IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Special Issue-3, April 2018 E-ISSN: 2347-2693

A Health Decision Support System for Disease Diagnosis based on Machine Learning via Big Data

S.Subbalakshmi[1*], M.Sumithra[2]

[1*,2] Dept. of Computer Science and Engineering, Meenakshi Sundarajan Engineering College, Anna University, Chennai, India

[*]*Corresponding Author: saisubba1996@gmail.com*

Available online at: www.ijcseonline.org

***Abstract*—** The usual method of health decision support system through regular database provides less efficient prediction. The analysis accuracy is reduced when the quality of medical data is incomplete. It is replaced by a health decision support system which uses big data and a framework called hadoop. The decision support system is used for implementing the healthcare with the help of Hadoop as it contains large amount of data. Hadoop is used to predict the disease based upon the symptoms. The patients are provided with the unique ID. The Patient's Health Record (PHR's) of the patient is stored in the public cloud and is encrypted by homomorphic encryption. When the PHR is needed, they are retrieved from the cloud by decrypting it with the key so, this results in providing the confidentiality to the data. This proposed system provides accurate information and is handy for doctors to diagnose the patients quickly.

***Keywords*—** Disease prediction, Machine learning, big data, Naïve Bayes, Hadoop, Health care, diagnosis

I. INTRODUCTION

A health decision support system predicts the disease, from big data through decision making done by machine learning. Big data is the huge amount of data which is collected from the globally available database. When regular database is used for the prediction, it becomes quite challenging, because when size of the data increases, the computational and the processing time increases. This is the major disadvantage of the existing system [1]. Due to this disadvantage, the diseases are not predicted on time by the doctors and the patients' health condition becomes even poorer. In order to overcome the challenge of processing time, this system is proposed to ensure that the diseases are predicted quickly on time by the doctors and this system is proved to be handier.

The existing system [1] uses relational database for storing the set of symptoms. It predicted rare diseases using the list of symptoms in the normal database. Disease Prediction was also carried on through systems which included wearable devices like AMON (A wearable multipara meter medical Monitoring and alert system)[3]and LOBIN(E-Textile and Wireless-Sensor-Network Based Platform for Healthcare Monitoring in Future Hospital Environments) [5] which is a combination of both electronic textiles and WSN.AMON[3] is a wearable belt like device which was worn on wrist of the patient and the disease or the panic of the affected were predicted or sensed by using sensors and the information was sent to the MMC(On Line Medical Mission Control) through GSM link. In LOBIN [5] the system included wearable fabrics and the location subsystem which were connected to Wireless Sensor Network (WSN) and WSN contains a gateway which is linked to the management subsystem connected through the IP network. In system [5] diseases are predicted by the sensors which are attached along with the fabrics and sends the information to the management subsystem. Although the systems [3] and [5] are wearable and portable they face many challenges. Even if any one of the sensor or the component is damaged, the systems [3] and [5] may not work. With the disadvantages faced by the systems [1],[3] and [5] this new system is proposed by using big data and a framework called hadoop for efficient health monitoring.

Section II contain the related work of the proposed system, Section III contain the some measures of this system, Section IV contain the architecture and essential steps of this proposed health monitoring system, section V explain the proposed methodology with flowchart, and section VIII concludes research work with future directions.

II. RELATED WORKS

Some of the related works for the health monitoring of the patients included: a system which analyzes the patients by using the physiological data [2], a system which included an alert mechanism [4] and a system which was monitoring patients continuously with personal care and proved to be energy-efficient [6].The health monitoring system [2] is based on anomaly detection and data mining. The system [2] included a flexible framework which had three phases. The first phase comprises of historical data that has both common

and uncommon data which is preprocessed and fed into the model building. The model building is the sub-block within the first phase of the framework. The model building has two components: Feature extraction and Risk component assessment. After model is built using feature extraction and risk component assessment the final data is used for final model which includes discrete risk levels. The second phase of the framework has three sub-phases: preprocessing, real-time classification and global risk. In this second phase, the real time classification has three components such as feature extraction, risk component assessment and risk evaluation. The third phase included real-time sensor data which included measured value for rate of heart beat, percentage of mercury content in millimeters and percentage of spO2.Here [2] both the historical data and the sensor data are used for monitoring the health conditions of the patients and finally the results are given for the sensor monitored diseases graphically using the risk levels. The framework diagram [2] is shown above: It tells about the working of the health monitoring system which is proposed in [2]. Each of the blocks and the sub blocks are clearly seen in diagram [2](Fig 1)

