BIG DATA ANALYTICS IN HEALTHCARE: BENEFITS, TECHNOLOGY AND CHALLENGES

Monal H. Jain[1], Mehul P. Barot[2]
[1]Research Scholar, Computer Engineering Department, LDRP-ITR, KSV University
[2]Assistant Professor, Computer Engineering Department, LDRP-ITR, KSV University

Abstract

Health care sector is growing tremendously from last few decades. Huge amount of data is being continuously generated by health care sector that has huge volume, enormous velocity, massive variety and complex values. This information is a form of "big data," so called for its sheer volume, its complexity, diversity, and timeliness. Also it comes from a variety of new sources as hospitals are now tend to implemented electronic health record (EHR) systems. These sources have strained the existing capabilities of existing conventional relational database management systems. Big Data solutions obtain more meaningful and knowledgeable information from these massive, heterogeneous and complex data sets. Healthcare stakeholders now have access to these promising new threads of knowledge. Pharmaceutical-industry experts, payers, and providers are now beginning to analyze big data to obtain insights. The technologies that able to deal with "Big Data" offer many opportunities to the health care sector. Big data analytics can be used for diagnosis of cancer, to improve cardiovascular care, in radiation oncology treatment. Predictive methodology can be used for diabetic data analysis.

Index Terms: Big Data, Healthcare, technology, benefits, challenges.

I. INTRODUCTION

Big Data is a term used to describe large collections of data (also known as datasets) that may be unstructured, and grow so large and quickly it is difficult to manage with regular database or statistics tools. Big Data is getting generated everyday by diverse segments of industries like business, finance, manufacturing, healthcare, education, research and development etc [4]. The traditional DBMS's and RDBMS's in market are incapable of storing such vast amount of data. We cannot take full advantage of the hidden knowledge and information from this data as the traditional data mining algorithms do not work effectively on this enormous data. So there is need of developing and using effective, innovative tools and technologies offered by Big Data [4]. The analytics of Big Data leads to any organization towards better decision making and strategic steps. Giant companies in sectors like retail, manufacture and government agencies are using Big Data to meet their business and strategic objectives. The Big data analytics also plays a vital role for small and medium size industries to capitalize their business [5].

A. Characteristics of Big Data

Big Data refers to collection of structured and structured data. It is mainly characterized by volume, velocity and variety which is as shown in below figure 1.

Figure1: Characteristics of Big Data[11]

Industry analyst Doug Laney originally coined the concept of Big data while referring to the challenge of data management. According to that, there are three important dimensions of the Big data concept known as 3 V's of Big Data which is illustrated in figure 2 below. Today, many organizations are gathering, storing, and analyzing huge amounts of data. These data is known as a Big data as it has Volume, Velocity and Variety [5].

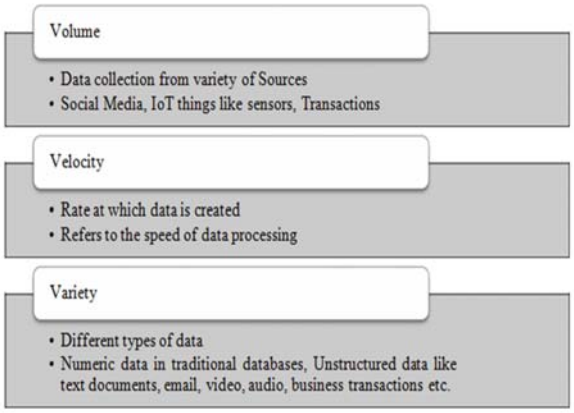

Figure2: Three V's of Big Data [5]

Data Volume: The Big word in Big Data itself defines the volume. At present the data existing is in petabytes and is supposed to increase to zettabytes in nearby future. Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can ace ss it.[12]

Data Velocity: Velocity in Big Data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of the incoming data but also speed at which the data flows and aggregated [12].

Data Variety: Data variety is the measure of the richness of the data representation-text, images, videos, audios etc. Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web pages, web Log files, social media sites, email documents [12].

Complexity: Complexity measures the degree of interconnectedness and interdependence in big data structure such that a small change in one or few elements can yield very large changes or small change that can ripple across or cascade through the system and substantially affects its behavior or no change at all[12].

