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. 4, Issue. 1, January 2015, pg.175 – 184

SURVEY ARTICLE

A SURVEY ON EARLY DETECTION AND PREDICTION OF LUNG CANCER

Neha Panpaliya[1], Neha Tadas[2], Surabhi Bobade[3], Rewti Aglawe[4], Akshay Gudadhe[5]

Department of Computer Engineering, Bapurao Deshmukh College of Engineering, Sevagram, Wardha, RTMNU, Nagpur, India

[1] panpaliyan786@gmail.com, [2] neha.tadas@gmail.com, [3] surabhi.bobade@gmail.com,

[4] rewtiaglawe@gmail.com, [5] akshaygudadhecompengg@gmail.com

Abstract:- Lung cancer is the leading cause of cancer death worldwide. The earlier detection of lung cancer is a challenging problem due to structure of cancer cell, where most of the cells are overlapped each other. For early detection and treatment stages image processing technique are widely used and for prediction of lung cancer, identification of genetic as well as environmental factors are very important in developing novel method of lung cancer prevention. In various cancer tumours such as lung cancer the time factor is very important to discover the abnormality issue in target images. Prediction of lung cancer we consider significant pattern and their corresponding weight age and score using decision tree algorithm. Using the significant pattern tool for lung cancer prediction system will develop. In this proposed system we use Histogram Equalization is used for preprocessing of images and feature extraction processes and neural network classifier to check the state of patient whether it is normal or abnormal. If the lung cancer is successfully detected and predicted in its early stages will reduce many treatment options and also reduce risk of invasive surgery and increase survival rate. Therefore lung cancer detection and prediction system will propose which is easy, cost effective and time saving. This will produce promising result for detection and prediction of lung cancer. Therefore early detection and prediction of lung cancer should play a vital role in the diagnosis process and also increase the survival rate of patient.

Keywords:- data mining, early detection, image processing, prediction, Lung cancer

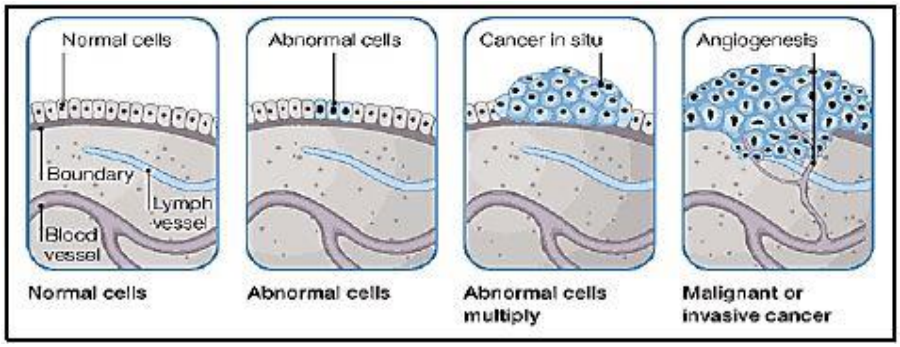
1. INTRODUCTION

Image processing techniques provide a good quality tool for improving the manual analysis. Image processing techniques are used in several areas such as military, space research, medical and many more. In this proposed system image processing techniques are used for image improvement in earlier detection and treatment stages. Image quality assessments as well as improvement are depending on the enhancement stage where pre-processing technique is used based on principal component analysis and Histogram Equalization. Classification is very important part of digital image analysis. It is computational

procedure that sort images in to groups according to their similarities. In proposed system Histogram Equalization is used for preprocessing of images and 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 will predict the survival rate of patient by extracted features.

In proposed early detection and prediction system we consider several significant patterns which are Smoking, Environment, Alcohol, Obesity, Chronic Lung Disease, Balance Diet, Mental trauma, Radiation Therapy, Tobacco, and Genetic Risk. Using this significant pattern the system will predict lung cancer. Cigarette smoking is the most important cause of lung cancer. Cigarette smoke contains more than 4,000 chemicals, many of which have been identified as causing cancer. A person who smokes more than one pack of cigarettes per day has a 20-25 times greater risk of developing lung cancer than someone who has never smoked. About 90% of lung cancers arise due to tobacco use. However, other factors, such as environment pollution mainly air; excessive alcohol may also be contributing for Lung Cancer. Lung cancer occurs for out-of-control cell growth and begins in one or both lungs. Lung cancer that spreads to the brain can cause difficulties with vision, weakness on one side of the body. Symptoms of primary lung cancers include cough, coughing up blood, chest pain, and shortness of breath.

