ISSN No:-2456-2165

Big Data-Awareness in Healthcare

Aishwarya HD , Pooja P
Department of Information Science and Engineering
School of Engineering and Technology, Jain University
Bangalore, India

Sowmya KN, Manjunath CR
Department of Information Science and Engineering
Department of Computer Science and Engineering
School of Engineering and Technology, Jain University
Bangalore, India

Abstract:- As the volume of information creating is expanding step by step in this web world, the term Big Data is turning into an extremely prominent popular expression in the present market. Enormous Data is used as a piece of different areas of the web world. Scientific researches these days looked to exceptionally monstrous information handling, which devour moderately an excessive amount of time and exertion. The prospect of Big Data can be utilized for better wellbeing planning. Its techniques can be utilized for medicinal services information investigation which helps in better basic leadership to expand the business esteem and client premium and to give eHealth administrations. Enormous Data strategies can connect to create frameworks for the early determination of sickness, investigation, repeats, tranquilize disclosure in malignancy. An attempt is made to analyse all the technologies of Big Data used in different stages of cancer treatment as an example to show how it can be used to create an awareness in healthcare.

Keywords:- *Big Data, Virtual Screening, Map Reduce.*

I. INTRODUCTION

Big Data is set of information that are so huge and complex that conventional data processing application software are lacking to deal with them. Challenges in Big Data include challenges of the data, storing the data, analysis of the data,search,sharing, transferring, visualization, querying, updatin g, information privacy and data sources. A number of concepts are associated with big data: originally it had only 3 concepts volume, variety, and velocity. Other concepts later attributed with big data are veracity (i.e., how much noise is in the data) and value.

Fig 1:- Concepts of Big Data

Big Data will definitely open new openings and enable breakthrough related to, among the others medicinal services information examination tending to alternate points of view: (i) what happened answered by the descriptive analysis, (ii) the reason why it happened answered by the diagnostic analysis, (iii) what will happen to be understood by the predictive analysis and (iv) detect how we can make it happen using the prescriptive analysis.

Fig 2:- Data patterns including hindsight, insight and foresight

In today's world cancer is one of the most deadliest disease. Thousands of people die as its victim. Among the types of cancer breast cancer is the one that usually occurs only in women. It is increasing day by day in the developing countries. This is because very less amount of data is available for its analysis. When you're educated that you have bosom malignancy, it's normal to ponder what may have caused this illness. However the reasons for bosom growth is obscure. Specialists only here and there know why one lady creates bosom growth and another doesn't, and most ladies who have bosom malignancy will never have the capacity to pinpoint a correct reason. The bosom growth is more likely to create in ladies with certain hazard factors than others. A danger factor is something that may grow the likelihood of getting a disorder. Some risk factors (absence of physical exercise, less than stellar eating routine, being overweight or stout, drinking liquor, radiations to the chest) are avoidable. Be that as it may, most hazard factors, (for example, having a family history of bosom malignancy) can't be maintained a strategic distance from.

A portion of the signs and side effects of bosom disease consolidate an alteration in the size or shape, a knot or locale that feels thicker than the straggling leftovers of the bosom, a change in skin surface, for example, puckering or dimpling,

ISSN No:-2456-2165

redness or rash on the skin or potentially around the areola, liquid that begins from the areola without squeezing, torment in your bosom or your armpit that is there all or constantly and swelling in your armpit or around the collarbone.
Big Data is a term utilized for a gathering of educational accumulations so broad and complex that it is difficult to process using conventional applications/instruments. It is the data surpassing Terabytes in measure. In light of the variety of information that it incorporates, enormous information dependably conveys various difficulties identifying with its volume and intricacy. A present overview says that 80% of the information made on the planet are unstructured. One test is the means by which these unstructured information can be organized, before we endeavour to comprehend and catch the most vital information. Another test is the way we can store it. A portion of the best advancements utilized to store and break down Big Data are Apache Hadoop, Microsoft HD Insight, NoSQL, Hive, Sqoop, Poly Base, Big data in EXCEL, Presto.

The healthcare industry is going to be more advanced with the emerging Big Data technology. Some of the big data applications in medical field are: Healthcare Intelligence

- Monitoring Patient Vital
- Fraud prevention and Detection
- Smoother Hospital Administration
- Big Data to Fight Cancer

Big Data is also considered in various domains of the internet world. Healthcare is also moving towards Big Data to take the advantages using its advanced tools and methodologies. Big Data also assists you in better health planning. as the years passed huge amount of data was being generated and hence more advanced technology like Big Data supporting any tools and technologies came into existence. To control the mortality of chronic diseases it is required to understand the causes, effects and risk factors of such diseases. For example the ability of Big Data to combine the heterogeneous datasets and through predominant predictive modelling, statistics and advanced algorithms it is possible to gain the insights of HATS(HIV/AIDS Tuberculosis and Silicosis).

