A novel m-Health system for epidemic tracking and prediction using Big Data and Electronic health record

Mohammed Bouziane Soussi, Mourad Hadjila, and Rachid Merzougui
STIC Lab., Dept. of Telecommunications Faculty of Technology, University of Tlemcen, Algeria
bouziane.soussi@gmail.com, mhadjila_2002@yahoo.fr, merzrachid@yahoo.fr

2020 4th International Symposium on Informatics and its Applications (ISIA) | 978-1-7281-9652-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISIA51297.2020.9416555

Abstract— **In modern cities, the investigation of an outbreak using the traditional systems is a time-consuming task that requires collecting and processing complex information in order to track the origin of the disease. The COVID-19 Virus shows the current system limitation. For this reason, establishing an early detection and prediction is essential to halt the spreading of epidemics in an earlier stage. The aim of our work is to design a system capable of detecting and forecast people carrying of an epidemic in the first days of incubation using geolocation of the patient's and big data.**

Index Terms—M-Health, Tracking, Prediction, Big data, Electronic Health Record, Epidemic.

I. INTRODUCTION

Recently, the world is witnessing a novel coronavirus "COVID-19" epidemic, which has rapidly affected almost all the countries and has left more than 469,500 dead from December 2019 to June 2020 [1]. The incubation period is an approximately 5.2 days with median of 14 days [2, 3] to appear first symptoms fever, cough, shortness of breath or difficulty breathing, tiredness, aches, runny nose and sore throat [4, 5]. Person to person transmission occurs high causes of spreading by close and direct contact or through droplets spread by sneezing or coughing from an infected individual [6, 7].

The COVID-19 is the perfect example, due to its long incubation period; it spreads silently within several days before showing the first symptom, to show the importance of a complete e-health system for an earlier detection and prevention of the epidemic. Due to the variety of the health data (e.g. Medical History, Blood analysis Lab Results, Medical images, Patient Details, wearable health devices... etc) all this data information stored in database is known as Electronic Health Records (EHR) [8]. Electronic health records are digital versions of a patient chart, but are a more detailed record of a patient's medical history. EHRs are designed to be shared with other medical providers so that authorized users can easily access a patient's records digitally from different health care practices. EHR is consistently explosive in volume and can store diverse types of data such as text, image, audio, or video. The traditional methods and management tools are not sufficient enough to manage and analyze so large and complex data sets. There are new and innovative Big Data tools that have ability of managing healthcare data.

Nowadays, the world is focusing on enhancing healthcare using ICT technologies (e-health) such as smart mobile phones (m-health) because these widespread devices allows people to be permanently connected to the internet on one hand and the integration of multiple sensors (heart-rate, step-counter, location sensor) on the other hand [9].

A complete e-health system requires a big database to store information about the health of each citizen. Big data provides a great opportunity for epidemiologists, and health policy experts to make data-driven judgments that will eventually develop the patient care.

C. Jason Wang at al [10] gives the use of technology Big Data Analytics and New Technology of Taiwan government to lunches a quarantine system for travelers arriving on their land; this application aims to check their travel during their stay in relation to the epidemiological map using the QR code. Kia Jahanbin [11] uses data originate from social networks specially Twitter (comments, photos and videos) about the spread of an outbreak to prevent and control epidemics. Rohan Pandey at al [12] uses machine learning and m-Health to disseminate information on the role of hygiene in the prevention of COVID-19 in the local language (Hindi) as audio-visual content. In [13] Tsung-Han Chen at al used massive data collected from three datasets, Google trends data, King Net national medical diagnosis and consultation records and Centers for Disease Control in Taiwan from 2010 to 2016 to explain their correlation in the aim to get prediction of flu trend. In [14] Najihah Ibrahim et al. introduced how to use machine learning to determinate the epidemic disease infected area by classification of epidemic diseases dissemination (population density, climate, geodemographic, clinical case, vaccination tracking, geo-mapping using social media).

