A SURVEY ON PREDICTION OF HEART DISEASES USING BIG DATA ALGORITHMS

MsS.Suguna[1], *Sakthi Sakunthala,N, S.Sanjana* [2], *S.S.Sanjhana* [3]
Assistant Professor, Department of Information Technology, Sri Krishna College of Technology, Kovaipudur, Coimbatore, India[1]

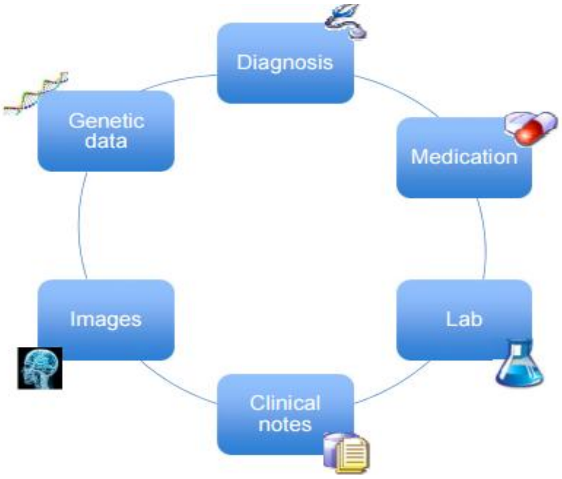
U.G Scholar, Department of Information Technology, Sri Krishna College of Technology, Kovaipudur, Coimbatore, India[2,3]

***ABSTRACT*- Big Data is a technology used to manage a large volume of both structured and unstructured data. It is difficult to process using traditional database and software techniques. Nowadays it is effectively used in all technologies to bring unique solution. Big Data has the ability to help healthcare organization to improve the prediction of heart diseases and make faster and more intelligent decisions. With the tremendously growing population, the doctors and experts available are not in proportion with the population. Symptoms of heart disease is very difficult to predict in the present world scenario. The main objective of this research project is predicting the heart disease risk level of a patient using Big Data algorithms. The main feature of Big Data is creating a Centralized System for both doctors and patients to login and view the data on Cloud. The hospital records maintained in big data is handled using Hadoop Map Reduce programming. A graphical representation of machine learning is used for easy view and to know the exact condition of the patient. This application can be implemented using cloud platform for accessing globally using any browsers in any part of the world.**

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

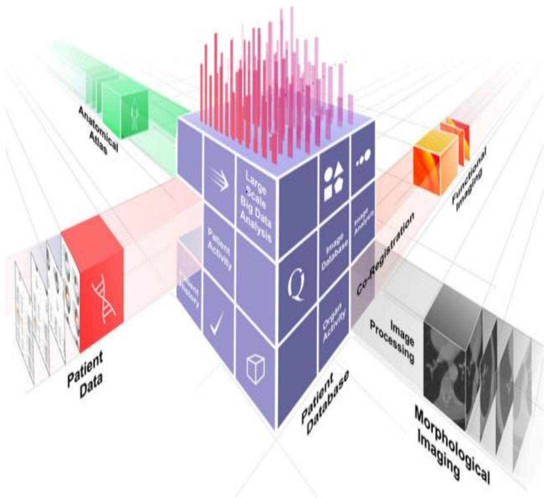
INTRODUCTION

This paper mainly introduces the characteristics of Big Data, health care data and some major issues of Big Data. Big data in health care is used to predict the diseases, analyze the symptoms, imp rove the diagnosis, provide medicine correctly for the patients to recover from heart diseases, and enhance the quality of care , lower the cost, improve the life span and to reduce the impact of death in advance. There are many organization and university have joined together to provide a solutions for big data in health care. These issues include Big Data benefits, its applications and opportunities in medicare and health care.Today many people in the world are affected by heart related diseases. Big data plays a major role in order to save the patients health and to reduce the death of heart patients. Apollo Hospital and the US-based Alive cor Inc collaborated together to invent the Mobile ECG 4(Electro Cardio Gram) which monitor the stroke and arrhythmia (irregular Heart Beat) screening through mobile devices. The sensors which are mounted on the mobile devices monitors the patient’s heart beat by simply rest it on their chest. The patient health information is automatically recorded through mobile devices in the form of ECG and then it is directly uploaded to the patient’s data bases.



