International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6046
***Article Received**: 26 November 2022 **Revised**: 24 December 2022 **Accepted**: 04 January 2023*

A Novel Machine Learning Framework for Tracing Covid Contact Details by Using Time Series Locational data & Prediction Techniques

[1]Dr. C.N.Ravi, [2] Yasmeen, [3] Dr. Kaja Masthan, [4] Rajesh Tulasi, [5] Duba Sriveni, [6]P. Shajahan

[1] Professor, Dept of CSE, CMR Engineering College, Hyderabad,Telangana
[2]Assistant Professor, Dept of CSIT, CVR College of Engineering,Hyderabad
[3]Assistant Professor, Dept of CSE , Sphoorthy Engineering College, Hyderabad
[4]Assistant Professor, Dept of CSE, Koneru Lakshmaiah Education Foundation, Guntur
[5]Assistant Professor, Dept of CSE, Research Scholar VTU, Belgaum, Karnataka
[6] Assistant Professor, Dept of CSE, Srinivasa Ramanujan Institute of Technology, Ananthapuramu

Abstract
It is difficult to prevent the spread of new infections in densely populated areas because they spread at a faster rate. One of the most commonly used techniques for this type of scenario is contact tracing, which involves locating the infected character and his close contacts after he has been infected. This is one of the most recent and effective methods that the health authorities have supported. We can see Machine Learning strategies that require some region information to efficiently implement contact tracing. Contact tracing is used by local governments and health authorities to halt the spread of rapidly spreading diseases. It is one of the locally focused methods that work well when the number of cases is small. As a result, we can say that it can be or is primarily used in rapidly transmitted diseases and newly emerging infections. Using cluster-based region identifications, the utility of touch tracing is investigated using nearest neighbour approaches and absolute deterministic simulations (MLDBSCAN). Emerging or re-emerging infectious diseases like SARS, Ebola, Lassa fever, tuberculosis, and, most recently, COVID-19 require extremely effective prevention methods and strategies.

Keywords: Machine learning, Clustering, MLDBSCAN, data analytics

I. Introduction

Many critical infections have emerged in the past, and in the present, a brand-new infection is emerging, resulting in a massive loss of human life, and we can expect to see this type of rising infection in the future as well. To maintain good control of such pandemics, a thorough examination of pandemics is required, as well as the implementation of effective techniques. Individual-to-character transmission, droplet spread (direct contact), and airborne transmission, as well as infected objects, food, animal-to-human verbal exchange, and trojan horse bites, are all ways in which humans spread disease (oblique contact). Infectious diseases can spread through direct or indirect contact, making everyone vulnerable to becoming ill. SARS, Ebola, Lassa fever, tuberculosis, and, at the moment, COVID 19[1] are examples of infectious diseases. These diseases required a quick response as well as specific manipulation techniques. We can break the transmission chain by using a powerful technique known as contact tracing in the early stages of infection spread. We will use existing technology to conduct powerful contact tracing.

Machine learning is one of the most widely used types of AI [2]. In order to aid decision-making, it analyses and discovers trends in massive data sets. Algorithms, which are a set of instructions for completing a set of tasks, are the building blocks in machine learning programmes [3]. The algorithms are designed to learn from data without human intervention. According to a detailed research into what machine learning [4] is the term "representation" refers to the classification of data in a format and language that a machine can understand. This section lays the groundwork for the following step, evaluation, which will determine whether or not the data classifications are useful. An algorithm goes through this learning process without needing to be programmed. Machine learning [5] can be supervised, unsupervised, semi-supervised, or reinforced. Figure 1 depicts some of the most popular real-world machine learning applications.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6046
***Article Received**: 26 November 2022 **Revised**: 24 December 2022 **Accepted**: 04 January 2023*

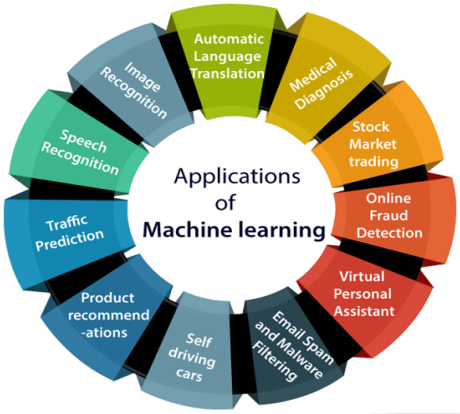

Fig.1 Machine Learning

We'll use one of the machine learning algorithms to track down infected people using their location information and shape unique clusters.