There is variety of sources from where the Big Data is generated. Social Media sources such as Facebook, Instagram, Twitter generates terabytes of data on every day. Machines such as desktop computer, laptop generates tremendous amount of data. Geospatial data is generated by cell phones and even from satellites. The IoT (Internet of Things) devices like sensors, pocket computers are also generating a massive amount of data. Data is also generated as an output of research projects like Large Hadron Collider (LHC) at CERN in Switzerland and France output an enormous amount of data - over 200 petabytes.

B. Big Data Analytics

Big Data analytics is the process of exploring huge data sets that may contain a variety of data types to reveal hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Big data analytics has emerged from two distinct concepts: big data and analytics. Several sectors are benefited from the Big Data analytics like The Financial Services Industry, The Automotive Industry, Supply Chain, Logistics, and Industrial Engineering, Retail, Health care, Entertainment etc. Combining Big Data with Analytics leads to any organization towards many tasks like determining root causes of collapses, issues and defects in near-real time, generating coupons based on the customer's buying habits at the point of sale, recalculating entire risk portfolios in minutes, detecting fraudulent behavior before it affects your organization etc.[5]

Pharmaceutical-industry experts, payers, and providers are now beginning to analyze big data to obtain insights. Big Data analytics could collectively help the industry address problems related to variability in healthcare quality and escalating healthcare spend. For instance, researchers can mine the data to see what treatments are most effective for particular conditions, identify patterns related to drug side effects or hospital readmissions, and gain other important information that can help patients and reduce costs.[1]

Many innovative companies in the private sector—both established players and new entrants—are building applications and analytical tools that help patients, physicians, and other healthcare stakeholders identify value and opportunities.[1]

C. Challenges in Big Data

- **Work on Reliable Data:**

With the explosion in the volume of available data, the challenge is how to separate the "signal" of "data" and "valuable information". Unfortunately, at this point, a lot of companies have difficulty identifying the right data and determining how to best use it. The fight against "spam data" and data quality is a crucial problem. Companies must think outside of the box and look for revenue models that are very different from the traditional business [13].

- **Data access:**

Data access and connectivity can be an obstacle. McKinsey survey shows that still a lot of data points aren't yet connected today, and companies often do not have the right platforms to manage the data across the enterprise [13].

- **Embedding increasingly complex data:**

If the Big Data was first concerned the "simple" data (tables of numbers, graphs …), the processed data is now more and more complex and varied: images, videos, representations of the physical world and the living world. It is, therefore, necessary to rethink and reinvent the big data tools and architectures to capture, store and analyze this data diversity [13].

- **Better integrate time variable:**

The time dimension is also an important challenge for the development of Big Data, both to analyze causalities in the long term than to treat accurate information in real time in a large data flow. Finally, the problem also arises in terms of storage. The volume of created data will exceed the storage capacities and will require careful selection [13].

- **Security:**

Keeping such vast lake of date secure is a big challenge itself. But if companies limit data access based on a user's need, make user authentication for every team and team member accessing the data and make a proper use of encryption on data, we can avoid a lot of problems [13].

- **The image of data:**

An algorithm is no longer enough. To look big data head on, the visual experience must be in line with the expectations and limits of a variety of audiences; data scientists, marketers, or HR professionals. Visualization experts are currently grappling with a challenge, both in the graphic rendering of the data and in the development of tools to access the information. That challenge is to use images to turn data into knowledge and break down technical barriers for the future of innovation and research [14].

- **Data quality:**

To start out on the big data adventure, we first need to identify and regularly "clean up" the data we want to work with. Given the immense volume and variety of data available, the quality of the data we use has to be a priority. Databases contain all sorts of errors—often human errors—that need to be rectified for the requisite standards to be met. On that basis, a data quality audit should be the first undertaking in any big data project. To offset the problem, we can also put data governance initiatives in place to assess the relevance of each item of data, and set up automatic correction tools [14].

- **Processing Data:**

Data are everywhere and are never presented in the same manner; each database has its own format. For some experts, the big issue in big data is more one of processing than of volume. Automated data management and processing solutions have become indispensable to promote the intelligence of the data, just like the possibility of analyzing the data in real time[14].

- **The humanity of data:**

The last, and unique, challenge faced by companies in relation to the uses they intend to make of their data is this: respecting the human behind the data. The individual is not a mere statistic. Above all, the purpose of the data must be to make a connection with the individual, and to intelligently guide us in the relationship between those who produce the data—the customers—and those who use it; the companies[14].