Today many people in the world are affected by heart related diseases. Big data plays a major role in order to save the patients health and to reduce the death of heart patients. Apollo Hospital and the US-based Alive cor Inc collaborated together to invent the Mobile ECG 4(Electro Cardio Gram) which monitor the stroke and arrhythmia (irregular Heart Beat) screening through mobile devices. The sensors which are mounted on the mobile devices monitors the patient’s heart beat by simply rest it on their chest. The patient health information is automatically recorded through mobile devices in the form of ECG and then it is directly uploaded to the patient’s data bases.

Here, all the information are collected from the mobile devices, Heart Patients data collected from various hospitals, Data collected from experts in treating cardiac diseases, Clinical information, records in the paper format are entered in to the digital format which is called EHR. The huge volume of data in the EHR is stored, processes and analyzed through big data by using Map reduce and HDFS. This data which is retrieved through big data analytics is helpful for patients, undergone training physicians, Doctors, Cardiac disease researchers etc.



I. CHARACTERISTICS OF BIG DATA

Volume – The name 'Big Data' itself is related to a size which is enormous. Size of data plays very crucial role in determining value out of data. It is mainly dependent upon volume of data. Here ***'Volume'*** is the only main characteristic which is considered in 'Big Data'.

Variety: Variety refers to heterogeneous sources and the nature of data, both structured and unstructured from key value web clicks and unstructured data from email messages, articles and streamed video and audio, etc

Velocity – *V*elocity refers to the speed of data that is generated accordingly. It shows how fast the data is generated and processed to meet the demands and determines the real potential in the data. The flow of data is huge and continuous.

Value: It is defined by the added-value that the collected data can bring. It refers to the value that the data adds to creating knowledge. There is some valuable information somewhere within the data.

Veracity*:* Veracity refers to the biases, noise and abnormality in data. It checks whether the data is being stored, and mined meaningful to the problem being analyzed. veracity in data analysis is the biggest challenge when compared to things like volume and velocity. It has scope to help in keeping the data clean and processes to keep ‘dirty data’ from accumulating in your systems.

II. HEART DISEASE PREDICTION

Heart disease diagnosis is depends upon the clinical data. Prediction of heart disease could assist the medical experts for heart disease prediction is being done within the patient's clinical data . The industry of healthcare is collecting the huge data from healthcare organizations, which actually need to get mined and discover all spied information for making accurate and enhanced decision . Heart is an important part of human body which helps to pump and purify the blood in body. The deficiency of blood circulation could be a cause for inactivity of heart, unbalancing brain, kidney failure and even instant death also. The human life is depending on the efficiently and proper working of heart. The heart disease system refers blood vessel and heart disease within it. The major risk factors that cause heart disease are smoking, overweight, cholesterol,blood pressure,diabetes, unhealthy diet and physical activity. Nowadays, most of the hospitals are sorting their own hospital information for managing the patient data or healthcare system. These systems are generating typical large data through text, images, charts and numbers. Using Naïve Bayes they proposed a heart disease prediction system and compared the results with Neural Network and Decision Tree algorithms. According to that method, thus this algorithm provides good prediction. However, these data are being rarely used for supporting the decision making system for clinic. There is more data and information are hidden which could be smartly untapped Technology Used

- **Apache Spark**: Apache Spark is one of the tool set from the big data stack technology. It is much faster in performance and also to perform coding in Apache Spark as compared to map reduce. RDD (the resilient distributed dataset), is used by lot of developers as a normal variable which nicely handles all the distributed computing work. It comes with other cool packages like Spark streaming, Spark sql, etc.

- **Spark SQL**: Spark SQL is a tool set from API that supports DataFrames which is similar to Python but this one runs over a full distributed dataset and hence does not have all the similar functions).

- **Parquet**: It is a columnar storage format available to any project in the Hadoop ecosystem, regardless of the choice of **data** processing framework, **data** model or programming language. The raw data files are parsed and stored in parquet format. This speeds up the aggregation queries and this columnar format helps in choosing only the columns that are needed and hence reduces disk input tremendously.