Persons after getting infected with the deadly virus will carry the virus with them and can spread to another one. So, there is a need to follow up these contacts and identify the contacts to whom the infected persons contacted. So, this will help in give medical treatment to people exposed to the person before the symptoms arises.

If a person becomes inflamed with the ailment and does not get checked, he is unfastened to head anywhere he wishes, permitting the disorder to unfold to different human-beings. In some other state of affairs, if he is screened and the outcomes indicate fake positivity, he is loose to travel approximately and unfold the aliment speedy. So, there is a want for contact tracing.

Different pandemics [6] are arising in the present situation the health care authorities need to be very careful to stop the spread of these infectious diseases. Effective strategies need to be implemented. In the past we can see some of the dangerous diseases which caused a huge loss to the humans without proper knowledge about how to contain the infections.

From those situations we have to learn and make efforts to effectively implement the different methods to stop spread of diseases. Most of the work is concentrated in two areas: First collection of the location data of the contacts and making a dataset and visualizing the data. Second using the machine learning algorithm i.e., DBSCAN, forming the clusters and tracing the most probable infected contacts. The objective of our work is to: Form the clusters (using safe distance) and to Find the infected contacts.

We have implemented this paper in python programming language. We used machine learning [7] to identify potential infected contacts of the known victim. In machine learning, clustering algorithms can be used to do this type of work. We have used the DBSCAN algorithm in our project. What these clustering algorithms do is they form clusters rom the data given and by using those clusters we can trace the potential carriers of the disease.

To properly apprehend the implementation info of the task, it is important that we first recognize what system getting to know is, what's clustering and how all the distinct clustering algorithms function after which we are able to be capable of recognize why we have chosen the DBSCAN set of rules in preference to different algorithms and the way we carried out it in our assignment particular.

Machine learning [8] is a branch of computer science and specifically a sub division of the artificial intelligence branch which focuses on large amounts of data and implements algorithms which try to imitate human behaviour and perform tasks as if they are being performed by humans without any input or instructions from the user (human being).

That is, by using machine learning algorithms we are able to train computers to perform tasks which require intelligence. The computers when implemented with these machine learning algorithms can actually learn from the previous work and all the data that is supplied to them and work on future data without the need to give them instructions on how to work with that future data.

In today's world, machine learning techniques are being used in every walk of the daily life and it is not an exaggeration to say that applications of machine learning can be found in almost all the applications of our daily life which have previously been using traditional programming to do the job and not only these, but previously unseen and impossible tasks are also being made possible by machine learning, such as in fields like health care, environmental conservation, animal species extinction prediction, and a lot of other simple daily life applications such as email classification, etc.

A neural network can be considered to be designed based on the neural networks [9] (networks of neurons) in our brain. Millions of these neurons form a network that helps us human beings in performing our day-to-day tasks and it can be said that these neural networks are responsible for our intelligence and cognitive abilities. These artificial neural networks are model upon these neural networks in our brain, in ways like, these artificial neural networks are very interlinked with each other and rapidly transfer information between each other and there are also several layers of these neurons and the information or data is passed through all these layers and there will be a final output layer where we will be able to see the output of all the computation that has been done.

These neural networks can then be tweaked little by little, to make them produce the desired result and therefore be trained properly [10]. This tweaking is done assigning

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6046
***Article Received**: 26 November 2022 **Revised**: 24 December 2022 **Accepted**: 04 January 2023*

weights to the nodes of the network and thereby affecting the data that is being passed through that node and with these tweaking of weights, the output is also changed and we can obtain our desired output by tweaking our way through these networks.

