A Survey on Mining Techniques for Early Lung Cancer Diagnoses

Juliet Rani Rajan, Assistant Professor
Department of Computer Science and Engineering
Jerusalem College of Engineering
Chennai, India

Dr.C.Chilambu Chelvan
Professor and Head
Department of Electronics and Instrumentation
RMD Engineering College, Chennai, India

Abstract **-With the huge growth in the volume of data today, there is an enhanced need to extract meaningful information from the data. Data mining contributes towards this and finds its application across various diverse domains such as in information technology, retail, stock markets, banking, and healthcare among others. The increase in population coupled with the growth in diseases has necessitated the inclusion of data mining in medical diagnosis to extract the underlying pattern. Of these, cancer is one of the widespread diseases that claim over 7 million lives every year and lung cancer accounts for 18% of these mortalities. Earlier researches and case studies indicate that the survival rate of the patients suffering from cancer is higher when the disease is diagnosed at an early stage.Lung cancer, a disease highly dependent on historical data for early diagnosis, has influenced researchers to pursue the data mining techniques for the pre-diagnosis process. The five year survival rate increases to 70% with the early detection at stage 1, when the tumor has not yet spread. Existing medical techniques like X-Ray, Computed Tomography (CT) scan, sputum cytology analysis and other imaging techniques not only require complex equipment and high cost but is also proven to be efficient only in stage 4, when the tumor has metastasized to other parts of the body.The proposed system involves the development of a data mining tool that will help in the classification of patients into the category that could potentially test positive for lung cancer in stage 1. Based on the pre-diagnosis results from the tool, the doctor can perform the diagnosis for the confirmation of tumor in the patient and initiate the treatment at an early stage thereby increasing the survival rate.**

Keywords—Artificial Neural Networks, Biomarkers, Data Mining, Pattern Evaluation

I. Introduction

Lung cancer is the leading cause of death in both men and women. Lung cancer is the killer disease in both men and women. In 1987, it exceeded breast cancer to become the leading cause of cancer deaths. The disease is characterized by the uncontrolled growth of cells. If it is not diagnosed and treated early, the tissues can be metastasized to other parts of the body such as the brain, bone, liver and adrenal gland.

As per the CancerCare, widely accepted tool for early lung cancer detection is not yet available **[1]**. There are two major types of lung cancer, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Over 85 percent of all lung cancers are non-small cell lung cancers, while about 13 percent contribute small cell lung cancers **[2]**. Staging lung cancer is based on whether the cancer is local or has spread from the lungs to the lymph nodes or other organs. Because the lungs are big, tumors can grow in them for a long time before they are found. Even when symptoms like coughing and fatigue occurs, people think that they are due to other causes. Because of this reason, early-stage of lung cancer (stages I and II) is difficult to detect. Most people with NSCLC are diagnosed only at stages III and IV.

Existing techniques like the chest X-Ray**[3]**, Computed Tomography (CT) scan**[4]**, sputum cytology, biopsy, bronchoscopy, needle aspirations, electronic nose**[5]** and others, not only require complex equipment and high cost but is also proven to be efficient only in stage 4, when the tumor has metastasized to other parts of the body. Also, it has been found that 0.4% of current cancers in US are due to the CT scans performed in the past and this may increase to as high as 1.5-2% as per the 2007 report **[6]**. The ionizing radiation emitted by the CT scan has the capability to damage the DNA which cannot be corrected by the cellular repair mechanism. Biomarker test is available for diagnosing cancers but there is no specific biomarker found so far for lung cancer **[7]** and researchers are still working on that. In spite of the available existing techniques, most of the time lung cancer is detected only after crossing stage 1. The lung cancer five-year survival rate (16.3%) is lower than many other leading cancer sites **[8]**.The five-year survival rate for lung cancer is 52.6% for cases detected early, when the disease is still localized (within lungs). However, only 15% of lung cancer is diagnosed at this early stage. For distant tumors, the five-year survival rate reduces

978-1-4673-6126-2/13/$31.00 ©2013 IEEE

significantly to 3.5%. So, over half of the people with lung cancer die within one year of being diagnosed.

As the volume of data is growing proportionally with the increase in population, there is a greater need to extract the knowledge from the data. Data mining contributes much towards this and finds its application in various diverse fields including the healthcare industry. Data mining is the process of sifting through historical data thus providing an insight into the patterns from large dataset and helps to incorporate the pattern in everyday activity. Data mining helps in medical diagnosis to extract the underlying pattern of the disease. Researchers are suggesting that applying data mining techniques in identifying effective pre-diagnosis of the disease can improve practitioner performance **[9]**. Lung cancer being a disease which is highly dependent on historical data can make use of data mining for its early detection .Researchers have been investigating on applying various data mining techniques on lung cancer dataset for early diagnosis of lung cancer.

