ISSN (e): 2250 – 3005 || Volume, 07 || Issue, 12 || December – 2017 ||
International Journal of Computational Engineering Research (IJCER)

Use Of Big Data Analytics In Lung Cancer Data Set

Sangram Keshari Swain[1]
[1]Associate Professor, Department of Computer Science & Engineering, School of Engineering & Technology, Bhubaneswar Campus Centurion University of Technology and Management, Odisha, India

ABSTRACT

Big data is a declaration used to recognize the database whose area is the potential of typical database software tools to store, organize and examine. Big data have demonstrated a new path towards the world. Cancer is a tumor of diseases involving abnormal cell growth with the possibility to occupy or disperse to other parts of the body. It has been a deadline for millions of people. Every cancer is particular and the medical treatment is complex. The data of each cancer patient are too large and it varies from one person to another person. This paper will traverse censorious tools and tackle that accelerate knowledge, locating along with the cancer study continuity.
Keyword: Big data, Lung cancer, Methodology

Date of Submission: 25-11-2017 Date of acceptance: 14-12-2017

I. INTRODUCTION

Cancer is a multifaceted disease. It is also familiar that a variable is present within a solitary patient. It is clear that actual alteration Collect mutually unlike unaccompanied cancer cells. Our existing mechanization for the cancer biomarker finding is normally proficient of identifying a snapshot, but not the energetic and long-term modification of the cancer landscape. Big data can block this barrier by assembling dissimilar molecular characteristic at the DNA, RNA, and protein and metabolite proportion. Bio data-mining has guided those working in the existence sciences to accept the information area, in order to assist the barrage of big data produced by next-peer biotechnologies, such as next-peer sequencing, proteomics, and metabolomics, as well as the arranged and disarranged medical and healthcare statistics from computerized well-being evidence.

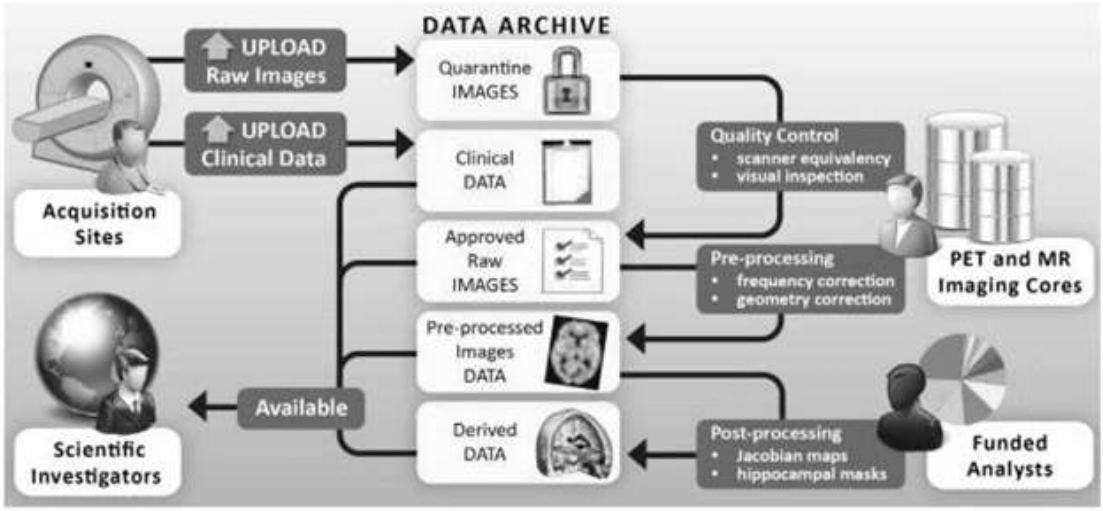


Figure 1: Process of Data flow

Big data definitely throws dare on cancer research in different condition. The trends of organizing different platforms and their findings create a junction among different sectors and measurement, which is ultimately a step towards an entire understanding of the disease. Nowadays, next-generation syncing, paramedics and metabolomics develop fast in different directions. However, they also create a difficult question regarding the

existing platform of using a single feature to measure cancer. A multifaceted approach to cancer research is urgently needed. On the other hand, there is a multitude of data, not only huge in volume and complicated in structure but also vast in dynamic scale and depth. We for sure that a large amount of loosely connected, naturally noisy and miscellaneous data may also be collected in the databases. All data do not contain useful data, we cannot judge whether the information is useful or junk. Thus, certain standards, protocols or extension may be needed to develop useful data in the same database.