D. Applications of Big Data Analytics

- **Healthcare:**

The big data is in extended use in the field of medicine and healthcare. It is a great help for even physicians to keep track of all the patients'

history. The link to the patient's history can be accessed only by the patient and his particular physician. A large number of medical devices are there which are big data oriented. Today data is used to such an extent that doctor prescribes the medicines without even visiting the patient by knowing the heartbeat and temperature through the heart and temperature monitoring watch fitted on the patient's hand that stays in a remote place.[15]

- **Public Sector:**

Big data provides a large range of facilities to the government sectors including the power investigation, deceit recognition, fitness interconnected exploration, economic promotion investigation and ecological fortification. Big data is even used to examine the food based infections by the FDA. Big data results are fast which outputs to quicker well-being. Also in the investigation of a huge volume of communal complaints uses the big data analytics. This same analytics are utilized in the course of health check statistics in urgency and resourcefully for quicker pronouncement manufacture and to become aware of mistrustful or falsified declarations.[15]

- **Insurance Services:**

Big data is the technology tool that is being used in the production to offer purchaser insights for see-through and simpler commodities, by finding out and foreseeing buyer behavior from side to side information obtained from internet websites including the social media as well as CCTV video recording. The big data as well enables for the better purchaser preservation from insurance agencies. In the claims administration, extrapolative big data business analytics has been utilized to provide more rapid service given that enormous quantity of information can be worked on particularly in the countersigning period [15].

- **Industrial and Natural Resources:**

The high demand of the natural sources on this earth is challenging the high volume as well as the velocity of big data. Similarly, a great quantity of data commencing the built-up industry is unexploited. The unused data avoids advanced eminence of merchandise, power competence, dependability, and improved income boundaries. In the natural wealth industry, big data enables for analytical modeling to sustain judgment creation that is used to consume and incorporate huge amounts of information from geographical information, graphical information, manuscript and chronological statistics [15].

- **Banking Zones and Fraud Detection:**

Big data is hugely used in the fraud detection in the banking sectors. In banking sectors as the big data is implemented, it finds out all the mischief tasks done. It detects the misuse of credit cards, misuse of debit cards, archival of inspection tracks, venture credit hazard treatment, business clarity, customer statistics alteration, public analytics for business, IT action analytics, and IT strategy fulfillment analytics [15].

- **Communication, Media and Entertainment industry:**

Organizations in this industry simultaneously analyze customer data along with behavioral data to create detailed customer profiles that can be used to create content for different target audiences, recommend content on demand, and measure content performance [16].

- **Transportation:**

Individual use of big data includes route planning to save on fuel and time, for travel arrangements in tourism etc. Private sector use of big data in transport includes revenue management, technological enhancements, logistics and for competitive advantage (by consolidating shipments and optimizing freight movement). Governments use of big data includes traffic control, route planning, intelligent transport systems, congestion management (by predicting traffic conditions).[16]

- **Energy and Utilities:**

Smart meter readers allow data to be collected almost every 15 minutes as opposed to once a day with the old meter readers. This granular data is being used to analyze consumption of utilities better which allows for improved customer feedback and better control of utilities use. In utility companies the use of big data also allows for better asset and workforce management which is useful for recognizing errors and correcting them as soon as possible before complete failure is experienced.

E. Big Data Analytics in Health Care

One of the characteristic that health care sector possesses is its data richness. With the development in diagnostic and treatment, health

care sector evolved so quickly in last few decades. There are many sources in this sector from where the data is generated. These data is undoubtedly in the form of Big Data. The data came from many sources and categorized as follows:

- **Web and social media data:** Data captured from Facebook, Twitter, LinkedIn, blogs, and the like. It can also include health plan websites, smart phone apps etc.

- **Machine-to-machine (M2M) device generated data:** readings from remote sensors, meters, and other devices.

- **Biometric data:** Data may in form of retinal scans, x-ray images, finger prints, genetics, handwriting, other medical images, blood pressure and other similar types of data.

- **Human-generated data:** In the form of unstructured and semi-structured data. Some of the examples are EMRs, Doctor's notes and paper documents.

- **Genomic Data:** data in the form of DNA sequences.

Big Data analytics in Healthcare is fundamentally a set of methodologies, procedures, frameworks and technologies which are used to transform raw data into meaningful as well as useful information. These set of information are used to make decision making tasks more effective whether they are strategic, tactical & operational. Figure3 depicted the key components playing a role in Big Data analytics for health care sector.

