented in Sect. 5. Section 5 discusses the experiment and results of the findings, while Sect. 6 draws conclusions based on the findings, stating the practical implications.

2 Review of related works

Prior to this work, several authors have published research findings on human mobility [1–7], pandemic and social restriction policies [8–13], and movement monitoring & pandemic tracking [14–19]. The authors in [1] reviewed different approaches and models for learning as well as analyzing mobility patterns through the application of different machine learning models, while [2] dwelt on the discovery of human mobility patterns from 4G cell-phone geographical location datasets recorded every second for

each participated user. The study in [3] reviewed present mobility data types (i.e., phone records, GPS trace, and social media posts), characteristics, sources, and deep learning approaches for their next-location prediction. Authors in [4–6] researched on human mobility patterns, while [7] leveraged on GPS trajectory datasets to propose new metric for quantifying the degree of similarity of trajectories recorded within the same time frame. Interestingly, social restriction policies were also introduced by governments of different countries around the World as a strategy for curtailing the spread of COVID-19. These restrictions include early movement restrictions [8], international travel restrictions [9, 10], and lockdown of infected cities to reduce infection cases outside of those cities [11]. Authors in [12] developed a model for measuring/quantifying the degree of potential effects on various intracity mobility restrictions on the spread of COVID-19, while [13] revealed that the more the airports, and elderly population in a city, the more likely for that city to have rise in COVID-19 cases. The use of GPS technology to quantify human mobility is investigated in [14].

Other researchers such as [15] proposes a reconstruction of the epidemic curves from the fractal interpolation point of view, [16] carries out a review of 63 scientific articles on geospatial and spatial-statistical analysis of the geographical dimension of the 2019 coronavirus disease (COVID-19) pandemic, [17] critiques recent studies that apply to machine learning and artificial intelligence technologies towards augmenting the researchers on multiple angles in tackling COVID-19 pandemic. Yet much research works is still anticipated because the virus is still trending.

This research work falls under the broad term of covid-19 contact tracing based on GPS Trajectory Data Mining [3]. Contact tracing is one of the import surveillance strategies for controlling the spread of a Pandemic disease such as COVID-19. It is a process of monitoring persons who have been exposed to another person infected with the covid-19 virus. Contact tracing involves identification, listing and following up of persons who encountered an infected person (i.e., a person who has been tested positive for COVID-19) to monitor their symptoms for the next 14 days (2 weeks).

During the past nine months, several research groups have developed a shared privacy minded protocols (i.e., TCNCoalition [18]) for seamless access of infection status across users of the different systems. These groups including the Tracetogether [19] team in Singapore, the Private Automated Contact Tracing (PACT) group [20] led by researchers at the Massachusetts Institute of Technology (MIT) in Cambridge as well as participants from Boston University, Covid-watch [21], CoEpi [22], and the largely European consortium Decentralized Privacy-preserving Proximity Tracing (DP-3 T) [23, 24]. These teams have embraced a basic concept based on Bluetooth technology [25]. A smartphone regularly broadcasts a random string of characters that serve as a pseudonym to other closer phones using Bluetooth low-energy specification for sending short bursts of data (i.e., anonymized Bluetooth proximity data). In this respect several countries such as the US, China, South Korea, Singapore, Australia, Israel, Germany, etc., have adopted the use of mobile Apps to support contact tracing of COVID-19 patient [26]. Similarly, Apple and Google recently collaborated to build internal Apps into their individual phone operating systems, with the support of the PACT group [27, 28].

In appraising the extant literature reviewed above, it is obvious that what is not known generally which this paper demonstrates clearly is that Bluetooth technology is not adequate for contact tracing on account of its inherent limitations, viz: requires pairing of smartphone devices within a specific range (10 m) which makes it more suitable for Social distancing awareness and notifications. As opposed to the above, our approach is based on GPS and Assisted-GPS technologies which work independent of other smartphone devices, to collect trajectory data and transmit to a cloud-based database. This means that every device logs its own geographical locations as well as the time stamps when the locations were captured. Whenever a user is reported as infected, his/her trajectories (sequences of locations and time) would be subjected to analysis. Hence, the pseudonymous trajectories of all infected users (i.e., COVID-19 patients) with every other user (i.e., non-COVID-19 patients) are required to check if they were in close contact with an infected individual for prompt action to be taken [29]. Although a few isolated studies [30, 31] and [32] applied trajectory datasets as well as Machine Learning tools for contact tracing of pandemic cases, yet, non-applied or utilized these tools to compute the Stay Points of COVID-19 patient or for the prediction of the next probable geographical location of a COVID-19 patient. Thus, a research gap exists in this important area which this paper purposes to close.

3 Materials

The study of human mobility and its patterns remains very significant in gaining insights on the spread of pandemic diseases such as in the case of COVID-19. This spread of pandemic diseases usually occur through person-to-person contacts resulting from Social interactions motivated by human mobility. Human movements are observed not only across space (location), but also through time. Therefore, it is required to record daily movement of individuals such as where they go and how long they stay at such places as well as different persons they have associated with.

Through such way, the movement path (i.e., trajectory) of every person can be described in daily space–time, i.e., left home 6.00 am, being at work by 7:30 am, left for the doctor's appointment at 11:30 am, got there at 12:00 noon, and so on.

Torsten Hagerstrand in Hugo et al. [6], enunciated the cannons of time geography in the early 1950's with due emphasis on the relevance of time and provided a conceptual and graphical representation of the trajectories of individuals over space and time. His basic idea was to consider space–time paths in a 3-dimentional space (3D space) and the vertical axis represents time. As depicted in [42], it means that space and time are inseparable in Hägerstand's time geography, hence, making the graphic representation of his ideas very clear if carefully studied.

Presently, owing to the evolution of ubiquitous computing such as Internet of Things, the Human mobility data such as Social media data, Mobile phone data, GPS trace, etc., has continued to grow at an unprecedented breath, depth, and scale. This growth has motivated researchers to strengthen Surveillance through digital traces (i.e., *digital footprints*) left by people while interacting with the cyber-physical spaces (i.e., wireless sensor networks and mobile/wearable devices). These data are usually generated by the data acquisition sensors and devices such as GPS-enabled Smartphones and the rest carried by People as well as communications devices such as Satellites (GPS), Radio Frequency Identification (RFID), CCTV, GSM Network, Infrared systems, Bluetooth, and so on [48].

Accordingly, leveraging the capacity to collect and analyze the "*digital footprints*" at community scale according to [49–53] has given rise to revealing the patterns of human (individual/group) behaviors, social interactions as well as community dynamics (eg., city hot spots, traffic jams, and significant locations).

This study tries to take advantage of this current (high rate) penetration of mobile and wearable devices such as smartphones as well as other GPS-embedded sensors to monitor, track, and record the geographical locations of moving objects (i.e., COVID-19 patients) in real-time. Furthermore, it is obvious to note that greater percent of people around the world today move about with at least one or more of these smart devices (i.e., smartphones, PDAs, and so on), in their hands, bags, or pockets. Consequently, this rampart possession of these mobile devices by people has increased the interest in collecting and analyzing their movement data both day and night.

A smartphone usually possesses extensive computing and communication capabilities such as high-speed Internet access using both Wi-Fi as well as mobile broadband. Presently, most, if not all, smartphones and other mobile devices have satellite navigation features, meaning that GPS on smartphones is no longer an emerging trend, but a must-have feature. This shows that in identifying and capturing mobile GPS receiver's user location-points also known as, the GPS and/or other positioning technologies are required [45].

Interestingly, with the GPS on mobile devices, massive amount of trajectory datasets can be generated and recorded in real-time, locally on the device, or to a server computer. Gang et al. [54] explained that, usually, trajectory datasets of mobile devices (i.e., smartphones, PDAs, etc.) approximately reflect the time series traces of their owners. According to [55], such time series data is very useful in analyzing movements of the user under monitoring.

In view of the above, it has become obvious that time and locations are paramount in human movement and activity monitoring. Hence, if the geographical locations of moving objects (say, COVID-19 patients) are recorded in real-time at regular spaced time interval, e.g., every 5 minutes, then the recorded geographical location points when logically linked, forms a trajectory, or set of trajectories which may contain vital information. The information obtained from the analysis of trajectory datasets can help in strategic decision making such as controlling, predicting, preventing the spread of infectious diseases such as COVID-19. Therefore, this study shows how people (i.e., COVID-19 patients) can be monitored through GPS, and their logged trajectory datasets be used to infer their movement patterns by applying Data Mining / Machine Learning algorithms.

3.1 Notations

This subsection presents some of the notations used in this paper and their meanings. The notations and their definitions are as presented in Table 1 below:

3.2 Problem statement/formulation

With the advent of COVID-19 in Nigeria, issues of contact tracing and monitoring became paramount. The recent trend of absconding COVID-19 patients from treatment/isolation centres further compounded and indeed undermined Government efforts to curtail the spread of COVID-19 virus.

Utmost to Government from the onset, was the need to maintain adequate communication with the infected persons for subsequent treatments. Given this disposition by some infected persons to abscond from the isolation centres without further trace, our paper as a pro-active research endeavor goes beyond mere contact tracing and monitoring to predicting the next probable geographical location of the COVID-19 patients. This is the problem that this paper set

Table 1 Notations

S/N	Notation	Definition
1	SMTPHONE	The set of smartphones
2	*smtphone*	A smartphone
3	USER	The set of smartphone users
4	*User*	A smartphone user, i.e., COVID-19 patient
5	*p*	The geographical location point
6	*TR*	The set of geographical location points, also denoted as trajectory
7	Ψ	The set of trajectories, also denoted as location log
8	LDT	The set of location logs, also denoted as location data table
9	O_{id}	The unique object identifier
10	*lat*	The latitude in decimal degree
11	*lon*	The longitude in decimal degree
12	*alt*	The altitude in decimal degree
13	*T*	The date and time lat, lon, and alt were recorded
14	*subTR*	The sub-Trajectory
15	*stp*	The Stay Point
16	T_{arv}	The arrival time on stay point
17	T_{dep}	The departure time on stay point
18	*STP*	The set of Stay Points
19	μ_{lat}	The mean latitude of a stay points
20	μ_{lon}	The mean longitude of a stay points
21	ΔT	The time interval between departure & arrival time on stay point
22	GEODIST (*a*, *b*)	The geographical distance between two location points, *a* and *b*
23	TIMEDIFF (*a*, *b*)	The time difference between two location points, *a* and *b*
24	Length(*X*)	The no of elements in *X*
25	*lut*	the land use type
26	*LUT*	The set of land use types
27	T_{min}	The minimum time threshold
28	D_{max}	The maximum distance threshold
29	ID, id	The unique identification
30	$\mathcal{M}$	The mapping function

out to address. The problem formulation is as set out below:

I. Given a set of m Smartphones, SMTPHONE = $\{smtphone_l | \ 1 \leq l \leq m\}$, and a set of n Smartphone-users (i.e., both COVID-19 patients and non-COVID-19 patients), USER = $\{user_u \mid 1 \leq u \leq n\}$ moving in (x, y)-plane such that a Smartphone(s), $smtphone_l$, is assigned to a Smartphone-user, $user_u$. Our aim is to read and record (using assigned Smartphone, $smtphone_l$, for instance), a sequence of date/time-stamped smartphone-user's geographical location points, $\{p_1, p_2, p_3, \ldots, p_i, \ldots\}$, in a log (also called location-log, and denoted with Ψ), as he/she moves about with the assigned Smartphone(s) in a particular geographical area.

II. Given a set of m location-logs, LDT = $\{\Psi_l | 1 \leq l \leq m\}$, known as location-datatable of moving persons (i.e., both COVID-19 patients and non-COVID-19 patients); where each location-log, denoted as, $\Psi_l = \{\langle O_{id} \rangle, TR_Z | 1 \leq Z \leq m\}$, is a set of trajectories,TR_Z, generated by an assigned Smartphone, with a unique object identifier, $\langle O_{id} \rangle$. Our aim is to determine the following:

i. the movement trace of COVID-19 patient in a geographical area?

ii. the geographical location(s) where COVID-19 patient spent some time?

iii. the relationships of COVID-19 patient with a person or group of persons in a geographical area and at given time interval?

iv. the prediction of the next location of COVID-19 patient given the geographical locations history?

3.3 A trajectory

In view of the concept of human mobility as described in the introduction to Sect. 3 above, a trajectory, TR, of a person's movement, can be represented as a chronological (continuous) sequence of multidimensional geographical location points, p_i, ordered according to the time, T_i, they were visited by the object (i.e., person),O, under observation. Accordingly, p_i is usually characterized by a coordinate system such as latitude (x_i), longitude (y_i), altitude (h_i), and time (T_i), i.e., $p_i = (x_i, y_i, h_i, T_i)$. The h-part, i.e., elevation, may be ignored in the trajectory dataset, because, the changes of the h-part are very small, hence insignificant, especially for trajectories recorded within cities, therefore, p_i, may become (x_i, y_i, T_i) in such case, where its base (x- and y- coordinates) represents spatial dimensions (geography) and the vertical line (T- coordinate), perpendicular to the base, represents time dimension [45].

Therefore, if given a moving object's Location Log, Ψ, which denotes a set of trajectories, $\{TR_Z | 1 \leq Z \leq \text{Length}(\Psi)\}$, where the moving object,O, is identified by a unique identifier $\langle O_{id} \rangle$, then, TR_Z can be formally represented as sequence of such quadruple tied with the unique object identifier $\langle O_{id} \rangle$ associated with their recording: $\{\langle O_{id} \rangle, (lat_{Z_i}, lon_{Z_i}, alt_{Z_i}, T_{Z_i}) | 1 \leq i \leq \text{Length}(TR_Z), T_{Z_i} < T_{Z_{(i+1)}}\}$; where (lat_{Z_i}, lon_{Z_i}) represents a spatial location $(location_{Z_i})$ visited by the moving object at timestamp, T_{Z_i}.

3.4 Study area

This study focuses on Oshodi-Isolo Local Government Area (LGA). This LGA which is formed by the second republic Governor of Lagos State, Alhaji Lateef Kayode Jakande is in the southern part of Lagos State, Nigeria (See red map in Fig. 1) and it constitutes 11 wards including Oshodi/Bolade, Orile Oshodi, Mafoluku, Sogunle, Sogunle/Alasia, Isolo, Ajao Estate, Ilasamaja, Okota, Ishagatedo, and Oke-Afa/Ejigbo. It has a population density of about 1,000,509 (in 2017) and area extent of approximately 45 square kilometres. The LGA lies within GPS coordinates, latitude 6° 58'N, longitude 3° 39'E, latitude 6° 50'N, and longitude 3° 35'E.

