A Critique on Heart Diseases Predictive Analytics Using Big Data Algorithms

K.Gayathri, N.Uma Maheswari, G.Mariammal

Abstract- A large volume of both structured and unstructured information is managed by the emerging technology big data. This information is complicated to practice using set records and software techniques. An elite solution is brought in all technologies by using them competently. To improve the prediction of heart diseases earlier and bring more intellectual decisions the big data is potential in healthcare organization. In the present world condition the doctors and experts available are very intricate to forecast the heart diseases. The heart attack has become a remarkable cause of the endless demise worldwide. Heart attack is essential to predict it at an earlier stage to standby the existence of individuals and it is the main source of demise. The primary purpose is to predict the risk level of a person using Big Data algorithms for the cardiac disease. Big Data is primarily designed to provide a national scheme for physicians and patients to login and view Cloud information. Hadoop Map Reduce programming is used to maintain the hospital details. The machine learning algorithms is used to view the precise condition of the patient in its graphical demonstration. Using cloud platform for accessing globally exploitation any browsers in any a part of the globe this application are often enforced.

Keywords- Big Data, Health Care, Diagnosis, Big Data Analytics, Hadoop, cloud platform, machine learning, Map reducing Algorithm

I. INTRODUCTION

This manuscript mainly introduces the peculiarity of and some most important issues of Big Data, health care data [1]. Big data understanding is used in disease prediction, symptom examination, detection, proper drug delivery, and improved quality of treatment and lifespan and a reduced effect of mortality to patients[8,9].[8,9]. Several universities and organizations have recently entered in providing a remedy for enormous expertise in healthcare. This manuscript contains the benefits of huge data about its applications and healthcare possibilities. Many people in the global area are now over-exaggerated by cardiac diseases[11]. The large information performs a significant function to save the lives of the patient and to decrease the deaths of cardiac patients. In collaboration with Apollo Hospital and Alive in the US, Mobile Electro Cardio Gram was invented to monitor the stroke and arrhythmia testing via cellular phones. The sensors mounted on the mobile devices placed on the patient's chest simply monitors the heartbeat of the patient [13, 14]. The health information of the patient are registered deliberately via mobile devices within the cardiogram type, and then clearly added to the records of the patient.

Today numerous individuals on the planet are influenced by heart related illnesses. Enormous information assumes a noteworthy job so as to spare the patients wellbeing and to lessen the demise of heart patients. Apollo Emergency clinic and the US-based Alive Corporation teamed up to invent the Portable Electro Cardio Gram which screens the stroke and arrhythmia screening through cell phones. The sensors which are mounted on the cell phones screens the patient's heart beat by basically lay it on their chest. The patient wellbeing data is naturally recorded through cell phones as cardiogram and afterward it is straightforwardly transferred to the patient's records.

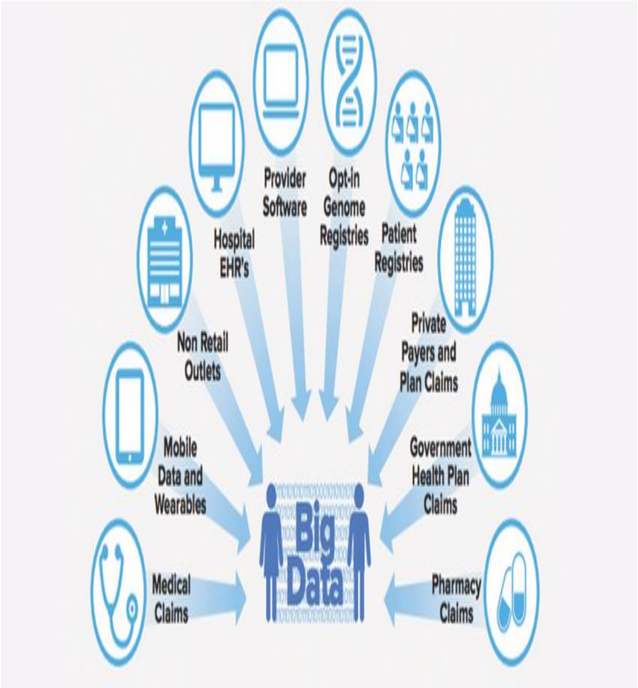


Fig 1. Big Data in Healthcare

All the records are gathered from the cell phones, the heart patient's information are gathered from different medical clinics. The information gathered from specialists is helpful in treating cardiovascular maladies, clinical data.

Revised Manuscript Received on July 18, 2019.