Breast cancer is one such disease where the enough amount of data is not available for its detection, prediction and analysis. Big Data can be used as one of the technology to create and awareness for this disease.

II. BIG DATA IN ONCOLOGY

There is an immense volume of instructive records gathering data on tumor genome, transcriptome, clinical data and a great deal in oncology. The probability of handling data from these collected data provoke an adjustment in harm calm treatment and result.

It is acknowledged by the researchers and doctors that change in the disease treatment relies upon a superior learning about malignancy science. The age of huge information in the disease comes chiefly from the high throughput advancements used to think about the "Omics" sciences. The as of late utilized innovation that permits DNA and RNA sequencing much quick and economically than the beforehand utilized strategy is the cutting edge sequencing (NGS), or high throughput sequencing. Roughly 4TB of crude information can be produced by an undertaking including around 10-20 whole genome sequencing as determined in [1]. This issue is the "information storm" issue. Transcriptomic examination is made of high-throughput strategies, like continuous quantitative PCR (qPCR), microarrays and RNA-Seq (NGS sequencing). Proteomics is the investigation of a particular proteome, containing information on the protein structure and the capacity, protein articulation profiling, their varieties and changes, meaning to know cell forms.

The National Institutes of Health (NIH) made a brain boggling wonder in 2006. The Cancer Genome Atlas (TCGA) Data Portal was started as a three-year pilot wander. Their essential point was to make a colossal and broad aggregation of data of changes that occur specifically illness makes. The target treatment relied upon the use of a drug that goals especially a genomic change. Another essential wander of NIH is the Roadmap Epigenomics Mapping Consortium. The essential purpose of this is to make an open resource of human epigenomic data. This empowered the scientists to make imperative disclosures and clinicians to upgrade the remedy, in perspective of the epigenetic features of tumors.

The enormous information venture Cancer LinQ, from ASCO, enabled the casualties and doctors to share the information about the medicines and the results. It makes a persistent cycle of discovering that begins and finishes with the patient itself. One critical note is that patients are anonymized in databases to ensure their character. Patients and their specialists can contribute and pick up from the aggregated information living in the database. The patient's treatment and result can be enhanced by the doctors with the assistance of CancerLinQ database.

Apollo stage pointed in making of more firm framework for institutionalizing long haul gathering of patients clinical history and information got from their natural examples.

III. ANALYSIS AND PROGNOSIS OF CANCER

Information Standards in diseases look into the have been advancing impressively. Specialized Standards and information organization are the necessities of the Diagnosis and Prognosis with enormous difficulties. Cancer, a destructive disease can be dissected with the incalculable genomic information and moral investigation with enormous information functionalities as its essence. Regularly NGS advancements creates the exome information for examination. A solitary Cancer patient's exome information ranges from 10 Gigabytes to

15 Gigabytes as specified in [2]. Big Data computational models must break down this biggest scope of Omics information.

Petabytes of information is required for tumor investigation. Different clinical and diagnostic databases with high dimensional scaling and elucidation are the requirements of this. These databases are libraries of data that are gathered from different logical investigations, high throughput computational examination furthermore, production literature. High dimensional data are mostly expected to be scaled by the prerequisite of Big data and it computational advancements. The examination strategies being joined with Big Data yields an outcome with the investigation and the forecast over the survival rate of the patient. This examination also includes suggestions for the line of treatment with the suitable drug, along with the side-effect and the toxicity of the prescribed drugs and the individual person correct, medication based on the past medical history.

The classification model is used to create the Omic data. On receiving new patient's data the Classifier uses the classification model and the patient information to classify whether the patient is suffering cancer or not. If the person is found to have cancer the personalized drug will be suggested by the inference engine by accessing the knowledge-base and the clinical information. If the person is already entering into the prescribed drug system then it checks whether the prescribed drug is toxic or not.

Big Data analytics thus transforms the way of solutions in healthcare with its efficient and dynamic analysis. The diagnosis is well run by the supervised principal component analysis and the Cox regression model explores the prognosis with its high interpretation over huge number of data.

IV. PREDICTION OF BREAST CANCER USING MAPREDUCE

The traditional data analytic may not be that capable to handle the huge amounts of the data. Because of this rapid growth in the data solutions need to be contemplated and taken in order to handle and extract value and knowledge from these data. Capability to expand the bits of knowledge from this varying and rapidly evolving information should be possessed by the decision makers. This can be done by utilizing the big data analytic which is the using of advanced analytic techniques on the big data using MapReduce approach. [3] paper examines to construct a high performance platform to efficiently analyse big SEER(Surveillance, Epidemiology and End Results) breast cancer data set using Map Reduce to find the recurrence of breast cancer for a breast cancer patient.