Yuanfang Chen [15] analyzed mobile big data shared by volunteers of Ebola outbreak area to evaluate the impact of the network to predict disease dynamics. Rajinder Sandhu Harsuminder et al. [16] used collected data from various sources (healthcare services, social media, individual users, etc) to classify patients as infected or uninfected by H1N1 virus according to WHO standard.

From the literature review, we have noticed a shortfall in tracking the origin of the outbreaks in real time. This motivates us to propose a solution capable of detecting the evolution of epidemic map and predict infected persons before the first signs appear.

This paper highlights the design paradigm of a new solution proposed to predict epidemic map in real time in order to track the origin of the disease. The solution consists of periodically collecting the geographic location of each citizen using mobile phone or smart e-health sensor. The collected data is stored in an electronic health record (EHR) that contains clinical data of each citizen, then using big data analytics techniques to detect the outbreak and track the infected person and/ or area by using historic GPS locations.

The remaining this document is organized as follows: Section I introduces some important concepts and summarizes the related works. Overall architecture is presented in Section II. The conclusion and discussions are presented in Section III.

II. OVERALL ARCHITECTURE

Based on the definition of the epidemic as an inflectional contagious disease affecting many persons at a particular time and spreading over a geographical area, there are some important parameters must be taken into account: timing, location and the way of spreading. For this reason, establishing an early detection and prediction systems is a crucial step to combat.

The proposed solution will help government and healthcare departments to track, forcast and prevent outbreaks spreading, the architecture is composed of smartphone, EHR and treatemnt of data (see Fig.1).

A. Mobile phone:

To achieve this goal, we will take the advantage of the mobile technology in terms of connectivity, simplicity and its ubiquity in our daily life.
First, each citizen shoud install an application "eHealth_App" on his smartphone, which will display a simple registration form so that the user could register to the server by filling in all the field Firstname, Second name, birth day and national identity number). The national identifier number is

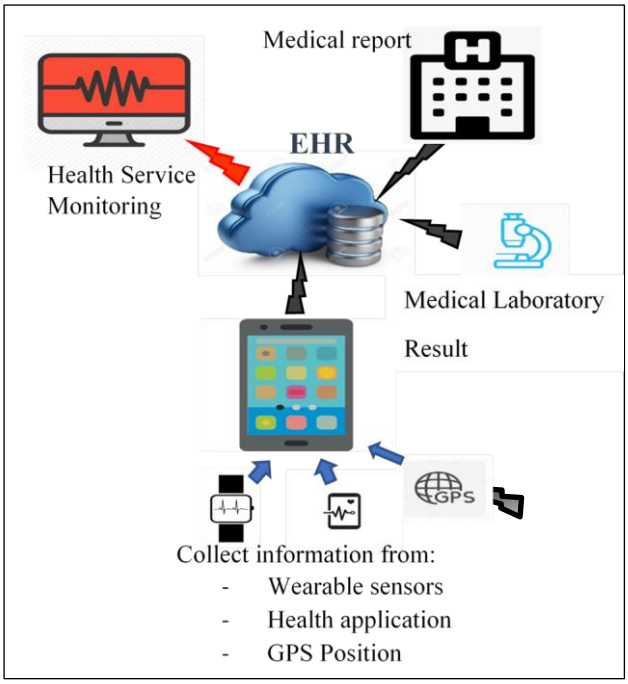

Figure 1. Overall architecture

used as a unique ID in our database and intended to verify user authenticity.

This application runs in background to continously monitor wearable sensor and a movement of the user and sends periodicly the signs life and geographic position GPS to the server.

The frequency of sending GPS positions i.e. its time resolution depends on the velocity of the mobile such that the app will send an update of the GPS position many times when the mobile moves rapidly and vice versa. This use keeps the precision of the location and optimal battery utilization. The flowchart of reading GPS position is presented in Fig.2.

B. eHealthServer

EHR is a database that can be used to build and manage an e-Health system- For our solution, the file of each patient in the database should store the GPS location sent by the mobile device. For privacy concern, this data should be protected and restricted to government healthcare employee only through a non disclosure agreement. Fig.3 presents EHR architecture where GPS location can be added for every patient files.
Traditional databases have set their limits with the volume of medical data, imaging, blood analysis and in addition, the promotion of wearable health devices accelerates the explosion of healthcare data [9, 17] and complex heterogeneous data [18], hence the need for migration to a NoSQL database.

