e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

# HEALTH STATUS PREDICTION USING BIG DATA ANALYTICS

**Pooja J P[*1], Yamini V[*2], Amrutha R[*3], Sadiq Vasim[*4], Radhika T V[*5]**

[*1,2,3,4]Student, Department of Information Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India.

[*5]Assistant Professor, Department of Information Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India.

## ABSTRACT

Big data has changed the way we manage, analyze, and leverage data across various industries. One of the most notable areas where data analytics is making big changes is healthcare. Health care is the prime most industry affecting the lives of people especially during the hard times of these pandemic. In fact, the healthcare analytics has a very great potential to reduce the costs of treatments, detect and avoid preventable diseases, predicting outbreaks of epidemics and hence improve the quality of life in general. The average human lifespan was greatly increasing the world, which is continuously posing new challenges to today's method of treatment. Health professionals, like business entrepreneurs, are capable of collecting massively huge amounts of data and are always on a look for the best strategies to use these numbers. A big data analytic platform which is cloud-enabled is one of the best ways to analyze the structured and unstructured data that are generated from healthcare management systems. Future Health Condition Prediction (FHCP) algorithm is used to predict the future health condition of the most correlated patients based on their current health status.

**Keywords**: Health, Prediction, Symptoms, Big Data Analytics, Machine Learning.

## I. INTRODUCTION

Information has been the key, from a very long time, for a better organization and for new developments along since life existed. The more the data that is being generated, handled, manipulated and the information that is retrieved and being used, we have, the more optimally we would be organizing ourselves to deliver the best outcomes using the most optimal methods and in the fastest, more accurate, reliable and efficient method of retrieval. Hence data collection is always the most important part of processing and working with data, for any existing organization. We can use this data that is collected, to process the data in order to obtain useful information, which can be used for the prediction of current trends in data and future events or technologies that can play a major role with data. As humans are becoming more aware of this fact, we have begun producing various kinds of data, in diverse categories, each and every day and collecting such data about everything, by introducing varied forms of technological developments to aid in this direction. Today, we are in a situation wherein we are bombarded with tons of data from almost all the directions of our life such as science, social activities, histories, discoveries, inventions, rediscoveries, work, health, etc. Hence, we can compare the present situation to the scenario of a data deluge. The technological advances have helped us to generate huge quantities of data, which are both useful and not useful, at times, that is exponentially growing day by day, where it has reached a stage that data has become unmanageable with currently available and being used technologies.

Big data has massively changed the way we observe, manage, analyze, and leverage data across various spanning industries across the world. One of the most notable areas where data analytics is making big changes is in the field of healthcare. In fact, the field of healthcare analytics has the potential to reduce the immense costs of treatment to very less price, to predict the outbreaks of random fast spreading epidemic; it avoids preventable diseases, and also improves the quality of all the lives living on this Earth, in general. The average human lifespan is increasing across the world population, which is posing several new challenges to today's health treatment delivery methods. Health professionals are similar to business entrepreneurs, in the perspective of the capability of collecting massive amounts of data and look for the best known strategies to use these data that are collected in such huge numbers. A cloud-enabled big data analytics platform is one of the best ways to analyze both, the structured and the unstructured data, that are generated from the various healthcare management systems. Future Health Condition Prediction (FHCP) algorithm is an algorithm that is used to predict the future health condition of the most correlated patients based on their current health status using machine learning models. Along with the usage of Decision Trees, Naïve Bayes and Random Forest.

The problem statement is to detect the possible illness with precision in humans based on the symptoms of illness using data analytics and algorithms. To predict suitable prescription of medicines based on the input of symptoms (future enhancement).

## II. SYSTEM ARCHITECTURE

System architecture is the conceptual model that defines the structure, behavior, and more views of a system. Patient's data plays the most crucial role in the system. There are various ways to collect such data. These data are used to train and test the model, which is used to predict the flu disease. The data collected refers to the various symptoms that are associated with disease flu. Electronic Health Record, Public Health Record and Clinical Data are collected through various methods and are stored in databases and then into the data warehouses. These collected and stored data undergoes different data preprocessing stages. The big data analytics section that provides insights are the descriptive, the diagnostic and the predictive section. The prescriptive section provides the foresight of the input collected data. This process results in improved outcomes. After employing all these pre processing techniques on data, the data would be easier, faster and more integrated to manipulate, edit, process, add, change as well as to delete data.

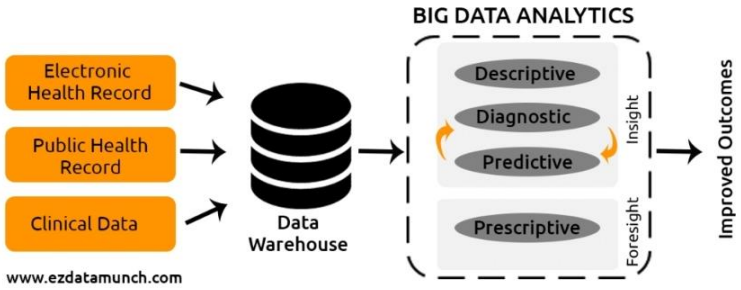


**Figure 1:** System Architecture Of Data Processing and Output

### A) Electronic Health Record

EHR is the digital version of a paper chart of patients. EHRs are real time data which is patient centered records that make the information available instantly, would be secured and accessed only by authorized users. With the help of EHRs, an organization can help build a future which is healthier for our nation.