This paper proposes a model for measuring if applying data mining techniques to lung cancer dataset can provide reliable performance in the detection of lung cancer at Stage I. The rest of the paper is organized as follows: Section 2 provides a literature survey on using data mining techniques which help health care professionals in the early diagnosis of lung cancer. Section 3 provides an overview of data mining and the various steps involved in data mining for the classification process. Section 4 discusses the proposed research model which is followed by a summary section.

II. Literature Survey

Over the last decade, many interesting techniques of data mining were proposed to detect various types of cancers. Few of the techniques are described below with their significance and limitations.

In **[10]**, Ahmed et al, implemented a model to diagnose lung cancer risk at an early stage using k-means clustering. The significant patterns are then discovered using Apriori Tid and Decision Tree algorithm. Apriori Tid, which is an extension of Apriori algorithm is one of the most influential algorithms to mine frequent item sets**[11]**. Apriori algorithm called two sub-processes which are Apriori-gen() and subset(), Apriori-gen() process produces a candidate for lung cancer, then use the Apriori property to delete those candidates of the non-frequent subsets. Once all the candidates gets generated, the database will be scanned and for each transaction, the Subset() sub procedure is used to identify all the candidate subsets, and make cumulative count for each of these candidates. Finally, all candidates met the minimum support form frequent item set L . The major drawback of this system is that it requires lot of database scans as the number of attributes increases. Also, it takes lot of time, space and memory for the candidate generation.

Oh et al. **[12]** proposed a method for predicting local failure in lung cancer post radiation therapy using Bayesian network. Bayesian networks encode the relations between variables using probability theory. They are used to predict an outcome and also to interpret the predictions based on the encoded relations. The attributes of the patient records are assigned to the nodes of the graph. The joint probability distribution function is then encoded by the network as per Equation 1.

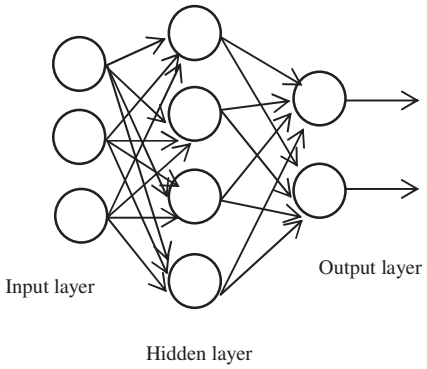
$$P(X) = \prod_{i=1}^{n} P(X_i | par(X_i)) \quad (1)$$

This Bayesian model makes use of both the physical attributes and the biological attribute, i.e, the blood biomarkers, for predicting the local failure. This method has outperformed **[10]** by including the biological parameters.

A novel approached was proposed by Palanisamy et al. in **[13]** for selecting the significant genes of leukemia cancer using k-means clustering algorithm. Clustering was basically carried out on the gene expression data. Here, the k value is fixed as 5, 10 and 15 and the classification is done for every k value separately and the accuracy is calculated. The k-means clustering always converges to a local minima. Classification accuracy depends on the starting cluster centroid selection which was one difficulty faced with the k-means algorithm. Bayesian method as proposed byOh et al. gave a better accuracy **[12]** because of the inclusion of physical parameters in combination with the biological parameters.

In 2011, Liu et al. implemented a classifier using Discrete Particle Swam Optimization(DPSO) **[14]** with new rule pruning procedure for detecting lung cancer and breast cancer. This is a slight modification of the Particle Swarm Optimization where DPSO does not make use of the velocity and the initial weight. With this new procedure, the average accuracy rates have improved. Also, the impact is more noteworthy on lung cancer data than it is on the breast cancer one. The major limitation of PSO and DPSO is that the algorithm has a limit on the number of iterations and in most situation it is not possible to know that the program has found an optimal solution.

Artificial neural networks (ANN) provide a powerful tool to help doctors to analyze, then model and make sense of complex clinical data across a wide range of medical applications **[15]**. It is a mathematical model developed on the basis of biological neural networks. Each neuron / node in the input layer represents each attribute of the patient dataset. The values from the input layer are then send to the nodes in the hidden layer along with the weight values where the learning actually takes place. After the learning process, the classification is done in the output layer. Figure 1 shows the structure of the basic network used for classification.