Figure 2. Flowchart of mobile phone application

Cassandra is designed to store and handle a huge volumes of data from multiple servers, it is open source, elastic scalable, column-oriented and has a Peer to Peer Architecture [19]. Cassandra is considered as a hybrid NoSQL database. It supports the column oriented database of BigTable Data model [20] introduced by Google, and a key-value

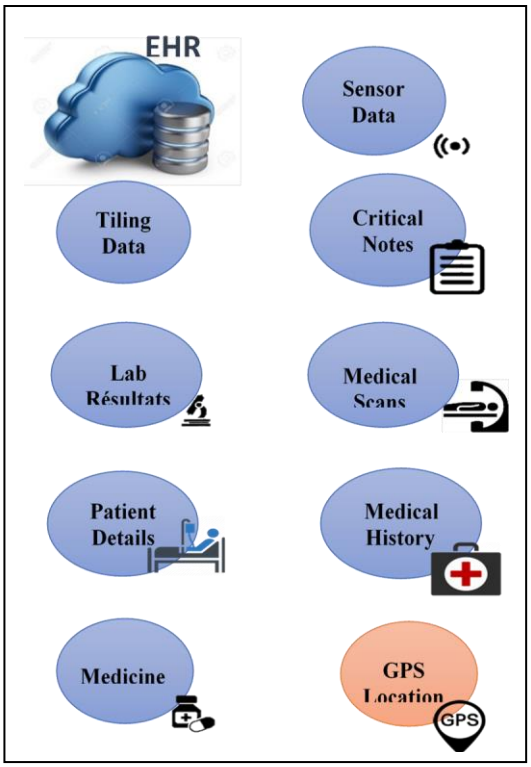

Figure 3. EHR architecture with GPS Location

and document database owned by Amazon's Dynamo [21]. Cassandra database is constructed from *keyspaces*, column families, columns, and rows, where each row has unique row key, the characteristic of key-value model inspired from Dynamo [21, 22, 23](see Fig.4 and Fig.5).

The collected information from medical report, medical labotory result and from application of each citizen is stored and managed by in a Cassandra databse system which is a perfect platform for mission-critical data.

Figure 6 illustrates an overview the EHR server database architecture

- *keyspace* = EHealthServer,
- *Column Family* 1= Personal information,
  - *Row Key i* = National Identity number *i* (NIN *i*),
  - *SuperColumn Name* = Personal identity,
    - *Column name 1*= F.Name
    - *Column name 2* = S.Name
    - *Column name 3* = Birth day
    - *Column name 4* = Gender
  - *SuperColumn Name* = GPS
    - *Column name* = Location
    - *Column value* = Row_key1
  - *SuperColumn Name* = Labo result
    - *Column name* = Result
    - *Column value* = Row_key2