- **Spark MLLib**: Machine Learning library from Spark is a type of algorithms in this library that are optimized to run over a distributed dataset. The main difference between this library and the other popular libraries are SciKit that run in a single process.

- **HDFS** : This is used for storing the raw files, storing the generated model and storing the results.

III. MACHINE LEARNING ALGORITHMS:

A. Naive Bayes Classifier:

The first method used for prediction of heart disease is Naïve Bayes classifier. 13 preprocessed attributes are used as input in this algorithm. In Naïve Bayes assumption all attributes are independent of each other, this significantly reduces the calculations . Naïve Bayes formula is given by

$$\underbrace{P(c\mid x)}_{\text{Posterior Probability}}=\frac{\overbrace{P(x\mid c)}^{\text{Likelihood}}\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

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

• fp(c|x) - posterior probability of class (target) given predictor (attribute).

• P(c) - prior probability of class, also called prior. It is the probability of observing a class in general.

• P (x|c) - likelihood which is the probability of predictor given class.

• P(x) - prior probability of predictor also called evidence.

Hence with the inputs a patient record of 13 attributes we can calculate posterior probability for all possible risk levels. Patient has the risk level for which the posterior probability is maximum. Training data set is used for calculation of class conditional probabilities . Given an attribute xi we can calculate P(xi|Cj) for class Cj. For this we can use basic definition of probability that is.

$$P(x_i|Cj) = \frac{\text{Number of times } x_i \text{ occurs in rows of training data set } X^t \text{ for clas}}{\text{Number of times Cj occurs in training data set } X^t}$$

xi € X where j=0,1,2,3,4. For the calculation of likelihood entire training dataset is used. However this method of calculation holds good only if variables are discrete in nature like , chest pain type etc. for patients record. In dataset exactly 5 attributes are mainly used i.e., age, cholesterol, resting blood pressure, thalach and oldpeak are continuous. Hence they use probability density function for initial approach, calculation of class conditional densities using assuming Normal distribution for all the continuous variables as shown

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

- Here σc2 - variance for variable x given class C.

- µc - mean for variable x given class C.

Using normal distribution method for age, cholesterol and thalach is an approximately holds good .Resting blood pressure and old peak do not fit into this distribution and this does not holds good. As a result, we get partially accurate results leading to low accuracy., we can use another approach to avoid dealing with distribution of variables, i.e. assuming the variables to be discrete. In such case, calculation of class conditional probabilities for these variables is done in the same way as done for other discrete variables. This assumption value holds good in this case since it contains large datasets and also leads to expected results and high accuracy.

B. Probabilistic Analysis and Classification (PAC)

The supervised machine learning algorithm derived from Naïve Bayes Algorithm are Probabilistic Analysis and Classification. It uses the concept of calculating the weighted average probability over the entire training data set {Xt}.It is formed over Naïve Bayes model to overcome the disadvantages of Naïve Bayes algorithm. One advantage is using discretization technique,it does complete reduction of continuous variables to discrete variables.Another advantage is due to complete conversion of continuous variable to discrete variable, Laplacian smoothing method that are used in Naïve Bayes Classification is not required, which in turn reduces unnecessary comparisons and unwanted instructions.

The main concept of this algorithm is to use weighted average calculation for all heart disease values until and unless we find an exact same tuple in the data set, thus in this case the risk level of tuple is assigned to the risk level of the inputs in the patients record. This case occurs very rare and so we have to use weighted average calculation for the entire data set and calculate the contribution of each and every value for that particular risk level and find the solutions for different contributions for entire data set are. In considering the entire data set we have used number of supporting tuples for various risk levels in the data set. This concept is similar to "Prior" in Naïve Bayes algorithm but in Naïve Bayes algorithm the prior probabilities give more weight to risk levels on the basis of their own values. In PAC it simultaneously reduces this weight which results in error, due to difference in percentage increase in numerator and denominator in the term αi. So to overcome this disadvantage, we multiply by normalizing factor to reduce this error and give accurate results. Finally the maximum term µi among all risk levels is returned as the risk level for the patient.