K.Gayathri, Assistant Professor, Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu, India.(email: gayathri1773@gmail.com)

Dr.N.Uma Maheswari, Professor, Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu, India.(email: numamahi@gmail.com)

Ms.G.Mariammal, Assistant Professor, Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu, India.(email: marisl_g1985@yahoo.com)

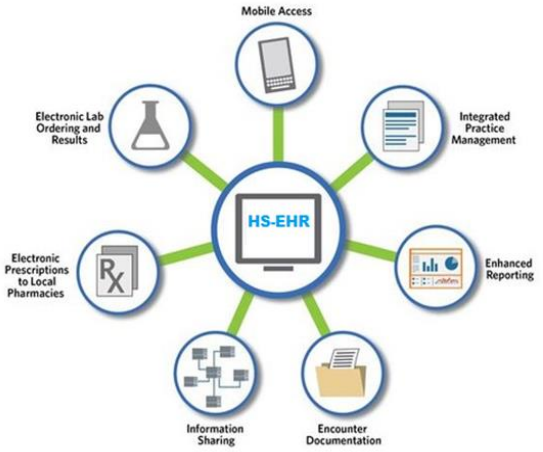

Mobile Access
Integrated Practice Management
Enhanced Reporting
Encounter Documentation
Information Sharing
Electronic Prescriptions to Local Pharmacies
Electronic Lab Ordering and Results
HS-EHR

Fig 2. Electronic Health Records

The report in the paper is entered in to the advanced configuration which is called EHR. The enormous volume of information in the EHR is placed away; it is formed and dissected through huge information by utilizing the HDFS [17]. This information which is gathered through vast in order of examination is useful for patients, experienced doctors, specialists and heart ailment analysts and so forth.

1.1 Big data Characteristics

Velocity:

It relates to the resulting data velocity and the information are huge and consistent. It demonstrates how fast the information is generated and processed to collect the load and determines the true potential of the information.

Volume:

Volume is the only main feature in big data, it is mainly dependent upon quantity of data and it is related to a size which is enormous. The size of data plays a very critical role in determining the value which is out of information.

Value:

Value is the collection of information that can bring more added-value to the information for creating knowledge. There is some precious information within the data.

Variety:

It refers to variety of sources and the nature of data, both structured and unstructured data from the key value, network and also unstructured data from emails, articles and streaming video and audio.

Veracity:

It refers to the noise and deviation in knowledge. It checks whether or not the info is being hold on and mined in meaningful manner to the matter being analyzed. Reliability in knowledge analytics is that the biggest challenge when put next to volume and speed. It’s scope maintaining the info, cleaning and process to stay the unclean knowledge that is gathered within the systems.

II. PREDICTION OF HEART DISEASES

Cardiac analysis relies on the clinical data of the patient, which may assist the medical professionals in forecasting cardiac disorders. Huge data is gathered from healthcare organisations by the health industry. It is essential that all the spied data is mined and discovered in order to create a better and accurate termination.

Heart is the key component of the human body that contributes to the pumping and purification of blood in the organ. The lack of blood flow could also lead to cardiovascular imbalance, brain disorders, renal failure and even instant mortality. The precise working of a heart relies on human existence. The cardiac syndrome relates to the heart's blood vessel. Smoking, obesity, cholesterol, body stress, a bad nutrition are the primary risks causing cardiovascular disease. Most of the hospitals classify their own hospital information for healthcare system patient records. Typical large-scale information is generated via manuscript, descriptions, charts and records. Using Naive Bayes the results have been compared to the Neural network and Decision Tree algorithms and a cardiovascular disease prediction system. It offers excellent prediction quality[19], according to the Naive Bayes method. However, this data is seldom used to maintain the clinical resolution system. The data and information is concealed due to the non-exploitation of the technology used.

2.1 Spark MLLib:

Machine learning library from Spark may be a variety of algorithms during this library that are optimized to run over a distributed dataset. The most distinction between this library and the other well-liked libraries are SciKit that run in a very solo method.

2.2 Apache Spark:

It is a single tool set from the big data stack knowledge and it is quicker in performance to execute coding in Apache Spark as compared to map reduce. All the distributed computing work is handled by the standard variable RDD (the resilient distributed dataset), which is used by lot of developers. It comes with other packages like Spark streaming, Spark SQL.

2.3 Spark SQL:

It is a tool set from API with the intention of supporting data frames which is related to Python but this one runs over a packed distributed dataset and therefore does not have all the related functions.