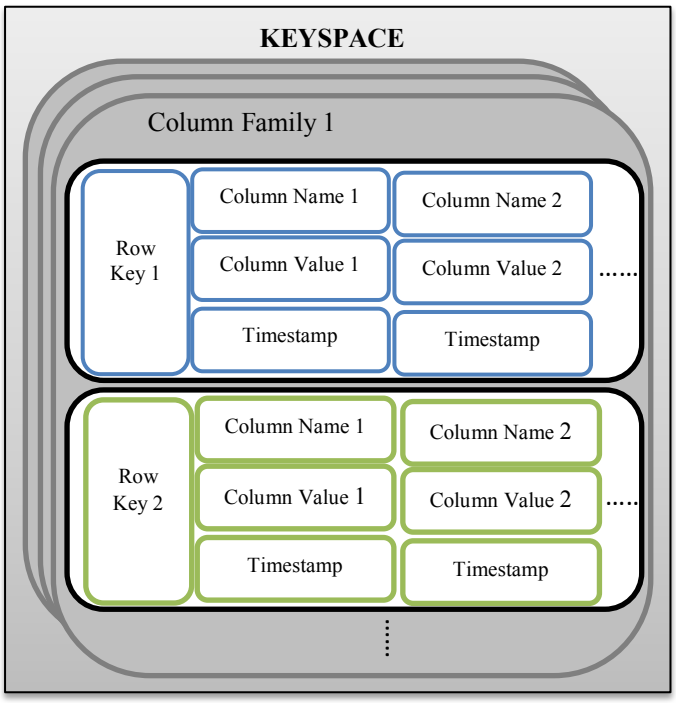

Figure 4. Cassandra Data Model

Relational Model	Cassandra Model
Database	Keyspace
Table	Column Family
Primary key	Row key
Column name	Column name
Column Value	Column Value

Figure 5. Relational Model vs. Cassandra Model

- *SuperColumn Name* = Medical
  - *Column name* = Report
  - *Column value* = Row_key3
- *Column Family* 2= Location,
  - Column name = Position
  - Column Value = Latitude,Longitude
  - Timestamp

Timestamp represents the time of registration '*yyyy-mm-dd HH:MM:SS [(+|-)NNNN]'*

C. *Treatment of data*

The spread of an epidemic or contagious infection depends on the methods of propagation. Contamination can be either by water as in the case of typhoid, cholera and polio; or by mosquitoes as in the case of yellow fever, dengue fever, malaria; or by air as in the case of H1N1 and even COVID-19.

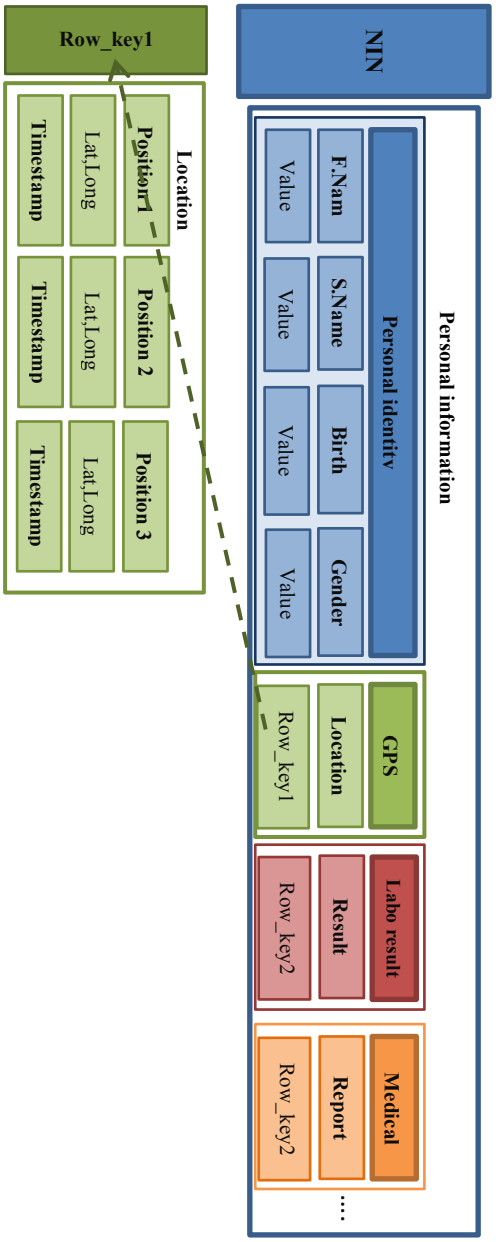

Figure 6. EHealthServer architecture

With these examples, notice that there are epidemics that:

- propagate in a locality,
- propagate by people

1- Prediction of an infected area:

The dataset used in this article is taken from the case of a cholera epidemic, which occurred from August, 7th to September, 6th 2018. The disease has affected 217 cases two of them have been died (announced by The Algerian ministry of Health). The confirmed cases have registered in six (06) provinces of the country. One water source tested positive from 21 sources have been tested for bacterial contamination in the infected areas (in six 06 cities) [24].

In our example, 04 first patients geographically separated (see Fig.7) show symptoms of cholera (i.e. acute watery diarrhea, severe dehydration, with or without vomiting) where each of them appears at the nearest medical center.