### B) Public Health Record

A PHR includes health information managed by the individual. A PHR can be defined as an electronic application using which individuals can access, manage, edit as well as share their health information. Here data of others (generally patients) can be accessed by others who are authorized, in a private, secure way through a confidential environment.

### C) Clinical Data

Clinical data will have information ranging from determinants of health and measures of health and health status to documentation of care delivery. These captured data serves a variety of purposes and would be stored in numerous databases across the various healthcare systems. Clinical data management is a critical process which leads to generation of high – quality, reliable and statistically sound data from clinical trials.

### D) Data Warehouse

Data warehouse has the repository of all the data related to the patients including patients' personal data along with the input of symptoms faced by the patient. The data required for the model is obtained through the API call. There are websites which provide the data of parameters of illness for a particular disease. These information are helpful in obtaining the actual real-time data in the required format in a very quick and easy manner. The data which is used to train the model is obtained from the personal data collected.

### E) Big Data Analytics

For a given type of machine learning model, the role of data is very significant. The machine learning model is always dependent on data and because the model used is trained upon this training data, data needs to be consistent and very precise. The data obtained for training data set using the API call would be processed before utilizing it to train the model, hence the current step would be called as pre-processing. Any missing value in the data set is handled with the help of an imputation method. Here the pre-processed data would be

used for diagnostic, description, predictive insights along with the prescriptive foresights, where these results aids in improved outcomes of the model.

**F) Machine learning models**

Machine learning is a superset of Deep Learning; it provides a wider perspective about the different types of machine learning models that can be used for different data and the use cases. At first, the dataset is imported. Data quality check is done, to check if null values are present, and these rows with null values are dropped, if found. Now features column is created. Now modeling of data is executed using the following steps: 1) The dataset is split using "train-test-split" function. 2) Then classification is applied on dataset 3) The model is now trained and tested 4) Now the accuracy score is found 5) Saving and using the ML model or Loading the model to predict the output for new test data which are patient's symptoms.

**G) Graphical User Interface (GUI)**

All the old school and modern technical aspects are in a complex manner, to be understood by a common person. People are only interested in knowing the illness that is predicted or the current health conditions that is faced by the person. Hence a simple, easy and understandable user interface is designed so that any layman can use it to know the future weather conditions. A simple visualization tool is also presented in the user interface, where the data of the symptoms and other user details can be visualized using the graphs. A user can also view the graphical representation of the probabilities of the disease that is being developed in their body. User just needs to input the various symptoms that are associated with the disease (here, flu). Hence, the GUI is user friendly, easy, simple in nature.

## III. DATA FLOW DIAGRAM

A data flow diagram shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and the various sub processes the data moves through. DFDs are built using standardized symbols and notation to describe various entities and their relationships. The first stage is for collection of Patient's data. The second phase is preprocessing of the collected data. The third phase is extracting and modeling the feature dataset. The fourth phase is training the model. The fifth phase is to take data as user input. The sixth phase is that of the prediction model, where the data set is given as input to the model and hence is used to test the model as well. The seventh and the last phase is to output the predictions obtained from the previous stage, to the user or the consumer, in the current phase (which is the last phase).

The data flow in a health condition prediction system is represented in the above diagram starting from the collection of raw patient's symptom data to the illness prediction as output. The illness data is collected from the website and various studies. The collected data is stored in CSV format. Followed by cleaning of the data, as well as analyzing the same, by applying Exploratory Data Analysis. Now the obtained resultant data is a preprocessed data that posses only clean values. Select those features or attributes which are required for the model. Train the model with the processed data with required features. Now different sets of data are used to train the different machine learning models. At last, the model that gives the maximum accuracy amongst all the other models would be selected. After training the model, the model would be tested with the user input data.

Here the input data is collected from API call to the model. From this API call, the required features would be obtained and would be given as input to the prediction model. The prediction model based on these inputs predicts the illness as output. The output data will be the prediction of whether the person has flu or not by the usage of prediction model.

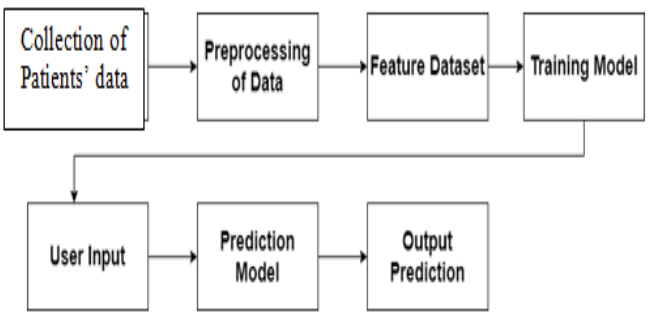


**Figure 2**: Data Flow Diagram for the system process

## IV. IMPLEMENTATION

HMTL is the standard mark-up language. HTML is used for creating web pages and web applications. With CSS and JavaScript, it creates a full working front end of any website. It was developed in 1980 by Tim Berner Lee at cern. Web browser use html to interpret and compose text, image, and other material into visual and audio web page.

PHP is the server side scripting language which is found embedded in HTML. PHP is used to manage databases, dynamic content, session tracking.PHP can also be used to build a complete functional e-commerce sites. It is integrated with a number of popular databases, including MySQL, PostgreSQL, Oracle, Sybase, Informix, and Microsoft SQL Server.