Other variations and possibilities from Naïve Bayes implementation are loading of training data and Big Data files, which have to be parsed in pre-initial step to convert continuous variables to discrete variables. This can also be used in one more form, that is conversion of continuous variables given by user to discrete form so that algorithm can read and can be processed.

Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes

Introduction to PAC Algorithm, Comparison of Algorithms and HDPS

PAC Algorithm:

```
PAC (ip.csv)
{
fp = ip.csv
fw = traing.csv
fq = make_discrete(fp)
while fq!=EOF
for each line in fq
for each line in fw
$\alpha i = \sum 1$ (for each matching attri)
13 Where i= diff risk levels
End For
$\beta i = \alpha i/SPi$ Where SPi is the number of supporting cases.
For each risk level
End For
$\mu i$= Normalizing Factor j x $\beta i$
r=max ($\mu i$)
Op r as the risk level
End While
}
maximize ($\mu i$)
{
Return index for which $\mu i$ is max for all i=0, 1,2,3,4
}
made_discrete(fq)
{
Assign continuous variables discrete values Vi by splitting into equal intervals with varying ranges.
Return the dataset which has maximum accuracy
}
```

IV. HADOOP MAP REDUCE PROGRAMMING FOR PROCESSING BIG DATA

This paper is a successful algorithm design for accurate prediction of heart disease risk level. This PAC algorithm is built using an existing machine learning algorithms which covers up the drawbacks of the existing algorithms and in turn increase the accuracy of prediction of disease risk level. Many hospitals and health care industries have huge amounts of patient data. With the tremendously growing population, the doctors and experts available are lack in proportion with the population in which Doctors may sometime fail to correctly diagnose the severity of the disease. Hadoop, a single node cluster is used to process Big Data. Map Reduce code is implemented for the designed algorithms.

Mapper: Inside Mapper function each line from input file is taken as input to map phase and is taken to different map-tasks in parallel, considering multi-node cluster as each node follows the same procedure in simultaneously. If there are N lines in input file and we have default M map tasks then number of lines processed by each map task , then the mapper function executes our algorithm on each and every map task of node and in each and every node in a multi-node cluster. Every time it takes single line from Big Data as input and processes machine learning algorithm to calculate risk level. But, here the line number is taken as key and entire line is taken as value. The risk level is supplied as a key to reducer and value is assigned to every r attribute the needs to evaluate with. The context file is the intermediate output given by mapper function as input to reducer function.

Reducer: The reducer shuffles the risk level provided by context file and sorts them according to the key values given to reducer function in ascending order and stores the sorted output in a file. The map-reduce jobs are used to process Big Data in both the types of algorithms. Various Map-Reduce functions are implemented to calculate the graphs for different attributes with number of people with and without disease. This can be employed for various population journals.

Logical View of Map Reduce Algorithm:

The Map and Reduce are the functions both defined with respect to structured data in (key, value) pairs. *Map* defines one pair of data with a type in one of data domain, and returns a list of pairs in a different domain of data

Map(k1,v1)→list(k2,v2)

The Map function is applied in parallel to every pair (key by k1) in the dataset containing inputs. This produces a list of pairs (keyed by k2) for each call. After that, the MapReduce framework method is done it collects all pairs with the same key (k2) from all lists of data and groups them together as clusters, creating one group for each key.

The Reduce function is then performed in parallel to each group, which produces a collection of values in the same domain:

Reduce(k2,list(v2))→list(v3)

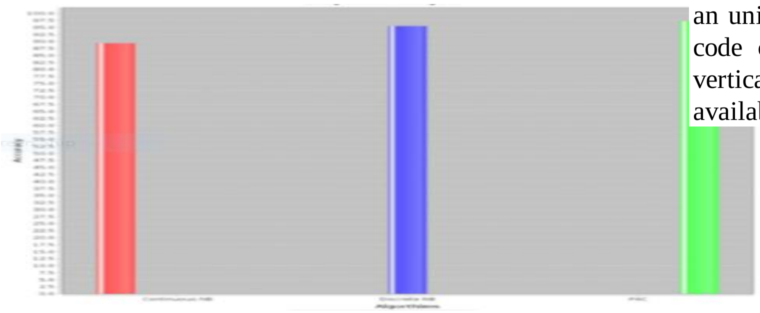
Reduce call typically produces either one value or an empty return, although one call is allowed to return more than one value. The returns all the calls and they are collected as the expected result list.