Some international corporation and institutions have made achievement to set guidelines to guarantee accuracy and endurance so as to authorize the suitability of the data. Electronic health records are becoming a better resource for databases. The information can be changed and together via telemedicine and mobile connectivity. Patients‟ information and data security become an issue and more security part needs to be in place. We are in the mid of medicine era, and precise decision making based on individual, detailed profiles is a pressing need. Tumor molecular profiling has enabled the subdivision of cancer to reconsider treatment system. Big data are exactly sailing in the same boat with correct medicine. Nowadays, more and more databases have been ingrained, e.g., the directory of actual variation in Cancer (COSMIC), which is the largest database of actual variation and their effects on human cancer. It is envisioned that in the years to come, we will step into a remodel era of applying big data to our cancer patients.

II. BIG DATA

There is a requirement of different approaches, techniques, tools & architectures in Big data for fixing new and old problems in a superior manner.

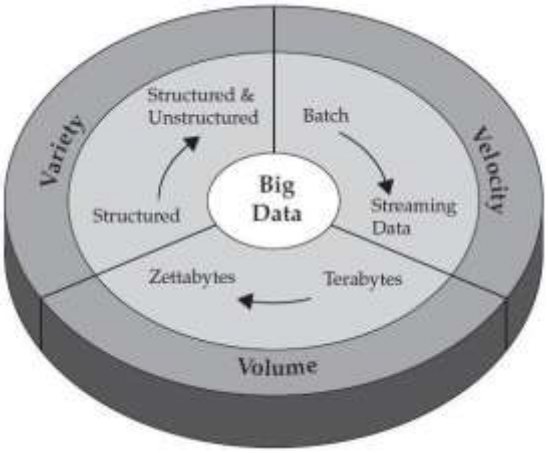


Figure 2: Big data

Key components for the advancement of Big Data are, the increase of storage space, the boost of processing power and availability of information. Big data is distributed storage system built on the Google File System. Big Data is the word used to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies.

III. LUNG CANCER

Lung cancer is the most common cancer among men from the Indian subcontinent and is the number one killer of men dying due to any cancer related affliction. For women the incidence of Lung Cancer ranks ninth among all other cancers. Considering all the cancers among men and women together, Lung Cancer ranks number four in India.

A. Symptoms of Testing Lung Cancer-

The most common symptoms of lung cancer are:

- A cough for three weeks or more
- A change in a cough you've had for a long time
- A chest infection that doesn't get better, or repeated chest infections
- Feeling breathless and wheezy for no reason
- Coughing up blood
- Chest or shoulder pain that doesn't get better
- A hoarse voice for three weeks or more.

Other possible symptoms are:
• Losing weight for no obvious reason
• Feeling extremely tired (fatigue)
• The ends of fingers change shape – they may become larger or rounded (clubbing).

B. Methods of Testing Lung Cancer-
1. Chest x-ray - If you haven't already had one, you will have a chest x-ray to check your lungs for anything that looks abnormal.
2. CT (computerised tomography) scan - Most people will have a CT scan. Depending on your symptoms, you may still have one even if your chest x-ray has not shown any signs of lung cancer. A CT scan takes a series of x-rays, which build up a three-dimensional picture of the inside of the body. The scan takes 10–30 minutes and is painless. It uses a small amount of radiation, which is very unlikely to harm you and will not harm anyone you come into contact with. CT scans can also be used to guide a biopsy, in which a small amount of tissue is taken to be examined under a microscope.
3. Bronchoscopy - A bronchoscopy is a test where a doctor or specially trained nurse looks at the insides of the airways (bronchus) and lungs. A tube called a bronchoscope is used and the test is carried out under local anaesthetic.
4. Lung biopsy - This test is done in the x-ray department, usually during a CT scan for guidance. You'll be given a local anaesthetic first to numb the area. The doctor asks you to hold your breath for a few seconds, while they pass a thin needle through the skin and into the lung. They check the CT or X-ray picture to make sure the needle is in the right place. The doctor removes a sample of cells from the tumour (biopsy). These are examined under a microscope for signs of cancer. The biopsy may be uncomfortable, but it only takes a few minutes.