As per the figure 3, data produces from variety of sources like hospitals, medical groups, payers or other data providers. These data first needed to be aggregated. Moreover, many processes like extraction, cleaning, conformation, transformation and loading are executed on data during this phase. Finally, some meaningful and useful information are generated which can be used by variety of users and purpose as shown in figure 3.

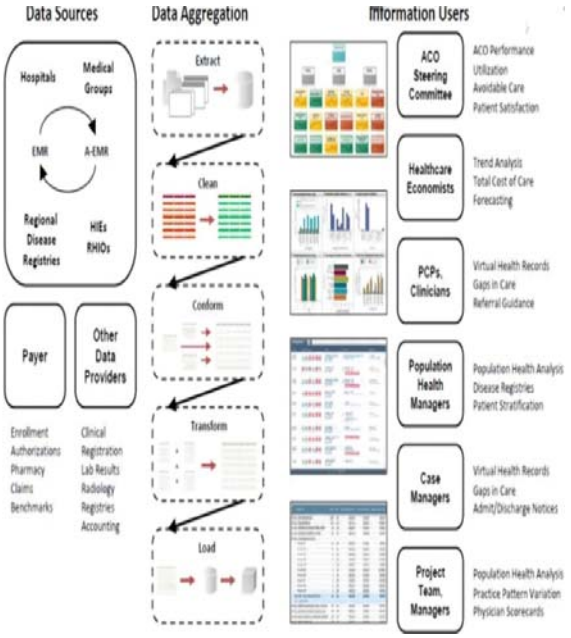

Figure 3: Big Data Analytics in Health Care: Key Components [5]

F. 5 V's of Big Data in Health Care

The 5 V's of Big Data relevant to Healthcare are[4]:

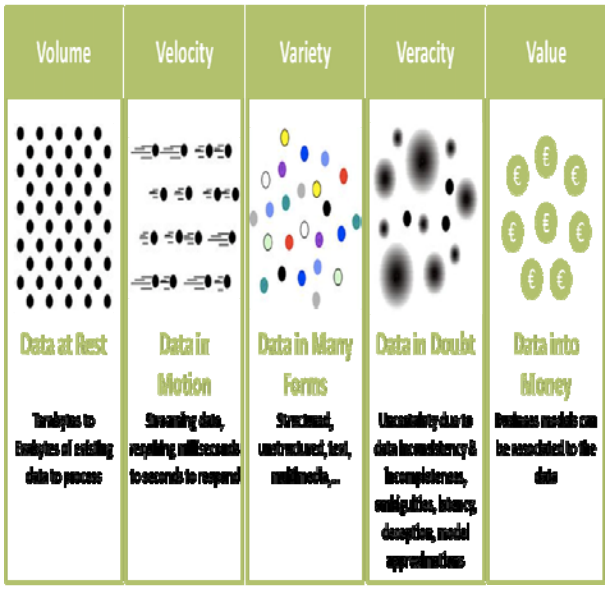

Figure 4: 5 V's of Big Data in Health Care[10]

Volume: As mentioned earlier, healthcare industry generates prodigious data at staggering rate. The report from EMC and the research firm IDC anticipates an overall increase in health data of 48 percent annually. According to the report the volume of healthcare data in 2013 was 153 Exabyte and it may increase to 2,314 Exabyte by 2020.

Variety: In the past, the emphasis had been on generating clinical data for patients with similar symptoms, storing and analyzing it to derive the most effective course of treatment for the admitted patient. Now the healthcare industry is focusing on complete healthcare, by providing an effective treatment through analysis of a patient's data from various other sources too. This refers to variety.

Variety: The varied health care data generally falls into one of the three categories i.e. structured, semi structured and unstructured. Generally the following data is gathered: clinical data from Clinical Decision Support systems (CDSS) (physician's notes, genomic data, behavioral data, data in Electronic Health Records (EHR), Electronic Medical Records (EMR)), machine generated sensor data, data from wearable devices, Medical Image data (from CT scan, MRI, X Ray's etc), medical claim related data, hospital's administrative data, national health register data, medicine and surgical instruments expiry date identification based on RFID data, social media data like twitter feeds, Facebook status, web pages, blogs, articles.

Velocity: It refers to the speed at which new data is generated and moves around. The sensor devices and wearable collect real time physiological data of patients at a rapid pace or velocity. This new data being generated every second poses a big challenge for data analysts. Social media data also adds to velocity as the users views, posts, feeds scale up in seconds to enormous amount in case of epidemics/national disasters.