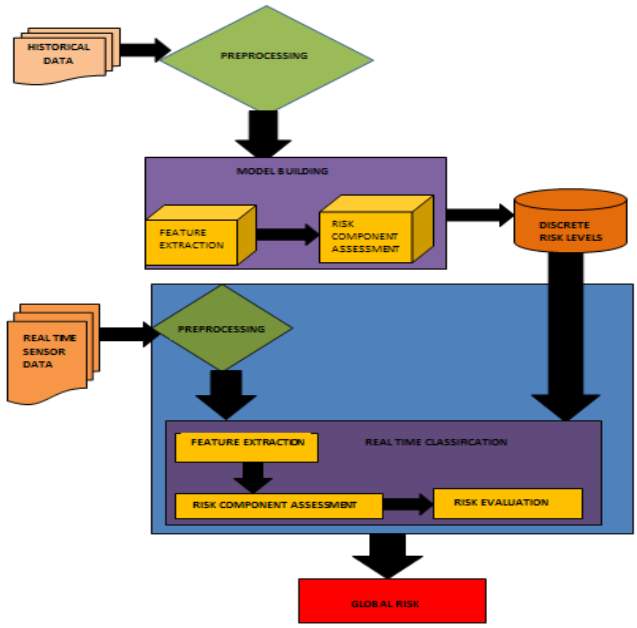

Figure. 1 Framework Architecture

In system [4] there are two scenarios for alert mechanism for monitoring the health services: Alert message transmission for not uploading physiological parameters on schedule and automatic notification of abnormal conditions. These two alert mechanisms were used to predict the people's conditions having the frontend as a mobile phone and the back end as a database. In first scenario, the patient does not upload any information such as blood pressure or ECG data. The alert message is automatically sent to the patient as an emergency alert. The patient neither receives the alert nor replies to that alert, due to some condition. Then the health care center takes into account that the patient may be under some risk and goes to the location of that patient and provides some necessary services. In the second scenario, the Bluetooth in the mobile phone of the patient measures the parameters such as blood pressure which causes dizziness and headache to the patient. If the values measured are abnormal, then an alert message is sent to the local officials and they perform the treatment either professionally or by sending some ambulance. The diagrammatic representation [4] of both scenarios of is shown below Fig 2(a) and Fig 2(b):

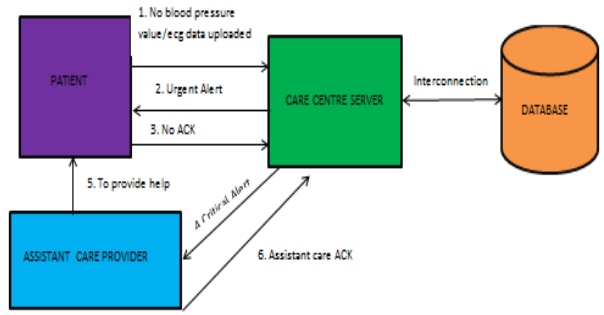

Figure. 2(a) Alert message transmission diagram for not uploading physiological parameters on schedule

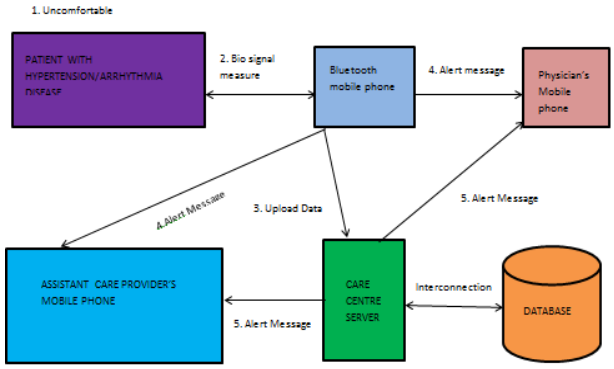

Figure 2(b) Alert message transmission diagram for automatic notification of abnormal conditions

In the energy-efficient system [6] ,the energy and the storage requirements of a health monitoring system such as heart rate, blood pressure, oxygen saturation, body temperature, blood glucose, accelerometer, ECG (electrocardiogram) and EEG (electroencephalogram) are quantified as the baseline of WBAN (Wireless Body Area Networks).The above mentioned eight parameters are used for the health monitoring in the system [6]. An accelerometer is placed in the wrist of the human body and a mobile phone is connected to the accelerometer through the Bluetooth. The measurements are sent to the hospital and the storage servers from the mobile phones via Bluetooth through the accelerometer. If the values are abnormal, then the necessary treatment is given to that patient by the doctors. Here even the energy consumption of the sensors are studied by the equation[6]:

$$E_{total} = E_s + E_t + E_c$$

Here,

E_{total} is the total energy consumption of the sensors.

