Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 4, April 2013, pg.1 – 6

RESEARCH ARTICLE

Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient

Ada[1], Rajneet Kaur[2]

[1]Student of masters of technology Computer Science, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

[2]Assistant Professor, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

[1]adathour@yahoo.com; [2]rosy.rajneet@gmail.com

Abstract— The early detection of lung cancer is a challenging problem, due to the structure of the cancer cells, where most of the cells are overlapped with each other. This paper presents the feature extraction process and neural network classifier to check the state of a patient in its early stage whether it is normal or abnormal. After that we predict the survival rate of a patient by extracted features.

Key Terms: - Neural Network Classifier; Data Mining; CT-Scan Images; Feature Extraction; Survival Rate

I. INTRODUCTION

Lung cancer is considered to be the main cause of cancer death worldwide, and it is difficult to detect in its early stages because symptoms appear only in the advanced stages causing the mortality rate to be the highest among all other types of cancer. More people die because of lung cancer than any other types of cancer such as breast, colon, and prostate cancers. There is significant evidence indicating that the early detection of lung cancer will decrease mortality rate [1]. There are many techniques to diagnose lung cancer, such as Chest Radiography (x-ray), computed Tomography (CT), Magnetic Resonance Imaging (MRI scan) and Sputum Cytology. However, most of these techniques are expensive and time consuming. In other words, most of these techniques are detecting the lung cancer in its advanced stages, where the patients' chance of survival is very low. Therefore, there is a great need for a new technology to diagnose the lung cancer in its early stages. Image processing and data mining techniques provide a good quality tool for improving the manual analysis [1].

II. MATERIAL AND METHODOLOGY [10]

A. Input Samples for the Study

We have collected the 300 CT-Scan images of lung cancer from the private hospital. The digitized images are stored in the DIACOM format with a resolution of 8 bits per plane.

B. Preprocessing of Images

Most of the pre-processing is done with the help of MATLAB software. Each image sample is scanned, and stored to a size of 512 X 512 pixels. Generally during the scanning, the quality of image is affected by different artifacts due to non-uniform intensity, variations, motions, shift, and noise [2]. Thus, the pre-processing of image aims at selectively removing the redundancy present in scanned images without affecting the details

which that play a key role in the diagnostic process. Hence, Histogram-Equalization becomes the important step in preprocessing. Therefore each image is preprocessed to improve its quality.

Fig1 Shows the Histogram Equalization on CT scan image

C. Features Extraction

Image features Extraction stage is an important stage that uses algorithms and techniques to detect and isolate various desired portions or shapes (features) of a given image. To predict the probability of lung cancer presence, the following two methods are used: binarization and GLCM, both methods are based on facts that strongly related to lung anatomy and information of lung CT imaging.

i. GLCM (Grey Level Co-occurrence Method) [4]

The GLCM is a tabulation of how often different combinations of pixel brightness values (grey levels) occur in an image. Firstly we create gray-level co-occurrence matrix from image using *graycomatrix* function in MATLAB. Then we normalize the GLCM using the following formula

$$P_{i,j} = \frac{V_{i,j}}{\sum_{i,j=0}^{N-1} V_{i,j}}$$

Where, i is the row number and j is the column number. From this we calculate texture measures from the GLCM.

The following features are extracted using this method-

- Contrast
- Energy
- Entropy
- Homogeneity
- Maximum Probability
- Correlation
- Cluster shade
- Cluster Prominence
- Dissimilarity
- Autocorrelation
- Sum variance
- Sum Entropy
- Difference Variance
- Difference Entropy
- Information Measure

ii. Binarization Approach [3]

Binarization approach has been applied for detection of cancer. In this we extract the **number of white pixels** and check them against some threshold to check the normal and abnormal lungs. If the number of the white pixels of a new image is less than the threshold, then it indicates that the image is normal. Otherwise, if the number of the white pixels is greater than the threshold, it indicates that the image in abnormal. The threshold value that is used in this research is 255.