Early prediction of lung cancer should play a pivotal role in the diagnosis process and for an effective preventive strategy. There are many techniques to diagnosis lung cancer, such as Chest Radiograph (x-ray), Computed Tomography (CT), Magnetic Resonance Imaging (MRI scan) and Sputum Cytology. However, most of these techniques are expensive and time consuming. Most of these techniques are detecting the lung cancer in its advanced stages, where the patient's 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 techniques provide a good quality tool for improving the manual analysis.



[7]

Figure1: The beginning of cancer

Using the significant pattern prediction tools for a lung cancer prediction system will develop. This lung cancer risk prediction system should prove helpful in detection of a person's predisposition for lung cancer. Therefore early prediction of lung cancer should play a pivotal role in the diagnosis process and for an effective preventive strategy.

There are many techniques to diagnose lung cancer such as CT-SCAN, MRI, X-RAY. These techniques are expensive and time consuming. 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. The proposed system can be used to detect lung cancer in early stages. The proposed early detection and prediction of lung cancer system which is easy, cost effective and time saving.

This proposed lung cancer detection and prediction system help to detect the lung cancer in its early stage and also to predict the lung cancer. Therefore the survival rate of patient will increase.

The purpose behind to designing this system is to predict and detect the lung cancer in its early stage on the basis on some factors and thresolding. We decrease the number of rule for testing in this system. This system reduce the time and cost required for various excessive medical test. The propose system is on web based due to this a rural site patient directly communicate with doctors and doctors will try to solve their questions (problems).

Objectives of this system are as follows:

- To decrease the number of rules for testing.
- To reduce the time and cost required for various excessive Medical Tests.
- To increase the accuracy of performance of Lung Cancer Prediction and Detection System.
- Use less number of attributes for prediction of Cancer.
- Early stage detection of cancer.
- Increasing the survivability of the patient more than 5 years.

In literature survey stage, we tried to get enough information related to our topic of research.

2. LITERATURE REVIEW

All researchers have aim to develop such a system which predict and detect the cancer in its early stages. Also tried to improve the accuracy of the Early Prediction and Detection system by preprocessing, segmentation feature extraction and classification techniques of extracted database. The major contributions of the research are summarized below.

[1]T. Sowmiya, M. Gopi, M. New Begin, L.Thomas Robinson - In this paper they described Cancer as the most dangerous diseases in the world. Lung cancer is one of the most dangerous cancer types in the world. These diseases can spread worldwide by uncontrolled cell growth in the tissues of the lung. Early detection of the cancer can save the life and survivability of the patients who affected by this diseases. In this paper we survey several aspects of data mining procedures which are used for lung cancer prediction for the patients. Data mining concepts is useful in lung cancer classification. We also reviewed the aspects of ant colony optimization (ACO) technique in data mining. Ant colony optimization helps in increasing or decreasing the disease prediction value of the diseases. This case study assorted data mining and ant colony optimization techniques for appropriate rule generation and classifications on diseases, which pilot to exact Lung cancer classifications. In additionally to, it provides basic framework for further improvement in medical diagnosis on lung cancer.

[2]Ada[1], Rajneet Kaur[2] (2013) - In this paper uses a computational procedure that sort the images into groups according to their similarities. In this paper Histogram Equalization is used for preprocessing of the images and 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. Experimental analysis is made with dataset to evaluate the performance of the different classifiers. The performance is based on the correct and incorrect classification of the classifier. In this paper Neural Network Algorithm is implemented using open source and its performance is compared to other classification algorithms. It shows the best results with highest TP Rate and lowest FP Rate and in case of correctly classification, it gives the 96.04% result as compare to other classifiers.

