Big Data Analysis for Prediction of Coronary Artery Disease

Prerna Jain [1], Amandeep Kaur[2]
Department Of Computer Science and Engineering,
Lovely Professional University, Jalandhar City 144411, India
[1] prernajain1717@gmail.com, [2] er.amandeep.kaur@gmail.com

***Abstract*— Heart disease is the major cause of the death. Medical treatment and diagnosis of the heart disease is the major factor to improve the death rate. The main risk factors of the heart disease are obesity, tobacco, alcohol consumption and age factor etc. To reduce the death rate, the predictive system will be design so that way of treatment and diagnosis will be improve. This paper presents a big data analysis for prediction of the Coronary Artery heart disease. The analysis of huge amount of a patient by using data mining and machine learning algorithms improves a hospital administration. As a huge amount of the data in increase in every field, so it is difficult to analysis, extract, manage and store a structured and unstructured data, so that the big data technologies and tools are used.**

***Keywords*—Big data, Healthcare, KNN algorithm, Genetic algorithm, Coronary Heart Disease;**

I. INTRODUCTION

Heart is the most sensitive organ. Coronary artery disease is the most common disease in today's world ,due to which death rate is increasing .According to the World health organization, 3.7 people worldwide die due to the Coronary artery heart disease(CAD).According to survey ,more than 10 million people in India die due to CAD. Coronary artery disease (CAD) is the heart disease due to which supply of blood to the heart artery gets slowdown [17] .A slowdown of blood supply to the artery causes myocardial infarction. According to the world health organization, it was predicted that more than 16 million Americans were suffering from Coronary artery heart disease. It was established that 50% of males and 30% of females over an age of 40 was suffering from CAD [18]. The major risk factors of CAD are high blood pressure, high blood cholesterol, obesity and diabetes etc. Heart disease can be reduced with the help of prediction system. The prediction systems would be a system which figure out the disease based on various symptoms and risk factors.

As a volume and different variety of the data is increasing day by day, so it is very difficult to manage, analyse and store a huge amount of data .The term big data has great impact of storing, managing and analysing huge amount of data. We can say that the big data is the technology through which we can able to extract the useful information from the data mining algorithm and then store, analyse and manage this knowledge-able data .The Big data is the data whose area and complexity require new architecture, techniques, algorithm and analysis to manage it, extract value and hidden knowledge from it.We can say that big data is the data which contains huge amount of data set. The data can be structured, unstructured and semi-structured. It is difficult to analyse unstructured and semi-structured data in a traditional database like DBMS so the big data tools and technologies are used. As a technology is increasing day by day so a data in every filed is increase in huge amount like in healthcare field it is difficult to manage patient records, to predicate a medicines for particular disease ,predicate a particular disease ,for better health planning and decision making. Big data technologies and tools allow extracting, managing and analysing a knowledge-able data from huge amount of dataset [1].

The medical data can be structured, unstructured and semi-structured. Structured data is data which is easy to analysis and captured and stored as discrete coded values. In case of medical filed, a structured data can be labs and medicine details. A structured and unstructured data can be analysis by using big data technology. The big data analyses are used in many fields like forecasting, medical, government, public sector and education etc. which improves services and applications. Healthcare is the greater origin of unstructured data. A healthcare data consists of electronic health record, clinical trials, health survey, laboratory testing data, genomic data and the data generated by the wearable sensors like ECG's ,MRI etc. These data can be analysis for the predication of future trends and improvement of trails and issues. An x-ray images, CT-scan, MRI and ultrasound etc. are unstructured data. By using big data technology and tools, these unstructured data can be analysed which improves a patient treatment with an effective cost. Big data have a great future scope which renew and recover health and HealthCare. The development of unstructured data in medical sector is increased promptly so there is demanding need to analysis, manage and store a data by using big data tools and techquies.Primarily, the big data tools are open –source .A most generally used tools are Hadoop, MonogoDB and NoSQl etc.[12].