Thus the MapReduce method changes a list of (key, value) pairs into a list of values. This approach is different from the typical functional programming,i.e map and reduce functionality, which accepts a list of arbitrary values and returns one single value and that combines all the values returned by map.

It is very necessary to have implementations of the map and reduce steps in order to implement MapReduce framework. Distributed implementations of MapReduce require a means of connecting the processes performing the Map and Reduce steps in each phase. This can be a distributed file system.

Graphical Analysis Results

The output of the project can be either a report of a single patient for form based input or graphical output if Big Data file is provided as input. A comparative study of machine 8learning algorithms explained above is made and an accuracy graph is plotted to determine the best algorithm for disease prediction. This includes multiple aspects of the study such as the total number of patients who have and do not have heart disease, number of patients of a particular age who have and do not have disease etc. all these aspects are shown in graphical format so that it is easier for the user to understand. Figure 1.6 shows the comparative study of the Machine Learning algorithms as explained in the paper Naïve Bayes for continuous variables (red), Naïve Bayes for discrete variables (blue) and PAC Algorithm (green).



Machine Learning Algorithms	Accuracy
NAÏVEBAYES CONTINUOUS VARIABLE	89.80%
NAÏVE BAYES DISCRETE VARIABLE	95.21%
PROBABILISTIC ANALYSIS	97.48%

V. BUILDING A CENTRALIZED SYSTEM AND DEPLOYMENT ON CLOUD PLATFORM- HDPS

Cloud computing is a process that involves distribution of computer network, where a program or an application may run on many connected computers at the same time. It specifically refers to a computer hardware machine or group of computer hardware machines which is commonly referred as a server connected through a communication network such as the Internet, also networks like (LAN) or (WAN). Any individual user who has permission to access the server can use the server's processing power to run an application, and also to store data or perform any other computing task. This project is developed on a Cloud Platform called Jelastic.

Jelastic is a platform which involves the characteristics of Platform-as-Infrastructure (PAI) cloud computing service which provides networks, servers, and storage solutions to software development clients, enterprise businesses, OEMs and web hosting providers. Some company has developed technologies for moving Java and PHP based on applications onto the cloud based platform.

It has an international hosting partners and data centers. The company can provide facilities like memory, CPU and disk space to satisfy the needs of the customer. The main competitors of Jelastic are Google App Engine, Amazon Elastic Beanstalk, Heroku, and Cloud Foundry. Jelastic is an unique platform that it does not have any limitations or code change requirements, and also it offers automated vertical scaling, application lifecycle management and availability from multiple host providers around the world.

VI. Conclusion

Health care related data's are huge in nature and they arrive from different birthplaces which are not suitable in structure or quality. Nowadays, the utilization of knowledge and experience of specialists and medical screening data of patients are collected in a database during the diagnosis process, which has been widely accepted.

Using Hadoop framework, node cluster is used for processing big data is one of the upcoming technologies.

Implementation of accurate machine learning algorithms are used to determine the heart disease possibility and comparison of algorithms is done to evaluate the accuracy using graphs. It is easier to understand the graphs and the user can also determine their risk level and to get the similar report. The project is globally accessible using cloud service and Big Data can be easily processed.