As specified in [3] to evacuate inadmissible cases the data were pre-processed. The final data set was constructed after utilizing the data cleansing and data preparation strategies. Finally, SEER dataset were analyzed for breast cancer recurrences happened in the initial 5 years after breast cancer treatment. The dataset were cleaned by handling missing values, noise, identifying and rectifying the inconsistencies by using Expectation maximization (EM) method. An arrangement of theories was produced by a proposed calculation and they were acquired through weighted greater part voting of the classes anticipated by the individual speculations. Occurrences drawn from an iteratively refreshed dissemination of the preparation information were utilized to produce the theories via preparing a trail classifier. This distribution was updated so that instances which were not classified properly by the previous hypothesis were more likely to be included in the training data of the next classifier.

Let the data set Dn={ (x1, y1) , (x2, y2) , ….., (xm, ym)} with label classification yi ∈ {Recurrence (R), Non-Recurrence (NR)}; xi ∈ X is the object or instance; The algorithm initialize all the records with weight, so that $D1\ i = 1\ n$ for all the examples in Dn, where t ∈ [1, T] and T is the total number of iterations. These weights are uniformly initialized before beginning the first iteration also, in every consecutive iteration they are updated. At each iteration, a base learner function is applied to the weighted form of the data which then returns an optimal weak hypothesis ht . The weak hypothesis is used to minimize the weighted error. At each iteration, the weak classifier is assigned a weight (αt). At the end of T iterations, the algorithm returns the final classifier H which is a weighted average of all the weak classifiers.

The Amazon EC2 was used to deploy the experiments and Weka software tool was utilized to experiment with this algorithm. The efficient estimation from incomplete data was accomplished by adopting Expectation maximization (EM) algorithm. In any inadequate dataset, there is indirect evidence about the probable estimations of the unobserved values.

The results of the experiment conducted in [3] concluded that for experimenting 2,20,811 instances and 17 attributes were used for determining the classification accuracy. It is specified that a confusion matrix where in 25,259 of 2,20,811 records were characterized vaguely. Among them 11,021 of the "Recurrence" cases were classified as "Non-Recurrence" and 14,270 of the "Non-Recurrence" were classified as "Recurrence".

The experimental results showed that the error rates were more accurate ones and were smaller in predicting the recurrence of the breast cancer.

V. CLASSIFICATION OF BREAST CANCER

The breast cancer is specific type of cancer in women only. It is curable but the survival rates are considered if it is diagnosed and prevented in the early stages. For diagnosis, prevention and prediction of the breast cancer it is significant that the available patient data must be analysed properly in order to extract the proper cause. Also, Machine Learning approaches which have been identified as the best ways to classify the breast

ISSN No:-2456-2165

cancer dataset are Support Vector Machine and Relevance Vector Machine.[4] paper investigates the Wisconsin Breast Cancer Data through Big Data approach using Hybrid SVM-RVM Model as classifier. [4] aims to develop a hybrid model based on the SVM and RVM and to analyse the performance of proposed model through Big Data analytics in the comparison to the nadve strategies.

The Support Vector Machine (SVM) is a non probabilistic binary and a non linear statistical tool that is based on the supervised learning. The data is analysed, the patterns are recognized and the data is classified based on the common attributes by using the kernel tricks. The function of SVM is to minimize the structural risk instead of the objective function. It minimizes the bound on generalization error for all the data which were not used during the training.

The Relevance Vector Machine (RVM) is a statistical tool for data classification based upon Bayesian estimation. It has been widely used on various types of cancers other than breast cancer and is known to deliver the optimal results by using only few training samples.

Wisconsin Breast Cancer dataset archived by UCI Machine Learning Repository is freely available dataset to study the breast cancer. The dataset is collection of multivariate data of fine needle aspirate (FNA) of a breast mass of donors obtained at University of Wincosin Hospitals.

The outcome of the results as specified in [4] considered two parameters viz. classification accuracy and implementation time. The results revealed that the SVM model was 93.10% accurate, RVM model was 94.01% accurate and the Hybrid SVM-RVM model was 96.41% accurate. The time taken by these models were specified by the results that is SVM took 18 micro seconds, RVM took 16 micro seconds and Hybrid SVM-RVM took 7.5 micro seconds.

Taking these results into consideration its was concluded that Hybrid SVM-RVM model outperformed the nadve SVM and RVM model.