Economou et al. has proposed a model that diagnoses pulmonary diseases such as tuberculosis, lung cancer , asthma, occupational disorders of lungs and others **[16]**. They also proved that feed forward networks, especially the back propagation network and the Kalman filter could give a better performance.

III. Proposed Method

In the last few years, the digital revolution has provided relatively inexpensive and available means to collect and store large amounts of patient data in databases containing rich medical information and made available through the Internet for Health services globally. Data mining techniques applied on these databases discover relationships and patterns that are helpful in studying the progression of disease**[17]**. The steps involved in data mining are :

1) Data Integration
   Heterogeneous data from various health organizations which are in different forms are collected from multiple sources and made into a common source.
2) Data Selection
   The dataset collected from various sources contain all sorts of data. Some of the data may be irrelevant for the mining process and also, some data contain a lot of missing information. Such data are discareded. Only those data relevant to the mining process are considered.
3) Data cleaning
   Some patient record contain errors, noise or missing information. Certain data are corrected and those that cannot be corrected are discarded. Fuzzy Self Organising Maps can also be used to filling the missing values **[17]**. Table 2 presents some of the attributes identified.

Table 2 Some of Lung Cancer Causes Attributes

Attribute	Type
Age	Numeric
Gender	Nominal
Height	Numeric
Weight	Numeric
Smoking habit	Nominal
Secondhand smoke	Nominal
Radon gas	Nominal
Asbestos	Nominal
Air pollution	Nominal
Radiation therapy to lungs	Nominal
HIV or AIDS	Nominal
Organ Transplant	Nominal
Women with HRT	Nominal

Table 3 and Table 4 represent some of the primary and secondary symptoms identified.

Table 3 Some of Lung Cancer Primary Symptoms Attributes

Attribute	Type
Chest pain	Nominal
Cough	Nominal
Coughing of blood	Nominal
Fatigue	Nominal
Losing weight without trying	Nominal
Loss of appetite	Nominal
Shortness of breathe	Nominal
Wheezing	Nominal

Table 4 Some of Lung Cancer Secondary Symptoms Attributes

Attributes	Type
Bone pain or tenderness	Nominal
Eyelid drooping	Nominal
Facial Paralysis	Nominal
Hoarseness or changing voice	Nominal
Joint pain	Nominal
Nail problems	Nominal
Shoulder pain	Nominal

Swallowing difficulty	Nominal
Swelling of face or arms	Nominal
Weakness	Nominal
Fever	Nominal

4) Data Transformation
The data obtained even after the cleaning process is not ready for mining. They have to converted to an appropriate form suitable for mining. Large values are normalized so that the calculations can be made faster. This can be achieved by making use of the following formula:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

After the normalization and the discretization process, the records of the patients are represented in the form of a matrix.

5) Data Classification

Interesting patterns are discovered in this step. Lot of techniques have been discussed in the literature survey. It has been found that the Artificial Neural Networks could yield better result when compared to other techniques. In **[16]**, Economou et al made use of the supervised learning methods but in **[18],** it has been proves that the unsupervised learning, to be more specific, Kohonen Self Organising Map (SOM) could yield better performance even in the case of missing data. A 'learnt' SOM can be used as an important visualization aid because it gives a complete picture of the data ,i.e, similar data items are automatically grouped together**[19]**. For this reason, this method can be used for the diagnoses of lung cancer at an early stage. With modification done to the learning rate and the neighborhood distance and the weight updation formula, the model can yield a better performance result.

IV. RESULTS AND DISCUSSIONS

The purpose of this survey is to examine the most effective method to extract knowledge and information from the existing lung cancer profile data. Several data mining techniques are used in the cancer and in allied areas. A brief survey of related work in the area of cancer is that the data involved here are not complete. Data cleaning is a challenging step involved here as the data collected from heterogeneous sources does not contain all the required attributes. Shidong et al. in **[19]** proposed a model using fuzzy self organizing map to fill in the missing data. Normally with increase in the training data, performance can be improved. Since SOM, has the capability of imitating the human brain, it could give a better classification accuracy during the pre-diagnoses process.

V. CONCLUSION

In this paper of research survey, data mining and pattern recognition techniques for early lung cancer detection is studied. The survey aims to come out of the techniques being used in cancer diagnosis and allied areas. The recommendations arising from this survey are: A comparision of different data mining techniques could produce an efficient algorithm for lung cancer classification for multiple classes.

