Lung Cancer Diagnosis using Data Mining Algorithms

Mr.S.Muruganandam, Ph.D Research Scholar, PG & Research Department of CS and CA,
Vivekanandha College of arts and Sciences For Women,
Tiruchengode, Namakkal.

Dr.S.Subbaiah
Associate Professor
Department of Computer Science,
Sri Krishana College of Arts and Science
Coimbatore .

Abstract—The Early detection of lung cancer can be used to cure the lung cancer completely. **The computer assisted diagnosis approaches which uses many computational algorithms have become more helpful to predict and diagnose the Lung Cancer more precisely. The recent statistics reports says that, the survival rate of lung cancer disease is only about 13 to 15 percentages. If malfunctioning cells are detected in the early stage, then the survival rate can be improved up to 50 percentages. The survival rate of Lung cancer affected person is based on the early detection of lung modules. In this paper we have studied different papers about diagnosing the lung cancer using data mining techniques.**

Keywords—***Data mining, Identification, Methodology, Mining tools, Mining techniques.***

I. INTRODUCTION

A. Cancer disease

Cancer is the abandoned growth of abnormal cells in the body. Cancer develops when the body's normal control mechanism stops working. Old cells do not die and cells grow out of control, forming new, abnormal cells. These extra cells may form a mass of tissue, called a tumor. Human body is made up of 100 million cells. Cancer cell can start when one of them starts to grow and multiply too much. This growth is called a tumour. Being tumours are limited growing, they are high impact on causing problems and put pressure on nearby tissues, such as the brain. Deaths due to Lung cancer are about 1.4 million per year worldwide.

B. Lung cancer

American Lung Cancer Society is publishing every year that Lung Cancer is a foremost cause of Mortality in the western world as proven by the striking geometric numbers. They specify that the 5-year survival rate of patients with lung cancer can be improved from an average of 14% up to 49% if the disease is diagnosed and treated

in its early stage. However, to extract this related hidden information is a critical step to their use. This reason motivates to use data mining capabilities for efficient knowledge extraction and find hidden lung. Mining the Medical images involves many processes.

Data Mining in medical is a promising area of computational intelligence applied to an automatically analyze patients records aiming at the detection of new knowledge useful for medical decision-making. Induced knowledge is projected not only to increase accurate diagnosis and successful disease treatment, but also to enhance safety by dropping errors.

The methods in this paper classify the digital X-ray chest films in two categories: normal and abnormal. The normal ones characteries a healthy patient. The abnormal category includes Type of lung cancer, we will use a common classification method namely SVMs & neural networks.

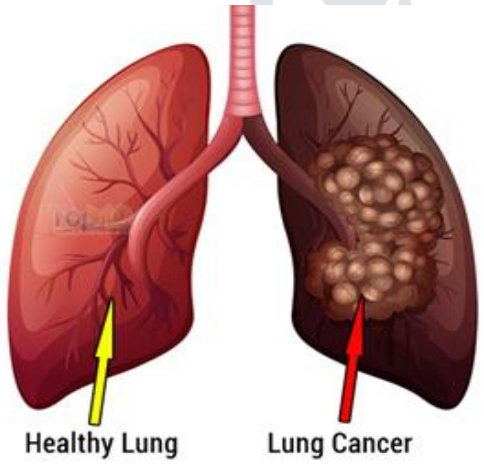


Figure1.1

I. DATA MINING APPRAOCH

A. Lung Cancer Diagnosis using Apriori Algorithm

Here the data mining techniques are applied for cancer diagnosis. The size and extent of the tumor are described using lung cancer pathologic staging. The patient's stages of cancer are predicted and help the doctor plan appropriate treatment. The surgery, biopsy is avoided which put a patient's health in danger.

The clinical information is only considered and to outplace the pathology report. The association between clinical information and pathology report is studied. The possibility of applying the clinical information is established to identify lung cancer using clinical information without surgery.

Data mining techniques are used to find an association between clinical and pathological report Association rules are if-then statement that discovers connection between unrelated data. Apriori algorithm is used to reduce computational effort. Find the frequent item sets: The sets of items that satisfy minimum support, confidence and lift. It uses the frequent item sets to generate association rules.

I. RELATED WORKS

1. A number of published studies also appear to lack an appropriate level of validation or testing. Among the better designed and validated studies it is clear that machine learning methods can be used to substantially [15–25%] improve the accuracy of predicting cancer susceptibility, recurrence and mortality. At a more fundamental level, it is also evident that machine learning is also helping to improve

our basic understanding of cancer development and progression.