II. Literature Survey

There have been many severe diseases that have evolved in the past that have caused significant loss of life, and these types of diseases are still emerging today and will continue to do so in the future. To keep these disorders under control A thorough examination of diseases, as well as preventive and control methods, is essential. There are a few diseases that are carried by humans, such as person to person transmission, droplet dissemination (direct touch), and airborne transmission, contaminated objects, food, animal to person contact, and bug bites (Indirect contact). Everyone is at risk of becoming unwell because infectious diseases can spread through direct or indirect contact. SARS, Ebola, Lassa fever, tuberculosis, and, most recently, COVID 19 [11] are examples of infectious diseases. These diseases necessitated a quick response and precise control techniques.

Since the 1980s, several new models and theories for contact tracing have been developed with the goal of finding the most successful models and studying contact tracing to reduce the spread of infectious illnesses [12]. And, as of today, there are a few unanswered questions or, to put it another way, new challenge. We will check and study the olden models and the new challenges that are arising in the field of contact tracing.

If the population has good immune response is then it is good, but if not immune then we need the vaccines and which is the late process [13]. The main problem we can find is the small false positivity can lead to large positivity. We will discuss different approaches to contact tracing through different papers and publications. Among all the papers the first paper for Contact tracing is a study by Hethcote and Yorke, at the time of spread of disease ganorrhea. They wrote the paper Gonorrhea transmission dynamics and control, citing the impact of contact tracing caused by a reduction in the effective transmissibility of infection [14]. Different models till date we can identify as individual based simulation models

As previously stated, mean numerical simulations are the most well-known robust algorithm to reformulating IBMs in terms of differential equations (ODEs). Rather than counting the number of members of particular kind, the ODE section defines the inferred relative frequencies (e.g., S, I, and R). Especially if the contact graph is densely clustered natively: If this is the case, the neighbour of an infected person is often already infected, and the spread of infection is slowed [15].

An enhance mean field approach, the pair estimator, was established in the 1990s by Japanese groups. When it comes to Contact Tracing modelling, pair assumption retains a key piece of information that mean field asymptotic discards: We know how effective it is that a neighbour of an infected person is infected as well. Few papers which are published earlier have already discussed this idea. Contact tracing is more efficacious in bunched populations than in small societies, according to. Iterative but each tracing are compared in, and "targeted CT," that is, Contact tracing concentrating on a risk group, is examined [16]. Recursive and targeted contact tracing were discovered to be especially effective.

The model calculates an infection tree, with nodes representing infected individuals and directed edges connecting infector and infected. We would rather have a forest than a tree as recoverable individuals are removed. It is now possible to represent contact tracing directly in the path of this process, as in IBMs. If a tree member is diagnosed, the adjacent nodes will indeed be tested and, if infected, isolated. The frequency of being infectious at a given time after infection is measured in order to address the nonlinear system [17]. This statistical likelihood is the central application that enables for the quick determination of an infection's effective reproduction number or species evenness. The removal rate can be assumed using heuristic arguments even in cases of high occurrence, and a modified average field equation has also been recommended. Several research articles look at branching process models from a various angles, keeping in mind generating operations for the degree distribution of randomly selected infectious agents. They do so by extending the traditional percolation-based analysis of epidemics on graph data to Contact tracing. Based on a branching-process formulation for CT, Tanaka 2020, and analysed data for COVID-19 [11][18] to estimate the fraction of asymptomatic cases.

III. Implementation

We have implemented this project using python3. Python is very flexible and easy to write, so we have chosen python to implement our project. To implement machine learning in python, we can use several machine learning libraries in python like sklearn, Tensorflow etc. We chose sklearn as it is straight forward and relatively easy to learn.