E_s is the sampling energy

E_t is the transmission energy

E_c is the on-sensor computation

E_s was derived in [6] as shown in the given equations [6]:

$$E_s = E_{ADC} * S$$

$$E_{ADC} < 4^{(ENOB+1)-9} pJ < 4^{(N-9)} pJ$$

$$S = f_s \left(\frac{1}{s}\right) * 60 \left(\frac{s}{min}\right) * 60 \left(\frac{min}{hr}\right) * 24 \left(\frac{hr}{day}\right)$$

Transmission energy [E_t] are derived by the equations[6] as:

$$E_t = \left(T_{send} * P_{send} + T_{standby} * P_{standby}\right) * C$$

$$C = f_t \left(\frac{1}{s}\right) * 60 \left(\frac{s}{min}\right) * 60 \left(\frac{min}{hr}\right) * 24 \left(\frac{hr}{day}\right)$$

T_{send} is a fixed value and measured as 2.6 milliseconds for a single packet transmission.

$T_{standby}$ depends on the transmission frequency (f_t):

$$T_{standby} = \frac{1}{f_t} - T_{send}$$

These related works [2],[4] and [6] were proved to be more efficient and provided mobile services. But since they used sensors for implementing, they may face challenges if the sensors are damaged. In order to overcome that challenge, a non-sensor health monitoring system using hadoop framework and big data is proposed in this paper.

III. METHODOLOGY

The admin preprocesses the set of data (symptoms) and then clusters them under a particular disease. Finally uploads the data into the system. When the patients approach the doctor, the list of symptoms are input into the health monitoring system and the disease is predicted from the set of symptoms gathered under a particular disease more quickly. After the disease is predicted, the patient health record (PHR) is generated and it is encrypted and stored into the cloud, so that the information is confidential. The architecture diagram of the proposed system is shown in the fig 3. The system has six modules for implementation:

- Analyzing big data
- Introducing hadoop framework
- Data preprocessing
- Data clustering
- Disease prediction
- Encryption of data

A. *Analyzing big data*

The large data set is taken from various sources. The sources are from the globally available databases. There is an actor called admin in this health monitoring system. The admin only analyzes the big data and further proceeds with data preprocessing and data clustering and finally uploads the data into the health monitoring system. A sample data set [7] is shown below in table 1:

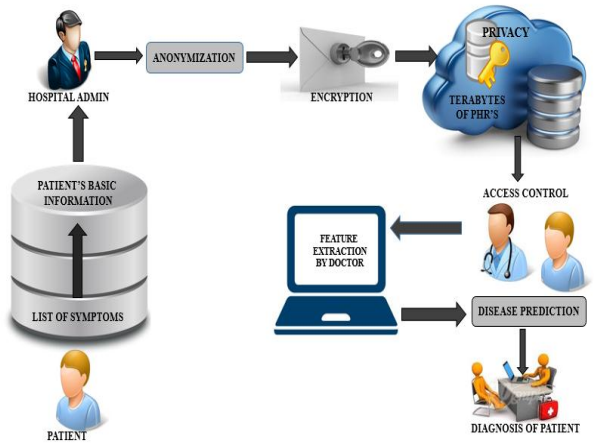

Figure 3. System Architecture

B. *Introducing Hadoop Framework*

Hadoop is a distributed framework, which is handled by two modules: HDFS (Hadoop Distributed File System) and Map Reduce. Hadoop Distributed File System will split the files into partitions and will store those split files distributed in the HBase(Hadoop Database).Here Hadoop is used since big data is processed quickly in short span of time. Map Reduce is a module in the hadoop framework architecture which reads each and every split files in the HDFS.The architecture of hadoop framework [8] is shown in the figure 4.