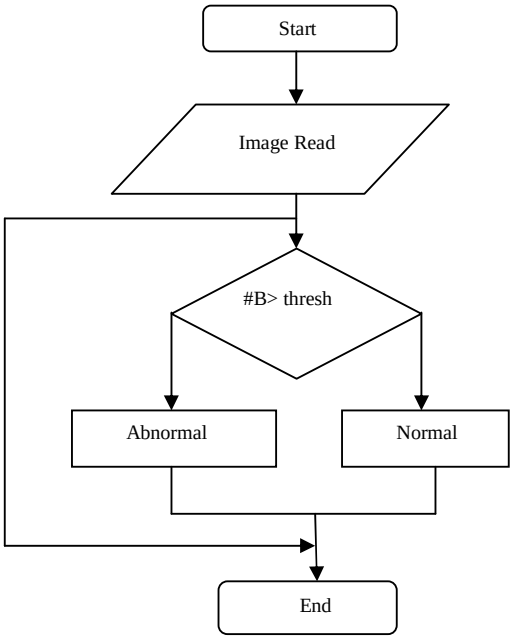

Fig.2 Binarization check method [3]

Combining Binarization and GLCM approaches together will lead us to take a decision whether the case is normal or abnormal.

D. PCA (Principle Component Analysis) [9]

PCA is to standardize the data in image. Real-world data sets usually exhibit relationships among their variables. These relationships are often linear, or at least approximately so, making them amenable to common analysis techniques. One such technique is principal component analysis ("PCA"), which rotates the original data to new coordinates, making the data as "flat" as possible.

The features extracted are passed through the PCA data mining for better classification. The following steps takes place in PCA:-

i. Calculate the mean and standard deviation of the features in the image using MATLAB.
ii. Subtract the sample mean from each observation, then dividing by the sample standard deviation. This centers and scales the data.
iii. Calculating the coefficients of the principal components and their respective variances is done by finding the Eigen functions of the sample covariance matrix.
iv. The matrix contains the coefficients for the principal components. The diagonal elements store the variance of the respective principal components. We can extract the diagonal.
v. The maximum variance in data results in maximum information content which is required for better classification.

E. Neural Network Classifier[5]

Especially, the neural network approach has been widely adopted in recent years. The neural network has several advantages, including its nonparametric nature, arbitrary decision boundary capability, easy adaptation to different types of data and input structures, fuzzy output values, and generalization for use with multiple images. Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions.(Actual biological neural networks are incomparably more complex.) Neural nets may use in classification problems (where the output is a categorical variable) or for regressions (where the output variable is continuous). The architecture of the neural network consists of three layers such as input layer, hidden layer and output layer. The nodes in the input layer linked with a number of nodes in the hidden layer. Each input node joined to each node in the hidden layer. The nodes in the hidden layer may connect to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables.

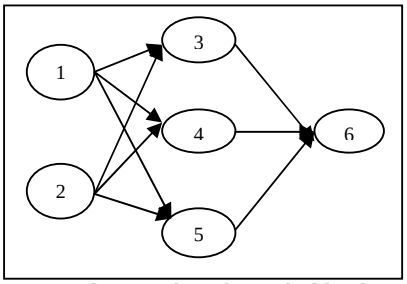

Fig. 3 A neural network with one hidden layer [5]

Steps Performed in Neural Network Classifier:-

- Create feed-forward back propagation network.
- Train neural network with the training samples and the group defined for it.
- The input image extracted PCA standardized data as the test samples, simulate the neural network to check whether the particular selected input sample has cancer or not.
- From the results of network and the samples trained in network classification rate is calculated using some mathematical formulas.

F. Survival Rate [6] [7]

Lung cancer survival rates are a measure of how many people remain alive with lung cancer after a certain amount of time. For example, a 5-year survival rate of 40% for a condition would mean that 40% of people, or 40 out of 100 people, would be alive after 5 years. When talking about lung cancer, physicians often use the term median survival as well. Median survival is the amount of time at which 50% of people with a condition will have died, and 50% are still alive [7].