Apache Spark is one of the popular data processing framework. Apche Spark can quickly perform processing tasks on huge data sets, and can also distribute data processing tasks across multiple computers, either on its own or with other types of distributed computing tools.

Apache Hive is an open source data warehouse software. It is used for writing, reading and managing large data set file. Where these files are stored directly in one of the two systems, ie. either the Apache Hadoop Distributed File System (HDFS) or any other data storage systems such as Apache HBase.

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two main entities, they are decision nodes and leaves.

Naïve Bayes algorithm is a supervised learning algorithm.It is based on Bayes theorem and is utilized for solving any type of classification problems. Naïve Bayes Classifier is one of the most simple and also the most effective Classification algorithms. This helps in building fast machine learning models which has the ability to make quick predictions.It includes the usage of high-dimensional training dataset.

Random Forest is a machine learning algorithm which belongs to supervised learning technique. It can be used for Classification as well as Regression problems. It has its concept based on ensemble learning, that is a process of combining multiple classifiers to solve any complex problem inorder to improve the performance of the model.

**Technical Aspect:**

The technical aspect of implementation of the project is realized using following Symptoms

THE PROJECT IS IMPLEMENTED BY USING APACHE TOMCAT SERVER

Apache Tomcat Server:

Tomcat server implements jsp and provide a pure java HTTP web server environment in which java code can run. Tomcat5.0 can support servlet 2.4 and jsp 2.0. With tomcat server project can run in more stable manner and it also offers extra level of security. As it is an open source, source code for server is readily available. Tomcat is connected to MySQL through JDBC connection. JDBC provides an abstraction layer between MySQL and tomcat, because of which project code does not need to be altered in order to communicate with multiple database.

## V. ADVANTAGES OF PROPOSED SYSTEM

There are many advantages of the system proposed. The following are the advantages of the proposed system, which are mentioned below:

1. The Electronic Health Records (EHR) helps the health care providers to conduct in-depth patient care analysis and understand the patient's illness..
2. The system enables real time monitoring and data-sharing, so necessary steps can be taken to treat patients.
3. The algorithm prevents human errors. So Big Data can be leveraged to analyze the users' data.
4. Usage of the system prevents the need of unnecessary ER visits to Hospitals.
5. Big data helps to personalize care and improve patient efficiency.
6. Big data helps promote hospitals by improving patient satisfaction and care efficiency.
7. The efforts of physician relationship management are done through tracking various physicians' preferences and their performances.
8. The system provides strong data security due to implementation using machine learning and cloud.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