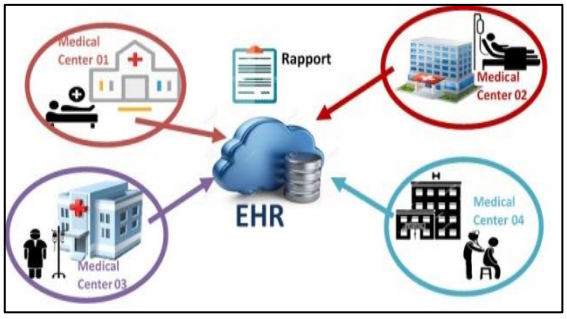

Figure 7. Four patients have same epidemic symptoms

So when they are positively tested to the disease (Vibrio Cholerae O1 or O139), the data should be filled in on our EHR platform to update, the patient files. Then the system will send an alarm to the healthcare services and government officials. The system will also track the movement of the patients throughout the incubation period of the disease (in our case, it is 07 days), the intersection of the GPS positions of the first patient gives us the most likely infected area (see Fig.8 and Fig.9).
In our case (in Algeria), the healthcare services carried out the investigation for the contaminated source in the six 06 provinces where there was presence of infected patients (an area of 22600 km2), without the use of system such ours, the investigation took one month to locate the source with a cost of two deaths.

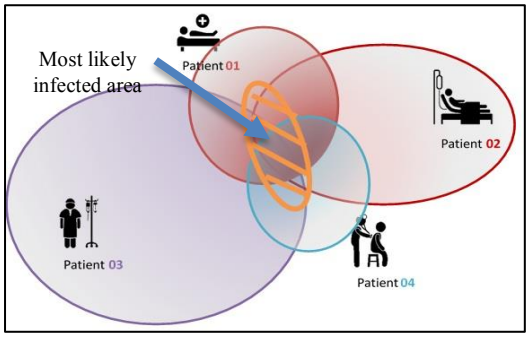

Figure 8. Prediction of an infected area

2- Prediction of an infected persons

In the case when the outbreak spreads by people (respiratory pathway), it isn't enough to determinate just infected area but also to track all people who have contact with the patient during incubation period of epidemic in order to get the potentially infected person before the first symptoms appear.