2. The digital x-ray chest films are storied in huge multimedia databases for a medical purpose. This multimedia database provides a great environment to apply some image recognition methods to extract the useful knowledge and then rules from the mentioned database. These rules that we could got using image recognition methods, will help the doctors to decide important decisions on a particular patient state.

3. It provides a Computer Aided Diagnosis System (CAD) for early detection of lung cancer nodules from the Chest Computer Tomography (CT) images. There are five main phases involved in the proposed CAD system. They are processing, extraction of lung region from chest, in computer tomography images, segmentation of lung region having feature extraction from the segmented region, classification of lung cancer as benign or malignant.

4. It surveys the different approaches used for lung cancer diagnosis. It provides the efficient way for early detection of lung cancer. It reduces the death rate and increases the survival rate.

5. It can actualize this model on bigger up and coming information set of patient to foresee proper treatment routines.

6. This survey presents, the domain of terms was the pair right, left, over which we expected a uniform distribution. In analyzing term frequencies in a thoracic lung cancer database, the TDDA technique led to the surprising discovery that primary thoracic lung cancer tumors appear in the right lung more often than the left lung, with a ratio of 3:2.

II. IDENTIFYING HOTSPOT IN LUNG CANCER

Data Hotspot Algorithm is an association rule mining algorithm which is used for recognizing hotspots. Identify the average survival time of the patient.[6]The association rule consists of left hand side or antecedent and right hand side or consequent. The consequent is fixed to the target attribute. It can be the average survival time of the patients.

The LHS or antecedent defines the segment characteristics for patients. It uses a greedy approach to construct the tree of rules in depth first fashion. The 13 patient attributes are considered, namely ‘age’, ‘birth place’, ‘diagnosis’, ’tumor’, ‘lymph node’, ‘surgery performed’, ‘nosurgery’, ‘radiation therapy’, ‘lymph node surgery’, ‘cancer stage’, ‘past history’ and ‘lymph node examined’.

A. Study Patient Attribute

Study the patient attributes that affects the survival time of the patient. Removal of Redundant Rules The 2-stage semi-manual procedure is used to remove the redundant rules when using HotSpot algorithm.

B. Classification Stage

To classify the lung cancer, by using the data mining, classification techniques like Sequential minimal optimization (SMO), J48 decision tree.

Naive Bayes, Logit boost etc., Once the classification is made, it is to be comparedwith the experimental results of the above classification techniques, and observe which one gives efficient and precise results.[6] Some data attributes like

age at diagnosis, gender, marital status, smoking, panparag, tobacco, area, business, exercise, symptoms, treatment, tumor size, cancer stage are used and includes the feature extraction attribute like the size of the nodule as classification attributes. The final outcome of this project is to detect the cancerous nodule as benign (or) malignant.

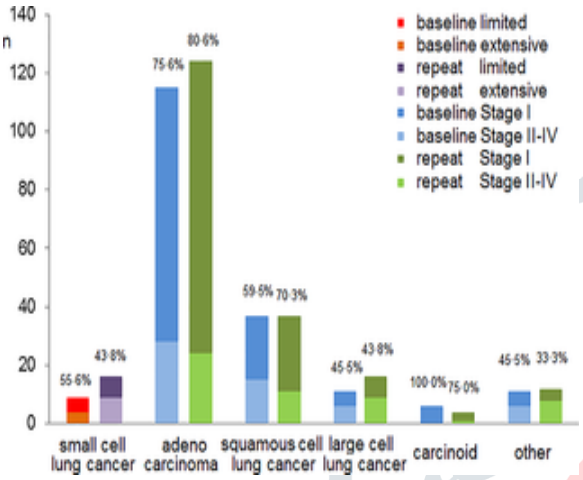
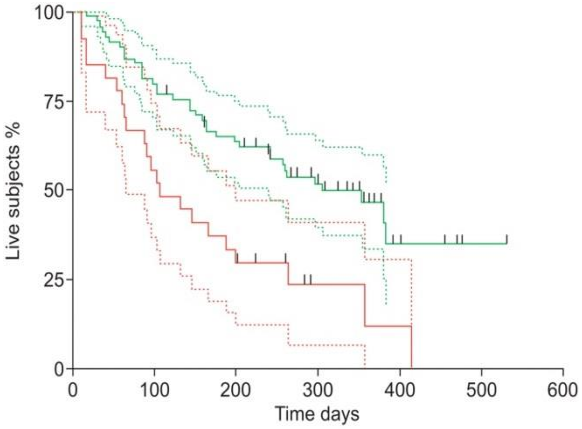


Figure 1.2

C. Feature Extraction Stage

Feature extraction is used to estimate the size of the tumor, to calculate the size of the tumor we need geometrical features like area, perimeter, irregularity index etc. The number of pixels having the values in the image array gives the area of the segmented tumor image. The number of boundary pixels in the tumor image is estimated as the perimeter of the tumor image.