2.4 HDFS:

This is used for storing the unrefined records, storing the generated pattern and storing the outcome.

2.5 Parquet:

It is a columnar storage format available to any development in the hadoop ecosystem, despite of the choice of data processing structure, data model or programming language. The raw data files are parsed and stored in strip format. It speeds up the aggregation queries and this columnar format helps in choosing only the columns that are desired and hence it reduces the disk input extremely.

III. ALGORITHMS USING MACHINE LEARNING

3.1 Naive Bayes Classifier

Naive Bayes classifier is the method used to predict heart disease. In Naive Bayes, all characteristics are considered independent, reducing calculations significantly[16]. The formula of Naive Bayes is

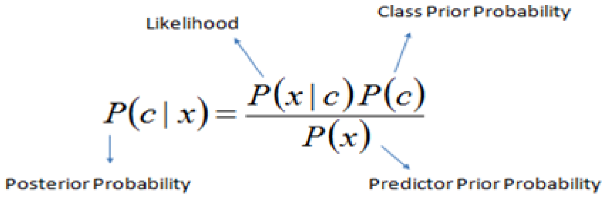
- fp(c|x) – the subsequent possibility of the analyst's class (target).
- P(c) – the class's previous possibility. It is the possibility of observing a common class.
- P (x|c) – the possibility of analyser in a class.
- P(x) – the previous probability of analyst.

Therefore with the inputs given a patient's documentation of a quantity of parameters subsequent probability for all achievable threat levels are computed. Patients have a amount of risk for which their possibility is highest. For computing the group of trained probabilities, training data set is used. We calculate P (xi|Cj) for Class Cj given an index xi. We can use this fundamental concept of probability as

$$P(x_i/C_j) = \frac{\text{Number of times } x_i \text{ occurs in rows of training dataset } X \text{ for class } C_j}{\text{Number of times } C_j \text{ occurs in training dataset } X} \qquad (1)$$

xi € X where j=0,1,2,3,4. Eqns (1) is used for the computation of possible complete training dataset. This calculation method only applies, however, if factors such as chest pain type for patient records are different in nature. Any five characteristics of age, cholesterol, relaxing blood pressure, thalac and oldpeak can be used precisely in the dataset. Consequently, for the main method, they use likelihood density feature to calculate skilled group densities with normal allocation for all incessant factors. Here,

- σc^2 - variance of x.
- μc – mean of x.


Likelihood
Class Prior Probability
$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Eqns (2) is trained for the use of age, cholesterol, and thalach normally distributed. Subsequent, blood pressure and the old peak are not good quality as a result of the moderately accurate result, mainly at low accuracy, so we have one further approach to avoid the distribution of variables, i.e. pretend to be the variables. The group of qualified probabilities are calculated in such cases in the same way as for other discrete variables. In this situation it includes big data sets and outcomes in predictable outcomes and high accuracy. This prediction valuation has a good value.

IV. PROCESS OF BIG DATA USING HADOOP & RESULTS

The algorithm can be constructed upon using an machine learning algorithm to increase the accuracy of the forecast of the risk of disease levels by covering the drawbacks of the achievable algorithms A large amount of data is available in many clinics and the fitness industry. The medical practitioner and the medical specialist available is extremely needed for the increasing inhabitants, as a proportion of the inhabitants in whom medical doctors may not diagnose severity of the disease properly. The Big Data approach is used with a given node array[15], Hadoop. For the algorithms designed, Map Reduce software is implemented.

4.1 Maping function

Each row of the entry document is drawn with the mapping feature. (3) In perspective of the multi-node group as each node continues the equal technique in the same time, join the chart section and is transferred to the distinct map segment in succession. If N traces in the input file are available and there are default M map segments, then the mapping function executes the algorithm in the node and in each node in the multi-node cluster with the help of each segment of the map. It requires a single Big Data row as entry and runs the machine learning algorithm to calculate the risk level every time. However, the range of the lines is used as the value is the key and the entire row. Each characteristic that needs an estimate will be allocated as a key reducer and fees will be assigned. The context file is the intermediate output provided as an entry to minimize feature by mapping function.

$$p(x=v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\, e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

4.2 Reduce Function

The reduce function transfers the threat level from the outlook file to the main attributes that are given to decrease the feature in the sequence of the mount and saves the ordered input into a file. The map reducing tasks are used in both algorithms for the implementation of Big Data. A range of map reduction features have been introduced to calculate graphs with and without disease for a number of characteristics. It can be used in a number of resident journals.