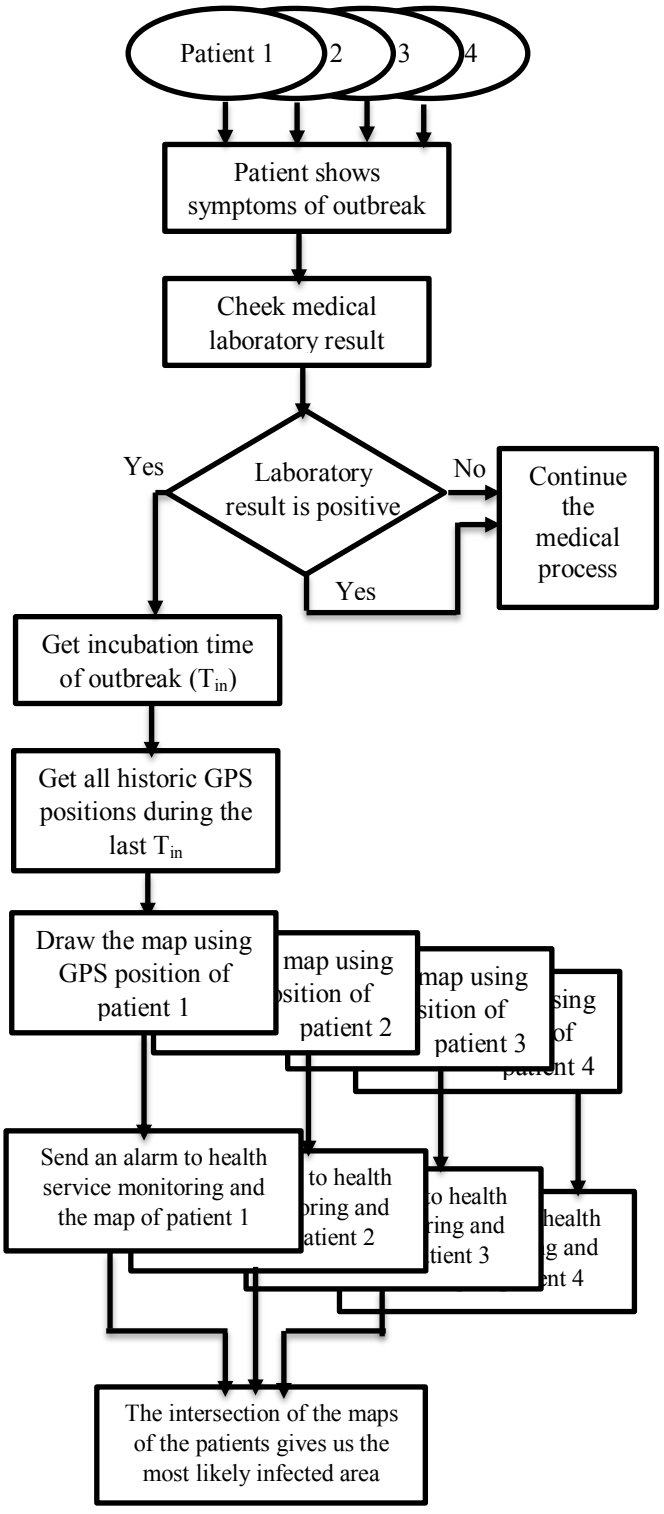

Figure 9. Flowchart to predict an infected area

In the contact, there are two type people (see Fig.10):

- Known people (family, friends…), that kind of person is easy to recognize
- Unknown people: the people traveling on bus, subway, people who go to commercial center, this category is the most dangerous

We take the example of COVID-19 outbreak, if we have a confirmed case. By using historic GPS position in the last 14 days (incubation time of COVID-19) of the patient, we compare this location with all GPS position in our EHealthServer of all citizens. The results will be classified based on:

- Same location in the same time
- How long they were together

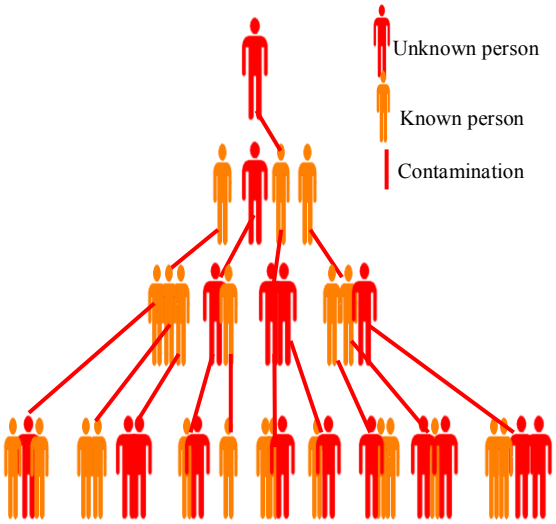

Figure 10. Outbreak method spreading

More the time to be together is long more the probability is high to be reach of disease and can add more criteria age gender and medical history all these parameters depend of the epidemic (see Fig.11). The list of people potentially infected will be send to healthcare service monitoring to quarantine and track their progress to stop spreading of the outbreak early.

III. CONCLUSION

With all the technological developments in the healthcare, the world is plunging into a big KO for more than a month, caused by the spread of the COVID-19 virus. Our proposed solution will deal efficiently with the most critical constraint, which is the timing of detecting an epidemic, it will have these advantages:

- Real-time: dissemination of information between medical centers and health services will be in real time
- Location: a better prediction of the infected geographical area and get real epidemic map
- Prediction: a better prediction of person potentially infected.

With all this information sent to government officials in real time and the list of all suspect people the healthcare service will have all information to identify and stop spreading of an epidemic early (better logistics, organization and appropriate medicines)

In this article, we tried to demonstrate the usefulness of adding the patient's GPS positions in EHR, in order to predict future epidemics, minimize the time of action and the rate of propagation, in order to save human life.