Veracity: It refers to trustworthiness of data. Analyzing such a voluminous, variable and fast paced data is a hurricane task. There is no scope for error especially in critical healthcare solutions where patient's life is at stake. The primary aim is to ensure the data is reliable. It is quiet difficult to ensure the reliability of unstructured data for e.g. usage of different terms/abbreviations for disease/symptoms, misinterpreted prescriptions due to bad handwriting, false/fake comments on social portals, improper readings from faulty machines/sensors, incomplete/inaccurate data filled by patients etc.

Value: It refers to the quality of data. Normally data from EMR's and EHR's are recognized as high value data. But it is hard to ascertain the value of data from social media. The effective analysis of high value data can lead to better quality, effective healthcare solutions and innovations.

II. TECHNOLOGY

A. Predictive Analysis[2]

Predictive model transforms large clinical data sets, or "big clinical data," into actionable knowledge.[3]

Predictive Analysis System Architecture

The architecture of predictive analysis system includes various phases like data collection, data warehousing, predictive analysis, processing analyzed reports. Figure 5 shows the complete architecture of proposed method.

Data Collection

The raw diabetic big data or data set is given as input to the system. The unstructured voluminous input data can be obtained from various Electronic Health Record (EHR) / Patient Health Record (PHR), Clinical systems and external sources (government sources, laboratories, pharmacies, insurance companies etc.), in various formats (flat files, .csv, tables, ASCII/text, etc.) and residing at various locations.

Data Warehousing

In this phase massive unstructured data warehoused into single unit in which, data from various sources is cleaned, accumulated and made ready for further processing. Integration of various EHRs can help in identifying the patterns for diabetes prediction system.

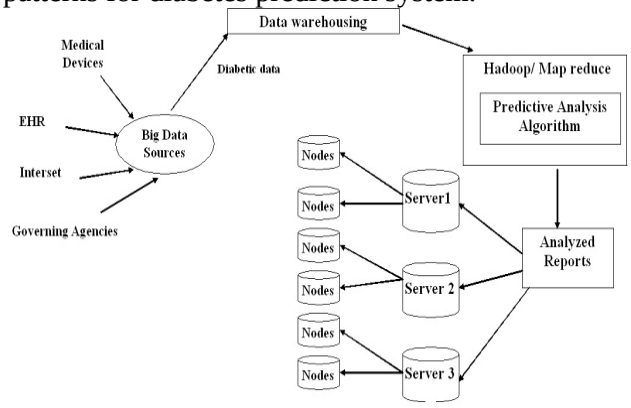

Figure 5: Architecture of the predictive analysis system-Health Care Applications

Predictive Analysis

Predictive analysis can help healthcare providers accurately expect and respond to the patient needs. It provides the ability to make financial and clinical decisions based on predictions made by the system. This system uses the predictive analysis algorithm in Hadoop/Map Reduce environment to predict and classify the type of DM, complications associated with it and the type of treatment to be provided.

Hadoop:

Hadoop is the open-source distributed data processing platform from Apache. Hadoop can serve the twin roles of data organizer and analytics tool. Hadoop has the potential to process extremely large amounts of health data mainly by allocating partitioned data sets to numerous servers like clusters, each of which solves different parts of the larger problem and then integrates them for the final result. Hadoop uses two main components to do its job:

Map/Reduce and Hadoop Distributed File System.

• Map/Reduce: Hadoop's implementation of Map/Reduce is based on programming models to process large data or datasets by dividing them into small blocks of tasks. Map/Reduce uses distributed algorithms, on a group of computers in a cluster, to process large datasets. It consists of two functions:

- ✓The Map ( ) function which resides on the master node and then divides the input data or task into smaller subtasks, which it then distributes to worker nodes that process the smaller tasks and pass the answers back to the master node. The subtasks are run in parallel on multiple computers.

- ✓The Reduce ( ) function collects the results of all the subtasks and combines them to produce an aggregated final result — which it returns as the answer to the original big query.

• Hadoop Distributed File System (HDFS): HDFS replicates the data blocks that reside on other computers in the data center (to ensure reliability) and manages the transfer of data to the various parts of the distributed system.

Pattern discovery:

For diabetic treatment it is necessary to test the patterns like, plasma glucose concentration, serum insulin, diastolic blood pressure, diabetes pedigree, Body Mass Index (BMI), age, number of times pregnant.