The Big data can be characterized as 5V's (Shown in Figure 1) which are explained as follows:

i) Volume: A volume of data is the data at rest. A volume of data is defined as data in massive amount. A volume of data is increase in terabytes to petabytes then petabytes to zettabytes and so on. In the healthcare sector, data of patients, hospital records are increase in huge amount .The big data technologies helps to capture and analysis a huge volume of data for better

978-1-5386-8025-4/18/$31.00 ©2018 IEEE
DOI 10.1109/ICCS.2018.00038

health planning and decision making, for better e-healthcare, for reduction of diseases and for improvement of patient treatment and also in hospital administration.

ii) Variety: A variety of data is said to be data in many forms. There are three forms of data: Structured, Unstructured and Semi-Structured. A type of data like text, video, audio, click streams and sensor data. As a unstructured and semi-structured are increased day by day in every filed ,so it is difficult to analysis and manage the data in traditional database so to overcome this limitation of traditional database, the big data tools and technologies are used. With a help of different data mining methods are data is categorized.

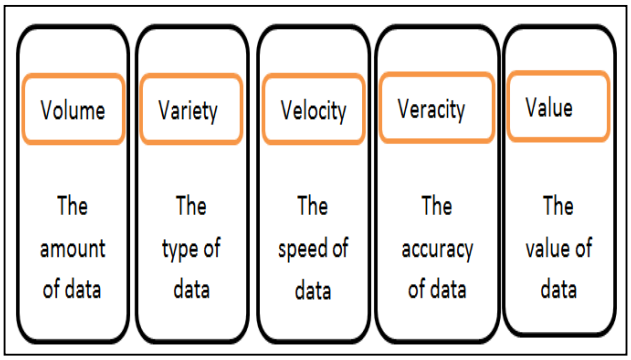

Figure. 1: 5V's of Big Data

iii) Velocity: Velocity of data refers to the data in motion. We can say the velocity is a speed at which data is generated, analysed, managed, transfer and store. To transfer a data from one place to another is depend on a velocity of data. In case of the healthcare sector, to transfer a patient information from one ward to another or from one hospital to another hospital is comes under velocity of big data.

iv) Veracity: Veracity of data refers to the data in doubt. We can say that veracity of the data is truth worthiness, accuracy and consistency of data. In the field of the healthcare, veracity of the big data depends on correctness and accuracy of patient records, health insurance and hospital administration data.

v) Value: Value of data denotes the value derived from the analysis of the data. Big data is of no use unless we can turn it into value [2].

Big data has processed in the form of pipeline which has explained as follows (shown in figure 2):

i) Data generation: It is a first step of data processing pipeline .In this phase data is generated or we can say extracted from different sources to analysis, manage and store.

A data is in database repository. In case of the healthcare sector, data sources of data include patient information, hospital administration information, diseases and medicines records.

ii) Data acquisition: A second step of processing pipeline is data acquisition .Data acquisition consists of three steps: Data collection, Data transportation and Data processing. In a data collection, the raw data is acquire by using different techniques like log files, sensors and web crawlers. From collection of data, the data is transported into data storage architecture for storing and analysis. After data transportation, data is pre-processed so that inconsistencies, noise and redundancy in data has been removed. A pre-processing can be done with a help of various techniques like integration, data cleaning and redundancy removal.

Figure. 2: Big data Pipeline

iii) Data storage: After the process of data acquisition, a data is store in database repository.

iv) Data analysis: A stored data is analysis by using different big data tools and techniques [2].

II.Literature Survey

In [1] author(s) had undertaken the security issues with big data. By using association rule mining, the author(s) surveyed about privacy prevention of data. By using association techniques like Apriori and function point analysis, a data is analysis, secured and protected. Big data denoted as effective computation and database transaction for extensive volume of the data. Data mining is a process of extracting knowledge-able information from huge amount of data set. For a ubiquitous taking out of data, association rule mining technique is used from interaction for database operations. In case of the big data techniques and tools, security plays a vital role. Security of data is a way of preventing data from unauthorized access. In case of security, Data mining algorithm provides high performance, complexity and reduce the cost, .As there is no effective security algorithm exists other than data mining algorithm which provides the security. The Apriori and function point are most effective data mining technique which provides security. According to performance analysis, it demonstrations that function point growth is more accurate technique then Apriori. Big data security is a biggest challenge in a distributed programming. So, association rule mining is best approach which provides security.