Classification efficiency of the breast cancer datasets is being improved by frequently using SVM and RVM techniques of Machine Learning. This research has explored a Hybrid SVM-RVM model through an innovative approach of the Big Data. The application of this model with Big Data to real time data may yield better and disease specific results. Also this model may be explored for other type of cancer as well.

VI. DRUG DISCOVERY FOR BREAST CANCER

Research on tranquilize revelation has accomplished a place where it has no other choice other than utilizing HPC and Big Data Processing Systems to accomplish its goals in sensible time allotments, Virtual Screening (VS) is considered as a champion among the most computationally thought and generous process which expect a vital part in illustrating new drugs and that must be done as snappy as conceivable keeping in mind the end goal to viably dock ligands in enormous the databases to a given protein receptor. Then again, a standout amongst the most hazardous illnesses on the planet is the bosom tumor; in excess of 1.5 million new cases are analyzed every year, with in excess of 400 thousands passings. As indicated in [5] the quantity of bosom disease cases and the demise because of this were most extreme in the year 1980 and 2010 in both the created and creating nations.

The medication inquire about for the bosom tumor should be possible by the above insights which are imperative. The way toward seeking and ID of medications is the medication disclosure and plan. The revelation of new medications depends on various methodologies or procedures and is additionally a to a great degree monotonous and extensive process. In this so as to achieve an exceptionally chose number of hopefuls a great many particles are must be chosen and prepared. Specialists in the pharmaceutical fields must pick between restricted choices to make the fundamental strides yet using computational strategies like Virtual Screening. The basic purpose of this technique i.e. Versus is to choose if a plan of ligands attach to a given protein receptor or not, if yes, scientist wind up familiar with how immovably.

Because the focus was on the breast cancer protein the receptor 4JLU was selected. The dataset was constructed by selecting 104 ligands randomly and the virtual screening was performed by Autodock Vina. The variation in the results is shown in the graph. Media was chosen as a separation point. All the ligands were converted into the finger print format (FPT). Open Babel, which is an open source chemical toolbox that permits transformations between chemical structure formats, was utilized to perform the change. The dataset was used to train five models, after the completion of the data construction phase. Calculations intended to be utilized as a part of the setting of enormous information over the stage Hadoop/Map Reduce were utilized.

As specified in [5] the classification results by utilization different machine learning algorithms in the context of Big Data were RF was 72%, NB was 68%, CNB was 63%, LRSGD was 65% and ANN MLP was 74%.

When these algorithms were compared with the precision of the ensemble of classifiers it was found that RF was 71%, NB was 68%, ANN MLP was 73% and EC was 78%.These results proved that we could reach upto 80% of the precision.

VII. AWARENESS IN HEALTHCARE

A great deal of difficulties tones of information exchange, stockpiling, calculation and investigation has been acquired by the exponential advancement of information in human services. For the social insurance utilization and applications, abundant

ISSN No:-2456-2165

patient information and authentic information, which encase rich and huge bits of knowledge that can be uncovered utilizing propelled devices and procedures and additionally most recent machine learning calculations. However, new enormous information investigation structure is required by the size and velocity of such awesome dimensional information.

One of the extreme leading and rising industry is the healthcare industry. With plenty of difficulties big data opportunities have been changed into the healthcare. The Big data here refers to the compilation of the data from different people. The data consists of diseases, varying symptoms, medicines, diet, exercises, prescription, lab reports, nurses and patients. With older tools and technologies this data could not be processed beneficially. Some of the new technologies that are discovered to overcome this problem are Hadoop(Distributed File System, Open Source, Low cost, Simple Coherency Model, Fault Tolerant etc),SAP-HANA(Column Storage, Compression Techniques, Parallelization, Data Locality, High Availability of data etc) and Spark(Event Driven Architecture, Better Parallelism, Resilient Distributed Datasets, Uses in-memory computing etc).

The healthcare industry generally has produced a lot of information, driven by record keeping, consistence and administrative necessities, and patient care. The present pattern is towards quick digitization of these a lot of information. Enormous Data in the social insurance alludes to electronic wellbeing informational indexes so tremendous and complex that they are hard to deal with the conventional programming or the equipment. Big Data is overpowering likewise due to its assorted variety of information composes and the speed at which it must be overseen. The total of data related to patient healthcare and well being make up “big data” in the healthcare industry. Enormous information investigation applications in social insurance exploit the blast in information to remove bits of knowledge for settling on very much educated choices, and as an exploration class are alluded to as, large information examination in human services. The objective was to describe the promise and potential of big data analytics in healthcare.