## VI. RESULTS

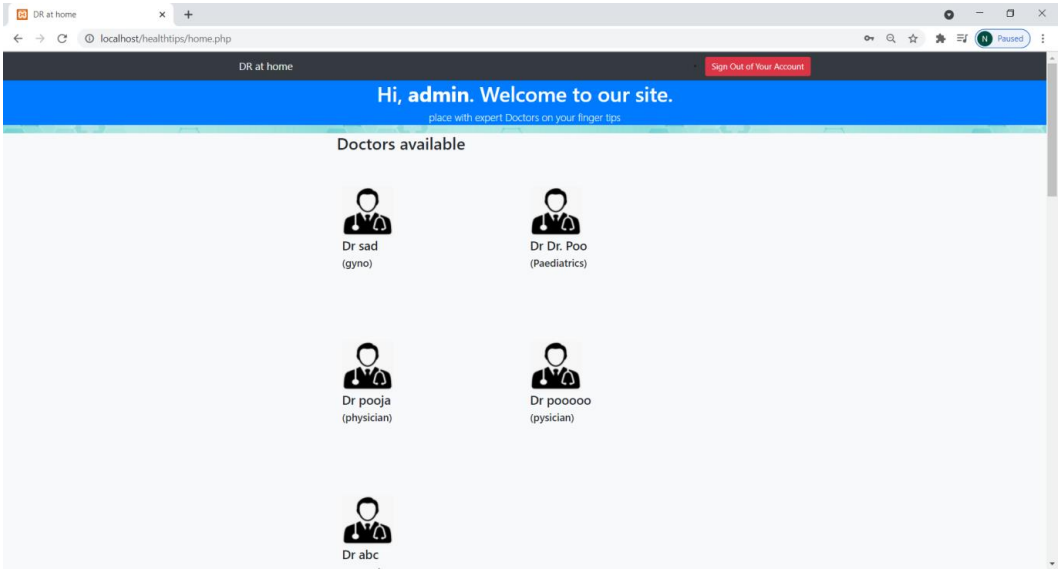


**Figure 3:** Web page with registered doctor's portal

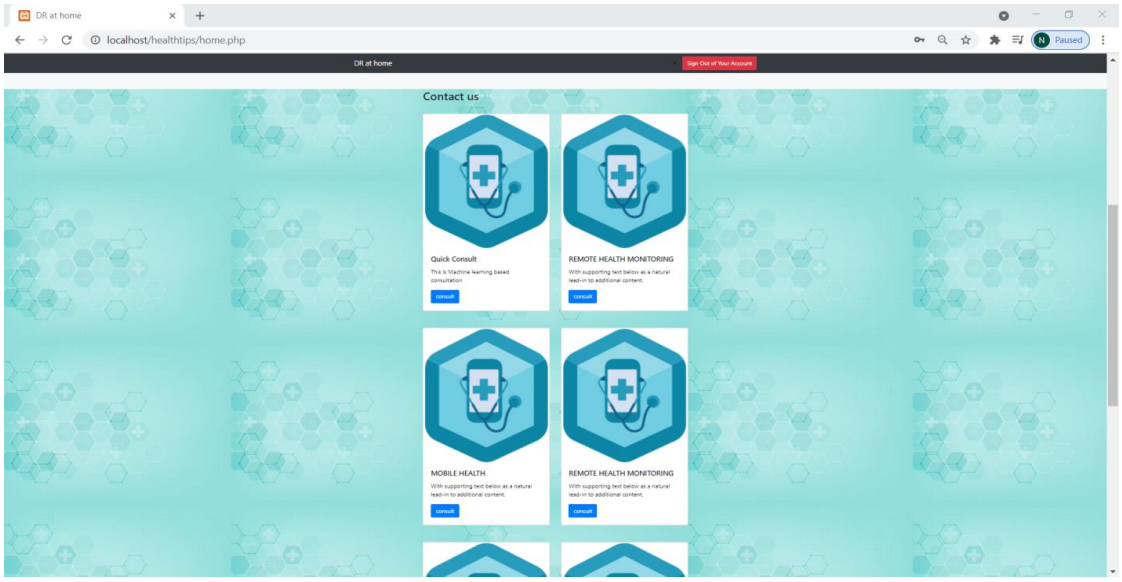


**Figure 4:** Web page with various telemedicine features

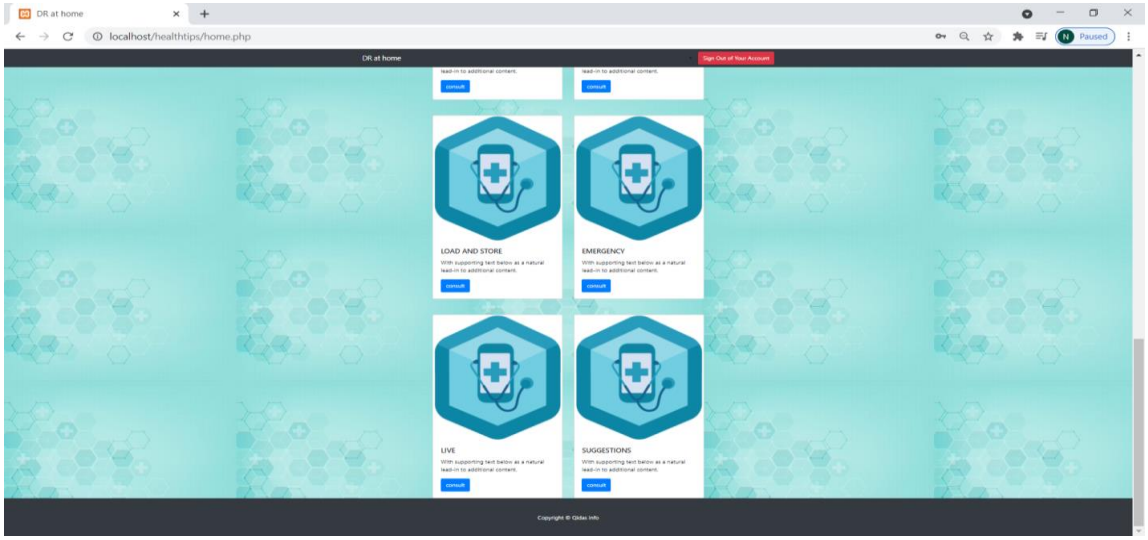


**Figure 5:** Continuation of Web page with various telemedicine features

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

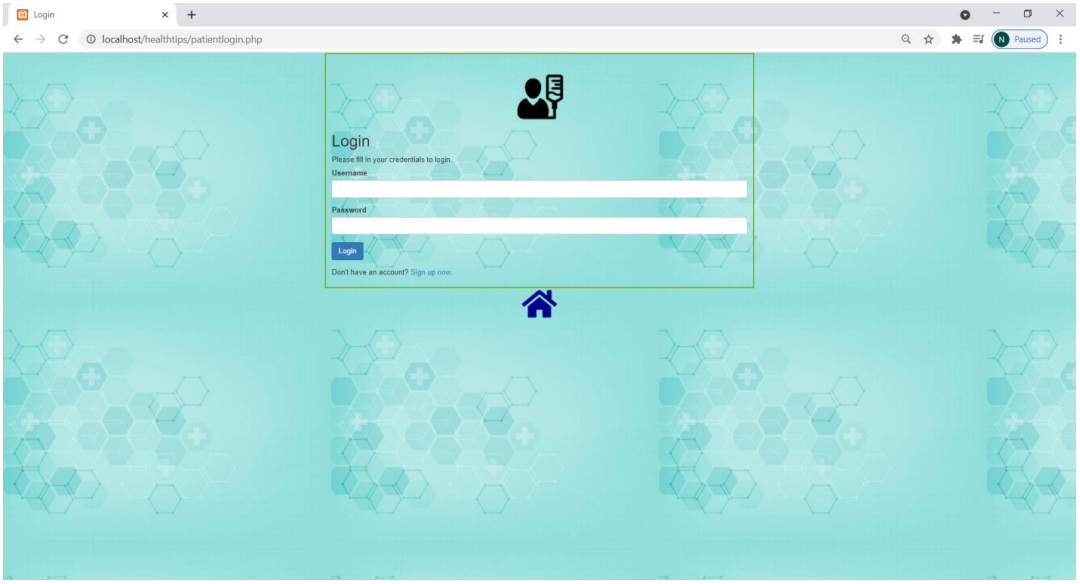


**Figure 6:** Login Portal for Patients

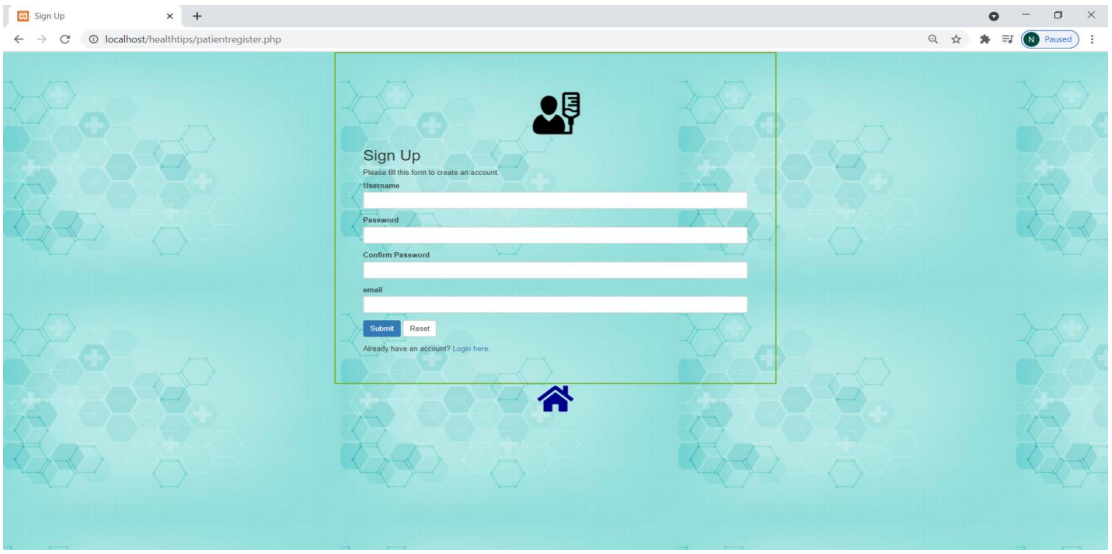


**Figure 7:** Sign up portal for patients

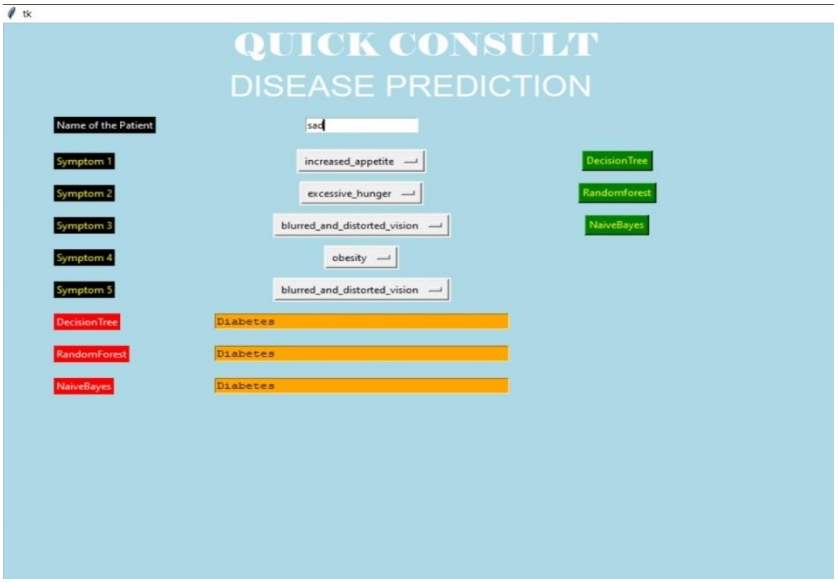


**Figure 8:** Prediction module

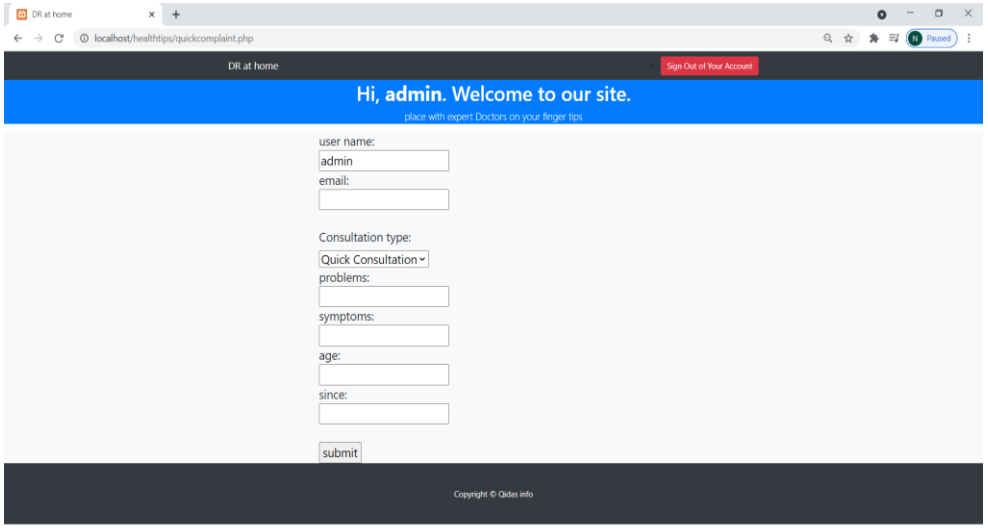


**Figure 9**: Web page for Patient details

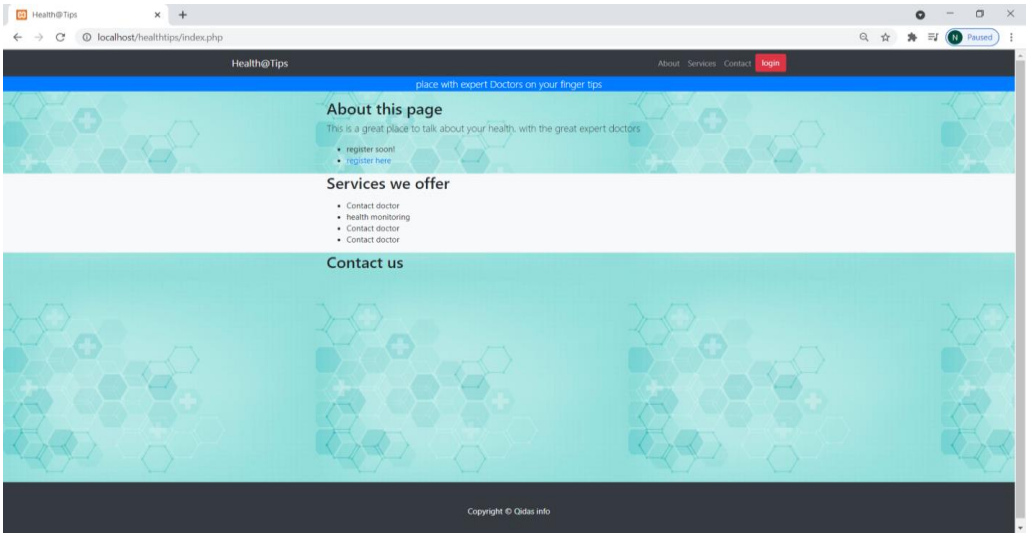


**Figure 10:** Web page describing about the project and about us

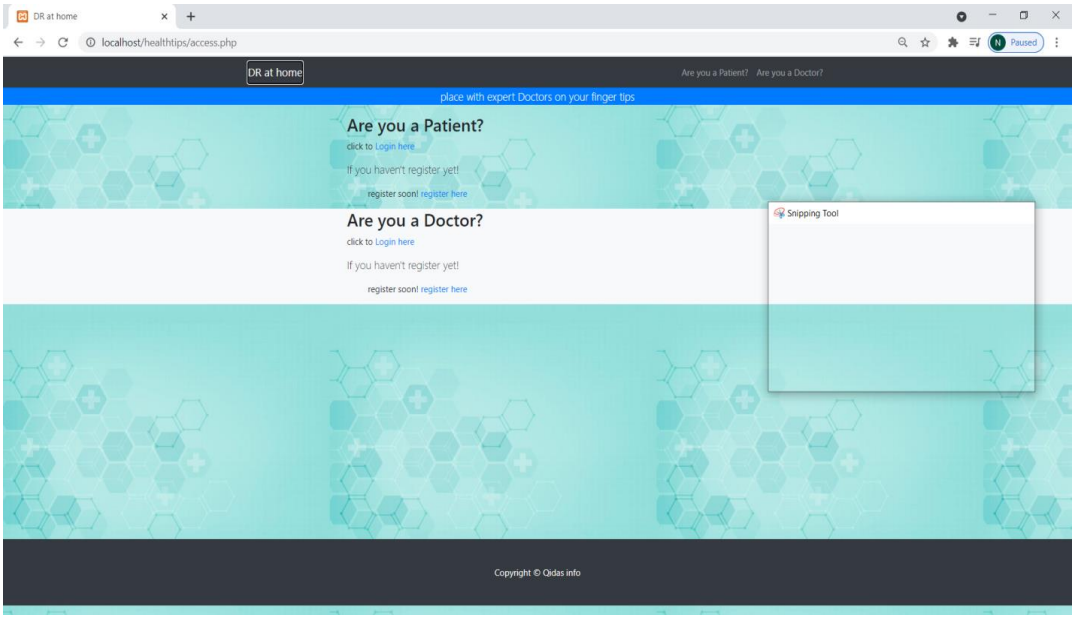


**Figure 11:** Registration portal for Doctors and Patients

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

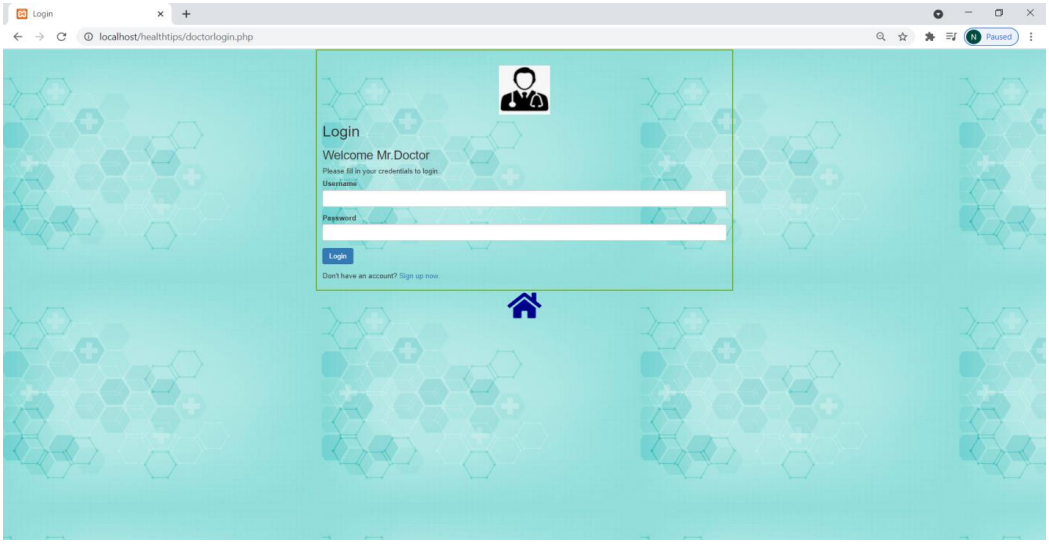