Table 1 Sample Dataset

Symptom name	Disease name	Co-occurs	Tfidf_score	Disease_id	Symptom_id	Doid_code	Doid_name
Aging, Premature	Acquired Immunode	3	10.39365447	D000163	D019588	DOID:635	Acquired Immunode
Aging, Premature	Breast Neoplasms	1	3.46455149	D001943	D019588	DOID:1612	Breast Cancer
Aging, Premature	Colonic Neoplasms	1	3.46455149	D003110	D019588	DOID:219	Colon Cancer
Aging, Premature	Skin Neoplasms	3	10.39365447	D012878	D019588	DOID:4159	Skin Cancer

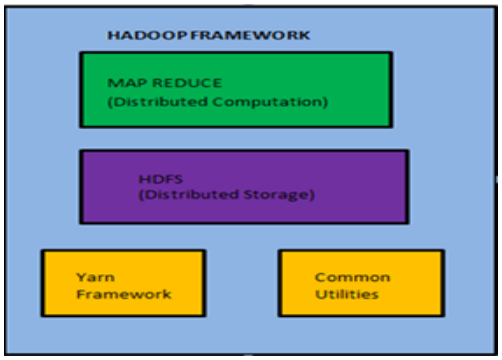

Figure 4. Hadoop Framework

C. Data Preprocessing

Data preprocessing is a method where the unwanted data from the data set is removed. The raw data is preprocessed by the admin and they are sent for the clustering. The data preprocessing helps in making the data set more efficient, since only the necessary information are found in the data set after the preprocessing of data. The mechanism of data preprocessing is shown in the fig 5.

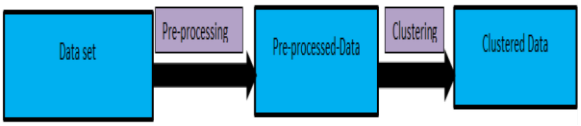

Figure 5. Data Preprocessing

D. Data Clustering

After the data has been pre-processed, they are clustered under a particular disease. For the clustering mechanism, Naïve Bayes algorithm has been used. The Naïve Bayes equation for a set of symptoms B_i belonging to a particular class or disease A is given by the equation [9] as:

$$P(A|B_1 \dots B_n) = P(A)\prod_i^n \frac{P(A|B_i)}{P(A)}$$

The set of symptoms are clustered under a particular disease as shown in the above equation. The derivation [10] is given for clustering and classification using naïve Bayes as follows: The conditional probability is given as:

$$P(C_K|x) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

Where,

$X=(x_1,\dots,x_n)$ representing some n features (independent variables),it assigns to this instance probabilities for each of K possible outcomes or classes C_k.

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

The conditional probability of Naïve Bayes is given as,

$$posterior = \frac{prior * likehood}{evidence}$$

The numerator is equivalent to the joint probability model which is represented by using the conditional probability as using the chain rule as

$$p(C_k, x_1, \dots, x_n) = p(x_1, \dots x_n, C_k)$$
$$= p(x_1|x_2, \dots, x_n, C_k)p(x_2, \dots, x_n, C_k)$$
$$= p(x_1|x_2, \dots x_n, C_k)p(x_2|x_3, \dots x_n, C_k)p(x_3, \dots, x_n, C_k)$$
$$= \dots\dots p(x_i|x_{i+1}, \dots C_k) = p(x_i|C_k)$$

Thus the joint model can be represented as,

$$p(x_i|x_{i+1}, \dots, x_n) \propto p(C_k, x_1, \dots, x_n)$$
$$\propto p(C_k)p(x_1|C_k)p(x_2|C_k)$$
$$\propto p(C_k)\prod_{i=1}^{n} p(z_i|C_k)$$

This means that under the above independence assumptions, the conditional distribution over the class variable C is:

$$p(C_k|x_1, \dots . x_n) = \frac{1}{z}p(C_k)\prod_{i=1}^{n} p(z_i|C_k)$$

Where the evidence is,
$Z = p(x) = \sum_k p(C_k)p(x|C_k)$ is a scaling factor depending on only $x_1,\dots,x_n$, that is a constant if the values of the variables are known.

E. Disease Prediction

After the data is preprocessed, clustered and classified, the diseases are predicted by the list of symptoms given by the patient to the doctor and finally the patient health record is generated.