The pattern discovery of predictive analysis must include the following:

• Association rule mining- Association between diabetic type and pages viewed (e.g. laboratory results)

• Clustering- clustering of similar patterns of usage, etc.

• Classification- Classification of health risk value by the level of patient health condition.

• Usage of statistics

• Application of pre-defined deductive rules across data

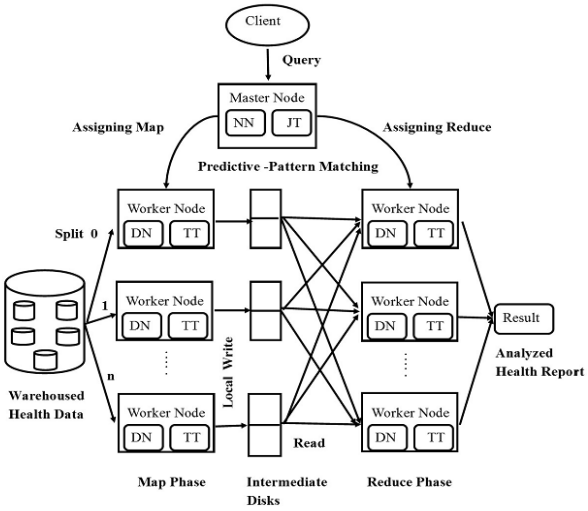

Figure 6: Predictive – Pattern Matching System

Predictive Pattern Matching:

Whenever the warehoused dataset was sent to Hadoop system, immediately the map reduce task is performed. In mapping phase, the Master Node splits large data into smaller tasks for numerous Worker Nodes. Figure 6 deploys the exact operation of predictive pattern matching system. The Master node is one consists of Name Node (NN) and Job Tracker (JT), which always employs the map and reduce task. The Worker Node or Slave Node receives the order from the Master Node, process the pattern matching task for diabetes data with the help of Data Node – Same Machine (DN) and Task Tracker (TT). The predictive matching is the process of comparing the analyzed threshold value with the obtained value. If the pattern

matching process was completed by all Worker Nodes based on the requirement, it was stored in intermediate disks. This process is known as local write. If the reduce task was initiated by Master Node, all other allocated Worker Nodes will read the processed data from intermediate disks. Based on the query received from Client through Master Node, the reduce task will be performed in Worker Node. The results obtained from the reduce phase will be distributed in various servers.

Processing analyzed reports

After the analysis of large diabetic data went through Hadoop, the final results are distributed over various server and replicated through several nodes depending on the geographical area. By employing proper electronic communication technology to exchange the information of individual patients among health care centers will leads to get proper treatment at right time in remote locations at low cost.

Benefits of this predictive analysis system

The diabetes may associate with severe diseases such as heart attacks, strokes, eye diseases and kidney diseases, etc. Analyzing the risk value by the level of patient health condition using above results of can be used by the physicians at remote locations to serve the people. Detecting diseases at earlier stages can help to be treated more easily and effectively. In developing countries such as India, it is mandatory to manage specific individual and population health and detecting health care fraud more quickly. The middle-income families can be with the high availability of medical facility at minimum cost. This system leads to the improved focus on every individual patient health. Thereby we can reduce and save our next generation from diabetic mellitus.

B. Machine Learning Methods[8]

Machine learning techniques are class of artificial intelligence (e.g., neural networks, decision trees, support vector machines), which are able to emulate human intelligence by learning the surrounding environment from the given input data and can detect nonlinear complex patterns in such data. Machine learning methods such as support vector machine, artificial neural networks, and deep learning can be used in radiation oncology.[9]

Neural Networks

Neural networks were extensively investigated to model post-radiation treatment outcomes for cases of lung injury and biochemical failure and rectal bleeding in prostate cancer.

Support Vector Machines

Support vector machines (SVMs), which are universal constructive learning procedures based on the statistical learning theory. For discrimination between patients who are at low risk versus patients who are at high risk of treatment, the main idea of SVM would be to separate these two classes with ‘hyper-planes’ that maximize the margin between them in the nonlinear feature space defined by an implicit kernel mapping .

Bayesian Networks

A BN provides graphical representation of the relationships between the variables represented as nodes in a directed acyclic graph (DAG), which encodes the presence and direction of relationship influence among the variables themselves and the clinical endpoint of interest. The relationship between parent and child nodes is modeled by conditional probabilities using Bayes chain rule. These methods are also robust for variable uncertainties and missing data, which would make them excellent candidates for clinical applications.