**Figure 12:** Login page for Doctor

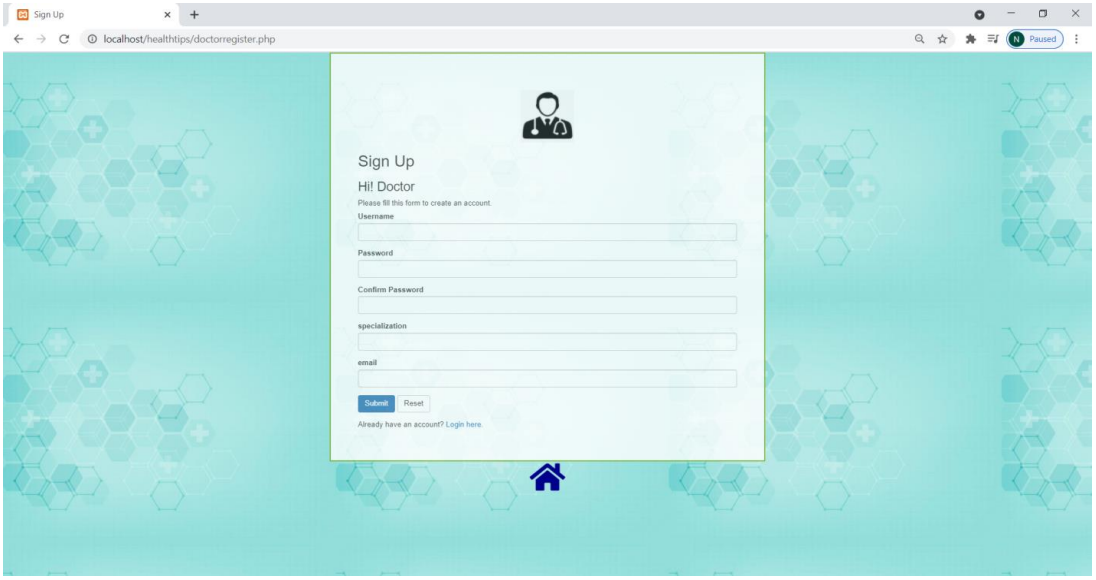


**Figure 13:** Sign up page for Doctor

## VII. CONCLUSION

Healthcare is a multi-dimensional system. It was established with the sole aim of preventing, diagnosis, and treatment of health-related issues in humans. The three major components of a healthcare system are: the health professionals (that includes physicians, surgeons, specialists, physiotherapists and nurses), health facilities (clinics, hospitals, labs and other diagnosis centers), and a financing institution that supports the former two. The health professionals belong to the any of the wide spread health sectors like dentistry, psychology, midwifery, nursing, physiotherapy; to name a few. Healthcare is required or is of utmost importance at several levels depending on the urgency and emergency of the ever changing situations. Professionals serve it as the first point of consultation, acute care which requires skilled professionals, advanced medical investigation