Fine contrast feature demonstrated a substantial degree of concordance with PET tumor staging (stage I, contrast <3.2591; stage II, 3.2591≤ contrast ≤4.2632; stage III, 4.2632< contrast ≤4.9345;
Stage IV: contrast > 4.9345.Furthermore a fine contrast above 4.2632 predicted tumors above stage II [6].

III. RESULTS

This section summarizes the results of our experiments. We first load the train samples and then test samples after that we input the image and extract the different features and then apply neural network classifier. On the basis of earlier steps we check the state of the patient and his survival rate and year.

A. For Abnormal Lungs

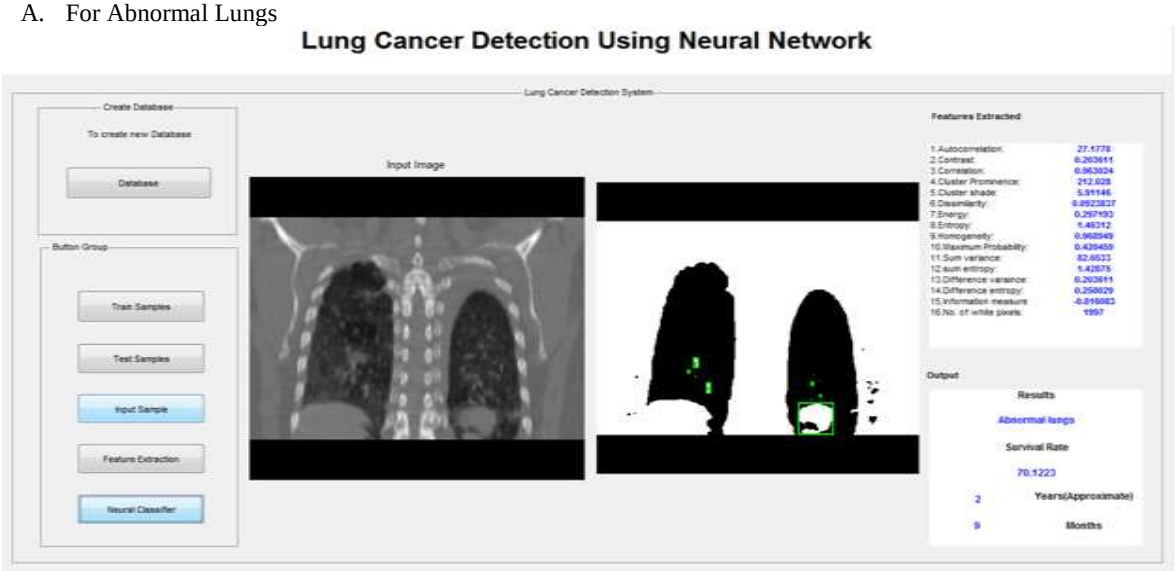

Fig. 4 Survival rate and year for abnormal lungs

In this we extracted the 16 features and on the basis of these features we predict the survival rate and survival year of the patient and on the basis of number of white pixels we find that lung is normal or abnormal.