- In [2] the author(s) undertaken that big data is the process of analysing non-traditional data by using different data mining and big data tools and techniques. Big data has processed into four steps. Firstly, data is generated or collected from different sources then that data is processed into ETL process of data mining. ETL is the process of extracting data from different sources then transported a data from one place to another and at last a data has loaded into database repository. After a data processing, data is analysis by using different data mining techniques and algorithm so that redundancy, inconsistency and inaccuracy has been removed from processed data by

using data filtering, data compression and error handling. The author(s) had undertaken the use of the big data in every field. Now days, in case of government work, big data is used for new implementation of rules and regulations, in case of healthcare, big data is used to improve a patient treatment, prediction of disease and medicine and improve the hospital administration etc. A big data is used widely but there are some major threat of the big data such as security, privacy and data quality. The data is not secured when it is accessed by some unauthorized person; in case of big data it is a biggest issue to protect data from unauthorized access. Moreover, other threat of the big data is data quality; sometimes it is difficult to remove consistencies, accuracy and redundancy.

In [3] author(s) had undertaken the impacts of the big data in healthcare sector. Big data is a kind of tools and technique which is use for better health planning. In Hospital administration, big data tools and techniques deal with analysis of hospital data. There are various areas where big data plays a vital role. In case of Clinical decision support system, big data tools and techniques help to deal with electronic health records and R&D etc. In case of image processing, big data analysis helps to analysis of X-ray, MRI etc. There are various challenges of the big data in healthcare like protection of the patient data, heterogeneous data and unequal real time processing.

In [4] the author(s) had undertaken that big data problems are very difficult to manage and analysis .In case of the business perspective, a data is divided into volume of data, variety of data and velocity of data. According to business perspective, business faces many problems in big data like in prediction of the power consumption, customer churn analysis, sentiment analysis, and call monitoring and fraud detection. Big data is classified into various types like shown in Table 1.

TABLE 1: BIG DATA CLASSIFICATION

Big data classification	Type
Analysis	Real time Analysis and Batch Analysis
Processing method	Predictive analysis and reporting analysis
Data frequency	On demand feeds, continuous feeds, real time feeds and time series
Data type	Meta data, historical data and transactional data
Content format	Structured, unstructured and semi-structured
Data sources	Web and social media, biometric data ,transaction data

In [5] author(s) had taken the analysis of big data by using different tools and techniques. Big data is data about data .In case of traditional Database. Big data is collection of data which contain structured, unstructured and semi-structured data. There are various challenges among big data like security and privacy of patient data. To analysis big data, various tools are used like MonogoDB, Hadoop, splunk etc. MonogoDB is the cross-platform document oriented database system use to analysis the semi-structured data .It is classified as NO-SQL database which provide high performance, high availability and easy scalability. Hadoop is the most widely used tool to analysis a big data. Hadoop is the batch processing platform which form group of clusters in distributed manner on commodity hardware. Data mining is a way of extraction of knowledge-able data from huge amount of data. There are various data mining techniques used for extraction, classification and storing of a big data. Cluster detection is the way of grouping of data. Cluster can be hierarchical clustering and non-hierarchical clustering. Clustering is undirected knowledge discovery. Memory based reasoning is a way of identification homogenous instances from historical records. It is a method of prediction of unknown samples. Decision based tree is a way of classification and prediction of data.

In [6] author(s) had undertaken that with a help of big data technology in healthcare, we can improve a patient treatment and better understand a source and reduction of diseases. A Framework of the big data in health consists of four layers: Data sources, big data Layer, Analytic layer and application layer. The author also describes use of various tools for big data analysis which can be found in Table 2.