REFERENCES

[1] L.C.Rogerio, M.Carolyn, "Progress in the treatment of Lung Cancer", Cancer Care Connect 2012 p3-6.

[2] "Lung Cancer: New Tools for Making Decisions About Treatment", Cancer Care Connect 2011

[3] Abdullah, A.A.; Mohamaddiah, H., "Development of cellular neural network algorithm for detecting lung cancer symptoms," Biomedical Engineering and Sciences (IECBES), 2010 IEEE EMBS Conference on , vol., no., pp.138,143, Nov. 30 2010-Dec. 2 2010

[4] Jia Tong; Wei Ying; Wu Cheng Dong, "A lung cancer lesions dectection scheme based on CT image," Signal Processing Systems (ICSPS), 2010 2nd International Conference on , vol.1, no., pp.V1-557,V1-560, 5-7 July 2010

[5] Ping Wang; Xing Chen; FengjuanXu; Deji Lu; Wei Cai; Kejing Ying; Yongqing Wang; Yanjie Hu, "Development of electronic nose for diagnosis of lung cancer at early atage," Information Technology and Applications in Biomedicine, 2008. ITAB 2008. International Conference on , vol., no., pp.588,591, 30-31 May 2008

[6]R.Smith, J.Lipson, R.Marcus, K.P.Kim, M.Mahesh, R.Gould, G.Berrington, DL.Miglioretti, "Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer", Arch Intern Med 2009 Dec 14 169(22): 2078-86

[7] JY Cho, HJ Sung, "Proteomic approaches in lung cancer biomarker development", Experts Rev Proteomics, 2009 Feb;6(1):27-42

[8] http://www.lung.org/lung-disease/lung-cancer/resources/facts-figures/lung-cancer-fact-sheet.html, American Lung Association

[9] R.Brian, W.Nancy, "Effects of computerized clinical decision support systems on practitioner performance and

patient outcomes: Methods of a decision-maker-research partnership systematic review", Implementation Science 2010, 5:12

[10] Kawsar Ahmed, Abdullah-Al-Emran, Tasnuba Jesmin, Roushney Fatima Mukti, Md Zamilur Rahman, Farzana Ahmed, "Early Detection of Lung Cancer Risk Using Data Mining", Asia Pacific Journal for Cancer Prevention, 14(1), 595-598.

[11] Yanxi Liu, "Study on Application of Apriori Algorithm in Data Mining," Computer Modeling and Simulation, 2010. ICCMS '10. Second International Conference on , vol.3, no., pp.111,114, 22-24 Jan. 2010

[12] Jung Hun Oh, Jeffrey Craft, Rawan Al-Lozi, Manushka Vaidhya, Yifan Meng, Joseph O Deasy, Jeffrey D Bradly and Issam El Naqa, " Predicting Local Failure in Lung Cancer Using Bayesian Networks, 2010 Ninth Conference on Machine Learning and Applications.

[13] Palanisamy, P.; Perumal; Thangavel, K.; Manavalan, R., "A novel approach to select significant genes of leukemia cancer data using K-Means clustering," Pattern Recognition, Informatics and Medical Engineering (PRIME), 2013 International Conference on , vol., no., pp.104,108, 21-22 Feb. 2013

[14]Yao Liu, Yuk Ting Chung, "Mining Cancer Data with Discrete Particle Swarm Optimization and Rule Pruning", IEEE Conference Proceedings, 2011.

[15]A.Qeethara, "Artificial Neural Networks in Medical Diagnosis", International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.

[16] G.P.K.Economou, C.Spiropoulos, N.M.Economoupoulos, N.Charakopos, D.Lymberopoulos, M.Spiliopoulou, E.Haralambopulu, C.E.Goutis, "Medical Diagnoses And Artificial Neural Networks: A Medical Expert System Applied to Pulmonary Diseases", IEEE Conference Proceeding 1994.

[17]Jitao Zhao; Ting Wang, "A general framework for medical data mining," Future Information Technology and Management Engineering (FITME), 2010 International Conference on , vol.2, no., pp.163,165, 9-10 Oct. 2010

[18] Shidong U, Qi XU, Xianfeng WU, "Fuzzy Self-Organizinf Maps for Data Mining with Incomplete Data Sets", International Conference on Computer Applications and System Modelling, 2010.

[19]O.Jason, A.Syed,"Data Mining Using Self-Organizing Kohonen maps: ATechnique for Effective Data Clustering &Visualisation", International Conference on Artificial Intelligence, June 28-July 1 1999, Las Vegas.