International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume3, Issue 5 (May '2017)

AN EARLY DIAGNOSIS AND DETECTION OF LUNG CANCER DISEASE USING DATA MINING AND MEDICAL IMAGE PROCESSING METHODS

Dr.R.Hemalatha,
Associate Professor,
Department of Computer Science,
Tiruppur Kumaran College for Women,
Tirupur,Tamilnadu,India.

K.Devipriya,
Assistant Professor,
Department of Computer Science,
Tiruppur Kumaran College for Women,
Tirupur,Tamilnadu,India.

L.Subathra devi,
Assistant Professor,
Department of Computer Science,
Tiruppur Kumaran College for Women,
Tirupur,Tamilnadu,India.

Abstract: As the amount of data is growing day by day, there is a high requirement to extract knowledge from the data. With the growth in population and disease, there is need to include data mining in the field of health care industry. Studies have shown that cancer is one of the widespread diseases leading to fatal death today. Among them, lung cancer and breast cancer accounts the most. It has been found that if the disease is being diagnosed at an early stage, the survival rate of the patient could be improved but most of the time the disease is being diagnosed at a later stage. Deaths due to Lung cancer are about 1.4 million per year worldwide. Therefore, identification of genetic as well as environmental factors is very important in developing novel methods of cancer prevention. However, this is a multi-layered problem. Therefore a cancer risk prediction system is here proposed which is easy, cost effective and time saving. This research paper analyzes how data mining techniques and image processing techniques are used for predicting and diagnosing major life threatening lung cancer disease.

Keywords*: Association rules, Clustering, Classification Cancer Survey , Data Mining , Medical Image Processing,*

I.INTRODUCTION

Recent days, Cancer is becoming one of the deadliest diseases in the world. Cell growth and development is one of the important metabolisms that are happening in our body all the time. Our body has many types of cells which undergo mitosis and grow several times in a particular time period maintain normal function. At times, this cell division got affected and creates an erroneous cell or an abnormal cell. These cells split further in an unordered way and affects or spread out in to other parts of the body. These cells which split unusually are called cancers. These cells can easily spread out in to other region of the same part or organ, or other parts of the body. Cancer cells can take place in almost all parts of the body and they are characterized mainly from the place and type of the cell.

Lung cancer is the second most common cancer, accounting for about one out of five malignancies in men and one out of nine in women. Unfortunately, over the past several years, while the incidence of lung cancer has gradually declined in men, it has been rising alarmingly in women. In 1940 only seven women in 100,000 developed the disease; today the rate is 42 in 100,000. And all the evidence points to smoking as the cause. As one specialist in the field reports, "How long it takes to get cancer depends on how many cigarettes you smoke a day." However, studies prove that quitting smoking does lower the risk. There are two major types of lung cancer: small cell lung cancer (SCLC)—which is also called oat cell cancer, because the cells resemble oat grains—and non-small cell lung cancer (NSCLC). The aggressiveness of the disease and treatment options depend on the type of tumor diagnosed. Because many types of lung cancer grow quickly and spread rapidly and because the lungs are vital organs, early detection and prompt treatment—usually surgery to remove the tumor—is critical.

Types of Lung Cancer

Non-Small Cell Lung Cancer (NSCLC)

Most lung cancers are classified as non-small cell lung cancer (NSCLC). About half of these are squamous cell carcinomas (SCC). SCC, sometimes called epidermoid carcinoma, is more prevalent in men and arises in the lining of the large air passageways, or bronchi. Another common type of NSCLC is adenocarcinoma, which occurs at the outer edges of the lung. A small percentage of NSCLC are large-cell carcinomas, which usually develop in the smaller bronchi. Non-small cell lung cancer that begins at the top of the lung sometimes spreads to the nerves and blood vessels leading to the arm. All three subtypes of NSCLC develop differently. Treatment are often based on the location of the particular cancer and its rate of spread.

- Squamous cell or epidermoid carcinomas usually occur in the bronchi in the center of the lungs, but about a third of them arise on the periphery. This type of NSCLC is more likely to cause ulcers in the bronchi and bleeding than the other forms. Typically, the cancer cells double every 180 days. Although it often invades nearby tissue, squamous cell

carcinoma is less likely to metastasize as soon as other types.

- Most adenocarcinomas begin in the middle of the lungs, but about 25 percent develop along the lung periphery. These tumors are small, and the cells double about every 180 days also. They are likely to metastasize early. The form known as bronchoalveolar adenocarcinoma develops in the alveoli and may spread through the airways to other parts of the lung.
- Large cell carcinomas are bulky tumors that usually develop on the organ's periphery; however, they can arise anywhere within the lung. The cells double about every 100 days and can invade the mediastinum during the course of the disease.

Small Cell Lung Cancer (SCLC)

About one in four malignancies involving the lungs are diagnosed as small cell lung cancer (SCLC). There are several types of SCLC or oat cell cancer, including a mix of small cell and other cell types. These cancers grow rapidly—doubling in cell number about every 30 days—and spread quickly to lymph nodes and other organs than the non-small cell type.