III. BENEFITS OF BIG DATA IN HEALTH CARE

1. For individual’s/patients[4]**:** Generally while deciding any line of treatment for a patient, historical data(of a set of similar patients) about the symptoms, drugs used, outcome/response of different patients is taken into account. With help of BDA, the move is towards formulating a personalised line of treatment for a patient based on his genomic data, location, weather, lifestyle, medical history, response to certain medicines, allergies, family history etc. When the genome data is fully explored some kind of relation can be established between the DNA and a particular disease. Then the specific line of treatment can be formulated for this subset of individuals. The patients will benefit in following ways:

- Correct and effective line of treatment.
- Better informed health related decisions.
- Preventive steps can be taken in time.
- Continuous health monitoring at patients place using wearable wireless devices.
- Designing personalised line of treatment for patient.
- Increase in life expectancy and quality.

2. For Hospitals[4]**:** By using effective BDA techniques on the data available the hospitals can reap following benefits:

- Predict the patients which are likely to stay for longer time or going to be readmitted after the treatment.
- Identifying patients that are at risk for hospitalization. Thus, health care providers could develop new healthcare plans to prevent hospitalization.
- Various questions can be answered by analysing the data using BDA tools and techniques like will a patient respond positively to a particular treatment?, Is a surgery required to be done on patient and will he/she respond to it?, Is the patient prone to catching disease after treatment?, what is his likelihood of getting affected from the same disease in near future?
- The hospital authorities can take better informed and managed administrative decisions. Like if no. of patients are not getting cured early and no. of readmissions are increasing because patients become ill again after treatment, then diagnose the root cause of problem, hire more competitive and experienced staff, invest in better drugs/instruments which can aid in effective treatment, increase the cleanliness of hospital, make the treatment more timely, engage more staff on floor, plan for more frequent post treatment follow ups etc

3. For Insurance Companies[4]**:** A large amount of expenditure is done by governments for giving medical claims to patients. By using BDA we can analyse, identify, predict and minimize the possible frauds related to medical claims.

4. For Pharmaceutical company[4]**:** By using BDA techniques effectively, the R&D can help to produce in a shorter time ,drugs/ instruments/tools that are most effective for treating a specific disease.

5. For government[4]**:** The government can use demographic data, historical data of disease outbreak, weather data, data from social media over disease keywords like cholera, flu etc. They can analyse this massive data to predict epidemics, by finding correlation between the weather and likely occurrence of disease. Therefore preventive measures can be used for avoiding the same. The BDA can help in improving the public health surveillance and speed up the response to disease outbreaks.

6. Analysis on data gives insight to health care providers to determine populations risk for illness[5].

7. With the advent of big data analytics , it is easy to capture , analyze and capture patients symptoms earlier to offer a preventive care in a better way.

IV. CHALLENGES OF BIG DATA IN HEALTH CARE

1. Unstructured data and provenance of data :As has been discussed before, the BDA process uses data from varied sources. Most of the data is unstructured data like medical prescriptions, blogs, tweets, status updates, and comments. We need to generate right metadata for this data and transform it into a structured format. The image and video data should be structured for semantic content and search. Provenance of data along with its metadata should be carried through data analysis process so that it is easier to track the processing steps in case of error.Some intelligent processing techniques should be devised to handle the data input from sensors and wearables in memory. This will help to filter/derive the meaningful data, which can then be stored on permanent storage. Therefore it will save space.

2. Missing or incomplete data: It is seen through practices that the patient tends to hide some of the personal facts or choices about his/her lifestyle while filling up forms or oral interview by physicians. When this data is stored

in digital format then many fields remain empty. Sometimes it happens that some of the fields carry wrong values. When analysis is done on the entire data such empty or wrong fields may or may not get processed. In both the cases they produce wrong results. If we are leaving some records as they are empty then our analysis is not on cumulative data. If we consider wrong value fields then the again the analysis is incorrect and unreliable. Such issues need to be addressed.

3. Quality of data: We need processes to ensure that data from sources is a valid data or not and is of good quality. Determining the validation and quality of data in case of social media data is another big challenge.