As mentioned above, we need a working python3 environment set up to implement this project. We used Jupyter notebook. Jupyter notebook has a neat and tidy approach and helps to easily visualize and understand the code, so we have chosen Jupyter notebook.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6046
***Article Received**: 26 November 2022 **Revised**: 24 December 2022 **Accepted**: 04 January 2023*

Our python environment was set up in a Linux machine. We can have python installed on any OS like Linux, windows 10 and mac OS. We just need to go to python.org and follow the instructions there for our specific operating system and we will have a working python environment within minutes. We have set up a virtual environment for python in our Linux machine and installed all the dependencies inside that virtual environment.

A virtual environment isolates the python environment inside the with the systemwide python environment. So, we will be able to use our python environment in however way we want without worrying about polluting the systemwide python environment and also potentially creating dependency and ownership issues by our use of pip (the pythons default package manager). Using pip systemwide to install python modules/libraries that are necessary for our project can be dangerous and we will have to manually delete those packages that we have downloaded with pip, if those packages are needed to be installed as dependencies for another systemwide package by the system default package manager. So, we always suggest you create a virtual python environment for any project that you plan to do, in order to save your system from going into dependency and ownership issues.

It will also be helpful to learn how to navigate and use Jupyter notebook, because it helps us to do our work efficiently and accurately. As with any other job, it is therefore always recommended to learn the tool that you use to do your work and also learn to use it efficiently and to it's full potential. Installing the python modules/libraries that are necessary for our project can be done using pip, which is the python package installer and it is a very convenient and handy tool to manage all the python libraries that are installed with our python environment on our system. We can also install packages without using pip and the process may be different for different operating systems. For example, in Linux, you will have some python libraries directly in the repositories of our distribution. When we are installing any python modules/libraries globally, we should prefer to use the distribution's package manager instead of pip as we have already discussed that using pip globally on a system can cause a lot of issues and can potentially damage the integrity of your system and can cause other problems.

Pip is a very convenient tool and it is operating system agnostic. In this section, we will explain in detail the software tools and methods that we have used to implement our project. First of all let us start with the computer type and OS.

The first few steps show how to read the data from the database and process I and also visualize it to get a better picture of the data that we are dealing with. This helps us in properly understanding the problem giving out an effective solution based on the inferences that we gather from the data pre-processing and visualization phase.

The next few steps correspond to the model generation and execution phase where we implement all the solutions that we have planned for the problem based on the inferences that we have gathered from the above data pre-processing and visualization phase. The algorithm is DBSCAN (Density Based spacial Clustering of Applications with Noise). This algorithm, as we can see from the title uses clustering with a density-based approach and this algorithm also can handle some noise within the data points supplied to the algorithm.

This module processes the data available in the dataset into proper format which can be given as input to the machine learning module.

This module also helps us visualize the data in the form of scatter plots and graphs which helps in observing trends and patterns in infection spreading.

In this module, we import the data from the dataset and put it in a Pandas data frame, we then process the data and also make a scatterplot representing the distribution of people at various location co-ordinates and this gives us a hint at potential infected contacts.

Procedure for implementation:

- Place the dataset in a location accessible from the python environment.
- Load the dataset into a Pandas data frame.
- Check how the data is formatted in the dataset and look for any anomalies.
- Now, after the data is properly formatted, display the head portion of the data to confirm.
- Now, plot a figure using matplotlib and give it appropriate parameters for fig size.
- Draw a scatterplot by considering the columns of the data in the data frame that was formatted earlier.
- Add legends to the scatterplot to improve readability of the scatterplot.