4.3 Map Reduce Algorithm -- Logical View

The Map and Reduce features are described in (key, value) pairs both with reverence towards structured information. Map describes a couple of information with a particular type in a data domain and produces a number of combinations in a separate data domain.

Map (key1, value1) →list (key2, value2) (4)

In Eqn (4), the Map function is used in the data set containing inputs, according to each pair (key by key1). For each call, this generates a set of combinations (key2). After that, the frame-technique Map Reduce gathers all rows of all information records using the same key (key2) and groups them in groups, thus generating each element for a single unit.

The Reduce function is then carried out simultaneously with each group, which generates a collection of values in the same field:

Reduce (key2, list (value2)) →list (value3)(5)

Reduce call generates one or no return value, although single call can yield more than one value. All calls are returned and the results are gathered as anticipated. Thus, by using Eqn (5), the Map Reduce technique will alter a (main, item) pair list into a valuation list. This approach is unlike classical function programming, i.e. the map and reduce functionality that takes a random value list and returns a single value and all the values that are returned by map together.

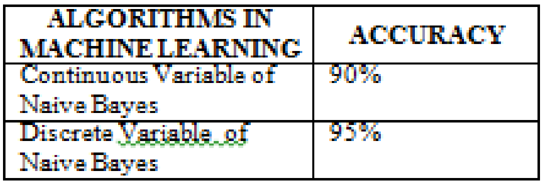
It is very important to implement the map and reduces steps in the way to implement Map Reduce framework. Distributed implementations of Map Reduce require the connecting of the processes performing the Map and Reduce steps in each phase. This can be a distributed file system.

4.4 Results using Graphical Analysis

The result of this task may only be a patient report depending on the initial data or graphic performance if large information files are an input document. Table I provides a comparison analysis of the above machine learning algorithms. The correct graph is presented to determine the finest algorithm for the forecast of diseases.

This involves several aspects of the research including the complete count of patients with and without cardiac diseases and the count of patients at a certain age. The graphic display of all these aspects is so easy for the user to understand. Figure 3 demonstrates the comparison analysis of the algorithms used in the Machine Learning as described in Naive Bayes document for the constant (Green) and naive (Violet) factors.

Table I. Comparative study and accurate values for Machine learning algorithms.



ALGORITHMS IN MACHINE LEARNING	ACCURACY
Continuous Variable of Naive Bayes	90%
Discrete Variable of Naive Bayes	95%

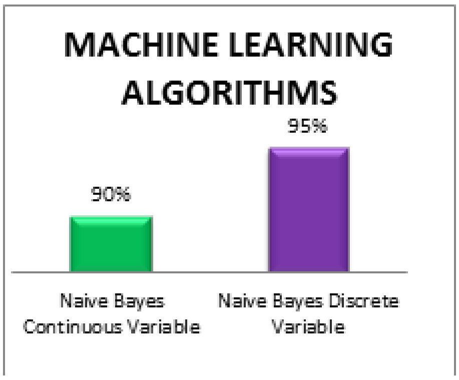


Fig 3. Comparative Charts of Machine Learning Algorithms

V. BUILDING A CENTRALIZED SYSTEM

Cloud computing is a scheme which divides the PC network by allowing a software or an implementation to work at the same time on a number of connected pcs. This applies in particular to the laptop hardware or to a set of hardware mechanisms, two of which are generally called servers linked to the Internet by a verbal interchange community. Any person authorized to enter the server can use server processing to operate the request or acquire any other computing function, or to retailer information. The task is created on a Jelastic Cloud Platform.

Jelastic is a system with the features of cloud-based system as Platform-as-infrastructure (PAI) providing networks, databases and storage solutions for consumers, OEMs and Internet service companies. Some companies have developed applied science in the cloud to update Java and PHP functions.