The residential land use (i.e., 35.87% of total land uses) is the dominant land area which includes, Oshodi, Isolo, Ilasamaja, Oke Afa, and Mafoluku. Next predominant land area is the Commercial land use (i.e., 11.27% of total land uses), found along major roads such as Agege motor road, Airport road, Oshodi-Apapa expressway. Another predominant land area is the Industrial land uses which accounts for 7.67% of total land area which includes Ilasamaja industrial area, Abimbola industrial area, Iyana Isolo industrial zone. And finally, the circulation land use which account for 10.10% of total land area includes major arterial roads within the local government such as Oshodi-Apapa Expressway, Agege Motor Road, Airport road, Cele Express link road. Other land use such as agriculture/open (i.e., 11.27% of total land uses), mixed use (i.e., 5.60%), and public/educational land use [57].

3.5 The proposed system architecture

The Fig. 2 presents our proposed high-level architecture for monitoring Movements of both Pandemic disease (i.e., COVID-19) patients and non-COVID-19 patients. This architecture is a conceptual framework integrating different functionalities of the system such as data capture, Communication/transmission, storage, processing, and visualization. It is a Client–Server architecture organized into three segments (from top to bottom). These include: the client (i.e., *GPS Location Data Capture and* (*Web*) *User Interface*), the server (i.e., *Data Pre-processing, Moving*

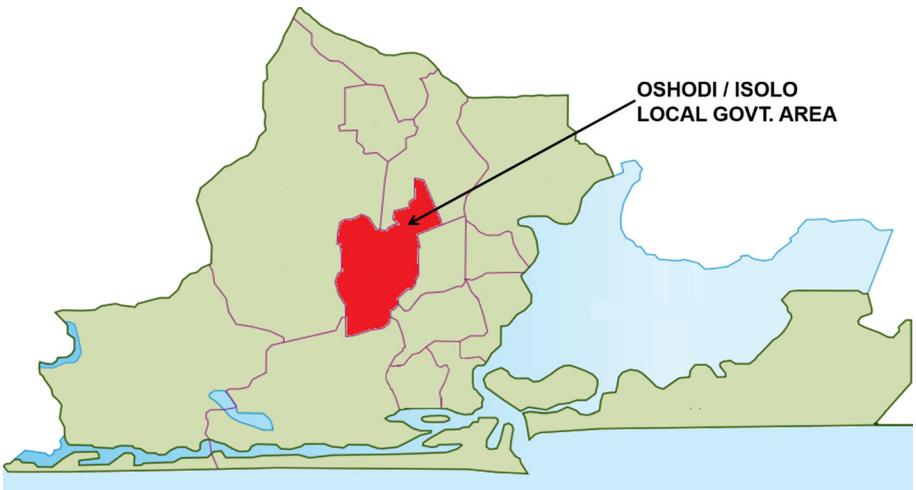


Fig. 1 Map of Lagos State (Nigeria) showing the Selected Study Area on Red: the OSHODI/ISOLO LGA [56]

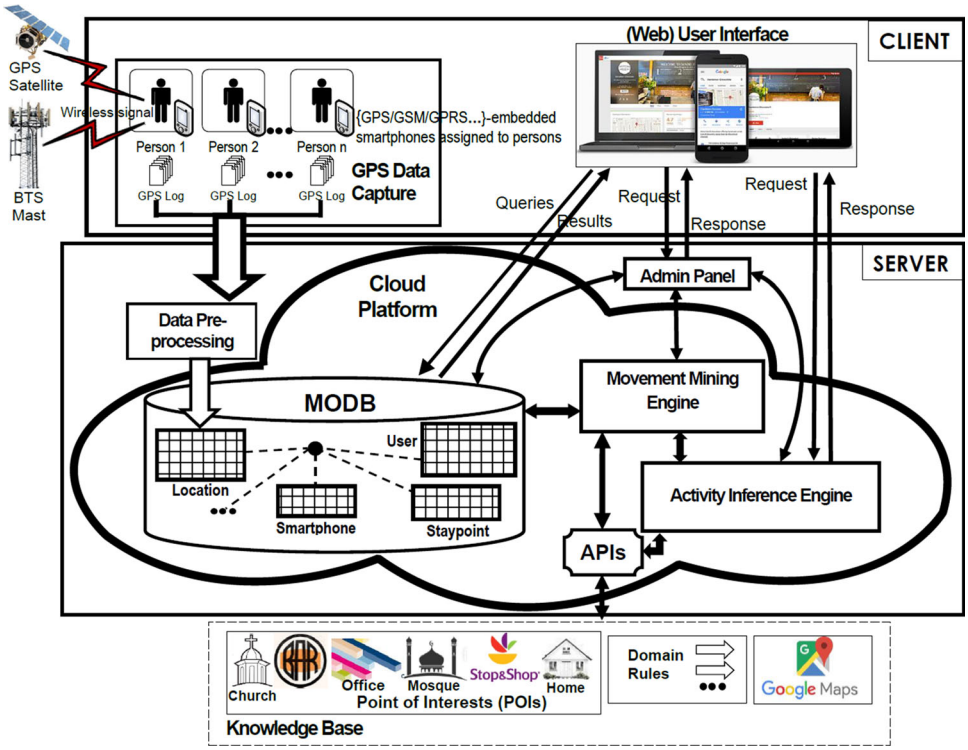


Fig. 2 The System Architecture

Object Database (*MODB*), *Admin panel, Movement Mining Engine, Activity Inference Engine, and the Application Programming Interfaces (APIs)*) which is connected to the knowledge base (i.e., Point *of Interests* (*POIs*), *Domain Rules, and Geographical Data*).

3.5.1 The client segment

This segment is composed of two modules, GPS Location Data Capture (LDC) and the (Web) User Interface for system administration. The GPS LDC which has already been developed in our earlier work [45] is reliant on a cellular network infrastructure consisting of Base-Transceiver Stations (BTS) and Mobile Switching Centre (MSC), a network of GPS satellites in space, and GPS-enabled android-based smartphones attached or assigned to individuals, in this case, COVID-19 and non-COVID-19 patients. Each smartphone is uniquely identifed by its International Mobile-station Equipment Identity (IMEI). Hence, geographical locations with their corresponding time stamps as well as all numerous users' information are captured & logged into the cloud based Moving Object Database (MODB) server for storage using a client–server model as shown in Fig. 2 above.

The second part of the Client segment is the (Web) User Interfaces module which is the interface that system users can utilize to manipulate the proposed system via the admin panel, movement mining engine, and the activity inference engine. A query can be issued from this segment to perform exploratory data mining. It can also allow system users to browse the MODB database, evaluate and/or visualize mined patterns. Below are some of the compulsory requirements and assumptions considered at the Client side prior to monitoring:

3.5.1.1 Compulsory requirements

i. The users (both COVID-19 and non-COVID-19 patients) must be duly registered with National Identity Management Commission (NIMC) Nigeria to have their respective unique National Identity Number (NIN).
ii. The users (both COVID-19 and non-COVID-19 patients) must have their respective SIM (Subscribers Identity Module) cards duly registered with Telecommunication local operator(s), i.e., Mobile Network Operators (MNOs).
iii. It is required that a registered phone belongs to one person since the SIM card is registered in one person's name (both COVID-19 and non-COVID-19 patients).

3.5.1.2 Assumptions

i. The users (both COVID-19 and non-COVID-19 patients) must respectively possess a functional smartphone with valid IMEI (International Mobile-station Equipment Identity) numbers.
ii. The SIM cards are always fixed on their respective smartphones as such connected.

iii. The Location manager of the individual mobile app (i.e., GPS LDC) must always be on.
iv. The users (both COVID-19 and non-COVID-19 patients) must always be with their respective smartphones.
v. The phones must not be switched off.
vi. All the point of interest (POI) must have been registered in an existing on-line database.

3.5.2 The Server segment:

This segment consists of four modules, Data Pre-processing, Moving Object Database (MODB), Admin panel, the Movement Mining Engine, Activity Inference Engine, and Application Programming Interfaces (APIs) to access the components of the knowledge base segment.

Data Pre-processing: This is where data cleaning, data transformations, as well as data normalization are performed.

Moving Object Database (MODB): This is the data storage & management component of the proposed system that stores data about Users, Smartphones, as well as geographical locations of users under monitoring. The MODB resides in the cloud storage infrastructure. It is housed and managed by MySQL Database Management System. The data in the MODB can be manipulated/ processed by individuals or programs such as data mining programs/Machine Learning algorithms.

Admin panel: The admin panel provides interface where users can manipulate the system as well as the MODB.

Movement Mining Engine: This is a set of functional programs for tasks such as, Stay Point clustering/estimation, semantic enrichment of Stay Points, association mining, and location prediction.

Activity Inference Engine: This is where the inference about the discovered patterns are made and subsequently, result will be visualized by the user.

Application Programming Interfaces (APIs): The APIs are used to access the components of the knowledge base segment such as the Points of interests (POIs), Google maps, etc.

3.5.3 The knowledge base

This segment contains the secondary data that will support the proposed system such as the Point of Interests (POIs), Domain Rules for guiding search of interesting patterns, and Geographical Data such as Google maps.

3.6 The ontology model

The Fig. 3 represents object types and conceptual ontology model containing trajectory, GPS-enabled device, Activity, POIs, and Stay. This ontology model presents the person under monitoring as well as their trajectory behaviors.

As shown in Fig. 3, the rectangular shapes depicts entitites while the arrows depicts relationships between two entities. Therefore, from the figure, a trajectory is composed of sub-trajectories such as a Stay and Mobility Patterns. A stay is associated with Activity and Point of Interest (POI). A Stay also has Time, Semantic characteristics, POI, and also activity. The relationships between two entities as presented in Fig. 3 above includes HasActivity, IsComposedOfStay, HasPattern, Is_a, and so on.

3.7 Trajectory datasets

The data sample used in this study are Trajectory datasets (i.e., Location datasets) which were generated using a sample of 15 participants (quantitative data) out of 31 participants drawn from Oshodi-Isolo, Local Government Area of Lagos State, Nigeria. The Trajectory datasets are datasets of geographical positions (i.e., latitude, longitude, altitude, time, etc.) of mobile devices such as smartphones or tablets logged in real-time. The details of participant's records were not revealed due to privacy issues. The tracking App was installed on individual Smartphones of the participants to facilitate experimental contact tracing. The participants willingly accepted to be involved in the study. Although about Eleven (11) participants had technical issues with their phones and could not report location data. Five (5) participants changed their mind for reasons best known to them. Fifteen (15) participants participated fully, though, with varied numbers of location datasets reported. A total of 157, 218 datasets were logged to the database by the 15 participants within three weeks period. All the participants are adults, thus, there were no ethical issues involved. The data generated through this process was recorded or logged in real-time as the users wanders about. A sample of recorded location history is as shown in subsequent section of this paper (i.e., Fig. 8).

The Tables 2 and 3 below describes the data dictionary/ structure for geographical Locations datasets and users' details (i.e., COVID-19/ non COVID-19 patient's information), respectively. The relationship between User ID from User table (i.e., Table 2) and Location ID from Location table (i.e., Table 3) is one-to-many, meaning that a User can generate many Location points.

Fig. 3 Ontology Model

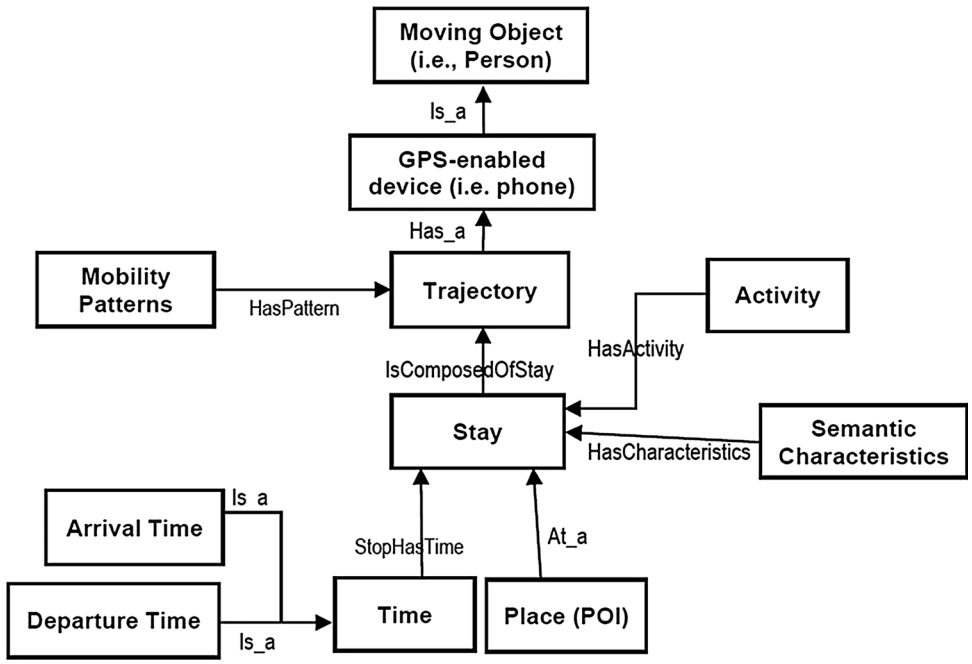


Table 2 Location – This data dictionary describes geographical Locations' information

S/N	Variable name	Description
1	Location ID	PK, unique identification (ID) no. of the location
2	User ID	unique national identification (ID) of the user (either COVID-19 or non-COVID-19 patients)
3	Phone ID	unique phone International Mobile-station Equipment Identity (IMEI) no
4	Lat	Latitude in decimal degrees
5	Lon	Longitude in decimal degrees
6	Alt	Altitude in meters
7	Timestamp	Date and Time of location capture
	Estimated Accuracy	Location estimated accuracy provided by GPS system in meters

Table 3 User – This data dictionary describes COVID-19/ non COVID-19 patient's information

S/N	Variable name	Description
1	User ID	PK, unique National Identification (ID) of the user, i.e., NIN
2	Surname	Surname of the user
3	Firstname	First name of the user
4	Othername	Other names of the user
5	Gender	User gender
6	Designation	Designation of the user
7	State	State of origin of user
8	Lga	Local Govt. Area where the user is from
9	DOB	Date of birth of user
10	Address	Address of user
11	Mobile No	Phone no(s) assigned to the user
12	Phone ID	IMEI of the user
13	Date	The date this record was registered
14	Health Status	Tested Negative / Positive

4 Methods

We present the approaches adopted by the research work in realizing the objectives of the proposed system within this section.