and treatment; and highly uncommon diagnostic or surgical procedures ie, primary care, secondary care, tertiary care and quaternary care respectively. There are few objectives and few drawbacks in present health systems which are described further to improve these issues and to make the healthcare available at the door steps or at your finger tips. Healthcare with big data together solve the above issues and helps us to consult a doctor and based on the data available it prescribes you with proper health care.

The web application developed predicts the illness experienced by the patient based on the machine learning techniques and big data analytics approach. Health status prediction using machine learning techniques is not an easy task because predicting does not always mean accurate illness because the person may face additional co morbidities. Machine learning techniques use few features and predict the current health condition but the

actual health status are dependent on multiple factors and also all those factors cannot be included in the machine learning algorithms and big data analytics approaches.

In any machine learning related model it is difficult to find the right algorithm for the right problem. The health status prediction application can be improvised by using other machine learning algorithms too. In this application only few features are predicted. Other features can also be added by training the model and integrating it with the current application. The accuracy of the system is about 96 percent when compared to already existing system.

**Future scope**

In the coming future, all healthcare organizations will adopt various big data analytics methods to achieve their business success, especially during the times of pandemic. This can also make marketing touch points smarter and provide a more integrated perspective. As a result healthcare marketers will be able to integrate large amounts of healthcare data that can provide insights to help and retain patients, provide the highest quality services.