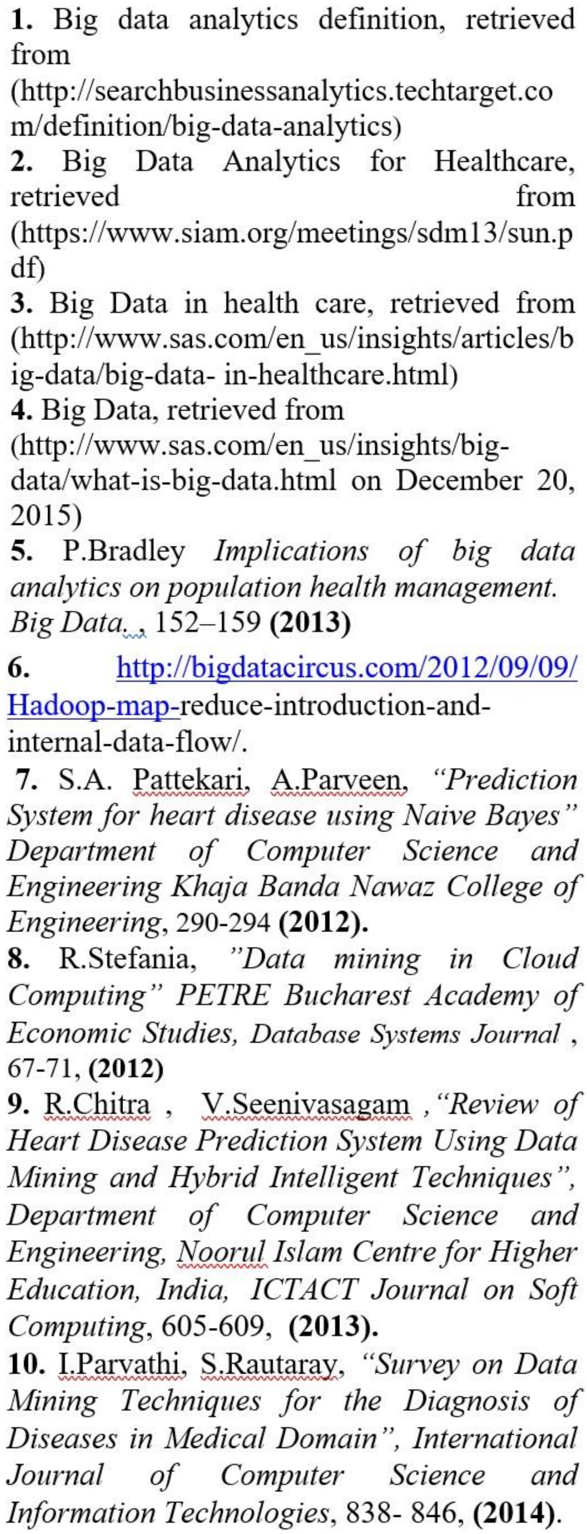
It has an international web hosting partner and data center. The company can provide services such as memory, CPU and disk area to meet the customer's needs. Google App Engine, Amazon Elastic Beanstalk, Heroku and Cloud Foundry are the major competitors of Jelastic. Jelastic is an only platform with a number of computerised vertical scaling and utility lifecycle panels that are accessible from several host vendors around the world and which now does not contain restrictions or code which seriously change requirements.

VI. CONCLUSION

The data relating to health care are extensive in nature and occur from one of a kind of place of origin that is no longer a desirable form or value. The use of facts and the familiarity of specialists and patient health statistics is now combined for the duration of a generally accepted analysis process. The Hadoop framework uses the node cluster for processing large records as one of the representations close to technology.

Execution of best computing device mastering algorithms are used to establish the heart ailment prospect and contrast of algorithms is executed to consider the truth the usage of graphs. It is less complicated to identify with the graphs and the consumer can additionally discover out their chance level and to get the associated information. The mission uses the cloud service on an international level and Big Data is easily handled.

VII REFERENCES



1. Big data analytics definition, retrieved from (http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics)

2. Big Data Analytics for Healthcare, retrieved from (https://www.siam.org/meetings/sdm13/sun.pdf)

3. Big Data in health care, retrieved from (http://www.sas.com/en_us/insights/articles/big-data/big-data- in-healthcare.html)