4.1 Data pre-processing

4.1.1 Testing for stationarity

When modeling time series data, stationarity is of great significance. Hence, the data must be tested for non-stationary. If the test is true, meaning that it is non-stationary, then it must be transformed (differenced) to stationary time series before modeling. Sometimes this test can be simply achieved by visualization of the time series graph which is not usually the best. Therefore, more accurate statistical methods such as unit root test, e.g., Augmented Dicky-Fuller (ADF) [93] and Johansen Tests [94, 95] may be applied. In this paper, we applied Johansen Tests (i.e., Tables 4 and 5), because of its capabilities in handling multivariate time series data. The test statistic and critical value in Trace test statistic and Eigenvalue test statistic tables were computed by passing Locations (i.e., Trajectory) dataset to our Python Programming codes.

The assumption of Johansen Tests about the time series data is through statistical hypothesis such as Null hypothesis which says there is no stationarity (H_0: No Stationarity) and the alternative hypothesis which says there is stationarity (H_1: Stationary). Therefore, the decision rule is: Reject H_0 if the test statistic (calculated) $\geq$ Table (-critical) value.

Johansen cointegration test using trace test statistic with 1% significance level.

Johansen cointegration test using maximum eigenvalue test statistic with 1% significance level.

The Trace test statistic values (1.044e + 04,2000.0, and 773.0) are greater than their corresponding critical values (35.46,19.93, and 6.635), so we can declare statistical significance and hence reject the null hypothesis (H_0) at a 99% confidence level, as the magnitude of the trace statistic is greater than the critical value, meaning that there is no cointegration, the alternative hypothesis is that there is cointegrating relationship. Therefore, the time series are not cointegrated.

Table 4 Trace test statistic

r_0	r_1	Test statistic	Critical value
0	1	1.044e + 04	35.460
1	2	2000.0	19.930
2	3	773.0	6.635

Table 5 Eigenvalue test statistic

r_0	r_1	Test statistic	Critical value
0	1	8444.0	25.860
1	2	1227.0	18.520
2	3	773.0	6.635

Similarly, the eigen statistics stores the eigenvalues in decreasing order of magnitude, they tell us how strongly cointegrated the series are or how strong is the tendency to mean revert. In our example, the eigen statistic for the null hypothesis, therefore, can be rejected at a 99% confidence level, since the test statistic values (8444.0,1227.0, and 773.0) are greater than their corresponding critical values (25.86.46,18.52, and 6.635). The above test shows that the series are cointegrated, hence the null hypothesis must be rejected and the alternate hypothesis to be accepted, meaning that our time series dataset is stationary. Therefore, we can proceed for the data modeling as presented in the subsequent sections.

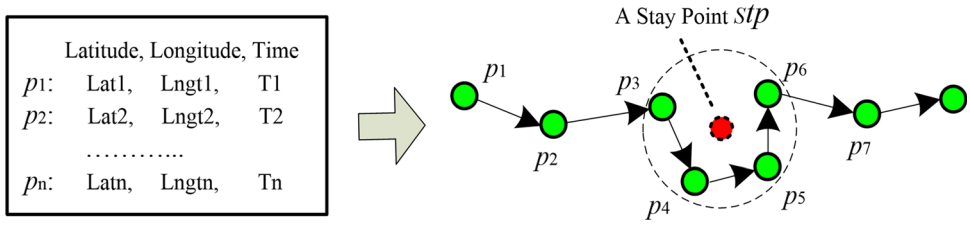
4.2 Mining movement patterns

This section demonstrates how patterns are mined from trajectory datasets (i.e., Location table) of both COVID-19 and non-COVID-19 patients. After the capturing and recording of the trajectory datasets, the data must be modeled and analyzed by applying data mining algorithm(s) to extract useful patterns. A pattern according to [58] is a way in which something happens, moves, develops, or is arranged. In the contest of a COVID-19 or a non-COVID-19 patient, trajectory patterns can be conceptualized as the representation of behaviors of his/her movement in both space (i.e., regions of space visited during movement) and time (i.e., the duration of movements). These include any recognizable spatial and temporal regularity or interesting relationship in a recorded moving object Trajectory dataset [59].

4.2.1 Stay points estimation

The estimation of geographical Stay Point, *stp*, from geographical location points, $\{p_i | 1 \leq i \leq n\}$, can be achieved by applying the Stay Points Estimation Algorithm following the process in Fig. 4 to cluster Stay Points of GPS phone user (i.e., COVID-19 patient).

Fig. 4 A Stay Point



This Algorithm is an improved DBSCAN (Density-Based Spatial Clustering of Application with Noise) [60–62] method based on the following definitions.

Definition 1- Geographical Distance and Time Difference Given two location points, p_{Z_i} and p_{Z_j}, the geographical distance between the two points, p_{Z_i} and p_{Z_j}, is denoted as $\text{GEODIST}(p_{Z_i}, p_{Z_j})$, while the time difference between the two points, p_{Z_i} and p_{Z_j}, is denoted as $\text{TIMEDIFF}\left(p_{Z_i}, p_{Z_j}\right) = \left|p_{Z_j} \cdot T - p_{Z_i} \cdot T\right|$.

Definition 2-Stay Point Given a trajectory,$TR_Z = \{\langle O_{id}\rangle, (lat_{Z_i}, lon_{Z_i}, alt_Z, T_{z_i}) | 1 \le i \le \text{Length}(TR_Z), T_{Z_i} < T_{Z_{(i+1)}}\} \subseteq \Psi$, a Stay Point, *stp*, stands for a geographical location where a user stayed over a time threshold, T_{min}, within a distance threshold of, D_{max}. In a moving object's trajectory (TR_Z), a Stay Point, *stp*, can be seen as a virtual location characterized by sub-trajectory, $\{SubTR_Z = \{p_{Z_i}, p_{Z_{(i+1)}}, \ldots, p_{Z_k}, \ldots p_{Z_j}\} \Rightarrow p_{Z_i} \rightarrow \ldots \rightarrow p_{Z_j}$, Thus, $\forall i \le k \le j$, $\text{GEODIST}(p_{Z_i}, p_{Z_k}) \le D_{max}$, $\text{GEODIST}(p_{Z_i}, p_{Z_{(j+1)}}) > D_{max}$, and $\text{TIMEDIFF}(p_{Z_i}, p_{Z_j}) > T_{min}$.

Therefore, $stp = ([lat, lon], T_{arv}, T_{dep})$, where $stp \cdot lat = \frac{\sum_{k=i}^{j} p_{Z_k} \cdot lat}{|subTR_Z|}$ and $stp \cdot lon = \frac{\sum_{k=i}^{j} p_{Z_k} \cdot lon}{|subTR_Z|}$ respectively stands for the average latitude (*lat*) and longitude (*lon*) of the collection, $SubTR_Z$.

This can as well be called the spatial mean centre (or centroid) and can be represented as $\left[\frac{\sum_{k=i}^{j} p_{Z_k} \cdot lat}{|subTR_Z|}, \frac{\sum_{k=i}^{j} p_{Z_k} \cdot lon}{|subTR_Z|}\right]$. Similarly, $stp \cdot T_{arv} = p_{Z_i} \cdot T$ and $stp \cdot T_{dep} = p_{Z_j} \cdot T$ respectively represent the user's arrival and departure time on Stay Point, *stp*.

4.2.2 Stay points annotation

A Stay Point, *stp*, does not have explicit meaning; thus, it is necessary to cluster Stay Points into Point of Interests sequences [63]. A Point of Interest (POI) may be a geo-referenced object or location like home, company, church, office, mosque, gym, and so on where a person may carry out a specific activity or might find useful/interesting. Therefore, every Stay Point, *stp*, must be associated with a POI or set of POIs as shown in Fig. 5 below.

Supposing a moving person (i.e., a COVID-19 or a non-COVID-19 patient) attached with a GPS-device moves from one location (i.e., source) to another (i.e., destination) at a particular time interval, the person is assumed to have stopped in the destination because he/she is attracted by the location. Hence, the geographical objects that could represent the goal of the stop are called POIs. It is also possible that the person may have stopped in different locations before getting to the destination (stop), where each stop may be associated with one or more POIs.

This study is more concerned about Stay Points which are places where someone or some persons have spent some time to perform an activity or some activities. This becomes relevant in the context of tracing the potential COVID-19 patient's contacts as the patient may have converged with other person(s) at the same location or different locations for related purposes. These Stay Points (computed from sub-trajectories) need to be annotated with environmental information such as the most probable visited POI category type and land use type (*lut*) surrounding them to infer the purpose of stopping at the Stay Point. To ascertain that a COVID-19 patient is prone to visiting any available POIs, geographical information sources such as Google or digital street maps could be used to gather background data of the Stay Point. Below is the description of the annotation process:

4.2.2.1 Annotation with land use types:

Stay Points are annotated with categories of land use types (*LUT*) to measure the topological correlation between each Stay Point and the semantic area using spatial join.

Let the set of land use types, $lut_i, 1 \le i \le \text{Length}(LUT)$, be represented as:

$$LUT = \left\{lut_1, lut_1, \ldots, lut_i, \ldots, lut_{\text{Length}(LUT)}\right\}$$

where $\left\{lut_1, lut_1, \ldots, lut_i, \ldots, lut_{\text{Length}(LUT)}\right\} \Rightarrow$ {commercial, Industry, Residential, Parks and recreation, institution, …}. The Algorithm 1 below describes how to choose the LUT.

Algorithm 1: ChooseLUT (SP_z, LUT): L_{Area}

INPUTS: Stay Point, $STP_z = \{stp_i \mid 1 \le i \le \text{LENGTH}(STP_z)\}$
Land use type, $LUT = \{lut_i \mid 1 \le i \le \text{LENGTH}(LUT)\}$

OUTPUT: Land use type decoded for each Stay Point, $L_{Area} = \{a_i \mid 1 \le i \le \text{LENGTH}(L_{Area})\}$

PROCEDURE:

1: *Begin*
2: $L_{Area} \leftarrow \{\}$
3: for all $stp_i = (\mu_{lat}, \mu_{lon}, T_{arv}, T_{dep}, \Delta T)$ **in** STP_Z
3.1: **if** $STP_Z \bigcap LUT$ **then**
3.1.1: $a_i = (stp_i \bullet \mu_{lat}, stp_i \bullet \mu_{lon}, stp_i \bullet T_{arv}, stp_i \bullet T_{dep}, stp_i \bullet lut)$
3.1.2: $L_{Area} \leftarrow L_{Area} \bigcup a_i$
3.2: **else**
3.3: **obtain nearest** LUT to stp_i
3.3.1: $a_i = (stp_i \bullet \mu_{lat}, stp_i \bullet \mu_{lon}, stp_i \bullet T_{arv}, stp_i \bullet T_{dep}, stp_i \bullet lut)$
3.3.2: $L_{Area} \leftarrow L_{Area} \bigcup a_i$
3.4: **end if**
3.5: **return** L_{Area}
4: end for
5: End Begin

4.2.2.2 Annotation with probable visited POI categories: Just like the annotation with *LUT*, this is annotation of Stay Points with different POI categories [64–66]. The list of POIs and their respective category names must have been collected and recorded. For a given Stay Point, STP_Z, the Algorithm (i.e., Algorithm 2) for retrieving probable visited POI category types can use the following inputs:

1. A set of computed Stay Points, $STP_Z = \{stp_i \mid 1 \le i \le \text{LENGTH}(STP_Z)\}$, where each$stp_i = (\mu_{lat}, \mu_{lon}, T_{arv}, T_{dep}, \Delta T)$; $(stp_i \cdot \mu_{lat}, stp_i \cdot \mu_{lon})$ represents the coordinates (i.e., mean of the latitudes and longitudes, respectively) of the Stay Point,stp_i; $stp_i \cdot T_{arv}$ denote the arrival time at the Stay Point,stp_i; $stp_i \cdot T_{dep}$ denote the departure time at the Stay Point,stp_i; and $stp_i \cdot \Delta T = |stp_i \cdot T_{dep} - stp_i \cdot T_{arv}|$ is the time interval stayed.
2. A set of Point of Interests, POIs = $\{poi_j \mid 1 \le j \le \text{LENGTH}(POIs)\}$, where, each POI, poi_j is associated with predefined categories,$Cat_j \in \{Cat_1, Cat_2, \ldots, Cat_j, \ldots, Cat_n\}$, as shown in Table 6, and it is defined as: $poi_j = \{lat_j, lon_j, Cat_j, Time_j, LST_j\}$ Where (lat_j, lon_j) is the geographical coordinate of the POI, Cat_j is a particular category type of POI, $Time_j$ is the commencement time of the POI, on the other hand, LST_j refers to expected least service time for each POI.
3. A set of user's characteristics such as Maximum Trekking Distance (MaxTD) and User Trekking Speed (UTSpeed) in a road network.