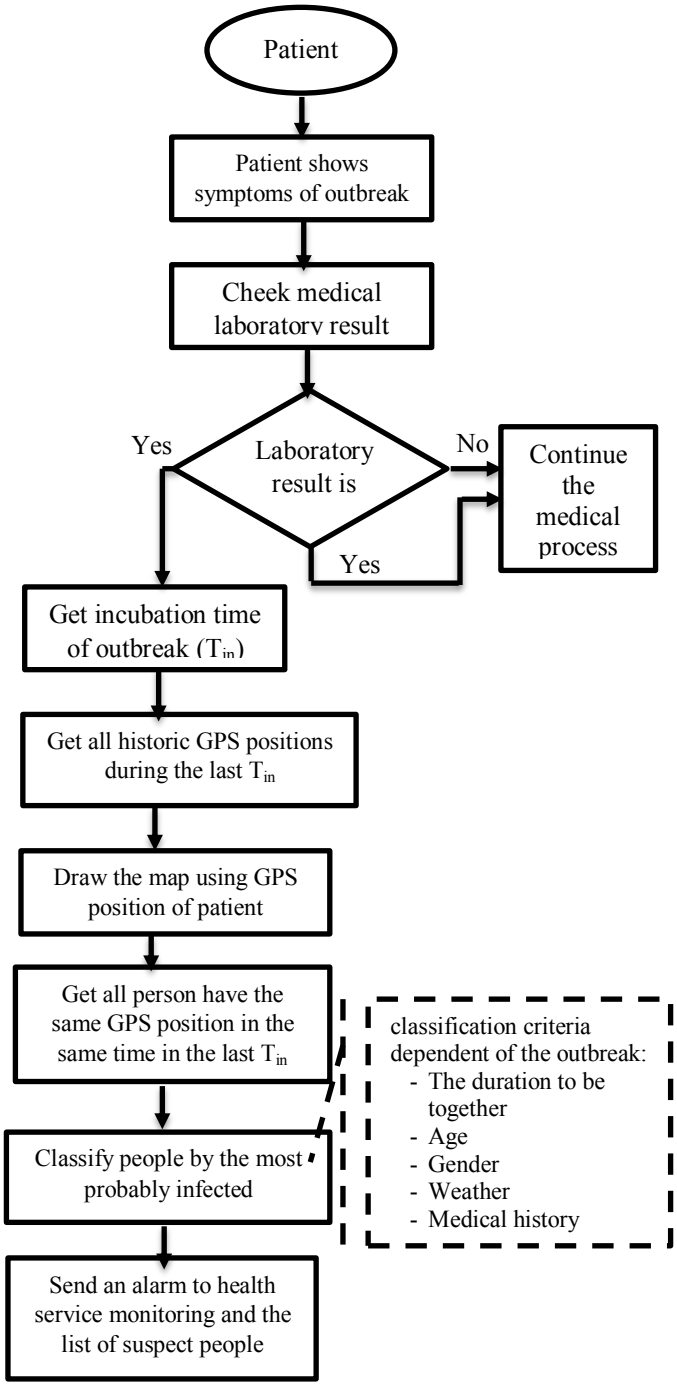

Figure 11. Flowchart to predict contaminated people

IV. REFERENCES

[1]https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200623-covid-19-sitrep-155.pdf?sfvrsn=ca01ebe_2 June 24, 2020

[2] W. Wang, J. Tang, F. Wei, Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China, J. Med. Virol. 92 (4) (2020) 441–447, https://doi.org/10.1002/jmv.25689.

[3] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, et al., Early transmission dynamics in wuhan, China, of novel coronavirus-infected pneumonia, N. Engl. J. Med. (2020), https://doi.org/10.1056/NEJMoa2001316.