F. Encryption

The PHR which is generated is encrypted using the homomorphic encryption.One of the main advantage of the homomorphic encryption is that the data can be encrypted without the secret key. The data is decrypted by the personal ID which is generated for a particular patient. Fig 6 shows how the data is encrypted:

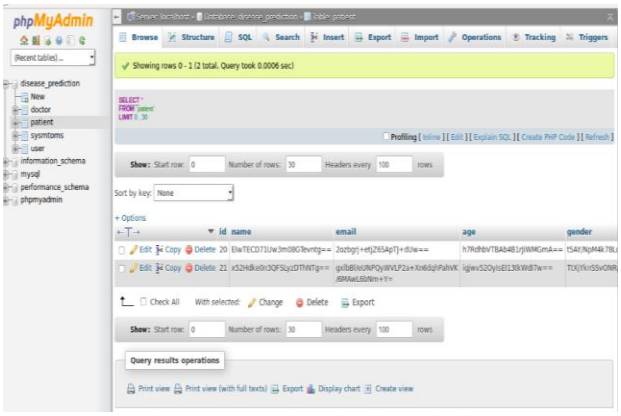

Figure 6. Encryption of patient's details

IV. RESULTS AND DISCUSSION

The implemented system generates the following output for monitoring the health of the patients. The application is in the form of a web page which is opened as shown in fig7.

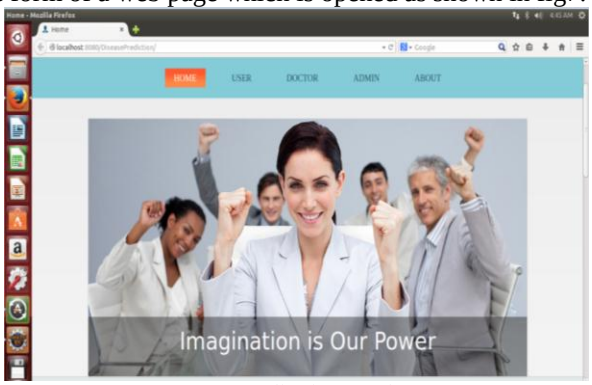

Figure 7. Application's Webpage

In this health monitoring system, an user can also register and use for monitoring his/her own health conditions by inputting their symptoms as shown in fig 8:

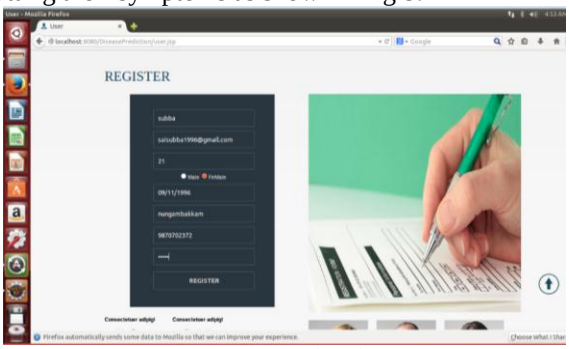

Figure 8. User's Registration Page

After registering the user can login directly into their account as shown in fig 9.

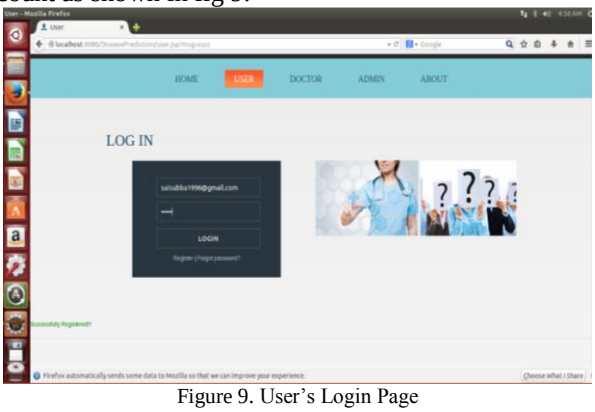

Figure 9. User's Login Page

Finally the user can search for the disease by inputting the symptoms as shown in fig 10.

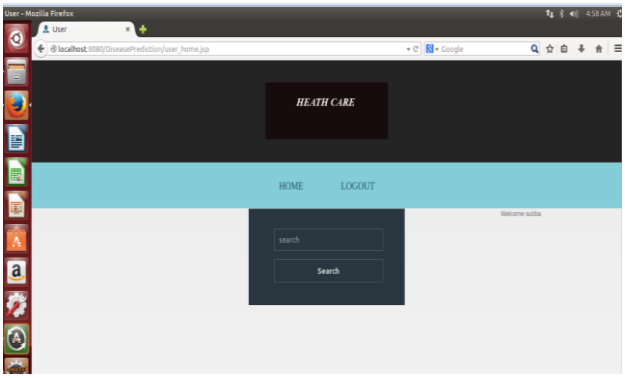

Figure 10. User's Searching Page

In this health monitoring system, the admin will add the doctors, through registration by login as shown in fig 11.