Number of patients at risk:

Time days	0	50	100	150	200	250	300	350	400
Good	69	64	56	51	43	37	26	18	6
Poor	27	22	16	12	8	7	3	3	2

Figure1.3

III. METHODOLOGY

I In order to facilitate medical decision makers evaluation and utilization of problem regarding healthcare resource of lung cancer patients. Traditional regression method in combination with modern data mining techniques uses to compare prediction power of different model with help of propensity scoring. Two algorithm decision tree and artificial neural networks have been applied to predict the model and to generate rules on large, public but complex insurance claim data file as a data mining method.

These help to analysis and discover variation in healthcare delivery pattern for lung cancer. Decision tree and artificial neural networks can combine and produce effective predictive result as compare Comparative analysis of data mining tools for lung cancer patients 35 Journal of Information & Communication Technology to stand alone application. This can help health care decision. Diagnosis of Cancer Stages

Radiation Therapy

Choose Treatment Chemo Therapy

Classify the best treatment of Survival longer period of time for lung cancer patient. Further it bring necessary information to doctors and physician to carry on their research, diagnosis and suitable treatment much more easily so data mining helps in this regard.

Now we can classify suitable treatment method using data mining techniques for lung cancer patient to survive longer period of time.

IV. COMPARISION OF DATA MINING TECHNIQUES

Different types of mining algorithms in the healthcare field have been proposed by different researchers in recent years. A particular algorithm may not be applied to all the applications due to complexity for appropriate data types of the algorithm.

Consequently the choice of an acceptable data mining algorithm depends on not only the purpose of an application, but also on the compatibility of the data set. The relative analyses of different data mining techniques and algorithms have been used by most of the researchers in medical data mining.

V. COMPARATIVE ANALYSIS OF DATA MINING TOOLS

Due to the extensive use and complexity involved in building data mining applications, a large number of data mining tools have been developed over the decades. Different tools use diverse algorithm base and techniques to carry out data mining tasks. Every tool has its own advantages and disadvantages.

The maturity and relevance of data mining algorithms requires the utilization of influential software tools. As the number of accessible tools continues to develop, the preference of the most suitable tool becomes increasingly tricky. Thus, a number of authors have proposed and/or used the multiplicity of data mining tools as presented. Even many tools have been developed in predicting lung cancer everyone have own their will power to came across cancer disease. Youths are mostly addicted to drugs . Thus the causes effect on health of drug users is a social problem these days and youth is in the falling in trap of drugs. There is need to aware the youth about the common problems with drug users.

REFERENCES

1. Cruz Joseph, A. and David S. Wishart, 2006. Applications of Machine Learning in Cancer Prediction and Prognosis, A Review – Cancer Informatics, 2: 59-77.

2.Naveenkumar, N. and G. Selvavinayagam, 2015. Mining Techniques for Clinical Expert System and Predicting and Treating Lung Cancer with Big Data, International Journal of Computer Science and Engineering Communications, ISSN:2347-8586, 3(3).

3. Chai S, Yang J, Cheng Y. The research of improved Apriori Algorithm for mining association rules. IEEE Conference on Service System and Service Management; Chengdu. 2007 Jun 9-11. p. 1–4

4. Agrawal A, Choudhary A. Identifying HotSpots in lung cancer data using

association rule mining. 11th IEEE International Conference on Data Mining Workshops (ICDMW); Vancouver, BC. 2011 Dec 11. p. 995–1002

5.Ms. Swati P. Tidke, Prof. Vrishali A., Chakkarwar, "Classification of Lung Tumor Using SVM", International Journal of Computational Engineering Research, Vol. 2, Issue 5, pp. 1254-1257, 2012.

6.V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra. 2013, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," International Journal of Computer Science and Information Technologies, Vol. 4 (1), 2013, 39 – 45.

7.Jeffrey A. Goldman, Wesley Chu, D. Stott Parker, Robert M. Goldman, "A Case History in a Lung Cancer Text Database".

8.Vikas Chaurasia "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability" International journal of Computer Science and Mobile Computing IJCSMC, Vol.3, Issue. 1, January 2014, pg. 10-22, ISSN: 2320-088X.

9. Reeti Yadav "Chemotheraphy Prediction of Cancer Patient by Using Data Mining Techniques" International Journal of Computer Applications (0975-8887), Volume 76-No.10, August 2013.