Further tests:
Some tests may be repeated during and after your treatment.
1. Mediastinoscopy - This test allows the doctor to look at the area in the middle of your chest called the mediastinum and the nearby lymph nodes. These are the first areas that lung cancer may spread to. The surgeon makes a small cut in the skin at the base of your neck and passes a tube like a telescope through the cut into your chest. The tube has a light and camera at the end and that magnifies the areas it looks at. The doctor can see any abnormal areas and take samples of the tissue and lymph nodes (biopsies) to check for cancer cells.
2. Thoracoscopy - This allows the doctor to look at the pleura and other structures around the lungs. You can have it done under a general anaesthetic. It can also be done with a local anaesthetic to numb the area and a sedative to make you drowsy. The surgeon makes a small cut in your chest wall and passes a tube called a thoracoscope (like the one we describe in a mediastinoscopy) into your chest. Your doctor can then take a biopsy of the pleura. Sometimes, doctors use a video camera to get a better view of the area surrounding the lung. This is called video-assisted thoracoscopy.
3. Endobronchial ultrasound scan (EBUS) - This test may be done instead of a mediastinoscopy or thoracoscopy. Some people may have this test instead of a bronchoscopy or a CT scan and biopsy. You can have it under a general anaesthetic, or using a mild sedative to help you to relax and feel drowsy. The doctor passes a bronchoscope, which has a small ultrasound probe on the end, down into your windpipe (trachea). An ultrasound uses sound waves that are converted into a picture by a computer.
4. Endoscopic ultrasound (EUS) - This is similar to an EBUS and is also sometimes done as an earlier test for lung cancer. While you are under a general anaesthetic or mild sedation, the doctor will pass a small, flexible tube (endoscope) through your mouth and into your gullet (oesophagus). An ultrasound probe on the end of the endoscope creates pictures of the area around the heart and lungs. It can show if any of the lymph nodes in the centre of the chest are enlarged.
5. PET/CT scan - This is a combination of a CT scan, which takes a series of x-rays to build up a three-dimensional picture, and a positron emission tomography (PET) scan. A PET scan uses low-dose radiation to measure the activity of cells in different parts of the body. PET/CT scans give more detailed information about the part of the body being scanned. You may have to travel to a specialist centre to have one.
6. MRI (magnetic resonance imaging) scan - This test uses magnetism to build up a detailed picture of areas of your body. The scanner is a powerful magnet, so you may be asked to complete and sign a checklist to make sure it's safe for you. The checklist asks about any metal implants you may have, such as a pacemaker, surgical clips or bone pins. You should also tell your doctor if you've ever worked with metal or in the metal industry, as very tiny fragments of metal can sometimes lodge in the body.
7. Abdominal ultrasound scan - Ultrasound uses sound waves to look at the liver and other parts of the body in the upper abdomen. Once you are lying comfortably on your back, a gel is spread onto the area to be scanned. A small device that produces sound waves is passed over the area and the sound waves are converted into a picture