## VIII. REFERENCES

[1] Prasan Kumar Sahoo, Suvendu Kumar Mohapatra, Shih-Lin WU,IEEE Member " Analyzing healthcare with big data with prediction for future health condition ", 2017 – IEEE Access

[2] Dharavath Ramesh, Pranshu Suraj, Lokendra Saini ,"Big data analytics in healthcare ", 2016 – Research Gate

[3] J. Archenaa , E.A.Mary Anita, " A Survey of big data analytics in healthcare and government"- Science Direct

[4] Safa Bahri, Nesrine Zoghlami, Mouraad Abed " Big data for healthcare, A Survey", 2018 – IEEE Access

[5] Wullianallur Raghupathi , Viju Raghupathi , " Big data analytics in health care : promise and potential, 2014 – Health Information Science and Systems

[6] Lidong Wang , Cheryl Ann Alexander, "Big data analytics in healthcare systems", 2019 – International Journal of Mathematical, Engineering and Management Science

[7] Revanth Sonnati, "Improving healthcare using big data analytics",2017 – International Journal of Scientific and Technology Research

[8] Priyanka K, Prof Nagarthana Kulennavar," A survey on big data analytics in health care", 2014 – International Journal of Computer Science and Information Technologies

[9] "Data Mining Application in Predicting Bank Loan Defaulters", International Journal of Innovative Technology and Exploring Engineering, 2020

[10] Jin Ma, Sung Chan Park, Jung Hun Shin, NamGyu Kim, Jerry H. Seo, Jong Suk Ruth Lee,Jeong Hwan Sa. "AI based intelligent system on the EDISON platform", Proceedings of the 2018Artificial Intelligence and Cloud Computing Conference on ZZZ - AICCC '18, 2018N

[11] Malamos, P.E. Barouchas, I.L. Tsirogiannis,A. Liopa-Tsakalidi, Th. Koromilas. "Estimation of Monthly FAO Penman-MonteithEvapo transpiration in GIS Environment, through a Geometry Independent Algorithm", Agriculture and Agricultural Science Procedia, 2015

[12] Ahmed mohamed. "A Data Mining Approach for the Prediction of Hepatitis C Virus protease Cleavage Sites", International Journal of Advanced Computer Science and Applications,2011

[13] An Lu, Wenbin Fang, Chang Xu, Shing-ChiCheung, Yu Liu. "Data-driven testing methodology for RFID systems", Frontiers of Computer Science in China, 2010

[14] Ameya Patil, Dipesh Rana, Sachin Vichare,Chinmay Raut. "Effective Authentication for Restricing Unauthorized User", 2018International Conference on Smart City and Emerging Technology (ICSCET), 2018

[15] Abdelmajed, Azza Kamal Ahmed. "AComparative Study of Locality PreservingProjection and Principle Component Analysis onClassification Performance Using LogisticRegression", Journal of Data Analysis andInformation Processing, 2016.

[16] Cheng-Ta Yeh, Mu-Chen Chen. "Applying Kansei Engineering and data mining to design door-to-door delivery service", Computers &Industrial Engineering, 2018

[17] Hang-Wai Law, Tak-Man Woo. "Quality controlinformation representation using object-oriented data models", International Journal of Computer Integrated Manufacturing, 2003

[18] Palli Suryachandra, P Venkata Subba Reddy."Comparison of machine learning algorithms forbreast cancer", 2016 International Conference on Inventive Computation Technologies (ICICT),2016

[19] Akshay Saji, Aldrin Peter, Anand Ajith, Fibin Mathew, K. K. Smitha. "Chapter 64 Assessment of Factors Causing Delays in Construction for Indian Residential Building", Springer Science and Business Media LLC, 2020Vatesh Pasrija, Praveen Ranjan Srivastava.

[20] "Evaluation of Software Quality using Choquet Integral Approach", International Journal of Fuzzy System Applications, 2013PublicationXu, P.. "Random forests and the data sparseness problem in language modelling", Computer Speech & Language, 200701

[21] Yunfeng Ma, Lu Li, Jun Yang. "A Gravitational Facility Location Problem based on Prize-Collecting Traveling Salesman Problem", 2012IEEE International Conference on Automation and Logistics, 2012

[22] Yeonkook J. Kim, Yoonhwan Oh, SunghoonPark, Sungzoon Cho, Hayoung Park. "Stratified Sampling Design Based on Data Mining", Healthcare Informatics Research, 2013

[23] Bhatt, Advait S.. "Comparative analysis of attribute selection measures used for attribute selection in decision tree induction", 2012International Conference on Radar Communication and Computing (ICRCC), 2012.

[24] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, Lin Wang. "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities",IEEE Access, 2017

[25] Marco Di Ciano, Domenico Morgese, AgostinoPalmitessa. "Chapter 3 Profiling Approach for the Interoperability of Command and Control Systems in Emergency Management: Pilot Scenario and Application", Springer Science and Business Media LLC, 2018Publication"Machine Learning and Data Mining in Aerospace Technology", Springer Science and Business Media LLC, 2020PublicationVinita Periwal, Jinuraj K Rajappan,

[26] Wullianallur Raghupathi, Viju Raghupathi , " Big data analytics in health care : promise and potential", 2014 https://www.sciencedirect.com/science/article/pii/S1319157817302938

[27] https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2014.0041

[28] https://www.hindawi.com/journals/bmri/2015/370194/