[4]Xu, Z., Shi, L., Wang, Y., Zhang, J., Huang, L., Zhang, C., Wang, F.-S. (2020). Pathological findings of COVID-19 associated with acute respiratory distress syndrome. The Lancet Respiratory Medicine February 17, 2020

[5]C.Lai, T.Shih, W.Ko, H.Tang, P.Hsueh 'Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges' International Journal of Antimicrobial Agents February 2020

[6] M.Holshue, C.DeBolt, S.Lindquist, K.H.Lofy, J.Wiesman, H.Bruce, C.Spitters, K.Ericson, S.Wilkerson, A.Tural, et al. First case of 2019 novel coronavirus in the united states. New England Journal of Medicine, 2020.

[7] D.S.Hui, E.I.Azhar, T.A.Madani, F.Ntoumi, R.Kock, O.Dar,G.Ippolito, et al '2019-ncov epidemic threat of novel coronaviruses to global health|the latest 2019' novel coronavirus outbreak in wuhan, china. International Journal of Infectious Diseases, 91:264-266, 2020.

[8] M. Kay, J. Santos, and M. Takane, "mHealth: New horizons for health through mobile technologies," *World Health Organ.*, vol. 64, no. 7, pp. 66–71, 2011.

[9] D. Garets and M. Davis, "Electronic medical records vs. electronic health records: yes, there is a difference," HIMSS Analytics, p. 2–14, 2006

[10] Wang, C. J., Ng, C. Y., & Brook, R. H. ' Response to COVID-19 in Taiwan Big Data Analytics, New Technology, and Proactive Testing' American Medical Association March 2020 doi: 10.1001/jama.2020.3151

[11] Jahanbin K, Rahmanian V. Using twitter and web news mining to predict COVID-19 outbreak. Asian Pac J Trop Med 2020; 13: doi:10.4103/1995-7645.279651.

[12] Rohan Pandey, V.Gautam, K.Bhagat, T.Sethi 'A Machine Learning Application for RaisingWASH Awareness in the Times of Covid-19 Pandemic' arXiv:2003.07074v1 [cs.CY] 16 Mar 2020

[13] T Chen, Y Chen, J Chen, F Chang, " Flu Trend Prediction Based on Massive Data Analysis" 2018 the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis p 304-308

[14] N Ibrahim, N Akhir, F Hafinaz- Hassan, "Predictive Analysis Effectiveness in Determining the Epidemic Disease Infected Area" The 2nd International Conference on Applied Science and Technology 2017 p 020064-1 6

[15] Y Chen, N Crespi, A M Ortiz, L Shu, "Reality Mining: A Prediction Algorithm for Disease Dynamics based on Mobile Big Data" Information Sciences (2016)

[16] R Sandhu, HK. Gill, SK. Sood, "Smart monitoring and controlling of pandemic influenza A (H1N1) using social network analysis and cloud computing" Journal of Computational Science, 2016.

[17] D. Garets and M. Davis, "Electronic medical records vs. electronic health records: yes, there is a difference," HIMSS Analytics, p. 2–14, 2006

[18] M. Chen, Y. Hao, Y. Li, D. Wu, and D. Huang, "Demo: LIVES: Learning through interactive video and emotion-aware system," in *Proc. ACM Mobihoc*, Hangzhou, China, Jun. 22–25, 2015.

[19] A. Lakshman and P. Malik, "Cassandra: a decentralized structured storage system," Operating Systems Review, vol. 44, no. 2,pp. 35–40, 2010.

[20] F. Chang, J. Dean, S. Ghemawat et al., "Bigtable: a distributed storage system for structured data," in Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI '06), pp. 205–218, 2006.

[21] G. DeCandia, D. Hastorun, M. Jampani et al., "Dynamo: amazon's highly available key-value store," in Proceedings of the 21st ACM Symposium on Operating Systems Principles (SOSP '07), pp. 205–220, ACM, October 2007.

[22] E. Hewitt, Cassandra—The Definitive Guide, O'Reilly, 1st edition, 2010.

[23] M. Klems, D. Bermbach, and R. Weinert, "A runtime quality measurement framework for cloud database service systems," in Proceedings of the 8th International Conference on the Quality of Information and Communications Technology (QUATIC '12), pp. 38–46, September 2012.

[24] https://www.who.int/csr/don/14-september-2018-cholera-algeria/en/ March 25,2020