VII. References

[1] http://www.rohitmenon.com/index.php/introducing-mapreduce-part-i/

[2] http://www.javacodegeeks.com/2013/08/writing-a-hadoop-mapreduce-task-in-java.html

[3] http://developer.yahoo.com/hadoop/tutorial/module4.html

[4] http://nxhoaf.wordpress.com/2013/01/04/hadoop-mapreduce-word-count-using-eclipse/

[5] http://bigdatacircus.com/2012/09/09/hadoop-map-reduce-introduction-and-internal-data-flow/

[6] https://www.harding.edu/fmccown/r/#barcharts

[7] http://stackoverflow.com/questions/19510656/how-to-upload-files-on-server-folder-using-js

[8] http://www.uniweimar.de/medien/webis/teaching/lecturenotes/machine-learning/unit-en-decision-treesalgorithms.pdf

[9] Prediction System for heart disease using Naïve Bayes *Shadab Adam Pattekari and Asma Parveen Department of Computer Science and Engineering Khaja Banda Nawaz College of Engineering.

[10] Data mining in Cloud Computing Ruxandra-Ştefania PETRE Bucharest Academy of Economic Studies ruxandra_stefania.petre@yahoo.com

[11] Review of Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques R.Chitra1and V.Seenivasagam2 1Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, India.

[12] Parvathi I, Siddharth Rautaray, ―Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain‖, International Journal of Computer Science and Information Technologies, Vol. 5 (1), 838-846, ISSN: 09759646, 2014.

[13] Dhanya P Varghese, Tintu P B, ―A Survey on Health Data using Data Mining Techniques‖, International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 07, e-ISSN: 2395-0056, p-ISSN: 2395-0072, Oct2015.

[14] Vahid Rafe, Roghayeh Hashemi Farhoud, ―A Survey on Data Mining Approaches in Medicine‖, International Research Journal of Applied and Basic Sciences, Vol 4 (1), ISSN 2251-838X, 2013.

[15] www.medicalnewstoday.com/articles/237191.php

[16] www.heart.org/idc/groups/ahamahpublic/@wcm/@sop/@smd/documents/downloadable/ucm480086.pdf

[17] T. Revathi, S. Jeevitha, ―Comparative Study on Heart Disease Prediction System Using Data Mining Techniques‖, Volume 4 Issue 7, ISSN (Online): 2319-7064, July 2015.

[18] Devendra Ratnaparkhi, Tushar Mahajan, Vishal Jadhav, ―Heart Disease Prediction System Using Data Mining Technique‖, International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 08, e-ISSN: 2395 -0056, p-ISSN: 2395-0072, Nov-2015.

[19] K.Manimekalai, ―Prediction of Heart Diseases using Data Mining Techniques‖, International Journal of Innovative Research in Computer andCommunication Engineering, Vol. 4, Issue 2, ISSN(Online):2320-9801, ISSN (Print):23209798, February 2016.

[20] Jyoti Rohilla, Preeti Gulia, ―Analysis of Data Mining Techniques for Diagnosing Heart Disease", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7, ISSN: 2277 128X, July 2015

[21] Survey on Data Mining Algorithms in Disease Prediction V.Kirubha1, S.Manju Priya2 1 Research Scholar, Department of Computer Science, Karpagam University,Coimbatore, Tamil Nadu, India 2Associate Professor, Department of Computer Science, Karpagam University, Coimbatore, Tamil Nadu, India.

[22] Prediction of Diseases using Big Data Analysis M Archana Bakare, Prof. R.V.Argiddi PG Student, Dept. of Computer Science and Engineering, WIT, Solapur, Maharashtra, India Associate Professor, Dept. of Computer Science and Engineering, WIT, Solapur, Maharashtra, India.

[23] Journal of the american college of cardiology vol. 69, no. 7, 2017 ª 2017 by the american college of

cardiology foundation published by elsevier issn-http://dx.doi.org/10.1016/j.jacc.2017.01.006

[24] Warner, D., 2013. Safe de-identification of big data is critical to health care. Health Inform. Manage.

[25] White, S.E., 2014. A review of big data in health care: Challenges and opportunities. Open Access Bioinform., 6: 13-18. DOI: 10.2147/OAB.S50519.

[26] Big Data Analytics for Healthcare Chandan K. Reddy Department of Computer Science Wayne State University Tutorial presentation at the SIAM International Conference on Data Mining, Austin, TX, 2013.
http://dmkd.cs.wayne.edu/TUTORIAL/Healthcare/
Jimeng Sun Healthcare Analytics Department IBM TJ Watson Research Center