A flowchart is nothing but the graphical representation of nay algorithms or procedures. This flow charts are mostly used in many fields in order to develop, enhance and understand the procedures through simple diagrams. It is similar to the activity diagram.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6046
***Article Received**: 26 November 2022 **Revised**: 24 December 2022 **Accepted**: 04 January 2023*

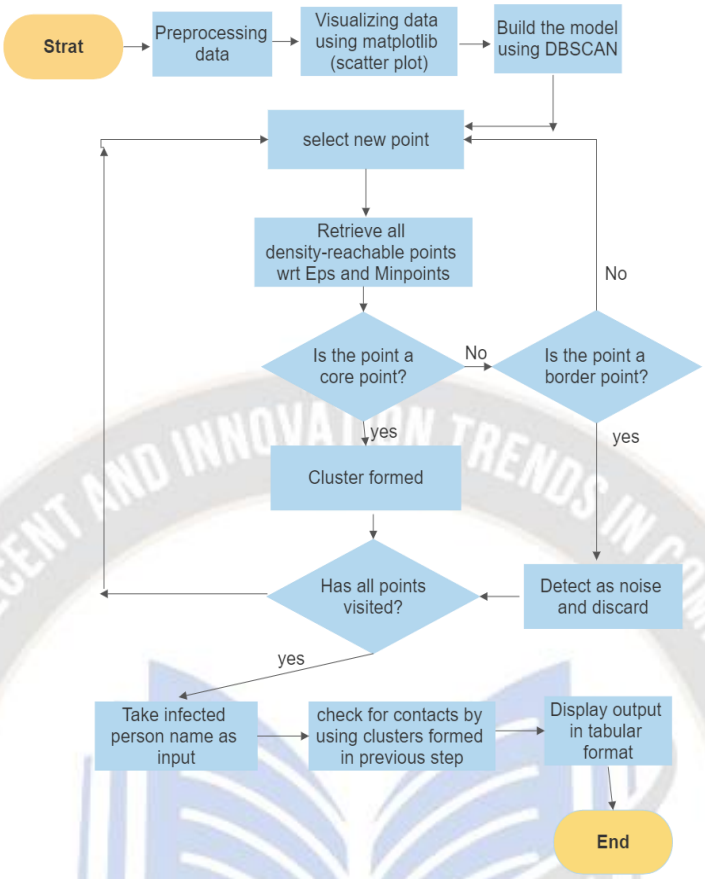

Fig.2 Implementation Flow

System architecture

The process of identifying the subsystems that make up the system, as well as the structure for subsystem interconnection, is known as system architecture design. The architectural design's purpose is to define the software system's general structure.

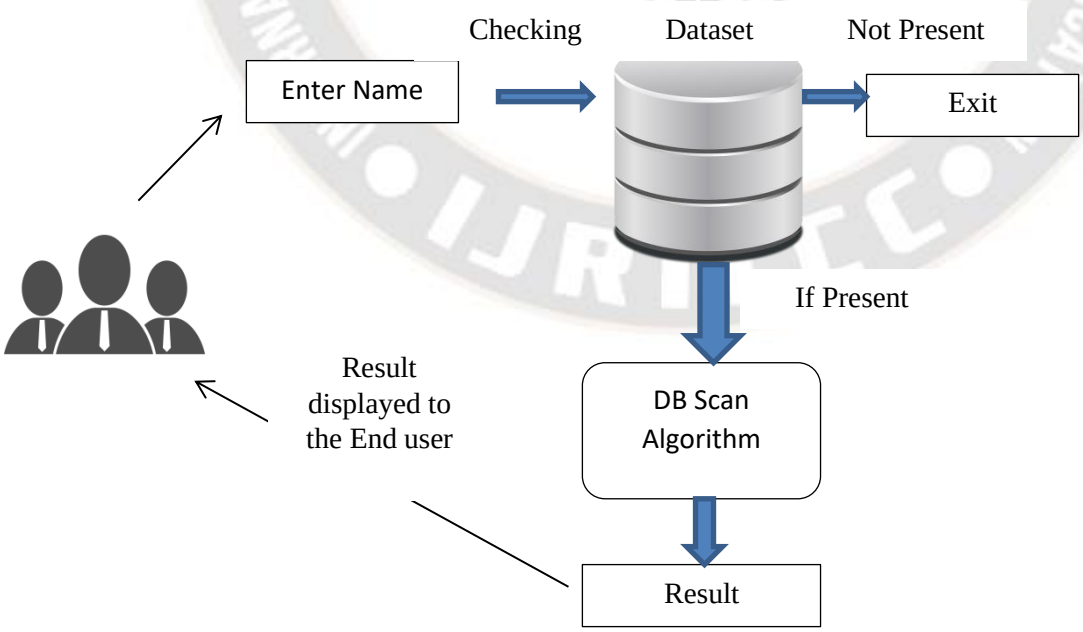

Fig.3 System architecture

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6046
***Article Received**: 26 November 2022 **Revised**: 24 December 2022 **Accepted**: 04 January 2023*