4. Big Data, retrieved from (http://www.sas.com/en_us/insights/big-data/what-is-big-data.html on December 20, 2015)

5. P.Bradley *Implications of big data analytics on population health management. Big Data.* , 152–159 **(2013)**

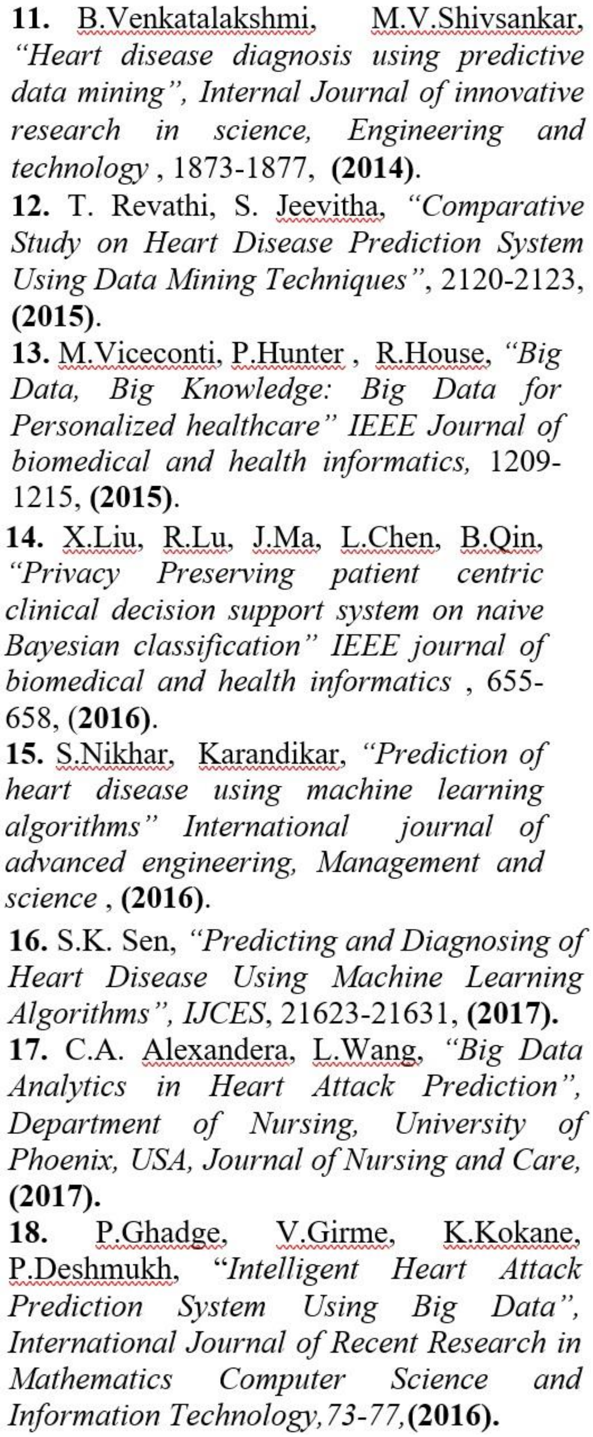
6. http://bigdatacircus.com/2012/09/09/Hadoop-map-reduce-introduction-and-internal-data-flow/.

7. S.A. Pattekari, A.Parveen, *"Prediction System for heart disease using Naive Bayes" Department of Computer Science and Engineering Khaja Banda Nawaz College of Engineering*, 290-294 **(2012).**

8. R.Stefania, *"Data mining in Cloud Computing" PETRE Bucharest Academy of Economic Studies, Database Systems Journal* , 67-71, **(2012)**

9. R.Chitra , V.Seenivasagam ,*"Review of Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques", Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, India, ICTACT Journal on Soft Computing*, 605-609, **(2013).**

10. I.Parvathi, S.Rautaray, *"Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain", International Journal of Computer Science and Information Technologies*, 838- 846, **(2014)**.



11. B.Venkatalakshmi, M.V.Shivsankar, *"Heart disease diagnosis using predictive data mining", Internal Journal of innovative research in science, Engineering and technology* , 1873-1877, **(2014)**.

12. T. Revathi, S. Jeevitha, *"Comparative Study on Heart Disease Prediction System Using Data Mining Techniques"*, 2120-2123, **(2015)**.

13. M.Viceconti, P.Hunter , R.House, *"Big Data, Big Knowledge: Big Data for Personalized healthcare" IEEE Journal of biomedical and health informatics,* 1209-1215, **(2015)**.

14. X.Liu, R.Lu, J.Ma, L.Chen, B.Qin, *"Privacy Preserving patient centric clinical decision support system on naive Bayesian classification" IEEE journal of biomedical and health informatics* , 655-658, **(2016)**.

15. S.Nikhar, Karandikar, *"Prediction of heart disease using machine learning algorithms" International journal of advanced engineering, Management and science* , **(2016)**.

16. S.K. Sen, *"Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms", IJCES*, 21623-21631, **(2017).**

17. C.A. Alexandera, L.Wang, *"Big Data Analytics in Heart Attack Prediction", Department of Nursing, University of Phoenix, USA, Journal of Nursing and Care,* **(2017).**

18. P.Ghadge, V.Girme, K.Kokane, P.Deshmukh, "*Intelligent Heart Attack Prediction System Using Big Data", International Journal of Recent Research in Mathematics Computer Science and Information Technology,73-77,***(2016).**