Risk Factors

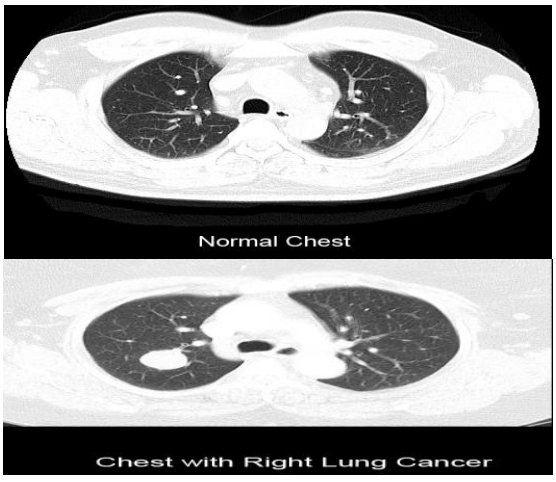
Men between 60 and 65 and women about 70 are at greater risk of having lung cancer.

Those who smoke have a risk of developing lung cancer that is 10 to 17 times higher than that of non-smokers. Women who smoke have a risk that is 5 to 10 times higher than that of women who do not smoke. The risk increases with the number of cigarettes smoked per day and the number of years the person has smoked. Some evidence suggests, however, that women and African-American men are more vulnerable

The risk of lung cancer for non-smokers who are exposed to smoke in the environment (known as second-hand, passive, or involuntary smoking), is as much as 30 percent higher than that of those who are not. The risk is even higher for exposure to side stream smoke (from the smoldering end of a cigarette) than for mainstream smoke (smoke that has been exhaled by the smoker). Industrial and atmospheric pollutants are responsible for a small percentage of lung cancer. For example, the risk of death from lung cancer is six to seven times greater for asbestos workers compared to the general population.

Early Detection

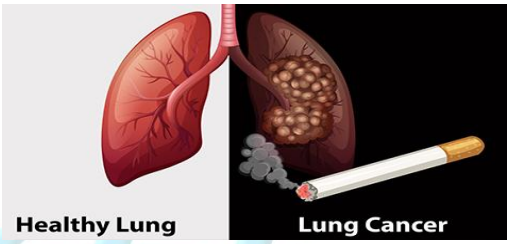
There are no routine screening tests for lung cancer. Detection at an early stage is possible with an x-ray or sputum analysis and some doctors order these tests, especially for people who smoke. However, there is no evidence that such attempts at lung cancer screening have a positive impact on treatment or survival. If a doctor suspects lung cancer, an x-ray is the first step in diagnosis.



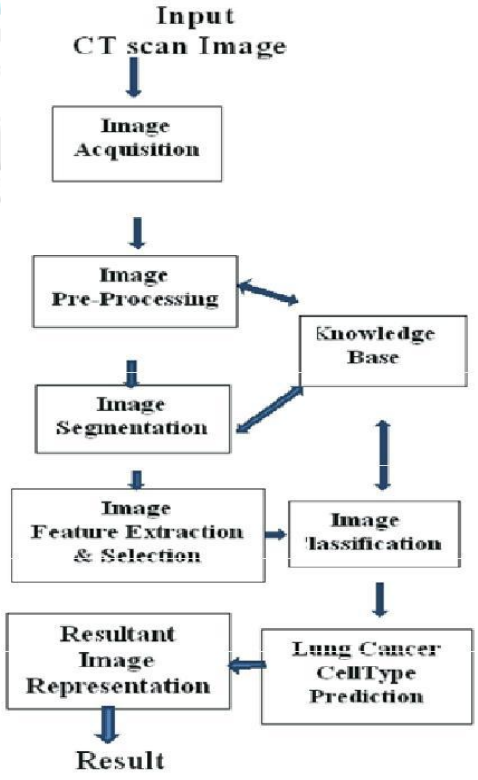
Signs & Symptoms

In many cases, symptoms do not appear until the cancer is quite advanced. However, by the time a tumor does cause changes within the lungs, the signs include:

- Difficulty breathing—stridor (a harsh sound with each breath), wheezing, labored breathing, shortness of breath (SOB);



- Coughing, possibly with blood in sputum;



II.DATA MINING STEPS

Data mining is a process of finding and extracting hidden pattern of correlation among the data which cannot be found by the normal statistical method. It is an iterative and interactive process. For the successful extraction of pattern, step by step procedure has to be followed.

A) Data Integration

The reports of the patients suffering from lung cancer are collected from various sources and integrated. Heterogeneous reports from different health care centers tend to give better result after the mining process.

B) Data Cleaning

The various causes and symptoms relevant to the mining process are retrieved from the heterogeneous reports thus generating the dataset required for the learning and testing .The dataset thus generated has a greater probability of containing missing information, erroneous data, noise or inconsistent data. Based on the domain knowledge, missing value for an attribute is filled. We ignore those records which has more than 40% of its value missing. In future, we can make use of SOM based fuzzy map model for data mining with incomplete dataset **[9]**. Table 1 presents some of the causes identified.

Table 1: Some 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

Symptoms of the patients are classified as primary and secondary symptoms. Table 2 and Table 3 present some of the primary and secondary symptoms identified.