B. For Normal Lungs

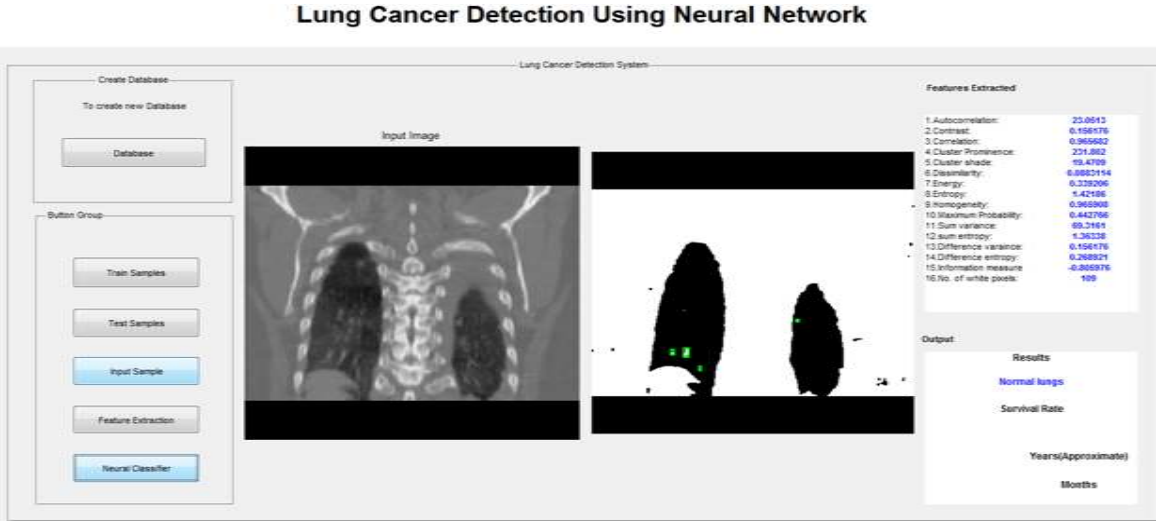

Fig. 5 No survival rate and year for normal lungs

In this we extracted the 16 features and on the basis of number of white pixels we find that lung is normal or abnormal. As this is a normal lung so we cannot predict its survival rate and year.

IV. CONCLUSION AND FUTURE WORK

A fast and effective method to detect the lung nodules, and separate the cancer images from other lung diseases like TB is becoming increasingly needed due to the fact that the incidence of lung cancer has risen dramatically in recent years and an early detection can save thousands of lives each year. In this research work, an attempt is made to detect the lung tumors from the cancer images and supportive tool is developed to check the normal and abnormal lungs and to predict survival rate and years of an abnormal patient so that we can save the lives of patients. In future same techniques can be applied to other type of cancer.

REFERENCES

[1] Almas Pathan, Bairu.K.saptalkar, "Detection and Classification of Lung Cancer Using Artificial Neural Network," International Journal on Advanced Computer Engineering and Communication Technology Vol-1 Issue:1 .

[2] Dr. S.A.PATIL, M. B. Kuchanur, " Lung Cancer Classification Using Image Processing," International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[3] Mokhled S. AL-TARAWNEH, "Lung Cancer Detection Using Image Processing Techniques," Leonardo Electronic Journal of Practices and Technologies Issue 20, January-June 2012, p. 147-158.

[4] Fritz Albregtsen, "Statistical Texture Measures Computed from Gray Level Coocurrence Matrices," International Journal of Computer Applications, November 5, 2008.

[5] Zakaria Suliman Zubi and Rema Asheibani Saad, "Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer," Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases, Libya, 2011.

[6] Balaji Ganeshan, Sandra Abaleke, Rupert C.D. Young, Christopher R. Chatwin, Kenneth A. Miles, "Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage," Cancer Imaging , v.10(1): 137–143, 2010 July 6.

[7] Lynne Eldridge MD. (2013, March 22). Lung Cancer Survival Rates by Type and Stage [Online]. Available: http://lungcancer.about.com/od/whatislungcancer/a/lungcancersurvivalrates.htm.

[8] Jaba Sheela L and Dr.V.Shanthi, "An Approach for Discretization and Feature Selection Of Continuous-Valued Attributes in Medical Images for Classification Learning," International Journal of Computer Theory and

Engineering, Vol. 1, No.2, June2009.
[9] Taranpreet Singh Ruprah, "Face Recognition Based on PCA Algorithm," Special Issue of International Journal of Computer Science & Informatics (IJCSI), 2231–5292, Vol.- II, Issue-1, 2.
[10] Guruprasad Bhat, Vidyadevi G Biradar , H Sarojadevi Nalini, "Artificial Neural Network based Cancer Cell Classification," Computer Engineering and Intelligent Systems, Vol 3, No.2, 2012.