The second paper of this same author is based on Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images. In this paper uses a hybrid technique baseon feature extraction and Principal Component Analysis (PCA).

[3]Dasu Vaman Ravi Prasad (2013) - In this paper image quality and accuracy is the core factors of this research, image quality assessment as well as improvement are depending on the enhancement stage where low pre-processing techniques is used based on Gabor filter within Gaussian rules. Following the segmentation principles, an enhanced region of the object of interest that is used as a basic foundation of feature extraction is obtained. Relying on general features, a normality comparison is made. In this research, the main detected features for accurate images comparison are pixels percentage and mask-labeling.

[4]S Vishukumar K. Patela and Pavan Shrivastavab (2012) - In this paper authors mostly focus on significant improvement in contrast of masses along with the suppression of background tissues is obtained by tuning the parameters of the proposed transformation function in the specified range. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of cancer, which improves the chances of survival for the patient. In this paper, authors proposed gabor filter for enhancement of medical images. It is a very good enhancement tool for medical images.

[5]Fatma Taher1,*, Naoufel Werghi1, Hussain Al-Ahmad1, Rachid Sammouda2 (2012) - This paper presents two segmentation methods, Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm, for segmenting sputum color images to detect the lung cancer in its early stages. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of lung cancer which will improve the chances of survival for the patient. However, the extreme variation in the gray level and the relative contrast among the images make the segmentation results less accurate, thus we applied a thresholding technique as a pre-processing step in all images to extract the nuclei and cytoplasm regions, because most of the quantitative procedures are based on the nuclear feature. The thresholding algorithm succeeded in extracting the nuclei and cytoplasm regions. Moreover, it succeeded in determining the best range of thresholding values. The HNN and FCM methods are designed to classify the image of *N* pixels among *M* classes. In this study, we used 1000 sputum color images to test both methods, and HNN has shown a better classification result than FCM, the HNN succeeded in extracting the nuclei and cy-toplasm regions. In this paper authors uses a rule based thresholding classifier as a pre-processing step. The thresh-olding classifier is succeeded in solving the problem of in-tensity variation and in detecting the nuclei and cytoplasm regions, it has the ability to mask all the debris cells and to determine the best rang of threshold values. Overall, the thresholding classifier has achieved a good accuracy of 98% with high value of sensitivity and specificity of 83% and 99% respectively.

3. METHODOLOGY

Diagnosis of lung cancer includes the following stages:

1. Images captured
2. Preprocessing of images
3. Image segmentation
4. Feature extraction
5. Principle component analysis
6. Neural network classifier
7. Diagnosis result
8. Prediction process

1. Images captured or collected :

Primarily, cancer and non-cancer patient's data or CT-Scan images will be collected from different diagnostic centers. The digitized images are stored in the DIACOM format with a resolution of 8 bits per plane [1].

2. Preprocessing of images :

The image Pre-processing stage in this system begins with image enhancement which aims to improve the interpretability or sensitivity of information included in them to provide better input for other programmed image processing techniques.

Image enhancement techniques can be divided into two wide types: Spatial domain methods and frequency domain methods. On the other hand, when image enhancement techniques are used as pre-processing tools for other image processing techniques, the quantifiable measures can determine which techniques are most suitable. In the image enhancement stage we will be using the Histogram Equalization.

The pre-processing of image aims for selective elimination of the redundancy in scanned images without affecting the original image, this play a vital role in the diagnosis of lung cancer. Therefore, Histogram- Equalization becomes the crucial step in preprocessing. Hence, each image is preprocessed to enhance its superiority.

3. Image Segmentation :

Image segmentation is a crucial process for most image analysis consequent tasks. Especially, most of the existing techniques for image description and recognition are highly depend on the segmentation results. Segmentation splits the image into its constituent regions or objects. Segmentation of medical images in 2D has many beneficial applications for the medical professional such as: visualization and

volume estimation of objects of concern, detection of oddities, tissue quantification and organization and many more.

The main objective of segmentation is to simplify and change the representation of the image into something that is more significant and easier to examine. Image segmentation is usually used to trace objects and borders such as lines, curves, etc. in images. More accurately, image segmentation is the process of allocating a label to every pixel in an image such that pixels with the same label share certain pictorial features.