Machine Learning based Clustering (MLDBSCAN)

MLDBSCAN is a density-based clustering algorithm. It can form cluster on spatial or geographical data based on the density of the data points at each location coordinates. This is different from other clustering algorithms as they don't necessarily use density-based clustering and differ significant with the DBSCAN algorithm in other ways as well such as the functionality and the working techniques.

We can say that DBSCAN is the perfect algorithm for our particular problem because of the following factors and influencing characteristics of our dataset. Our data consists of a few columns which tell us about the identity of the person along with the the location coordinates of the person. Our database contains these specific values, so as to convey all the useful information that is used in contact tracing i.e., the name (which is the identity of the person) and their location coordinates at different dates and different time stamps. All this data is highly valuable in determining the contacts of any specific person and help contain the spread of any infection.

IV. Results & Discussion

Data visualization is a key aspect of consideration when developing any project which involves a lot of data. Data visualization helps us infer a lot of things from the data that we are dealing with. Data visualization is a process where we try to visualize the data in the form of a graph, a chart or any other graphical representation.

Our data consists of the location coordinates; hence, we chose to visualize our data in the form of a scatterplot. Our scatterplot consists of data points scattered across the plane with latitude as x axis and longitude as y axis. This helps us to visualize all the points that a person has been and all the contacts that a person has had.

A scatterplot uses data points scattered across a 2d plane to visualize the data and to describe the connection between two variables. In our purpose, the scatterplot, while is very helpful, cannot display the whole of data as it is just a 2d plot. Ideally we can also add a z axis which represents the time of day and can therefore help to pin point the contact between any two given individuals.

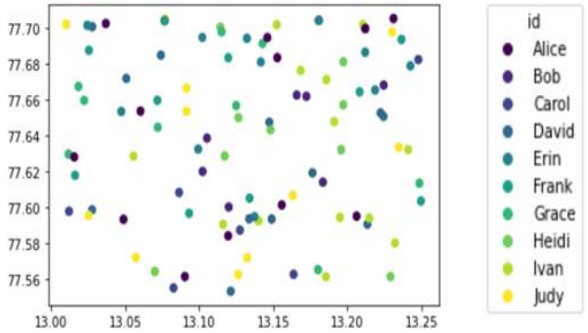

Fig.4 Matplotlib scatterplot with legends

We have provide three parameters epsilon (Eps), minimum points(Minpts), metric.

- **Epsilon** (Eps): Two points are considered neighbours if the distance between the two points is below the threshold epsilon.
- **Minimum points**: The minimum number of neighbours a given point should have in order to be classified as a core point. **It's important to note that the point itself is included in the minimum number of samples.**
- **metric**: The metric to use when calculating distance between instances in a feature array (i.e. euclidean distance).

The algorithm forms clusters based on the above given metrics. The total numbers of clusters formed are showed below with their respective numbering and color.

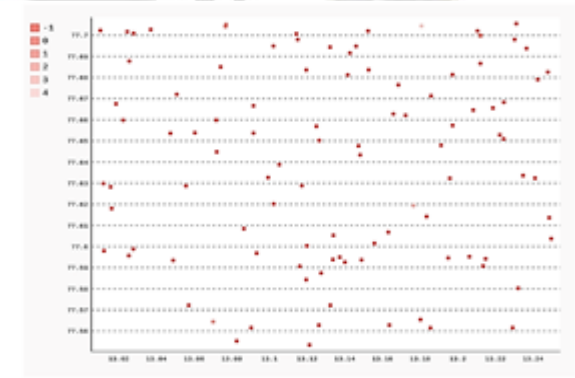

Fig.5. Scatterplot of six clusters.