The Algorithm, i.e., Algorithm 2, returns as output, the most probable POI category type as the place where activity was performed. Therefore, to obtain the POI category type for every Stay Point, the following three phases are followed as shown in the Algorithm 2 below:

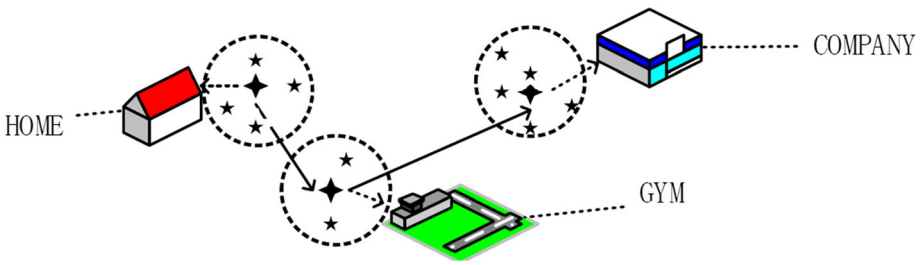


Fig. 5 Sketch of Point of Interests

Algorithm 2: POITypeAnnotation(STP_z, *POIs*, *MaxTD*, *PASpeed*): $Prob_{cat}$

INPUTS: StayPoint, $STP_Z = \{stp_i \mid 1 \leq i \leq \text{LENGTH}(STP_Z)\}$
Point of Interests, $POIs = \{poi_j\ 1 \leq j \leq \text{LENGTH}(POIs)\}$
MaxTD, PASpeed

OUTPUT: POI category type probability for each Stay Point, $Prob_{cat} = \{stp_i \bullet cat | 1 \leq i \leq \text{LENGHT}(Prob_{cat})\}$

PROCEDURE:

1: *Begin*
// **Initialization**
2: Set $slist\_pois \leftarrow \{\}; slist\_pois \bullet stp_i \leftarrow \{\};$
3: Set $slist\_ppois \leftarrow \{\}; slist\_ppois.stp_i \leftarrow \{\};$
4: Set $Prob_{cat} = \{\};$
5: Set $i \leftarrow 1\ ; j \leftarrow 1; stp_i \leftarrow \{\}; poi_j \leftarrow \{\};$
// **Phase 1:** Selecting all the reachable POIs from the computed Stay Points, stp_i
6: for each $stp_i = (\mu_{lat}, \mu_{lon}, T_{arv}, T_{dep}, lut) \in STP_Z$ **do**
6.1: **for all** $poi_j \subseteq POIs$ **do**
6.1.1: **while** $j \leq LENGTH(POIs)$ **do**
6.1.1.1: **if** (GEODIST$(stp_i, poi_j) \leq MaxTD$)
6.1.1.2: **and** $(stp_i \bullet Time\_duration \subseteq poi_j \bullet Opening\_Time)$ **then**
6.1.1.2.1: $slist\_poi \bullet stp_i \leftarrow slist\_pois \bullet stp_i \cup poi_j$
6.1.1.3: **endif**
6.1.1.4: $j = j + 1;$
6.1.2: **end while**
6.2: **end (for-loop)**
6.3: $slist\_pois \leftarrow slist\_pois \cup slist\_pois \bullet stp_i$
7: end (for-loop)
// **Phase 2:** Discovering the probable POIs
8: for each $slist\_pois \bullet stp_i$ in $slist\_pois$ **do**
8.1: **for all** poi_k in $slist\_pois.stp_i$ **do**
8.1.1: **while** $k \leq LENGTH(slist\_pois \bullet stp_i)$ **do**
8.1.1.1: **If** $(poi_k \bullet Revert\_Time \leq stp_i \bullet Time\_period_j)$ **then**
8.1.1.1.1: $slist\_ppois \bullet stp_i \leftarrow slist\_ppois \bullet stp_i \cup poi_k$
8.1.1.2: **end if**
8.1.1.3: $k = k + 1;$
8.1.2: **end while**
8.2: **end (for-loop)**
8.3: $slist\_ppois \leftarrow slist\_ppois \cup slist\_ppois \bullet stp_i$
9: end (for-loop)
// **Phase 3:** Computing the probability for each POI category type, cat_j
10: for each $slist\_ppois \bullet stp_i$ in $slist\_ppois$ **do**
10.1: **for all** $poi_k \subseteq slist\_ppois \bullet stp_i$ **do**
10.1.1: $POI_{cat} = \{\mathcal{M}(poi_k) = cat_j \mid 1 \leq j \leq \text{LENGTH}(POI_{cat})\}$
10.1.2: **for each** $cat_j \subseteq POI_{cat}$
10.1.2.1: **while** $j \leq \text{LENGTH}(POI_{cat})$
10.1.2.1.1: $distance = MIN\{\text{GEODIST}(stp_i, poi_k) | \forall\ poi_k \text{ in } cat_j \subseteq POIs\}$
10.1.2.1.2: $mass = \text{LENGTH}(cat_j)$
10.1.2.1.3: $Prob \leftarrow Prob \cup (cat_j, \frac{mass}{distance^2})$ // see eqn. 5
10.1.2.1.4: $j = j + 1;$
10.1.2.2: **end while**
10.1.3: **end (for-loop)**
10.1.4: $stp_i \bullet cat \leftarrow \text{MAX}(Prob \bullet cat_j)$
10.2: **end (for-loop)**
10.3: $Prob_{cat} \leftarrow Prob_{cat} \cup stp_i.cat$
11: end (for-loop)
12: Return $Prob_{cat}$
13: End Begin

Table 6 POIs categories

S/N	POI category ID	POI category name	POIs details
1	Cat 1	Financial services	Banks, insurance, ATM, etc
2	Cat 2	Education	Schools, colleges, university, polytechnics, etc
…	…	…	…
j	Cat j	Residentials/home	Hotels, hostels, homes,
…	…	…	…
n	Cat *n*	Worship houses	Mosques, churches, temple

Phase 1: Select all the reachable POIs from the computed Stay Points,stp_i, (lines 6–7) by considering the following two conditions:

i. the POI should be within a Trekking distance from the Stay Point, meaning that it should be close to the Stay Point place with a certain spatial range defined by a threshold, MaxTD, i.e., a trek from Stay Point,stp_i, to a POI, poi_j.
ii. the POI commences operation and is available during the stay in the Stay Point. This implies that the period of stay in each Stay Point must match with the commencement time of the POIs. Consequently, a Stay Point in the closure of POI need not to be matched with that POI, meaning that, a Stay Point at 11 pm need not to be matched with, say, a restaurant or a church but with a hotel. Hence, POIs are chosen for a Stay Point if they can be reached by trekking and their operating time,*operating_Time*, intersects with, the Stay Point time-period, *Time_period*.

Phase 2: Discover the probable POIs (lines 8–9)- to achieve this, users require some time to go to the POI from Stay Point and back to Stay Point putting into consideration the Least Service Time (LST) at the POI. Therefore, if stp_and_poi_Time is the time a person requires to cover the distance, d, between Stay Point, stp_i and the poi_j, then 2(stp_and_poi_Time) is the time a person requires to go to POI from stp_i, and back to stp_i:

$$\text{Revert Time} = 2(\text{stp\_and\_poi\_Time}) + \text{LST} \quad (1)$$

Hence, stp_and_poi_Time can be computed as:

$$\text{stp\_and\_poi\_Time} = d/PASpeed \quad (2)$$

where PASpeed is a person's average speed on the road network.

Phase 3 Compute the probability for each POI category type (lines 10–12)- After discovering the probable visited POIs, poi_k, from stp_i, next is the computation of the probability for each POI category type, Cat_j, i.e., $P(stp_i, cat_j) = f(GeoDist(stp_i, cat_j))$, being visited from Stay Point, stp_i.

The above expression implies that for each Stay Point, stp_i, scan for all the POIs, poi_k, surrounding the stp_i. In doing so, take note of the following: (a) keep to, a set Maximum Trekking Diatance, *MaxTD*, from the Stay Point, stp_i, to any POI, poi_j. (b) ensure that the period of stay at stp_i, is within, a set operating time, *operating_Time*, of the POIs, poi_j. At the end, i.e., if $(stp_i \cdot operating\_Time \subseteq poi_j \cdot Opening\_Time)$ then store the selected list of POIs surrounding each Stay Point, stp_i, in a variable, *slist_pois*, considering the aforementioned (a) and (b) conditions.

To compute the probable POIs, *ppois*, from the selected list of POIs, *slist_pois*, pick each of the POIs, poi_k, surrounding the Stay Point, stp_i, i.e., $stp_i \cdot poi_k$, check if the Revert Time, *Revert_time*, is less than the Stay Point time period, *Time_period*, i.e., if $(poi_k \cdot Revert\_Time \leq stp_i \cdot Time\_period_j)$ then store in the selected list of probable POIs, *slist_ppois*.

Finally, once the probable visited POIs are successfully stored in the selected list, *slist_ppois*, the Algorithm computes the probability for each POI category type. This is achieved by applying a method based on the Newton's First Law of Migration: The Gravity Model, which is a model derived from Newton's Law of Gravitational attraction between any two celestial masses. This Gravity Model has been adapted to this study for the purposes of estimating the degree of spatial interaction or movement between any two places. This degree is proportional to the masses and inversely proportional to the square distance between them, as represented in Eq. 3 below:

$$GravityLaw = (mass_1 * mass_2)/distance^2 \quad (3)$$

This Gravity Model is used in this study to infer the degree of relationship between a Stay Point, stp_i and every POI,poi_k, associated with the Stay Point,stp_i. From the above Gravity Model,$mass_1$, denotes the Stay Point,stp_i—to which 1 is assigned, and $mass_2 \Rightarrow mass$ denotes the number of probable visited POIs in each category, where the distance,*distance*, is the minimum distance among all the distances of POIs associated to the same category type.

In view of the above, we provide a probability of POI categories, instead of a single POI. This implies that the POIs related to the same category type are assigned the same probability of being visited.

More formally, for every Stay Point,stp_i, we determine the probability, P, of category type, cat_j, relative to the Stay Point using Eq. 4 below, $P(stp_i, cat_j) \Rightarrow GravityLaw$, represented as:

$$P(stp_i, cat_j) = \frac{|\{poi_k \in slist\_ppoi(stp_i) | \mathcal{M}(poi_k) = cat_j\}|}{\left[min\{\text{GeoDist}(stp_i, poi_k) | \forall poi_k \text{in} cat_j \subseteq POIs\}\right]^2} \quad (4)$$

Where

$$distance = min\{\text{GeoDist}(stp_i, poi_k) | \forall poi_k \text{in} cat_j \subseteq POIs\}$$

$$mass = length(cat_j) = |\{poi_k \in slist\_ppoi(stp_i) | \mu(poi_k) = cat_j\}|$$

Therefore:

$$GravityLaw \Rightarrow P(stp_i, cat_j) = mass/distance^2 \quad (5)$$

4.2.3 Mining association

Given a set of location-logs, LDT = $\{\Psi_l^- \bigcup \Psi^+ | 1 \le l \le m\}$, known as location-datatable of moving persons (both COVID-19 and non-COVID-19 patients); where each location-log, denoted as, $\Psi_l^- = \{\langle O_{id}\rangle, TR_Z | 1 \le Z \le m\}$ or $\Psi^+ = \{\langle O_{id}\rangle, TR_Z | 1 \le Z \le m\}$ is a set of trajectories,TR_Z, generated by a Smartphone, with a unique object identifier, $\langle O_{id}\rangle$.

If $\{\Psi_l^-\}_{l=1}^m = \{\Psi_1^-, \Psi_2^-, \ldots, \Psi_l^-, \ldots, \Psi_m^-\}$ are mutually exclusive set of recorded location-logs of non-COVID-19 patients and Ψ^+ is any recorded location-log of a COVID-19 patient associated with (or caused by) the location-logs of non-COVID-19 patients, $\Psi_1^-, \Psi_2^-, \ldots, \Psi_l^-, \ldots, \Psi_m^-$, then $\Psi^+ \bigcap \Psi_1^-, \Psi^+ \bigcap \Psi_2^-, \ldots, \Psi^+ \bigcap \Psi_l^-, \ldots, \Psi^+ \bigcap \Psi_m^-$, are also mutually exclusive.

If given $\Psi^+ \subseteq \{\bigcup \Psi_l^-\}_{l=1}^m$;

Therefore

$$\Psi^+ = \left(\Psi^+ \bigcap \Psi_1^-\right) \bigcup \left(\Psi^+ \bigcap \Psi_2^-\right) \bigcup \ldots \bigcup \left(\Psi^+ \bigcap \Psi_l^-\right) \bigcup \ldots \bigcup \left(\Psi^+ \bigcap \Psi_m^-\right)$$

In summary, $\Psi^+ = \Psi^+ \bigcap \{\bigcup \Psi_l^-\}_{l=1}^m$

Then our aim is to mine the relationship(s) of COVID-19 patient with non-COVID-19 patients in an environment at a given time, say,$T > 0$. This can be achieved by applying set theory, using the intersection, i.e.,$\Psi^+ \bigcap \Psi_l^- = \{x | x \in \Psi^+\}$, where,$1 \le l \le m$,$\Psi^+$ is a set of trajectories belonging to COVID-19 patient, $\bigcap$ is their intersection, $x = (lat, lon, T)$ is the element(s) that they have in common such as geographical location point(s). Hence, if there is an intersection ($\bigcap$) of COVID-19 patient with either one or more than one of the non-COVID-19 patients, then those involved should be isolated. Consequently, if $\Psi_l^- \cap \Psi^+ = \{\} = \phi$, then it will be concluded that, Ψ^+ and $\{\Psi_l^-\}_{l=1}^m$ are disjoint or mutually exclusive, hence they have no common element(s)/ geographical location point in common. The Algorithm 3 below describes the procedure described above:

Algorithm 3: DetectContacts $(\{\Psi_l^-\}_{l=1}^m, \Psi^+, \theta_{dist})$: $contact_l$

INPUTS: $\mathrm{LDT} = \{\Psi_l^- \cup \Psi^+ \,|\, 1 \le l \le m\}$ denote location-datatable of moving persons (i.e., non-COVID-19 patients and COVID-19 patients).
where Ψ^+ denote a recorded location-log of a Covid-19 patient and $\{\Psi_l^-\}_{l=1}^m$ is a set of recorded location-logs of a non-COVID-19 patient
θ_{dist} denote the distance threshold.

OUTPUT: Contacts of a Covid-19 patient, $contacts_l$

PROCEDURE:

1: *Begin*
2: $contacts_l \leftarrow \{\}$
3: for each $\Psi_l^- = \{\langle O_{id}\rangle, TR_Z^- \mid 1 \le Z \le \mathrm{LENGTH}(\Psi_l^-)\} \in \mathrm{LDT}$
3.1: **if** $\Psi_l^- \cap \Psi^+ = x$ and $\mathrm{GEODIST}(\Psi_l^- \bullet \langle O_{id}\rangle, \Psi^+ \bullet \langle O_{id}\rangle) \le \theta_{dist}$ **then**
3.1.1: $intersection_l \leftarrow \{\Psi_l^- \bullet \langle O_{id}\rangle, x = (lat, lon, T)\}$;
3.1.2: $contacts_l \leftarrow contacts_l \cup intersection_l$
3.2: **else**
3.3: **end if**
4: end (for-loop)
5: return $contacts_l$
6: End Begin

4.3 Proposed Machine Learning Models

This paper proposes six (6) predictive machine learning models for predicting geographical locations of Covid-19 patients. The models considered as presented in the following subsequent subsections include Multiple Linear Regression (MLR) [67, 68], k-Nearest Neighbor (kNN) [69, 74], Decision Tree Regression (DTR) [69–73], Random Forest Regression (RFR) [75–79], Gradient Boosting Regression (GBR) [80, 81], and eXtreme Gradient Boosting regression (XGBR) [82, 83].