C) Data Transformation

The attributes identified has to be transformed into form that is understandable by both human and the machine. Some of the parameters like age, height, weight are normalized for computational efficiency by using the following formula:

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

The attributes with nominal values are then converted into numeric or discrete variables. After the normalization and the discretization process, the records of the patients are represented in the form of a matrix.

(2) $$\begin{pmatrix} x_{11} & x_{12} \cdots x_{1n} \\ x_{21} & x_{22} \cdots x_{2n} \\ & \vdots \\ x_{p1} & x_{p2} \cdots x_{pn} \end{pmatrix}$$

Where p is the total number of training data and n is the number of attributes identified.

The dataset is then divided into 2 parts such that 80% of the data are used for the learning purpose and the remaining 20% of the data are used for the testing purpose.

Table 2: Some of Lung Cancer Primary Causes 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 3: Some of Lung Cancer Secondary Causes 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

III.CONCLUSION

This research paper is intended to explore the recent research on early diagnosis of Lung Cancer using Data mining and Medical image Processing techniques. Because in recent researches on early diagnosis of Lung cancer is viewed into two major areas, First the implementation of suitable and efficient Data mining algorithms for classification, clustering, prediction etc., Second processing of Lung cancer images that are mainly

obtained from CT scan, PET scan and X-ray methods. Most of the systems have been intended to achieve the maximum accurate values with less false positive value.

Lung cancer is the one of the deadliest disease in the world that affects more number of people around the world and it is constantly increasing. To discover the Lung cancer cells,the process of finding the disease plays a vital role. Discovery and Prediction of the Lung cancer in the starting stage is very much essential to cure the disease. For this purpose and to getprecise results, the work has been divided into following steps: ImageEnhancement stage, Image Segmentation stage and Features Extraction stage and classification.

REFERENCE

[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]. Taher Fatma and RachidSammouda, 2010. Artificial neural network and fuzzy clustering methods in segmenting sputum color images for lung cancer diagnosis, Intl. Conf. Signal Processing, pp: 513-520.

[4]. Krishnaiahet, V., 2013. Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques, International Journal of Computer Science and Information Technologies, 4 (1).

[5]. Rajan Juliet R and Jefrin J. Prakash, 2013. Early Diagnosis of Lung Cancer using a Mining Tool, International Journal of Emerging Trends in Computer Science, Special issue,

[6]. Ramachandran, P., N. Girija and T. Bhuvaneswari, 2014. Early Detection and Prevention of Cancer using Data Mining Techniques, International Journal of Computer Applications, 97(13).

[7]. Song Dansheng, Tatyana A. Zhukov and Olga Markov, 2012. Prognosis of stage i lung cancer patients through quantitative analysis of centrosomal features, IEEE.

[8]. Kumar Anita, 2015. A Study on Cancer Perpetuation Using the Classification Algorithms, International Journal of Recent Research in Mathematics Computer Science and Information Technology, 2(1): 96-99.

[9]. Deoskar Patag, Dr.Divakar Singh and Dr.Anju Singh, 2013. An Efficient Support Based Ant Colony Optimization Technique for Lung Cancer Data, International Journal of Advance Research in Computer and Communication.

[10]. Manzubi Zakariasuli and RemaAsheibaniSaad, 2014. Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer, Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases, pp: 32-37.

[11]. Dey Monali and SiddharthSwarupRautaray, 2014. Study and Analysis of Data mining Algorithms for Healthcare Decision Support System, International Journal of Computer Science and Information Technologies, 5(1): 470-477.

[12]. Zulpe Nitish and VrushenPawar, 2012. GLCM Textural Features for Brain Tumor Classification, International Journal of Computer Science Issues, 9(3): 3.

[13]. Gebejes A. and R. Huertas, 2013. Texture Characterization based on Gray-Level Co-occurrence Matrix, Conference of Infromatics and Management Sciences.

[14]. Ada and Rajneet Kumar, 2013. Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images, International Journal of Advanced Research in Computer Science and Software Engineering,ISSN:2277 128X, 3(3).

[15]. Mahersia, H., M. Zaroug and L. Gabralla, 2015. Lung Cancer Detection on CT Scan Images: A Review on the Analysis Techniques, International Journal of Advanced Research in Artificial Intelligence, 4(4).

[16]. Anam Quadri Rashida Shujaee and Nishat Khan, 2016. Review on Lung cancer detection using image processingtechnique, International Journal of Engineering Sciences & Research Technology, ISSN: 2277-9655.

[17]. AL-Tarawneh Mokhled, S., 2012. Lung cancer detection using image processing techniques, Leonardo Electronic Journal of Practices and Technologies, ISSN 1583-1078, 20: 147-158.

[18]. Lingayat Nitin S. and Manoj R. Tarambale, 2013. A Computer Based Feature Extraction of Lung Nodule in Chest X-Ray Image, International Journal of Bioscience, Biochemistry and Bioinformatics, 3(6).

[19]. Sharma Disha and Gagandeep Jindal, 2011.Computer Aided Diagnosis System for Detection of Lung Cancer in CT scan Images, International Journal of Computer and Electrical Engineering, 3(5).