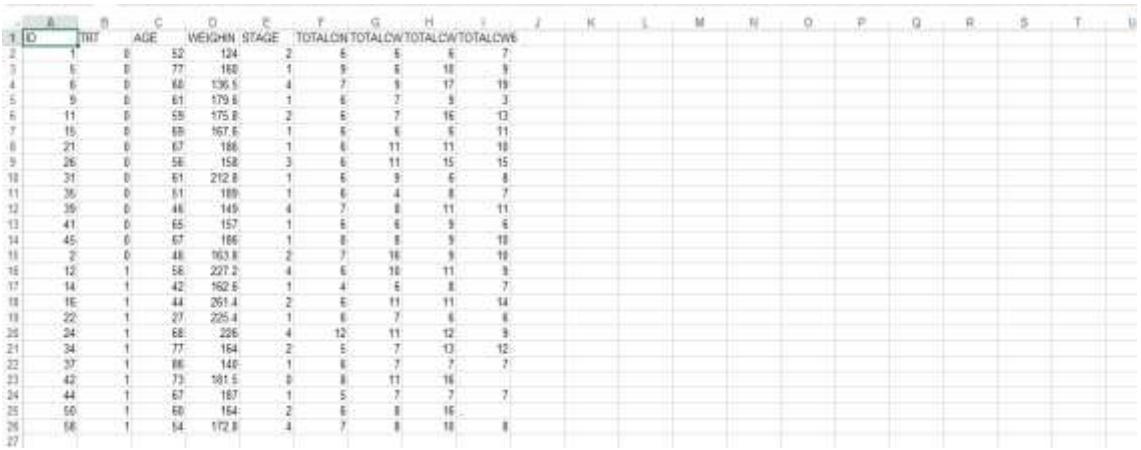
by a computer. The test only takes a few minutes. An ultrasound scan is also sometimes used to look at the lymph nodes in the neck.
8. Bone scan - This test shows abnormal areas of bone. You have a small amount of a mildly radioactive substance injected into a vein. The level of radioactivity used in the scan is very small and doesn't cause any harm. You wait for 2–3 hours after the injection before you have the scan, which may take an hour. Abnormal bone absorbs more radioactivity than normal bone and shows up on the scan pictures.
9. Lung function tests - If your treatment involves surgery or radiotherapy to try to get rid of the cancer, your doctor will arrange breathing or exercise tests for you. You have these to see how well your lungs are working.

B. Methodology of Big Data Analytics
Step 1:
Building the concept statement you need the use of big data analytics established on the "4Vs".
Step 2: Plan
1. Why big data analytics is used?
2. Why is it important?
3. What is the problem that is addressed?
4. Backdrop Material
Step 3: Methodology
1. Selection of variable
2. Outcome & insight
3. ETL and data revolution
4. Propositions
5. Platform/tool collection
6. Conceptual model
7. Data collection
8. Clustering, allocation, Analytic techniques -Association, etc.
Step 4: Deployment
1. Evaluation & validation
2. Testing

IV. CANCER DATA SETS

The dataset available consists of 40 patients who are suffering from lung cancer to match between the positive/negative effects of two chemotherapy treatments in prolonging durability time. The variables used in the data set are:

1. Survival Time: The duration time in days after the treatment
2. A variable indicator:
1: stands for censored cases
0: complete uncensored case
3. Age: Patient's age at diagnosis
4. Weight: Patient's weight
5. Stage: The general medical situation at analysis of lung cancer on a range of 0 to 100
5. Time_diag_study: Days between analysis of lung cancer and cure
6. Tumor Type: Tumor types
Squamous Small
Adeno Large
7. Treatment: Types of treatment
Standard treatment
Experimental treatment



ID	TRT	AGE	WEIGHIN	STAGE	TOTALCW	TOTALCW	TOTALCW	TOTALCW6
1	0	52	124	2	5	5	6	7
5	0	77	160	1	9	5	10	9
6	0	60	136.5	4	7	9	17	19
9	0	61	179.6	1	6	7	9	3
11	0	59	175.8	2	5	7	16	13
15	0	69	167.6	1	6	6	6	11
21	0	67	186	1	6	11	11	10
26	0	58	158	3	6	11	15	15
31	0	61	212.8	1	6	9	6	8
35	0	51	189	1	6	4	8	7
39	0	46	149	4	7	8	11	11
41	0	65	157	1	6	6	9	6
45	0	67	186	1	8	8	9	10
2	0	48	163.8	2	7	10	9	10
12	1	58	227.2	4	6	10	11	9
14	1	42	162.6	1	4	6	8	7
16	1	44	261.4	2	6	11	11	14
22	1	27	225.4	1	6	7	6	6
24	1	68	226	4	12	11	12	9
34	1	77	164	2	5	7	13	12
37	1	66	140	1	6	7	7	7
42	1	73	181.5	0	8	11	16	
44	1	67	167	1	5	7	7	7
50	1	60	164	2	6	8	16	
56	1	54	172.8	4	7	8	10	8