a. Multiple Linear Regression (MLR) is a mathematical model that the dependent/target variable$Y$, can be predicted from the knowledge of one or more independent variables denoted as$\boldsymbol{X} = \{x_i\}_{i=1}^{\mathrm{k}}$. The model,$f$, that maps, $\boldsymbol{X} \rightarrow Y$, always give space for a random error,ϵ, i.e.,$Y = f(\boldsymbol{X}) + \epsilon$. The estimate of,$Y$, represented as,$\widehat{Y}$, is a function of, $\boldsymbol{X}$, i.e., $\widehat{Y} = \widehat{f}(\boldsymbol{X})$ implies prediction,$\boldsymbol{X}$, and $\widehat{f}$ is the prediction function. However, the best possible model in regression is estimated by minimizing the expected squared error term [67, 68, 84]. The Loss functions are metrics that compare the predicted values $(\widehat{Y})$ to the actual value (Y), hence, the loss becomes $\left(Y - \widehat{Y}\right)$.
b. The k-Nearest Neighbor (kNN) is a supervised machine learning model which belongs to a class of learners known as *lazy learners*. When given a training tuple(sample) or training data, $\mathcal{D}_n = \{(x_k, y_k)\}_{k=1}^n$, a *lazy learner* simply stores it in the memory (or does only a little/minor processing) and waits until it is given a test tuple/data. It is only when it sees the test tuple does it perform generalization to classify the tuple based on its similarity to the stored training tuples. [69, 74],
c. A Decision Tree Regression (DTR) is a supervised machine learning model that uses a tree-like graph as well as Tree based learning algorithms to predict the value of a target variable by learning simple decision rules inferred from the data features [107]. Given a training set or tuple ,$\boldsymbol{X} = \{\boldsymbol{x_i}\}_{\boldsymbol{i=I}}^{\mathrm{n}}$, the mapping or function,$y = f(\boldsymbol{X})$, can predict the associated class label, y. This mapping is usually represented as a classification rules (i.e., "*if … then … else …*") [69–73].
d. Random Forest Regression (RFR) is a typical ensemble machine learning model that can perform multivariate non-linear regression, by combining the performance (predictions/classifications) of several decision tree algorithms into a single model to make a more accurate and better prediction/classification [111–114]. Given a training data,$\mathcal{D}_n$, of independent random variables, such that,$\mathcal{D}_n = \{(x_k, y_k)\}_{k=1}^{\mathrm{n}}$, where x_i, represents an input predictor random vector that matches a random response, $y_i \in \mathbb{R}$. Let $\boldsymbol{X} = \{x_i\}_{i=1}^{\mathrm{n}} \in \mathbb{R}^n$ be the complete predictor space, such that every dimension, $1, 2, \ldots, n$, represents a distinct predictor. $\mathcal{D}_n$, is used to estimate the regression function, f, that maps, $\boldsymbol{X} \rightarrow \mathbb{R}$ in such a way that as the number of

observation response pair in,$\mathcal{D}_n$, approaches infinity, the squared error between the estimated regression function,$\widehat{f}$, and the observed response values approaches, 0 [75–79].

e. GBR (Gradient Boosting Regression) [80, 81] is a machine learning algorithm meant for regression problems. It produces an improved prediction strength of poor prediction model based on ensemble of a set of weak decision trees. It applies stage-wise approach just like every other boosting model and generalizes the model by allowing an optimization of an arbitrary differentiable loss function. When given a training data,$\mathcal{D}_n = \{(x_k, y_k)\}_{k=1}^{n}$, the aim is to find the approximation $\widehat{\mathcal{F}}(x)$ to a function that maximizes the expected value of some specific loss function$\mathcal{L}(y, \mathcal{F}(x))$.

f. XGBR (eXtreme Gradient Boosting regression) [82, 83] is a scalable machine learning algorithm meant for tree boosting in prediction problems. It is best known for its optimum performance in solving complex regression problems using supervised learning approach. Its parallel and distributed computing features speeds up the model learning as well as prevents overfitting problems.

4.4 Model performance evaluation

The Mean Squared Error (*MSE*), Mean Absolute Error (*MAE*), and Coefficient of determination (R^2) are the evaluation criteria applied in this study to measure the performances of the predicted and actual values of geographical locations of COVID-19 patients. As a regression problem, the error metric objective of MSE and MAE is to minimize the errors between predicted and actual values while R^2 which gives some information about the goodness of fit of a model is to maximize. A perfect fit would result in an R^2 value of 1 and a very good fit near 1, meaning that error between actual and predicted data is very small. These error measure equations (i.e., *MSE*, *MAE*, and R^2) are as presented below in Eqs. 6–8.

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(d_k - y_k)^2 \tag{6}$$

$$MAE = \frac{1}{n}\sum_{k=1}^{n}(|d_k - y_k|) \tag{7}$$

$$R^2 = 1 - \left[\frac{\sum_{k=1}^{n}(d_k - y_k)^2}{\sum_{k=1}^{n}(d_k - \overline{d})^2}\right] \tag{8}$$

where d_k is the observed/desired/target output for training data set$k$, y_k is the computed/predicted output of the considered unit for training dataset k, n is the number of all training data set, and $\overline{d}$ the mean (average) value of the observed (Desired/target) outputs.

4.5 Geographical location prediction

This section demonstrated how to predict the next location of COVID-19 patient. The qualitative data used in this study is the trajectory data or geographical location history of the COVID-19 patient that comprises of *latitude, longitude, date/time, and device identification*. In predictive modeling, the unknown quantity is usually called a *target* (Y) and the supplementary facts are called *inputs* (X). The *inputs* and *target*, $\{X, Y\}$, typically represent measurements of an observable phenomenon, in our own case, it is the trajectory data or geographical location history of both COVID-19 patient and non-COVID-19 patients as described in Tables 2, above:

4.5.1 Model variables selection

The identification, selection procedures and characterization of the dependent and independent variables that were used for the development of the Stay Point/location predictive model are shown below.

4.5.1.1 Dependent variables The dependent variable is the element which the model will allow to estimate/explain. In a Stay Point estimation context, that element is the location predictor, i.e., place of stay which has latitude longitude, and altitude, see Table 7 below.

4.5.1.2 Independent variables The independent variables selected in this study are as shown in Table 8 below:

4.5.2 Model development:

To solve the fundamental problem in prediction of the next location of COVID-19 patient, a mathematical relationship between the inputs (Time) denoted by $\boldsymbol{X}$ and the target (location) outputs denoted by Y is constructed using the multi-instance multi-label learning equation [85, 86], $Y = f(\boldsymbol{X})$.

Consequently, given a set of m-length training examples (time series) of the form, $S_m = \left\{(\boldsymbol{X}_i, Y_i)_{1 \le i \le \mathrm{m}}\right\}$, such that,

Table 7 Dependent variables

S/N	Variables	Description
1	*lat*:	The latitude, in decimal degrees
2	*lon*:	The longitude, in decimal degrees

Table 8 Independent variables

S/N	Variables	Description
1	Location ID	The unique identification (ID) no. of the location
2	Phone ID	The International Mobile-station Equipment Identity (IMEI) of the Phone that generated the location data
3	User ID	The user identification number
4	Yr	The year of location capture as a string
5	Mth	The month of the location capture
6	Day	The day of the location capture
7	Hr	The hour of the location capture
8	Min	The minutes of the location capture
9	Sec	The seconds of the location capture
10	Hvd	Computed Haversine distances of *lat* and *lon*

$\boldsymbol{X}_i \subseteq X$, is the input vector of the i-th instance depicted as set of instances, $\{x_{i,1}, x_{i,2}, \cdots, x_{i,j}, \cdots, x_{i,n_i}\}$, where $x_{ij} \in X(j = 1, 2, \cdots, n_i)$, and then $Y_i \subseteq Y$ is its corresponding desired outputs or labels given by $\{y_{i,1}, y_{i,2}, \cdots, y_{i,k}, \cdots, y_{i,l_i}\}$, where n_i is the number of instances in X_i, and l_i is the number of labels in Y_i, then, $Y = f(\boldsymbol{X})$;

$$\Rightarrow [lat, lon] = \text{f (LocationID, PhoneID, UserID, Yr, Mth, Day, Hr, Min, Sec)} \tag{9}$$

Such that

$Y_i = \{lat_{i,1}, lon_{i,2}\}$, and

$X_i = \{\text{LocationID}_{i,1}, \text{PhoneID}_{i,2}, \text{UserID}_{i,3}, \text{Yr}_{i,4}, \text{Mth}_{i,5}, \text{Day}_{i,6}, \text{Hr}_{i,7}, \text{Min}_{i,8}, \text{Sec}_{i,9}\}$

Where $Y_i \subseteq Y$ and $X_i \subseteq X$.

To make a prediction, where the prediction period length is defined as Z, $Z \in \{1, 2, \cdots\}$, Therefore, Machine Learning models (i.e., MLR, kNN, DTR, RFR, GBR, and XGBR) are used to realize the time-series prediction task and give a new future time series, $S_{m+Z} = \left\{(\boldsymbol{X}_{m+k}, Y_{m+k})_{1 \le k \le Z,}\right\}$.

Consequently, due to the curvature nature of the earth surface, the distances (degrees) of the latitudes and longitudes values varies based on the positions of an object on the earth [87–90]. Hence, Haversine formula was applied to manage the conversion from coordinate to metric. With Haversine formula (Eq. 7), we computed the Haversine distance, *Hvd*, between two location points, $< p_i, p_j >$, on the earth's surface based on the central angle and the radius. The Haversine formula [88] is shown below:

$$Hvd(p_i, p_j) = 2R \times arcsin\sqrt{sin^2\left(\frac{lat_j - lat_i}{2}\right) + \cos(lat_i) \times \cos(lat_j) \times sin^2\left(\frac{lon_j - lon_i}{2}\right)} \tag{10}$$

where, R = 6371 km, represents the radius of the Earth, lat_i and lat_j represent latitude of p_i and p_j, respectively; similarly, lon_i and lon_j, represent the longitudes of of p_i and p_j, respectively. Therefore, $\boldsymbol{X}_i$, becomes:

$\boldsymbol{X}_i$ = {LocationID $_{i,1}$, PhoneID $_{i,2}$, UserID $_{i,3}$, Yr $_{i,4}$, Mth $_{i,5}$, Day $_{i,6}$, Hr $_{i,7}$, Min $_{i,8}$, Sec $_{i,9}$, Hvd $_{i,10}$}.

This mathematical relation (i.e., Eq. 9) is known as a *predictive model*. Once established, a learning algorithm (i.e., supervised Machine Learning model[91, 92]) seeks to learn the mapping function,$f : \boldsymbol{X} \rightarrow Y$, that predicts label for test example, i.e.,$\boldsymbol{X}_i \approx f(Y_i), \forall 1 \le i \le n$, where X is the input space and Y is the output space. Therefore through this, an estimate of an unknown target value, Y_i = {$lat_{i,1}$, $lon_{i,2}$}, of Covid-19 patients given a (new) set of input measurements, $\boldsymbol{X}_i$, will be produced.

The predictions were based on the movement Trajectory Datasets generated using mobile devices such as smartphones, see Fig. 6. In building the machine learning models, the model datasets are often partitioned into two sets of data, (i.e., after pre-processing), in our case it is partitioned into 80% for training and 20% for testing.

The first part, the training data set, is a set of previously observed input and target measurements, or *cases* used to build and validate the initial model. The validation process is used by the modeling algorithm to adjust the initial model to make it more general and less tied to the idiosyncrasies of the training set. The cases of the training data are assumed to be representative of future (unobserved) input and target measurements, hence, a predictive model assumes all possible input and target combinations are recorded in the training data.

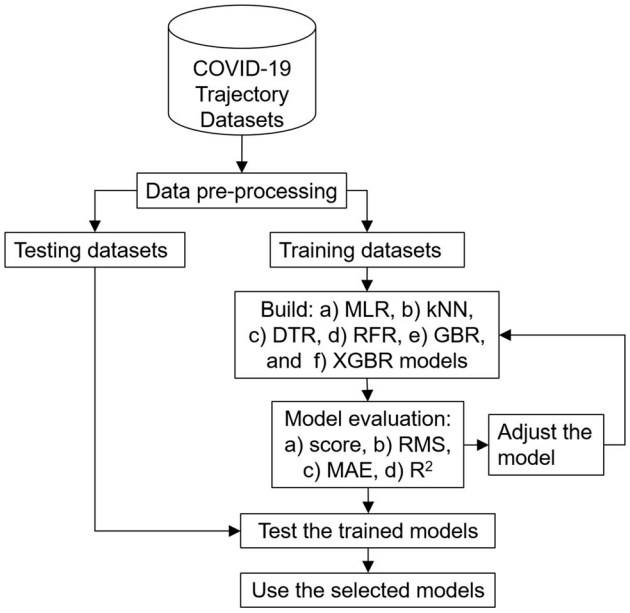


Fig. 6 Location prediction flow diagram

The second part, the test set, is used to gauge the likely effectiveness of the model when applied to unseen data. These two sets of data are necessary because once data has been used for one step in the process, it can no longer be used for the next step because the information it contains has been learned and already become part of the model; therefore, it cannot be used to correct or judge.

5 Experiments, results, and discussions

5.1 Experiments

In this study, we demonstrated the monitoring of COVID-19 patient using our earlier developed mobile App [45]. This application was installed on all participating android-based smartphones assigned to users (i.e., both COVID-19 patient and non-COVID-19 patients) under monitoring for tracking, capturing, and recording of their data/time-stamped locations. Users were selected using simple random sampling technique and in the same geographical area (Oshodi-Isolo, Local Government Area of Lagos State, Nigeria). Every user is identified by his/her Subscribers Identity Module (SIM) card on the assigned device as well as National Identity Number (NIN) assigned by National Identity Management Commission (NIMC), Nigeria.

The installed mobile application on the user's assigned phones runs automatically at the phone's background immediately they are turned on. The date/time-stamp location data of all users under monitoring were captured by the installed mobile App using GPS and Assisted-GPS (i.e., BTS mast/ mobile phone network) technologies. The mobile App creates a location-log (Ψ) on each assigned GPS-enabled smartphone. As the user moves about in a geographical area, the mobile App captures his/her geographical locations, periodically (i.e., at every 30 s interval) and first stores (logs) them locally on its created location-log before transmitting them over the wireless network (if internet is available) to the backend database server (MySql Server) hosted in the Cloud. Consequently, if there is no internet, the phone keeps its log until there is internet connection.

The experiment was performed on Intel Core i7 CPU machine with the specifications such as 32 GB RAM, 1 TB HDD, 17 inches screen, and 4 GB dedicated graphics. Python Jupyter notebook was used for writing, building, and execution of python codes for Trajectory exploratory data analytics and Machine learning modeling. MySQL Server was used for the storage of Location Data table and User details.

5.2 Results

The results of this study are presented in three folds, namely, the GPS Location Data Capture App (Fig. 7), the backend view of the Location Data Table (LDT) with recorded sample of geographical location datasets (Fig. 8), and sample Trajectories and Stay Point plots (Fig. 9).