Here we can see that, a total of 6 clusters have been formed and are numbered from -1 to 4.

Now, we will take four inputs and see thir respective output clusters and their contacts:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6046
***Article Received**: 26 November 2022 **Revised**: 24 December 2022 **Accepted**: 04 January 2023*

case 1, Bob :

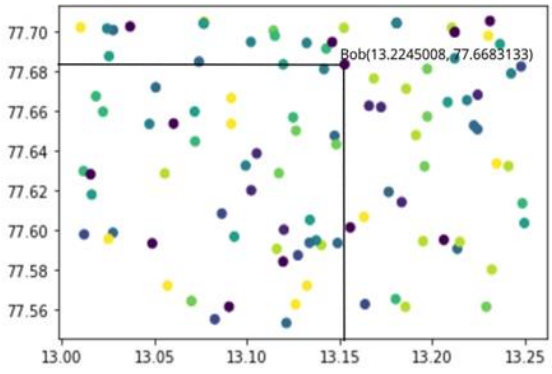

Fig.6. showing bob's position in scatterplot.

Case 2, Judy:

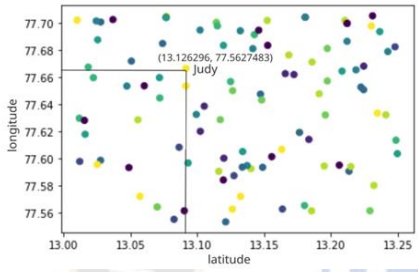

Fig.7 Showing judy's position in scatterplot.

Case 3, Heidi :

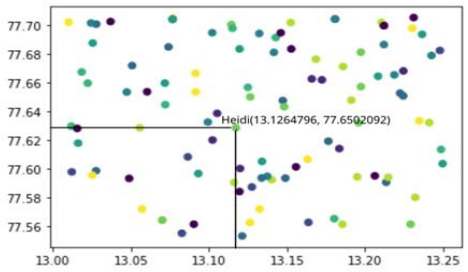

Fig.8 Showing heidi's position in scatterplot.

Case 4, Ivan :

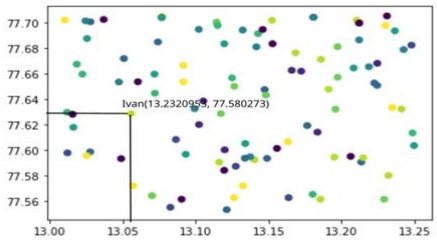

Fig.9. Showing ivan's position in scatterplot

V. Conclusion

Mathematical models are useful in anticipating new transmittable diseases because they allow stakeholders to plan for future public health situations before they occur. However, because accurate data is often scarce in such situations, long-term dynamics forecasts are frequently correlated with large confidence intervals. Our model answers the fundamental and ambitious question of whether rapid and complete contact tracing is appropriate for identifying infectious diseases. Contact tracing in public health is more complicated because it relies on the relative chronology of events and the monitoring of known contacts using machine learning algorithms. For contact tracing to be an effective health promotion measure, most secondary cases must be separated later before they become contagious. Several digital contact tracking designs have emerged, and related app applications have been chosen by governments worldwide. Bluetooth has been identified as the most promising wireless technology for implementing the contact tracing service, particularly its power-saving variant, Bluetooth Low Energy (BLE).

References

[1] African Union, Africa Centres for Disease Control and Prevention.Guidance on Contact Tracing for COVID-19 Pandemic – Africa CDC. 2020. Available from https://africacdc.org/download/guidance-on-contact-tracing-for-covid-19-pandemic/Google Scholar.

[2] Agarwal, Manju & Bhadauria, Archana. (2012). Modeling H1N1 flu epidemic with contact tracing and quarantine. International Journal of Biomathematics. 5. 38-57. 10.1142/S1793524511001805.