Figure 3: Cancer data set

V. KNOWLEDGE DISCOVERY FROM BIG DATA

A. Processing and Cleaning Data

It is always very important to check whether your data match your business objectives. If it does not, there is a lot of questions to be addressed like:

- What are the viable proxies?
- Are there outlets that need to be taken in the report?
- Does the data consist of bias? Are there removed values?

There are a number of methods that can be used to impute or fill in removing values, such as mean interpolation, Kalman, filter, and ARMA. The quality of your data will highly affect your test results.

B. Analyzing and Visualizing the Data

Noe is the time to analyze the processed data and then the data need to be visually inspected for arranging, trending, and clustering. By the use of visualization tool the examination of relationships and building hypotheses need to be done. There are a number of powerful tools available such as box plots, scatter plots, visual aids, stacked bar charts, heat maps, and line graphs.

C. Storing of Data

Different methods can be used to facilitate Kohonen Self-Organizing Maps for visualization, counting clustering, pattern recognition, market basket inquiry, K-Means, principal component study, hierarchical clustering, multi-dimensional scaling, and factor study. The organizations they focus on leveraging to store and mine their data productively have an important competitive advantage over their rivals. Because of the same they can gain important insights and react quickly to enlarge their business in an approach that was not possible in absence of predictive analytics.

D. Building a Model

Wide variety of models need to be considered which will provide different perspective data. neural networks, ARIMA, decision trees, regressions, SVM, Naïve Bayes classifier, and discriminant analysis are some possible models. Avoid overfitting as you may find more than one model to solve one particular problem. Understand not only the probable errors, but also the most common ones. Be sure to document and interact the presumptions and results clearly. Set guidelines to control against making the most serious of false inferences.

E. Generation of Results and Optimization

There are many relevant methods, such as linear and quadratic programming, minimum squares solvers, and differential equation solvers (PDE, ODE) which are used to authorize objective function to achieve actionable results. You may find more than one model to solve one particular problem with the description of the enhanced function (linear, quadratic, or discontinuous) and restraints on the variables (linear or not). The objective is how to use valuable business choices to produce results.

F. Verification of Results

Once the above processes are done, then you need to allow time to produce results. It is required to allow the results which are meeting the initial objective of the the business.

G. Selecting the Right Tools

You may find a bunch of software products to help in data investigation process. There is a requirement of having different benchmarks to see such as performance, scalability, data source conformability, reliability, and comfort of distribution. During selection of a data analysis tool, the following points need to be taken care of:

1. Whether the tool having limitation in memory size?
2. Check whether the software is giving information to users of data failures.
3. What next if user input data is not applicable?
4. Check the size of the problem; whether it is showing a descriptive message to make the user aware about what is happening?
5. Can the analytic be recycled inside the database?
6. Do the tool support cascade data?
7. What is the development as well as the objective deployment environment in terms of achievement and technology?
8. Weather the analytics can be treated securely?
9. Is the analytic software advanced for a deployment platform?
10. Does it take dominance of multi-core servers and can the computation be complemented?
11. Does it support MapReduce (which will be essential for Hadoop)?
12. Does it use industry standard original language to clarify embedding in your web, Linux or Windows application and deployment?
13. Does it require any framework to support the formation? If so, what is the additional hardware, software, and maintaining costs?
14. Has it been tested on all platforms? If not, the computational results can be kind of different and origin differences in analytical results.
15. What does the setup solution look like?