5.2.1 GPS location data capture (GPS LDC) app

The end-product of our developed mobile app (GPS LDC) is the Android Application Package (.apk) which contains the binaries of the mobile application. The apk file of GPS LDC was installed on every participating device that has the following specifications- Operating System (i.e., Android 2.2 or higher), Mobile data connection (3G Network), Functional GPS and Wifi. The apk file was deployed to all participating mobile devices either through *Bluetooth* or *Over-wire (Universal Serial Bus)*. Consequently, having successfully installed the.apk (i.e., compiled app, MOD.-apk) of the application on mobile device, the application automatically connects to the database (MODB) and user's locations are continuously logged. The application runs on the background, and therefore do not interfere with other applications or user's experience. A sample of captured and logged Coordinates are as shown on the screens in Fig. 7(a) and (b) respectively. Figure 7 shows the sample geographical location datasets logged by mobile devices running installed MOD.apk.

5.2.2 Trajectory and stay point plots

Figure 9 is a panoramic representation of the trajectory of participating user(s), the Point of Interest in a specific

Fig. 7 Mobile App User interface

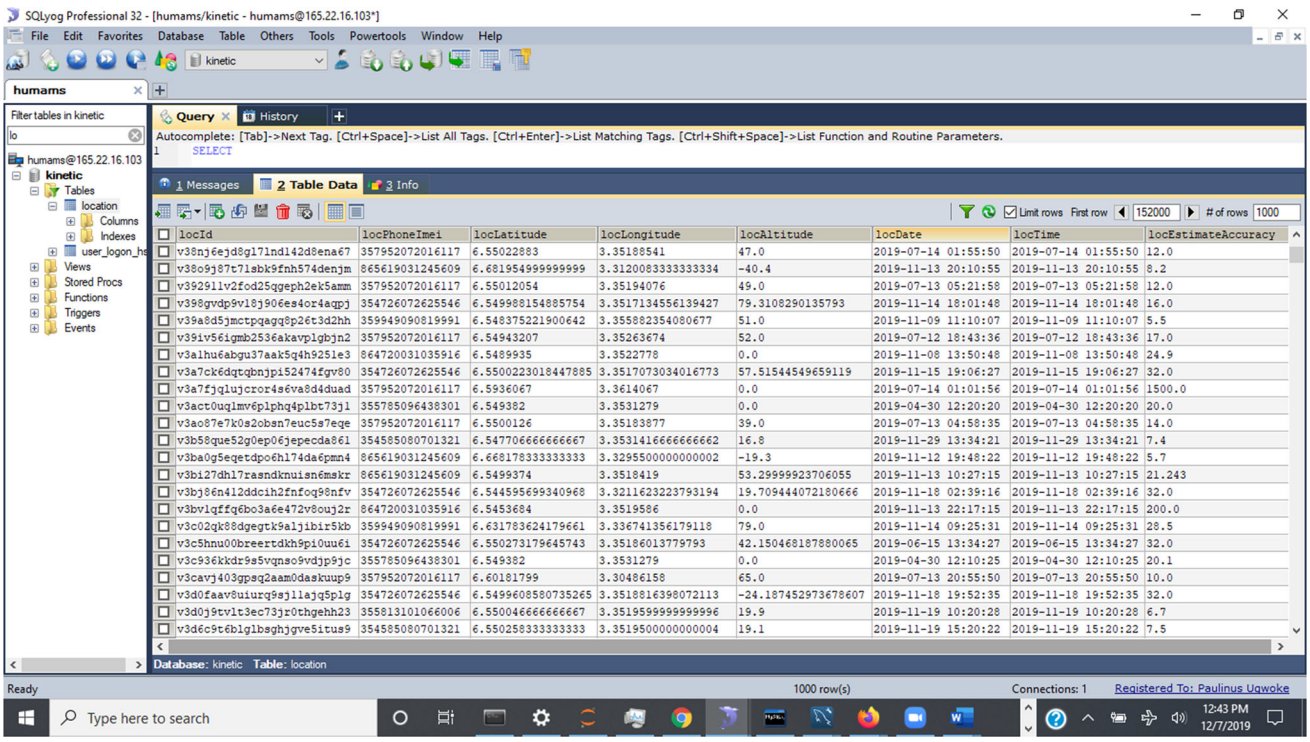


Fig. 8 The backend view of Location Data Table (LDT) with sample Location datasets

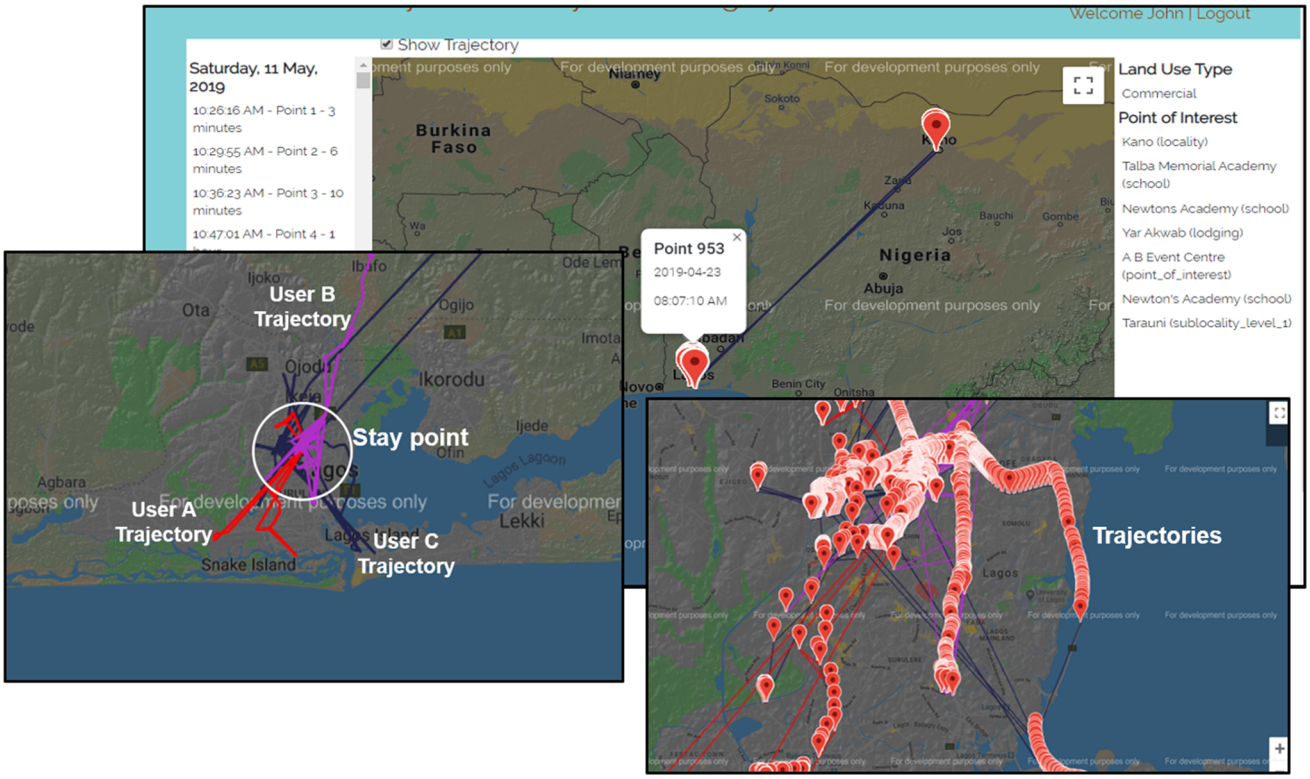


Fig. 9 Trajectories and Stay Point plots

location such as landmarks that a person might find useful within Stay Points. From Fig. 9 below, the Stay Point is described as 'Point 953' as depicted in the figure. Also, on the right panel of Fig. 9, is the screen showing the Point of Interest within the Stay Point at the given date and time. The activity base are clusters of stay points of user(s) where we can define their points of interest. The clusters below show the activity base of 3 users (User A Trajectory, User B Trajectory, and User C Trajectory) selected.

5.2.3 Model learning performance

One of the objectives of the current study is to establish the efficiency of Machine Learning models (i.e., MLR, kNN, DTR, RFR, GBR, and XGBR) in predicting the next probable location of a COVID-19 patient. The Machine Learning models as discussed in Sect. 4.3 have been implemented on the generated COVID-19 Trajectory Datasets from 15 participants. Statistical methods (i.e., Mean Absolute Error (*MAE*), Mean Squared Error (*MSE*), R-squared (R^2), and the Model score value know as *Accuracy*) have been adopted as Key Performance Indicators (KPIs) to evaluate the prediction performances of the implemented Machine Learning models. The KPI metrics is as summarized in the Table 9 below**.** Consequently, the analysis of KPIs is as presented in bar charts in Figs. 10 and 11. Also, Fig. 12 presents the prediction graphs (i.e., Actual versus Predicted Latitude and Longitude against time) of the proposed Machine Learning models for MLR, kNN, DTR, RFR, GBR, and XGBR.

5.3 Discussions

Based on the Model performance metrics (Table 9) and the bar chart plots (see Figs. 10. and 11 above), of the proposed Machine Learning algorithms (i.e., MLR, kNN, DTR, RFR, GBR, and XGBR), we found that MLR has the highest MAE metric of 0.014007 followed by kNN = 0.012491, DTR = 0.010432, RFR = 0.010175, GBR = 0.009691, and the least XGBR = 0.009417. Similarly, kNN has the highest MSE metric = 0.005023 followed by GBR = 0.001534, XGBR = 0.001144, RFR = 0.000904, DTR = 0.000874, and the least MLR = 0.000537. The R-squared metrics which are the goodness-of-fit measure for regression models shows that MLR has the highest value of 0.906832, followed by DTR = 0.852567, RFR = 0.852567, XGBR = 0.820242, GBR = 0.765858, and the least kNN = 0.032041. Consequently, DTR has the highest Model score (accuracy) of 1.000000 followed by RFR = 0.999993, XGBR = 0.999929, kNN = 0.999564, GBR = 0.999396, and the least MLR = 0.945642. Therefore, from the models Key Performance Indicator (KPI) metrics, it can

Table 9 Models key performance indicator (KPI) metrics

Metrics ID	MLR metrics	kNN metrics	DTR metrics	RFR metrics	GBR metrics	XGBR metrics
ACCURACY	0.945642	0.999564	1.000000	0.999993	0.999396	0.999929
MAE	0.014007	0.012491	0.010432	0.010175	0.009691	0.009417
MSE	0.000537	0.005023	0.000874	0.000904	0.001534	0.001144
R^2	0.906832	0.032041	0.852567	0.850667	0.765858	0.820242

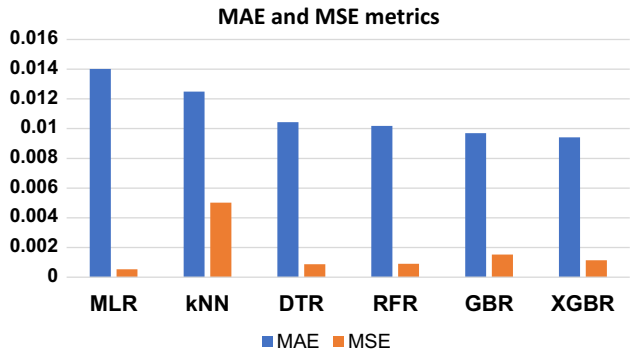


Fig. 10 MAE and MSE metrics chart

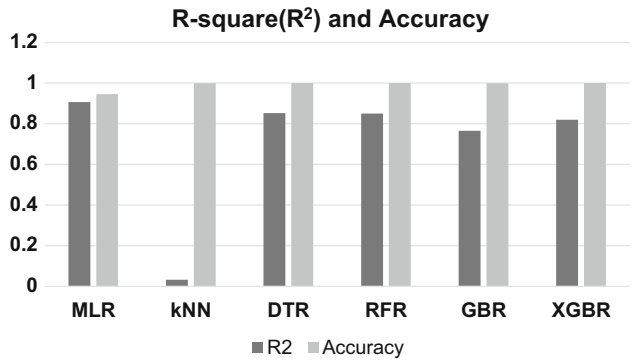


Fig. 11 R-square (R^2) and Accuracy chart

be deduced that DTR has an appreciable performance since the model score/accuracy surpasses the rest of the proposed models with a score of one (1.000000) which implies 100% accuracy. Although MLR performed very well in R-square indicator compared with the rest of the models but it has the worst model score/accuracy. Similarly, kNN has the worst performance indicators in MSE and R-square values, i.e., 0.005023 and 0.032041 respectively, compared with the rest of the proposed Machine Learning models.

Thus, it was observed that MLR and kNN models implemented on COVID-19 Trajectory Datasets are not good enough for predicting the next probable geographical location of a COVID-19 patient. We cannot forget to acknowledge the performances of three ensemble decision tree models (i.e., RFR, GBR, and XGBR) that have generally performed above average and next to DTR. Therefore, the prediction graphs in Fig. 12(a)–(f), can be further used to validate the performance metrics of all the proposed models through visualization of their individual plots. In summary, we conclude that DTR algorithm is more efficient in predicting the next probable location of a COVID-19 patient compare to MLR, kNN, DTR, GBR, and XGBR.

6 Conclusion

This paper presented a framework for monitoring movements of pandemic disease carrier based on GPS trajectory datasets. It carefully described a concise procedure to monitor the geographical movements of COVID-19 patient using assigned GPS-enabled smartphone. It further presented the contact tracing algorithm as well as clustering algorithms for discovering Stay Points (i.e., hot spots). The prediction of next location of COVID-19 patients using Machine Learning models as discussed in Sect. 4.3 have been implemented on the generated COVID-19 Trajectory Datasets from 15 participants was also reported.