TABLE 2 VARIOUS TOOLS OF BIG DATA ANALYSIS

Name	Developer	Characteristics
No SQL	Various	Data storage, Data Retrieval, Bigger data handling capability, No schema, Lesser server Cost
Hadoop	Apache Software Foundation	Data Processing, Open Source, distributed data replication, data and analysis co-location ,reliable error handling
Spark	Apache Software Foundation	Data Processing, Fault Tolerance ,Dynamic in nature ,real time stream processing, In memory computing,
Watson	IBM	Decision support systems, Cognitive system

In [7] author(s) describes the transformative of data in healthcare brought a lot of objection in a points of data transfer, data storage, computation and analysis. A patient records and hospital historical information can be process manage analysis and store by using machine learning and advanced tools. Big data is most popularly used in business intelligence by processing dementedly huge amount of dataset. Big data is widely considered in a various sectors like healthcare, public sector, retail, manufacture, and personal location data shown in Table 3. The author addresses big data plan consists of three core models: data input, analytic model and decision support tools through which a business value will be improve and increase. The author had undertaken various case studies like Diabetic Link, Brigham Women's Hospital, Asthma polis and Wearable monitors. These case studies improve a healthcare data management in different countries and hospitals.

TABLE 3 VARIOUS SECTOR WHERE BIG DATA TECHNOLOGIES IS USED

Sector	Main usage
Healthcare	Medical and medicines decision support systems, analysis the patient records, predication of the future medicines and diseases, improve the hospital administration, improve and manage the health insurance cases.
Public Sector	Generating clearness by available data, identify the requirements, automatic systems with decision making capabilities to reduce the risk, manage and increase the business plans and performance
Retail	Improve the delivery techniques, improve the work plans and optimize the use of the resources, helps to analysis the store performance, online marketing, product design and research
Manufacturing	Supply chain management ,Value chain management ,Sales management ,online marketing
Personal location data	Disaster management, City management and planning

In [8] author(s) describes the analysis and data archival of HIV/AIDS disease by using technique of K-means clustering algorithm and big data tool Mongo DB.As a number of patients, diseases, doctors and hospitals are increasing data by data ,so it is difficult to analysis the healthcare data. So, to manage, extract and store data –The big data tools and techniques are used. In this paper, author describes a way of analysis of HIV AIDS data with a help of Mongo DB tool. According to world health survey, it was observed that 37 million people are suffer from HIV and in every year death rate is increasing due to HIV-AIDS. It is difficult task to extract ,manage and store the huge amount of the HIV-AIDS data .So ,Mongo DB tool is used for manage ,analysis ,extract and store the huge amount of the data .Big data tool Mongo DB improves the way of treatment ,reduce a cost of treatment ,reduce time and cost. The author had undertaken data mining –Clustering techniques to extract, manage and store the HIV-AIDS data. Clustering is a way of classification and grouping of homogenous data .By using k-means clustering, the author analysis a huge amount of HIV-AIDS dataset.

In [9] author(s) describes analysis and data archival of mental health data by using genetic algorithm and big data tool Mongo DB. Since it is difficult to analysis the mental disease data, so the big data tools and techniques are used. In this paper, the author describes analysis and archival of mental health data by using Mongo DB tool. The author(s) had undertaken Genetic algorithm to extract knowledge-able data. Genetic algorithm is based on optimal solution by randomly selected genetic operators. It is based search-based optimization, genetics and natural selection. The algorithm is based survival of fittest .A algorithm begins with selection process of randomly dataset. A dataset is selected and gathered. After selection process, dataset is crossed over with a gene of another set. After cross over, gene mutation is done for permanent modification. This process involves deletion, insertion and reordering.

In [10] author(s) had undertaken the design and identification of drug for breast cancer by using machine learning, virtual screening, map reduce and mahout. In this paper, the author undertaken the drug discovery and design is the process of searching and identification of particular elements of drug for the particular diseases. By using machine learning algorithm in big data analysis, the drug will be designed for breast cancer .With a help of big analysis techniques like Map Reduce and Mahout. The map reduce is based on map() and reduce() function. Hadoop is an open source platform for storing and analysing huge amount of data in a distributed form of clusters of commodity hardware. Mahout provides java platform but don't provide user interface. Big data analysis and extraction is based on different data mining and machine learning algorithms. Like random forest algorithm, scalable map reduce random forest algorithm, naïve Bayes and complementary naïve Bayes, multilayer perceptron and logistic regression classifier algorithm.