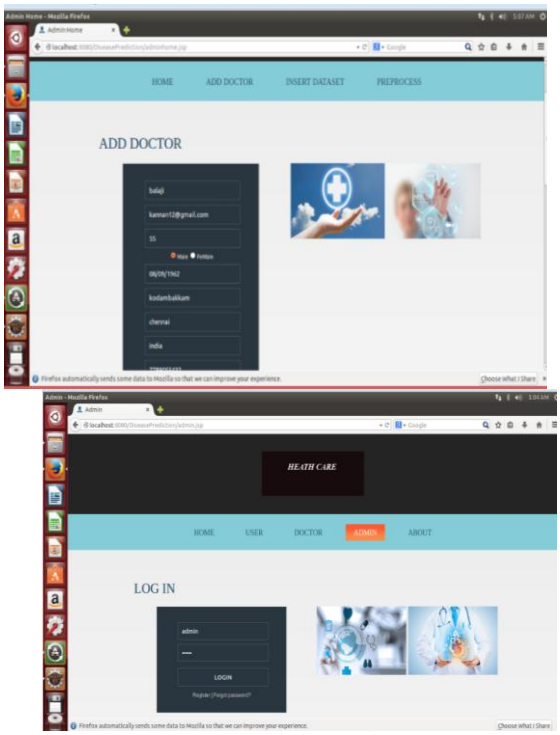

Figure 11. Admin's Login Page

Admin adds doctor as shown in fig 12.

Figure 12. Admin Adding Doctor

Admin inserts data as shown in fig 13.

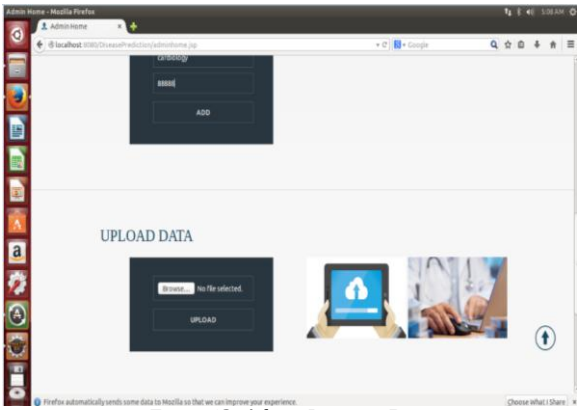

Figure 13. Admin Inserting Data

Now when the admin uploads the dataset, they are stored in the database as shown in fig 14.

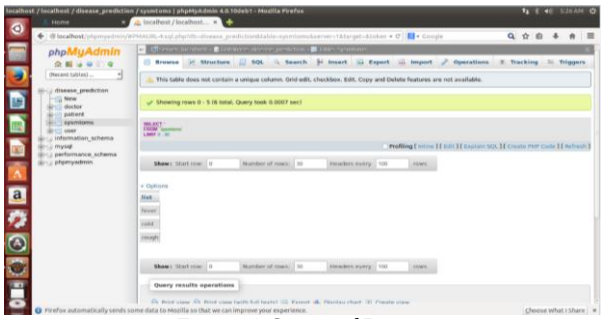

Figure 14. Storage of Dataset

The doctors who are added logins to the health monitoring system. When the patients visit the hospital, the doctor enters the patient details and their symptoms and diseases as shown in fig 15.

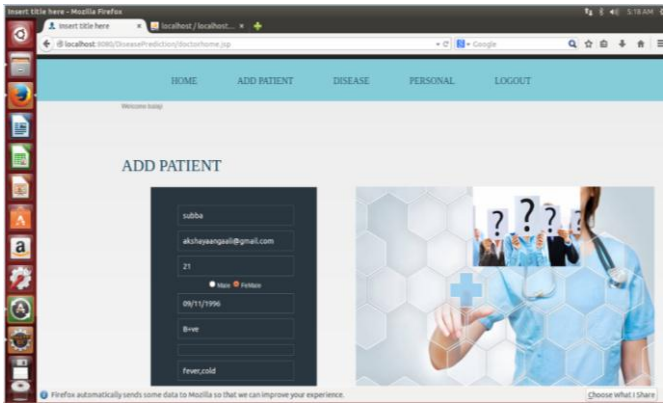

Figure. 15 Doctor Adding Patient

A personal ID is generated for the patient and using the personal ID, the doctor can directly retrieve the PHR of the patient.