From the results obtained in this study, it is evident that our methodology is an obvious improvement over other/similar studies of this nature such as [18–25] that relied on the use of Bluetooth technology with its inherent limitations as previously noted in the literature. The Apps arrived at in other/similar studies are like black box, the implementation steps or process followed in arriving at the Apps were not revealed, meaning that privacy cannot be guaranteed. Hence, the Apps are relied on blind trust as they lack transparency and cannot be reproduced. In contrast, our study relied on the GPS and Assisted GPS as the underlying technology with the distinctive property of anonymity, independence, and its non-inversive nature regarding collating and transmitting of trajectory data to a cloud-based database as well as mining movement patterns of a COVID-19 patients. This is however circumscribed within the ambit of the Government data policy. This paper demonstrated a scientific process that can be followed to reproduce the results obtained. This implies that the

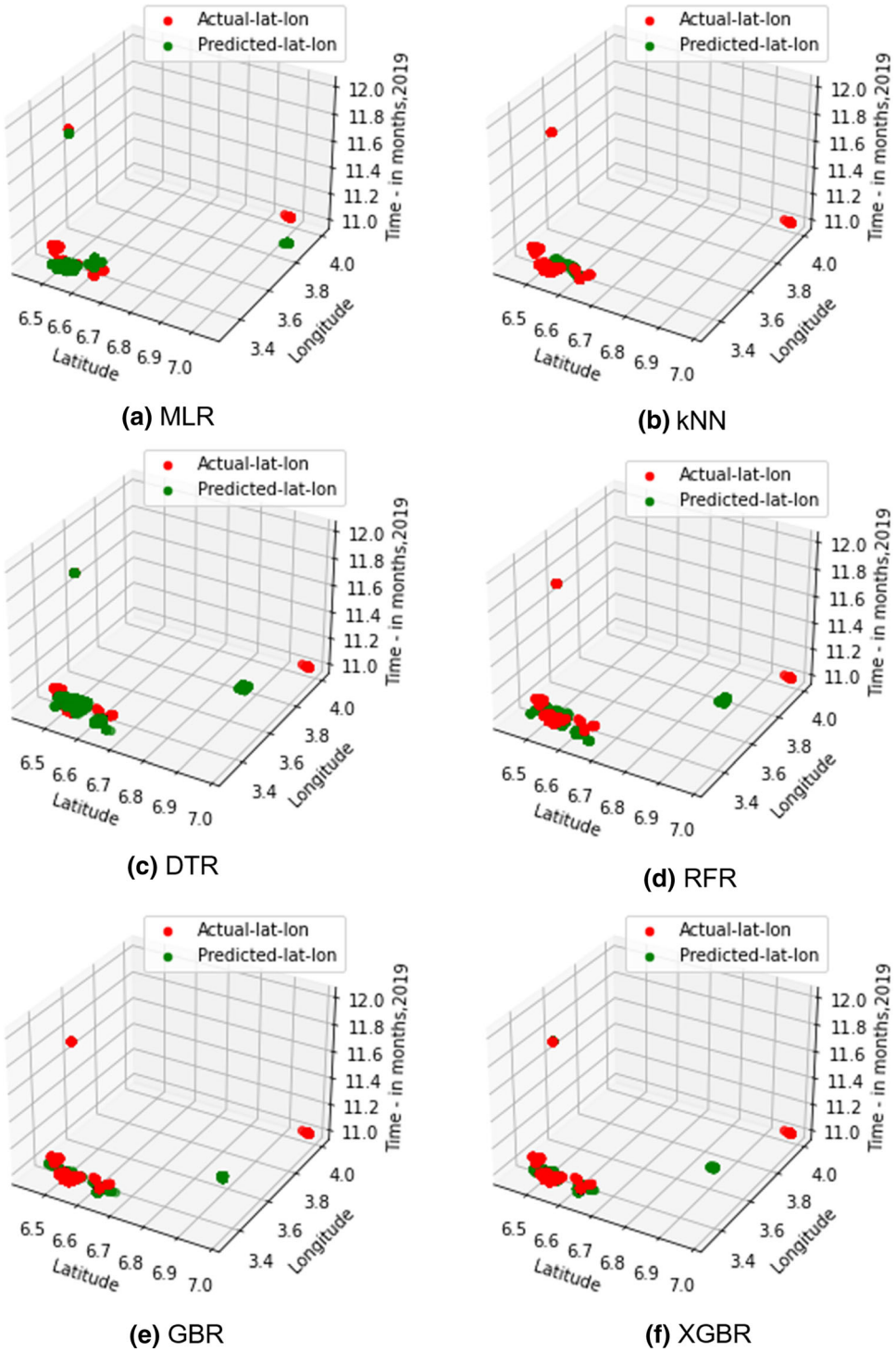


Fig. 12 Model Predictions: Actual vs Predicted Latitude and Longitude for [**a** MLR, **b** kNN, **c** DTR, **d** RFR, **e** GBR, **f** XGBR]

proposed framework in the paper will serve as a reference model for any researcher or software developer who might be interested in developing or implementing similar software solution(s) or App that is capable of tracking, recording, and analyzing date/time-stamped locations of a COVID-19 patient. Our future work will consider implementing Deep leaning Algorithms on Trajectory datasets of COVID-19 patients for the purpose of predicting their next geographical locations. Some of the limitations of this study include but not limited to:

a. Privacy issues which is very difficult in convincing the respondents to partake in the study.
b. Finance issues, as a real-time, online system, needs to be hosted life and requires finances.

c. The Lack of local expertise in Machine learning models, requires the researcher to travel to other Countries in other to acquire the requisite skill/knowledge.
d. Absence of online documented Point of Interest (POIs) database with regards to unit of analysis.

Funding The author(s) received no funding/financial support from any external source for the research, authorship, and/or publication of this article.

References

1. Toch, E., Lerner, B., Ben-Zion, E., & Ben-Gal, I. (2019). Analyzing large-scale human mobility data: A survey of machine learning methods and applications. *Knowledge and Information Systems, 58*(3), 501–523. https://doi.org/10.1007/s10115-018-1186
2. Zhao, C., Zeng, A., & Yeung, C. H. (2021). Characteristics of human mobility patterns revealed by high-frequency cell-phone position data. *EPJ Data Science, 10*, 5. https://doi.org/10.1140/epjds/s13688-021-00261-2
3. Luca, M.D., Barlacchi, G., Lepri, B., & Pappalardo, L. (2020). Deep learning for human mobility: a survey on data and models. https://arxiv.org/abs/2012.02825v1; Accessed on March 07, 2021.
4. Xia, F., Wang, J., Kong, X., Wang, Z., Li, J., & Liu, C. (2018). Exploring human mobility patterns in urban scenarios: A trajectory data perspective. *IEEE Communications Magazine, 56*(3), 142–149. https://doi.org/10.1109/MCOM.2018.1700242
5. Wang, J., Kong, X., Xia, F., & Sun, L. (2019). Urban human mobility: Data-driven modeling and prediction. *ACM SIGKDD Explorations Newsletter*. https://doi.org/10.1145/3331651.3331653
6. Hugo, B., Marc, B., Gourab, G., Charlotte, R. J., Maxime, L., Thomas, L., Ronaldo, M., Jose, J. R., Filippo, S., & Marcello, T. (2017). Human mobility: Models and applications. *Physics Reports, 734*, 1–74. https://doi.org/10.1016/j.physrep.2018.01.001
7. Huihan, L. (2020). Spatio-temporal analysis and simulation of human trajectories in urban environments. B.Sc. Thesis, Department of Computer Science, Wellesley College, May 6, 2020, https://repository.wellesley.edu/islandora/object/ir%3A1217/datastream/PDF/download; Accessed on March 13, 2021.
8. Wang, S., Liu, Y., & Hu, T. (2020). Examining the change of human mobility adherent to social restriction policies and its effect on COVID-19 cases in Australia. *International Journal of Environmental Research and Public Health, 17*(21), 7930. https://doi.org/10.3390/ijerph17217930
9. Zhang, C., Qian, L. X., & Hu, J. Q. (2020). COVID-19 pandemic with human mobility across countries. *Journal of the Operations Research Society of China*. https://doi.org/10.1007/s40305-020-00317-6
10. Gunthe, S. S., & Patra, S. S. (2020). Impact of international travel dynamics on domestic spread of 2019-nCoV in India: origin-based risk assessment in importation of infected travellers. *Global Health, 16*, 45. https://doi.org/10.1186/s12992-020-00575-2
11. Fang, H., Wang, L., & Yang, Y. (2020). Human mobility restrictions and the spread of the Novel Coronavirus (2019-nCoV) in China. *Journal of Public Economics, 191*, 104272. https://doi.org/10.1016/j.jpubeco.2020.104272
12. Zhou, Y., Xu, R., Hu, D., Yue, Y., Li, Q., & Xia, J. (2020). Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data. Lancet Digit Health, https://pubmed.ncbi.nlm.nih.gov/32835199/; Accessed on March 20, 2021.
13. Oztig, L. I., & Askin, O. E. (2020). Human mobility and coronavirus disease 2019 (COVID-19): A negative binomial regression analysis. *Public Health, 185*, 364–367. https://doi.org/10.1016/j.puhe.2020.07.002;AccessedonMarch20,2021
14. Maged, N. K. B., & Estella, M. G. (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. *International Journal of Health Geographics*. https://doi.org/10.1186/s12942-020-00202-8
15. Cristina-Maria, P., & Bogdan-Radu, N. (2020). An analysis of Covid-19 spread based on Fractal interpolation and Fractal Dimension. Available at: https://doi.org/10.1016/j.scitotenv.2020.140033; Accessed on March 13, 2021.
16. Ivan F.P., & Lawal, B. (2020). Spatial analysis and GIS in the study of Covid-19. A review. Available at: https://www.sciencedirect.com/science/article/pii/S0960077920304562; Accessed on March 13, 2021.
17. Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic. A review. [Online] Available at: https://www.sciencedirect.com/science/article/pii/S0960077920304562; Accessed on March 13, 2021.
18. Niyogi, S., Petrie, J., Leibrand, S., Gallagher, J., Eder, M., Szabo, Z., Danezis, G., Miers, I., de Valence, H., Reusche, D. (2020). TCN Protocol:Temporary Contact Numbers Protocol. [online] Available at: https://github.com/TCNCoalition/TCN; Accessed on March 3, 2021.
19. Tracetogether (2020). Trace Together. [Online] Available at: https://www.tracetogether.gov.sg/; Accessed on March 13, 2021.
20. PACT (2020). Private automated contact tracing. Available at: https://pact.mit.edu/wp-content/uploads/2020/04/The-PACT-protocol-specification-ver-0.1.pdf; Accessed on July 13, 2020.
21. Covid Watch (2020). Together, we have the power to stop COVID-19. [Online] Available at: https://covid-watch.org/; Accessed on March 20, 2021.
22. CoEpi (2020). CoEpi: Community epidemiology in action. [online] Available at: https://www.coepi.org/; Accessed on March 20, 2021.
23. Troncoso, C., Payer, M., Hubaux, J.-P., Salathé, M., Larus, J., Bugnion, E., Lueks, W., Stadler, T., Pyrgelis, A., Antonioli, D., Barman, L., Chatel, S., Paterson, K., Čapkun, S., Basin, D., Beutel, J., Jackson, D., Roeschlin, M., Leu, P., Preneel, B., Nigel, S., Aysajan, A., Gürses, S., Veale, M., Cremers, C., Backes, M., Tippenhauer, O.N., Binns, R., Cattuto, C., Barrat, A., Fiore, D., Barbosa, M., Oliveira, R., & Pereira, J. (2020). Decentralized privacy-preserving proximity tracing. [Online] Available at: https://arxiv.org/ftp/arxiv/papers/2005/2005.12273.pdf; Accessed on March 20, 2021.
24. Carmela, T. (2020). "Decentralized privacy-preserving proximity tracing: Simplified overview. April 8, 2020; [online] Available at: https://github.com/DP-3T/documents/blob/master/DP3T%20-%20Simplified%20Three%20Page%20Brief.pdf.
25. Bluetooth (2020). Bluetooth Technology. [online] Available at: https://www.bluetooth.com/learn-about-bluetooth/bluetooth-technology/; Accessed on March 18, 2021.
26. Alagappan, S. (2020). A basic guide to contact tracing. The SciTech Scoop, June 30, 2020; [Online] Available at: https://