4. Technical challenges:

- Data aggregation from different database systems is also a challenge in BDA. It can be made easier by devising certain standard database design practices meant for a specific domain like healthcare, financial sector etc.We need to come up with a lot more technological standards and protocols for different data systems to integrate seamlessly.
- The traditional algorithms for data mining processes or analysis have to be scaled up to handle the big volume of data. Another aspect is parallelisation of algorithms, the processors speed has come to a point beyond which it's hard to increase. Therefore the trend is moving towards multi-core processors. In such a scenario we need statistical algorithms which can be parallelised, else there computing performance will decrease when they handle complex big volume data. Apart from this scaling complex query processing techniques to terabytes, while enabling interactive response times is another big problem.
- An analysis is useful only when a non technical person is able to understand and interpret it. Considering the volume and variety of data used in BDA, it is very hard to depict it visually in a more understandable and easier way. Going one step further than this, a user should be able to perform the repeated analysis with different set of assumptions, data sets and parameters. It will help the user to better understand the analysis process and verify whether the system works in a desired way or not.
- With market flooded with so many open source and proprietary platforms and tolls for BDA, we need careful evaluation to use the best platform and tool for our problem.

5. Data Security: As more and more data is digitized, the role of securing data is gaining importance. Large no. of people are not willing to share their personal data with a fear of security breach. In case they are assured of full security then this problem can be tackled. We need strict Government policies and norms which should be adhered to regarding what data can be shared and what not. In addition to this, strong technological hardware and software level security measures should be implemented to discourage the hackers and malicious people.

6. Lack of Experts: There is a dearth of qualified and experienced data scientists in the world. We need to create an expertise in the field of data science, so that in future we have data scientists who can help to turn the promises of big data into reality.

V. CONCLUSION

Thus Health care sector is growing tremendously now. Huge amount of data is being generated known as big data from different sources such as electronic health records, electronic machine records, sensors, wearable devices etc. This data can be analyzed using different methodology like machine learning methods for oncology treatment, algorithms such as predictive analysis method for diabetic data analytics. While analyzing and collecting big data different challenges such as unstructured data, missing data, privacy etc. need to be addressed. By analyzing big data of health care sector effective line of treatment can be provided to patients, their treatment cost can be reduced, preventive steps can be taken so that disease cannot get worse.

VI. REFERENCES

[1] Peter Groves, Basel Kayyali, David Knott, Steve Van Kuiken,(2013), The

'Big Data' revolution in health care:Accelerating value and innovation.p.9.

[2] Dr Saravana kumar N M , Eswari T , Sampath P,Lavanya S(2015), Predictive Methodology for Diabetic Data Analysis in Big Data, 2nd International Symposium on Big Data and Cloud Computing.

[3] Gang Luo(2016), PredicT-ML: a tool for automating machine learning model building with big clinical data, Health Information Science and Systems, Vol. 4, Issue 5.

[4] Jasleen Kaur Bains (2016), Big Data Analytics in Healthcare- Its Benefits, Phases and Challenges , International Journal of Advanced Research in Computer Science and Software Engineering , Vol. 6, Issue 4.

[5] Sanskruti Patel , Atul Patel(2016), Big Data Revolution in Health Care Sector: Opportunities, Challenges and Technological Advancements, International Journal of Information Sciences and Techniques, Vol. 6, No. ½.

[6] Manju.K.K, Mrs.Srinitya.G(2016), Analysis and Prognosis of Cancer with Big Data Analytics, International Journal for Research in Applied Science & Engineering Technology, Vol. 4, Issue I.

[7] John S. Rumsfeld, Karen E. Joynt,Thomas M. Maddox(2016), Big data analytics to improve cardiovascular care: promise and challenges, Nature.

[8] Issam El Naqa(2016), Perspectives on making big data analytics work for oncology, Methods.

[9] Jean-Emmanuel Bibault, Philippe Giraud, Anita Burgun (2016), Big Data and machine learning in radiation oncology: State of the art and future prospects, Cancer Letters.

[10] https://www.google.co.in/search?q=5+v%27s+of+big+data&espv=2&source=lnms&tbm=isch&sa=X&ved=0ahUKEwi_odfQgcDTAhVKuY8KHf8QBxQQ_AUIBigB&biw=1280&bih=694#imgrc=QVy3hcQK43XQpM:

[11] https://www.google.co.in/search?q=characteristics+of+big+data

[12] https://www.slideshare.net/HarshMishra3/harsh-big-data-seminar-report

[13] https://dzone.com/articles/challenges-of-bigdata

[14] https://blog.outscale.com/en/the-five-major-challenges-of-big-data

[15] https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/

[16] https://www.simplilearn.com/big-data-applications-in-industries-article