In [11] author(s) had undertaken the analysis of big data of physiotherapy data by using visual analysis. As a technology and healthcare sector is increasing day by day, so an

amount of data and medical devices are increase. It is very complex and difficult task to analysis and stores huge amount of MRI, ultrasound and digital microscopy data. So, with a help of visual analysis, unstructured data is analysis straightforwardly. Visualization is a way of transform of data and information into graphical visual information. In this paper author(s) had undertaken analysis of the physiotherapy dataset for better diagnosis and treatment of the patients.

In [12] author(s) had undertaken the analysis of multi-diseases like cancerous, contagious and diabetes mellitus by using Hadoop. Big data analysis is used in many fields like forecasting, medical and education etc. which improves services and applications. Healthcare is the greater source of unstructured data. Healthcare data consists of electronic health record, clinical trials, health survey, laboratory testing data, genomic data and the data generated by the wearable sensors like ECG's, gyros etc. these data can be analysis for the predication of the future trends and improvement trails and issues.

In [13] author(s) had undertaken the analysis of big data to predicate stage of Diabetes Mellitus by using big data tool Hadoop and data mining algorithm like decision tree. The Author proposed architecture of prediction system of diabetes. The diabetes is a most common disease in today's world. It is difficult to analysis a data so with a help of big data and data mining tools and techniques. Data will be analysis and stored. Hadoop is an open source platform which distributes data in form of clusters of commodity hardware. Hadoop is based on HDFS and map reduce.HDFS is a database repository which stores huge amount of data in form of clusters. Map reduce is based on map() and reduce() function. Based on data mining technique-Decision tree, data can be classified and extracted.

In [14] author(s) had taken the applications of the big data analysis in field of healthcare. The author(s) had undertaken how the healthcare is improving with a help of the big data tools and techniques (shown in Table 4).

Table 4 Areas where the big data is used in healthcare

Various Places	Description
Clinical actions	Big data analysis determines the powerful way of diagnoses of the patient treatment through the clinical research from huge amount of the data sets
Research and development field	With the help of the predication and classification of the historical records of the patients and diseases, we are able to design the new drugs, tools and algorithms for better treatment of the particular diseases.
Remote controlling	Obtain and analyses the movement of the data form one hospital to other hospital.
Fraud detection and analysis	Manage identity and analyze fraud records and trails in the clinical data, patient records and hospital administration.

In [16] author(s) had undertaken a framework and methodology of big data analysis in the healthcare sector. With help of predicative analysis, the healthcare organization accuracy will be increase and improve. With the help of the predictive analysis, financial and clinical decisions issues can be predicted with a help of system. Advantages of data analysis in healthcare are:

- Will be completely transforming future of healthcare, medicines and hospital administration.
- Improve healthcare sector effectively by improve way of treatment or care.
- Reduce the treatment cost.
- Ability to extract information from different sources like genetic data, clinical records etc.
- Helps in disaster management.

Big data analysis has potential to detect diseases at preceding stage, so that the precautions of that disease would be taken as soon as possible. There are various places where big data analysis improve and use for healthcare shown in Table 4[16].

In [18] author(s) had undertaken the survey according to world health organization of the Coronary artery disease. This paper describes various data mining techniques which help in prediction of heart disease at early stage so that reduce a rate of increase of heart diseases. Data mining helps to develop medical system which is able to detect heart disease using large data set of patients. There are various data mining techniques are naïve Bayes, genetic algorthrim ,artificial intelligence, decision tree, clustering algorithm like k –means clustering algorithm. The healthcare filed has a huge amount of data which is difficult to mine and analysis for detection of heart disease. A massive of amount of the data is mined or extracted from the records of patient suffering from heart disease, and then extracted data is examined .Data mining is the way of the extracting the useful information called knowledge and big data is the massive amount of the data. These are two different terms but task carried out of data mining and big data are similar .These involves extraction of the huge amount of data, handling the data, analysis of data and store a data. Association rule, Neural network and genetic algorithm are used detect a heart disease. Bayesian classification and association rule is used to predict a coronary artery disease. Intelligence heart disease prediction system is designed by using decision tree, naive Bayes and neural network. Heart attack prediction system is design by using k-means algorthrim and MAFIA algorithm.