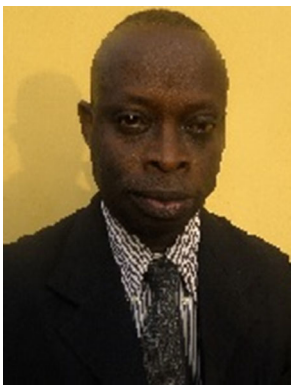
medium.com/the-scitech-scoop/a-basic-guide-to-contact-tracing-e190b4deecaf; Accessed on March 20, 2021.
27. Albergotti, R. (2020). "Apple and google launch coronavirus exposure software. *The Washington Post*, WP Company, 20 May 2020; [Online] Available at: http://ww.washingtonpost.com/technology/2020/05/20/apple-google-api-launch/; Accessed on March 20, 2021.
28. Wang, J. (2020). Apple and Google roll out COVID-19 exposure notifications through public health apps. The Android Police; May 20, 2020; [Online] Available at: https://www.androidpolice.com/2020/05/20/apple-and-google-are-working-together-to-fight-coronavirus-with-a-new-contact-tracing-tool/ ; Accessed on March 20, 2021.
29. Yves-Alexandre, de M., Florimond, H., Andrea, G., & Florent, G. (2020). Blogpost: Evaluating COVID-19 contact tracing apps? Here are 8 privacy questions we think you should ask. [Online] Available at: https://cpg.doc.ic.ac.uk/blog/pdf/evaluating-contact-tracing-apps-here-are-8-privacy-questions-we-think-you-should-ask.pdf; Accessed on March 20, 2021.
30. Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic. A review. [Online] Available at: https://www.sciencedirect.com/science/article/pii/S0960077920304562; Accessed on July 13, 2020.
31. Chuansai, Z., Wen, Y., Jun, W., Haiyong, X., Yong, J., Xinmin, W., Qiuzi, H.W., & Pingwen, Z. (2020). Detecting suspected epidemic cases using trajectory big data. CSIAM Transactions on Applied Mathematics, 1, 186–206. [Online] Available at: https://arxiv.org/abs/2004.00908 ; Accessed on March 20, 2021.
32. Chimmula, V. K. R., & Zhan, L. (2020). "*Time series forecasting of COVID-19 transmission in Canada using LSTM networks*. *Chaos, Solitons and Fractals*. https://doi.org/10.1016/j.chaos.2020.109864
33. Khan, U., Mehta, R., Arif, M. A., & Lakhani, O. (2020). Pandemics of the past: A narrative review. *Journal of the Pakistan Medical Association, 70*(Suppl 3), 34–37. https://doi.org/10.5455/JPMA.11
34. Miquel, P. (2008). A dictionary of epidemiology. Fifth Edition, [Online] Available at: http://www.academia.dk/BiologiskAntropologi/Epidemiologi/PDF/Dictionary_of_Epidemiology__5th_Ed.pdf; Accessed on March 20, 2021.
35. Vincent, C. C., Susanna, K. P. L., Patrick, C. Y. W., & Kwok, Y. Y. (2007). Severe acute respiratory syndrome coronavirus as an agent ofemerging and reemerging infection. *Clinical Microbiology Review, American Society for Microbiology, 20*(4), 660–694.
36. Wikipedia (2021). Wuhan. [Online] Available at: https://en.wikipedia.org/wiki/Wuhan; Accessed on March 23, 2021.
37. Wikipedia (2021). World Health Organization. [Online] Available at: https://en.wikipedia.org/wiki/World_Health_Organization; Accessed on March 20, 2021.
38. Wikipedia (2021). Public health emergency of international concern. [Online] Available at: https://en.wikipedia.org/wiki/Public_Health_Emergency_of_International_Concern; Accessed on March 21, 2020.
39. WHO (2021). Questions and answers. [Online] Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses; Accessed on March 18, 2021.
40. WHO (2021). WHO Coronavirus Disease (COVID-19) Dashboard. [Online] Available at: https://covid19.who.int/; Accessed on March 15, 2021.
41. Pulse (2021). 8 states where coronavirus patients have escaped. [Online] Available at: https://www.pulse.ng/news/local/8-states-where-coronavirus-patients-have-escaped/b2xy7f0 ; Accessed on March 18, 2021.
42. Kraak, M. (2003). The space-time cube revisited from a geovisualization perspective, The International Cartographic Association (ICA). In Proceedings of the 21st International Cartographic Conference (ICC); Durban, South Africa, August10–16.
43. Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6*(2), 94–98.
44. Enrique, H.-O., Pietro, M., Carlos, T. C., & Cano, J.-C. (2018). Evaluating how smartphone contact tracing technology can reduce the spread of infectious diseases: The case of COVID-19. *IEEE Access, 8*, 99083–99097.
45. Ugwoke, P. O., Inyiama, H. C., & Ikekeonwu, G. A. M. (2014). Real-time human trajectory dataset capture model (RT-HTDCM) using GPS and assisted-GPS technologies: African perspective. *The Journal of Information Engineering and Applications, 4*(10), 55–76.
46. Buchanan, B. & Miller, T. (2017). Machine learning for policy-makers- what it is and why it matters. The cyber security project, Belfer Center for Science and International Affairs, Harvard Kennedy School, 79 JFK Street, Cambridge; June 2017; [Online] Available at: https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf; Accessed on March 18, 2021.
47. Mitchell, T. M. (1997). *Machine Learning* (1st ed.). New York: McGraw-Hill Education.
48. Tanuja Vand Govindarajulu, P. (2016). A survey on trajectory data mining. International Journal of Computer Science and Security (IJCSS) 10(5) [Online] Available at: https://www.cscjournals.org/manuscript/Journals/IJCSS/Volume10/Issue5/IJCSS-1297.pdf Accessed on March 18, 2021.
49. Leonardi, P. M. (2020). COVID-19 and the new technologies of organizing: Digital exhaust, digital footprints, and artificial intelligence in the wake of remote work. *Journal of Management Studies*. https://doi.org/10.1111/joms.12648
50. Zhang, D., Guo, B., Li, B. (2010). Extracting social and community intelligence from digital footprints: An emerging research area. pp. 4–18, Springer-Verlag, Berlin Heidelberg.
51. Zhang, D., Guo, B., & Yu, Z. (2011). Social and community intelligence. *IEEE Computer, 44*(7), 21–28.
52. Guo, B., Zhang, D., Yu, Z., & Calabrese, F. (2011). From Digital Footprints to Social and Community Intelligence. ACM Workshop, UbiCamp'11, Beijing, China, September 17–21.
53. Zhang, D., Wang, Z., Guo, B., Yu, Z. (2012). Social and community intelligence: technology and trends. IEEE Computer Society, pp. 12–16.
54. Gang, P., Quande, Q., Wangsheng, Z., Shijian, L., & Zhaohui, W. (2013). Trace analysis and mining for smart cities: Issues, methods, and applications. IEEE Communications Magazine, pp. 120–126.
55. Andrienko, N., Andrienko, G., Pelekis, N., & Spaccapietra, S. (2008). (2008); Basic concept of movement data. In F. Giannoti & D. Pedreschi (Eds.), *Mobility, data mining and privacy-geographic knowledge discovery* (pp. 15–38). Berlin: Springer Verlag.
56. Wikipedia, Oshodi Isolo. [online] Available at: https://en.wikipedia.org/wiki/Oshodi-Isolo, 2017; Accessed on March 18, 2021.
57. Olatunde-Aremu, F. T., & Akinpelu, A. (2017). urban crime and safety: a case of some selected gated neighborhoods in Oshodi/Apapa local government area, Lagos State. *International Journal of Social Science and Development Policy, 3*(2), 42–53.
58. Košice, S., & Košice, S. (1999). Knowledge discovery in databases: A comparison of different comparison of different views. *Journal of Information and Organizational Sciences, 23*(2), 95–102.

59. Dodge, S., Weibel, R., & Lautenschütz, A.-K. (2008). Towards a taxonomy of movement patterns. *Information Visualization, 7*, 240–252.
60. Martin, E., Hans-Peter, K., Jörg, S., Xiaowei, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of 2nd international conference on knowledge discovery and data mining (KDD-96); [Online] Available at: http://www.di.unipi.it/~coppola/didattica/ccp0506/papers/kdd-96.pdf ; Accessed on March 18, 2021.
61. Kdnugget, Density based spatial clustering applications noise-dbscan, [online] Available at: https://www.kdnuggets.com/2017/10/density-based-spatial-clustering-applications-noise-dbscan.html ; Accessed on March 18, 2021.
62. Boeing, G. (2018). Clustering to reduce spatial data set size. Computer Science, Cornell University, 21 march, 2018, [Online] Available at: https://arxiv.org/abs/1803.08101v1; Accessed on March 22, 2021.
63. Xiaopeng, C., Dianxi, S., Banghui, Z., & Fan, L. (2016). Periodic pattern mining based on GPS trajectories. Atlantis Press, 2016 International Symposium on Advances in Electrical, Electronics and Computer Engineering (ISAEECE 2016), [Online] Available at: https://www.atlantis-press.com/proceedings/isaeece-16/25852862; Accessed on March 18, 2021.
64. Mousavi, A., Zadeh, A. S., Akbari, M., & Hunter, A. (2017). A New Ontology-Based Approach for Human Activity Recognition from GPS Data. *Journal of AI and Data Mining, 5*(2), 197–210.
65. Barbara, F., Paolo, C., Chiara, R., & Laura, S. (2013). Inferring human activities from GPS tracks. In ACM, UrbComp'13: Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, vol. 5, pp. 1–8 https://doi.org/10.1145/2505821.2505830
66. Yu, Z., & Xiaofang, Z. (2011). *Computing with spatial trajectories*. Berlin: Springer.
67. Bee, R., & Bee, F. (1999). Managing information and statistics. Chartered Institute of Personnel and Development, CIPD House, Camp Road London SW19 4UX.
68. jmp, Fitting multiple regression model, [online] Available at: https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-multiple-regression/fitting-multiple-regression-model.html ; Accessed on March 12, 2021.
69. Jiawei, H., Micheline, K., Jian, P. (2012). Data Mining: Concepts and Techniques. 3nd Ed; Morgan Kaufmann Publishers (an imprint of Elsevier), 225 Wyman Street, Waltham, MA 02451, USA, 2012.
70. Sakr, S., Elshawi, R., Ahmed, A. M., Qureshi, W. T., Brawner, C. A., Keteyian, S. J., Blaha, M. J., & Al-Mallah, M. H. (2017). Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project. *BMC Medical Informatics and Decision Making, 17*(1), 174.
71. Lior, R., & Oded, M. (2005). Top-down induction of decision trees classfiers- A survey. *IEEE Transactions on Systems, Man, and Cybernetics- Part C: Applications and Reviews, 35*(4), 476–487.
72. Ayon, D. (2016). Machine learning algorithms: A review. *International Journal of Computer Science and Informationa Technologies (IJCSIT), 7*(3), 2016.
73. Geeksforgeeks, Random forest. [online] Available at: https://dsc-spidal.github.io/harp/docs/examples/rf/; Accessed on March 18, 2021.
74. Raschka, S. (2018). STAT 474: Machine Learning. Lecture Notes, Department of Statistics, University of Wisconsin-Madison, 2018; [Online] Available at: http://stat.wisc.edu/_sraschka/teaching/stat479-fs2018/; Accessed on March 12, 2021.
75. Geeksforgeeks, Random forest, [online] Available at: https://dsc-spidal.github.io/harp/docs/examples/rf/; Accessed on March 12, 2021.
76. Verikas, A., Vaiciukynas, E., Gelzinis, A., Parker, J., & Olsson, M. C. (2016). Electromyographic patterns during golf swing: activation sequence profiling and prediction of shot effectiveness. *Sensors (Basel), 16*(4), 592. https://doi.org/10.3390/s16040592
77. Te, H., Dongxiang, J., Qi, Z., Lei, W., & Kai, Y. (2018). Comparison of random forest, artificial neural networks, and support vector machine for intelligent diagnosis of rotating machinery. *Transactions of the Institute of Measurement and Control, Sage Journals*. https://doi.org/10.1177/0142331217708242
78. Rodriguez-Galiano, V. F., Sanchez-Castillo, M., Dash, J., Atkinson, P. M., & Ojeda-Zujar, J. (2016). Modelling interannual variation in the spring and autumn land surface phenology of the European forest. *Biogeosciences, 13*, 3305–3317.
79. Towardsdatascience, Support vector machine. [online] Available at: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989; Accessed on March 12, 2021.
80. Yanru, Z., & Ali, H. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C, 58*, 308–324. https://doi.org/10.1016/j.trc.2015.02.019
81. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. https://doi.org/10.3389/fnbot.2013.00021
82. Tianqi, C., & Carlos, G. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD, international conference on knowledge discovery and data mining; August 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785, Accessed on March 12, 2021.
83. Brownlee, J. (2018). XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn. https://pdf-drive.com/pdf/Jason20Brownlee20-20XGBoost20with20Python.201.10.pdf; Accessed on March 12, 2021.
84. Kees, B. (2018). Quantifying uncertainty of random forest predictions: a digital soil mapping case study. An M.Sc. Thesis, Wageningen University and Research Centre, Netherlands, April 2018.
85. Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access, 7*, 53040–53065.
86. Zhou, Z.-H., Zhang, M.-L., Huang, S.-J., Li, Y.-F. (2012). Multi-instance multi-label learning. Artificial Intelligence, 176(1), 2291–2320, [Online] Available at: https://arxiv.org/abs/0808.3231v4; Accessed on March 12, 2021.
87. Loroy, J. (2016). Detecting user's habits using GPS data. An M.Sc. Thesis; Computer Science Department, UCL, Universite Catholique de Louvain, France; [Online] Available at: https://dial.uclouvain.be/memoire/ucl/fr/object/thesis:4610/datastream/PDF_01/view; Accessed on March 12, 2021.
88. Luo, T., Zheng, X., Xu, G., Fu, K., & Ren, W. (2017). An Improved DBSCAN Algorithm to Detect Stops in Individual Trajectories. *ISPRS International Journal of Geo-Information;* 2017, *6*; [Online] Available at: https://www.mdpi.com/2220-9964/6/3/63.
89. Symmetry, What is geolocation or geocoding. [online] Available at: https://www.symmetry.com/resources/payroll-news/2018/05/30/what-is-geolocation-or-geocoding ; Accessed on March 12, 2021.
90. Pinterest, [online] Available at: https://www.pinterest.ph/pin/564005553318904886/; Accessed on March 12, 2021.
91. Towardsdatascience: Machine learning types and algorithms. [online] Available at: https://towardsdatascience.com/machine-learning-types-and-algorithms-d8b79545a6ec; Accessed on March 12, 2021.
92. Towardsdatascience: Types of machine learning algorithms you should know, [online] Available at: https://towardsdatascience.

com/types-of-machine-learning-algorithms-you-should-know-953a08248861 ; Accessed on March 12, 2021.
93. Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimates for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*, 427–431. https://doi.org/10.2307/2286348
94. Dwyer G. P. (2015). The Johansen tests for cointegration. April 2015, http://www.jerrydwyer.com/pdf/Clemson/Cointegration.pdf; Accessed on March 23, 2021.
95. Abutu, U. O., & Agbede, E. A. (2015). Government expenditure and economic growth in Nigeria: A cointegration and error correction modelling. Munich Personal RePEc Archive (MPRA); Paper No. 69676, July 18, 2015; Available online at: https://mpra.ub.uni-muenchen.de/69676/; Accessed on March 23, 2021.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paulinus O. Ugwoke obtained M.Sc. (Computer Science) and M.Sc. (Statistics) degrees both from the University of Lagos, Nigeria; B.Sc. (Combined Hons, Computer Science & Statistics) from the University of Nigeria, Nsukka, Nigeria. He is a Doctoral Student of the Department of Computer Science, University of Nigeria, Nsukka, Nigeria. His research interests cut across Knowledge Discovery in Large Databases, Modelling and Simulation, emerging technology solutions for Smart sustainable cities such as Internet of Things (IoTs), Big Data, Blockchain Technologies as well as Artificial Intelligence & Machine Learning algorithms. He has some papers in some learned journals. He is currently the Head of Department of Research, Education, and Training at the Digital Bridge Institute (*International Centre for Information & Communications Technology Studies*), Nigerian Communications Commission, Abuja, Nigeria.



Francis S. Bakpo is a Professor in the Department of Computer Science University of Nigeria, Nsukka, Nigeria. He received his M.Sc. degree in Computer Science and Engineering from Kazakh National Technical University, Almaty (formerly, USSR) in 1994 and Doctorate degree in Computer Engineering in 2008 from Enugu State University of Science and Technology, Agbani. He joined the Department of Computer Science, University of Nigeria, Nsukka, Nigeria in 1995 and was progressed from the rank of lecturer II to Professor in 2010.

Collins N. Udanor obtained a B.Eng (Hons) in Computer Science and Engineering from Enugu State University of Science & Technology, Enugu, a M.Sc in Computer Science and a Ph.D in Electronic Engineering from the University of Nigeria, Nsukka, Nigeria respectively. He joined the services of University of Nigeria, Nsukka, Nigeria in 1999 and is currently a Senior Lecturer in Computer Science, University of Nigeria, Nsukka.

Matthew C. Okoronkwo obtained HND in Systems Science Institute of Management and Technology (IMT), Enugu, Nigeria. PGD Computer Science Anambra State University of Technology (ASUTECH), Enugu, Nigeria. PGD Finance/Banking, University of Nigeria, Nsukka, Nigeria, MSc. Computer Science Nnamdi Azikiwe University, Awka, Nigeria and PhD in Information Technology, Ebonyi State University (EBSU), Nigeria. He is currently a Senior Lecturer in the Department of Computer Science University of Nigeria, Nsukka, Nigeria.

Authors and Affiliations

Paulinus O. Ugwoke[1,2] · Francis S. Bakpo[1] · Collins N. Udanor[1] · Matthew C. Okoronkwo[1]

✉ Paulinus O. Ugwoke
okeyugwoke@yahoo.com

[1] Department of Computer Science, Faculty of Physical Sciences, University of Nigeria, Nsukka, Nigeria

[2] Department of Research, Education, and Training, Digital Bridge Institute (International Centre for Information & Communications Technology Studies), Nigerian Communications Commission, Abuja, Nigeria