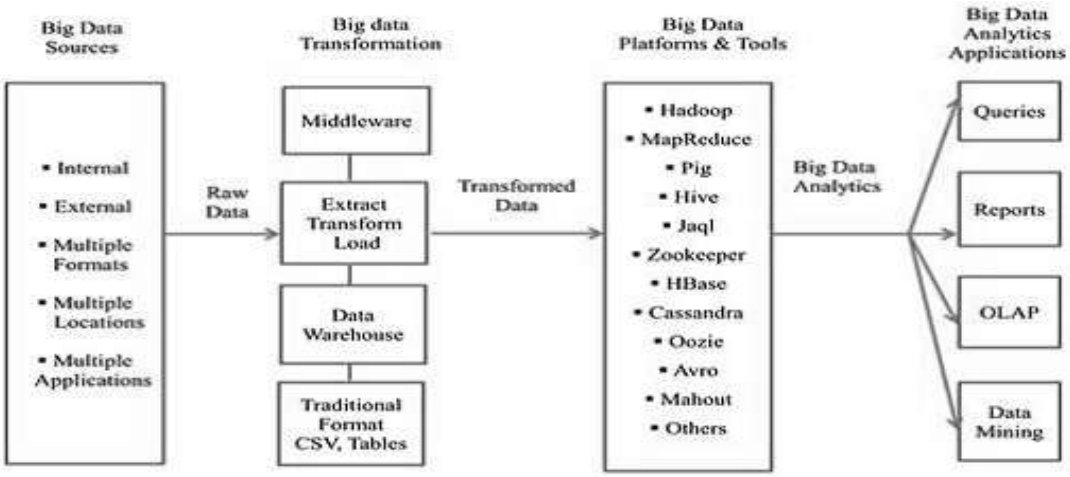


Figure 4: An applied conceptual architecture of big data analytics

VI. CONCLUSION

Predictability is an attribute of the data processing and it is not a component of the model. You can use predictive analytics to go above hardly improving the ability of your current processes. You can create new convenience or products based on the vision you accumulated from the data. While this process seems difficult, there are mature, commercially-available tools that have been approved, tried, and in management, like Rogue Wave Software IMSL Numerical Libraries, to help companies appliance all six steps in this process.

REFERENCES

[1]. Raghupathi W: Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity. Edited by Kudyba S. Taylor & Francis; 2010:211–223.
[2]. Burghard C: Big Data and Analytics Key to Accountable Care Success.IDC Health Insights; 2012.
[3]. Dembosky A: "Data Prescription for Better Healthcare."Financial Times, December 12, 2012, p. 19; 2012. Available from: http://www.ft.com/intl/cms/ s/2/55cbca5a-4333-11e2-aa8f00144feabdc0.html#axzz2W9cuwajK.
[4]. Feldman B, Martin EM, Skotnes T: "Big Data in Healthcare Hype and Hope." October 2012. Dr. Bonnie 360; 2012. http://www.west-info.eu/files/big-data-in-healthcare.pdf.
[5]. Fernandes L, O ″Connor M, Weaver V: Big data, bigger outcomes. J AHIMA2012:38–42.
[6]. IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the healthcare industry; 2013. http://ihealthtran.com/wordpress/2013/03/iht% C2%B2-releases-big-data-research-report-download-today/.

[7]. Frost & Sullivan: Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations. http://www.emc.com/ collateral/analyst-reports/frost-sullivan-reducing-information-technology-complexities-ar.pdf.
[8]. Bian J, Topaloglu U, Yu F, Yu F: Towards Large-scale Twitter Mining for Drug- Related Adverse Events.Maui, Hawaii: SHB; 2012.
[9]. Raghupathi W, Raghupathi V: An Overview of Health Analytics. Working paper; 2013.
[10]. I know Data Analytics for Healthcare: Creating Understanding from Big Data. http://info.ikanow.com/Portals/163225/docs/data-analytics-for-healthcare.pdf.

International Journal of Computational Engineering Research (IJCER) is UGC approved Journal with Sl. No. 4627, Journal no. 47631.

Sangram Keshari Swain "Use Of Big Data Analytics In Lung Cancer Data Set.” IInternational Journal of Computational Engineering Research (IJCER), vol. 7, no. 12, 2017, pp. 01-07.