.

III. Conclusion

In this paper, we have undertaken analysis of big data for prediction of coronary artery heart disease by using data mining and machine learning algorithm that is K-nearest algorithm and genetic algorithm. A data set of patient is collected from different sources, then pre-process and predicts a heart disease based on symptoms and risk factors. From this paper, we can conclude that through this predicted system we are able to predict any disease at early stage and this system is able to be use in medical field for physician to easily analysis and prediction of particular heart disease.

IV. Future Scope

For Future work, we are able predict any disease and reduce the data set of patient by improve an accuracy of KNN and genetic algorithm. This improves hospital administration, patient re-admission in hospital and improves a way of diagnosis and treatment.

References

[1] P. A. Patel, "A Survey Paper on Security Issue with Big Data on Association Rule Mining," no. March, pp. 108–109, 2017.

[2] S. Bajaj and R. Johari, "Big data: A boon or bane - The big question," *Proc. - 2016 2nd Int. Conf. Comput. Intell. Commun. Technol. CICT 2016*, pp. 106–110, 2016.

[3] D. K. Thara, B. G. Premasudha, V. R. Ram, and R. Suma, "Impact of big data in healthcare: A survey," *Proc. 2016 2nd Int. Conf. Contemp. Comput. Informatics, IC3I 2016*, vol. 5, pp. 729–735, 2016.

[4] S. J. Divakar Mysore, Shrikant Khupat, "Big data architecture and patterns, Part 1: Introduction to big data classification and architecture," pp. 1–7, 2013.

[5] V. Bhardwaj and R. Johari, "Big data analysis: Issues and challenges," *2015 Int. Conf. Electr. Electron. Signals, Commun. Optim.*, pp. 1–6, 2015.

[6] M. Sheeran and R. Steele, "A framework for big data technology in health and healthcare," *2017 IEEE 8th Annu. Ubiquitous Comput. Electron. Mob. Commun. Conf.*, pp. 401–407, 2017.

[7] J. A. Patel and P. Sharma, "Big data for Better Health Planning," *Adv. Eng. Technol. Res. (ICAETR), 2014 Int. Conf. IEEE.*, pp. 0–4, 2014.

[8] P. Dhaka and R. Johari, "HCAB: HealthCare Analysis and Data Archival using Big Data Tool," 2016.

[9] P. Dhaka and R. Johari, "Big Data Application : Study and Archival of Mental Health Data , using MongoDB," pp. 3228–3232, 2016.

[10] R. M. Constantine and M. Batouche, "Drug discovery for breast cancer based on big data analytics techniques," *2015 5th Int. Conf. Inf. Commun. Technol. Access.*, pp. 1–6, 2015.

[11] S. V Dugani and S. Dixit, "Physiotherapy data analysis of big data in healthcare applications," *2017 Int. Conf. Innov. Mech. Ind. Appl.*, no. Icimia, pp. 506–511, 2017.

[12] S. A. Sofi, "improvement in Healthcare," pp. 1–9, 2015.

[13] S. T. Prasad, S. Sangavi, A. Deepa, F. Sairabanu, and R. Ragasudha, "Diabetic data analysis in big data with predictive method," *2017 Int. Conf. Algorithms, Methodol. Model. Appl. Emerg. Technol.*, pp. 1–4, 2017.

[14] G. Srini and H. T. C. Global, "Applications of Big Data Analytics and Applications of Big Data Analytics and," no. March 2015, pp. 9–13, 2016.

[15] https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

[16] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, 2014.

[17] https://www.healthline.com/health/coronary-artery-disease#risks

[18] M. K. Homoud, "Coronary Artery Disease," pp. 1–13, 2008.