[3] Yadav, Latika & Maurya, Poonam & Maurya, Neelesh. (2023). New Onset of Health Complications in Patient after COVID-19 Recovery. Sustainability, Agri, Food and Environmental Research (ISSN: 0719-3726). 10.13140/RG.2.2.36236.03201.

[4] Angeletti, S., Benvenuto, D., Bianchi, M., Giovanetti, M., Pascarella, S., & Ciccozzi, M. (2020). COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. *Journal of medical virology*, *92*(6), 584–588. https://doi.org/10.1002/jmv.25719

[5] Allam, Z., & Jones, D. S. (2020). On the Coronavirus (COVID-19) Outbreak and the Smart City Network: Universal Data Sharing Standards Coupled with Artificial Intelligence (AI) to Benefit Urban Health Monitoring and Management. *Healthcare (Basel, Switzerland)*, *8*(1), 46. https://doi.org/10.3390/healthcare8010046

[6] Tsehay Admassu Assegie (2021). An Optimized KNN Model for Signature-Based Malware Detection. International Journal of Computer Engineering in Research Trends.8(0032).46-49.

[7] Balduini, M., Brambilla, M., Della Valle, E., Marazzi, C., Arabghalizi, T., Rahdari, B., & Vescovi, M. (2019). Models and practices in urban data science at scale. Big Data Research, 17, 66-84.

[8] Begun, M., Newall, A. T., Marks, G. B., & Wood, J. G. (2013). Contact tracing of tuberculosis: a systematic

review of transmission modelling studies. *PloS one*, *8*(9), e72470. https://doi.org/10.1371/journal.pone.0072470

[9] Breban, R., Riou, J., & Fontanet, A. (2013). Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk. The Lancet, 382(9893), 694-699.

[10] Sushma Joshi, S.M Joshi.(2019). Phishing Urls Detection Using Machine Learning Techniques. International Journal of Computer Engineering in Research Trends.6 (6).326-333.

[11] Gonzalez-Dias, P., Lee, E. K., Sorgi, S., de Lima, D. S., Urbanski, A. H., Silveira, E. L., & Nakaya, H. I. (2020). Methods for predicting vaccine immunogenicity and reactogenicity. Human Vaccines & Immunotherapeutics, 16(2), 269-276.

[12] Lin, L., Lu, L., Cao, W., & Li, T. (2020). Hypothesis for potential pathogenesis of SARS-CoV-2 infection–a review of immune changes in patients with viral pneumonia. Emerging microbes & infections, 9(1), 727-732.

[13] Lippi, G., & Plebani, M. (2020). Laboratory abnormalities in patients with COVID-2019 infection. Clinical Chemistry and Laboratory Medicine (CCLM), 1(ahead-of-print).

[14] Y.Yashasree,K.Venkatesh Sharma (2020). Creditcard Fraud Detection and Classification Using Machine Learning Based Classifiers. International Journal of Computer Engineering in Research Trends.7(9).1-8.

[15] Narin, A., Kaya, C., & Pamuk, Z. (2021). Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern analysis and applications : PAA*, *24*(3), 1207–1220. https://doi.org/10.1007/s10044-021-00984-y

[16] Hussan, M.I. & Dorepalli, Saidulu & Anitha, P. & Manikandan, A. & Naresh, P.. (2022). Object Detection and Recognition in Real Time Using Deep Learning for Visually Impaired People. International Journal of Electrical and Electronics Research. 10. 80-86. 10.37391/ijeer.100205.

[17] Ndiaye, B. M., Tendeng, L., & Seck, D. (2020). Analysis of the COVID-19 pandemic by SIR model and machine learning technics for forecasting. arXiv preprint arXiv:2004.01574.

[18] V. Krishna, Y. D. Solomon Raju, C. V. Raghavendran, P. Naresh and A. Rajesh, "Identification of Nutritional Deficiencies in Crops Using Machine Learning and Image Processing Techniques," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 925-929, doi: 10.1109/ICIEM54221.2022.9853072.