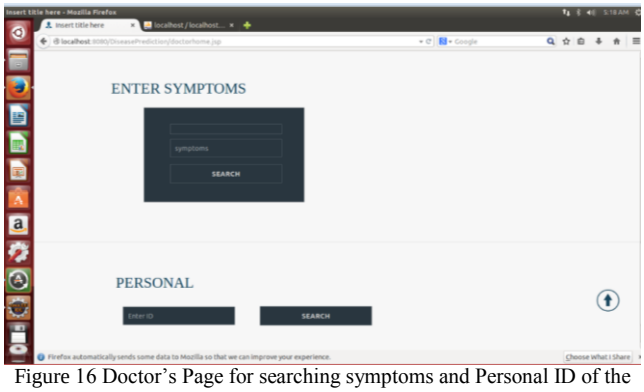

Figure 16 Doctor's Page for searching symptoms and Personal ID of the patients

The PHR finally generated is encrypted and stored in the cloud and can be retrieved later using the personal ID as shown in fig 16, which is generated. This is the generated result using the health monitoring system and it predicts the disease more quickly within a short span of time, since hadoop framework is used for processing the big data. It overcomes the challenges faced by[1] which used SVM(Support Vector Machine) and relational database for storing the details.

V. CONCLUSION AND FUTURE SCOPE

There are many technologies arising worldwide for monitoring health and disease prediction. This implemented health decision support system using hadoop framework and big data is efficient in finding the results quickly, but still it is facing an issue that it can process big data only up to size of peta bytes of data. If the size exceeds, the system may face challenges. So let the system which is further implemented in future may have storage limitation of big data exceeding petabytes.

REFERENCES

[1] Marc Pi˜nol, Rui Alves, Ivan Teixid´o, Jordi Mateo, Francesc Solsona, Ester Vilapriny´ o." Rare Disease Discovery: an optimized disease ranking system" IEEE Transactions on Industrial Informatics vol no1551-3203 (c) 2016 IEEE.

[2] Daniele Apiletti, Elena Baralis, Giulia Bruno, and Tania Cerquitelli." REAL-TIME ANALYSIS OF PHYSIOLOGICAL DATA TO SUPPORT MEDICAL APPLICATIONS", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 13, NO. 3, MAY 2009

[3] Urs An liker, Jamie A. Ward, "AMON: A WEARABLE MULTIPARAMETERMEDICAL MONITORING AND ALERT SYSTEM". IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 8, NO. 4, DECEMBER 2004

[4] Ren-Guey Lee, Kuei-Chien Chen, Chun-Chieh Hsiao, and Chwan-Lu Tseng, "A MOBILE CARE SYSTEM WITH ALERT MECHANISM". IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 11, NO. 5, SEPTEMBER 2007

[5] Gregorio Lopez, Vıctor Custodio, and Jose Ignacio Moreno, "LOBIN: E-TEXTILE AND WIRELESS-SENSOR-NETWORK-BASED PLATFORM FOR HEALTHCARE MONITORINGIN

FUTURE HOSPITAL ENVIRONMENTS", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 14, NO. 6, NOVEMBER 2010

[6] Arsalan Mohsen Nia, Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K. Jha, "ENERGY-EFFICIENT LONG-TERM CONTINUOUS PERSONALHEALTH MONITORING", IEEE Transactions on Multi−Scale Computing Systems, 2332−7766 (c) 2015.

[7] https://github.com/dhimmel/hsdn/blob/gh-pages/data/symptoms-DO.tsv

[8] https://www.tutorialspoint.com/hadoop/hadoop_introduction.htm

[9] Yunjing An, Shutao Sun, Shujuan Wang," Naive Bayes Classifiers for Music Emotion Classification Based on Lyrics" IEEE ICIS 2017, May 24-26, 2017, Wuhan, China

[10] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

AUTHORS PROFILE

Ms S.Subbalakshmi is a student of Meenakshi Sundarajan Engineering College, who is currently pursuing her degree in Bachelors of Engineering in the stream of Computer Science and Engineering.

Ms. M.Sumithra is an assistant professor in Meenakshi Sundarajan Engineering College in the branch of Computer Science and Engineering.