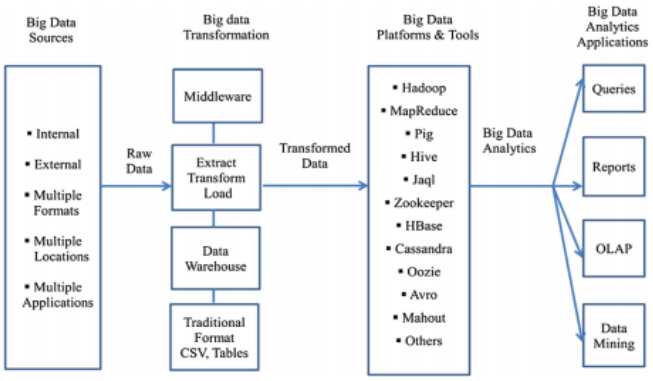

Fig. 3. An applied conceptual architecture of big data analytics

From the above research we get to know the nascent field of big data analytics in healthcare, discussions of the advantages, outlines and the architectural framework and the methodology. They provided an expansive diagram of huge information investigation for the social insurance analysts and specialists. Enormous information examination in human services is developing into a promising field for giving understanding from huge informational collections and enhancing results while decreasing expenses. Its potential is extraordinary; however there remain difficulties to overcome.

One of the primary techniques which is considered as the fundamental strategies to always take care of the expanding demand on IT assets which is forced by enormous datasets is the MapReduce. This is on the grounds that the greatly parallel and conveyed execution on expansive PC hubs is empowered by the high versatility of MapReduce worldview. It is an intense programming and basic model that encourages simple advancement of adaptable parallel applications to process enormous measures of information on extensive bunches of ware machines. Utilizing this the developer require not need to stress over the execution subtle elements of parallelism and adaptation to non-critical failure rather he just needs to think as t how to embrace the issue to the model.

Considering all the above factors we get a clear idea that a disease in order to be detected, analyzed or prevented must have sufficient amount of data. Usually these data if available they exists in large sets. Using proper technology like Big Data, we can create an awareness in healthcare.

VIII. CONCLUSION

Nowadays people are more cognizant about their prosperity and are set up to contribute huge measure of their compensation towards social protection to achieve better prosperity organizing workplaces. As the present world is a business world and everything is found in the business perspective, the restorative administrations ventures are in like manner moving in the method for expanding predominant business regards and points of interest by giving better social protection workplaces in light of a honest to goodness worry for the clients. An effort is made to pass on various points of interest those human administrations systems can make by using Big Data advancement. The extended number of usages in human administrations using Big Data is making the cases for Big Data advancement.

The Big Data techniques and abilities are utilized for social insurance information investigation, giving eHealth administrations, to create frameworks for the early diagnosis of the illnesses, to avoid and control ceaseless infections and furthermore to give protection and security to customized medicines.

Till now all the researches were done as to how Big Data can help only a particular stage of a disease.This paper gives knowledge to the upcoming scientists and the physicians as to how Big Data can be used in every stages of a disease and hence create an awareness in healthcare.

REFERENCES

[1]. Carmela Dantas Barbosa, "Challenges with Big Data in Oncology", Journal of Orthopedic Oncology , 2016.

[2]. Manju.K.K, Mrs.Srinitya.G, "Analysis and Prognosis of Cancer with Big Data Analytics", International Journal in Applied Science and Engineering Technology, January 2016.

[3]. Umesh D.R , B.Ramachandra, "Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach" , International Journal of Computer Applications (0975-8887) Volume 150-No.7, September 2016.

[4]. Dr.Savita Kumari Sheoran, "Breast Cancer Classification Using Big Data Approach", Indian Journal of Research, Volume-7 , January-2018.

[5]. Rostom Mennour, Mohamed Batouche, "Drug Discovery for Breast Cancer Based on Big Data Analytics Techniques",IEEE,2016.

[6]. Prof.Jigna Ashish Patel, Dr Priyanka Sharma, "Big Data/or Better Health Planning", IEEE International Conference on Advances in Engineering and Technology Research (ICAETR-2014) , August 2014.

[7]. Wullianallur Raghupathi, Viju Raghupathi, "Big Data Analytics in Healthcare:Promise and Potential", Health Information Science and Systems,2014.

[8]. David Becker, Trish Dunn King, Bill McMullen, "Big Data, Big Data Quality Problem", IEEE International Conference on Big Data (Big Data),2015.

[9]. Thara D K, Dr.Premasudha B G, Ravi Ram V, Suma R, "Impact of Big Data in Healthcare:A Survey", 2nd International Conference on Contemporary Computing and Informatics, 2016.