The outcome of image segmentation is a set of segments that collectively cover the entire image, or a set of edges extracted from the image i.e. edge detection. In a given region all pixels are similar relating to some distinctive or computed property, such as texture, intensity or color. With respect to the same characteristics adjacent regions are significantly different.

One of two basic properties of intensity values Segmentation algorithms are based on: discontinuity and similarity. In the first group we partition the image based on abrupt changes in intensity, such as edges in an image. The next group is based on segregating the image into regions that are alike according to a predefined criterion. Histogram thresholding methodology comes under this group.

4. Feature Extraction:

Image features Extraction stage is a crucial stage that uses algorithms and methods to detect and separate various preferred portions or shapes of an inputted image.

The following two methods are used to predict the probability of lung cancer presence: binarization and GLCM, both methods are based on facts that strongly related to lung anatomy and information of lung CT imaging.

4. A. Binarization Approach

For detection of cancer binarization approach has been applied for detection of cancer. In binarization we extract the number of white pixels and check them against some threshold to check the normal and abnormal lung cells. After this process the condition is check whether number of white pixels of a new image is less than the threshold then it indicates that the image is normal, or else if the amount of the white pixels is greater than the threshold, it specifies that the image in abnormal.

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

4. B. GLCM (Grey Level Co-occurrence Method)

The GLCM is a process of tabulating different combinations of pixel brightness values called as grey levels which occurs in an image. In this first step is to create gray-level co-occurrence matrix from image in MATLAB.

In second step 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 compute texture events from the GLCM [1].

4. C. Masking Approach

Inside lungs masses are appeared as white connected areas inside ROI (lungs), masking approach depends on this. As they increase the percent of cancer presence increases. Also combining Binarization and Masking approaches together will help us to take a decision on whether the case is normal or abnormal according to the mentioned assumptions in the previous two approaches, we can make a conclusion that if image has number of black pixels greater than white pixels then that image is normal or otherwise we can say that the image is abnormal.

5. PCA (Principle Component Analysis)

PCA is a technique to normalize the data in image. Real-world data sets generally display associations among their variables. These associations are frequently linear, or at least practically so, making them agreeable to common examination techniques. One such technique is principal component analysis ("PCA"), which rotates the original data to new coordinates, making the data as "even" as possible. The features mined are delivered to the PCA data mining for better sorting [1].

The following steps takes place in PCA:-

i. Calculating the mean and standard deviation of the features in the image.

ii. Subtracting the sample mean from each observation, and then dividing by the sample standard deviation. This scales and centers the data.

iii. Then we calculate the coefficients of the principal components and their relevant changes are done by finding the Eigen function of the sample covariance matrix.

iv. This matrix holds the coefficients for the principal constituents. The diagonal elements store the modification of the relevant principal constituents. We can mine the diagonal.

v. The maximum variance in data results in maximum information content which is required for better classification [1].

6. Neural Network Classifier

Supervised feed-forward back-propagation neural network ensemble used as a classifier tool. Neural network contrasts in different means from traditional classifiers like Bayesian and k – nearest neighbor classifiers. Linearity of data is one of the major variances. Other existing classifiers like Bayesian and k – nearest neighbor entails linear data to work properly. But neural network works as well for non-linear data because it is simulated on the reflection of biological neurons and network of neurons.

Training the neural network with wide range of input data will increase the detection accuracy, in other words the system will get biased with a small set of data or large set of similar data. Hence neural network classifier needs a large set of data for training and also it is time consuming to train to reach the stable state. But once it is trained it works as fast and quick as biological neural network by transmitting signals as fast as electrical signals.

Input layer, internal hidden layer and output layer are the three layers of the architecture of the neural network. The nodes in the input layer are linked with a number of nodes in the internal hidden layer. Each input node connected to each node in the internal hidden layer. The nodes in the internal hidden layer may connect to nodes in another internal hidden layer, or to an output layer. And the output layer consists of one or more response variables.

Following are the general Steps performed in Neural Network Classifier:-

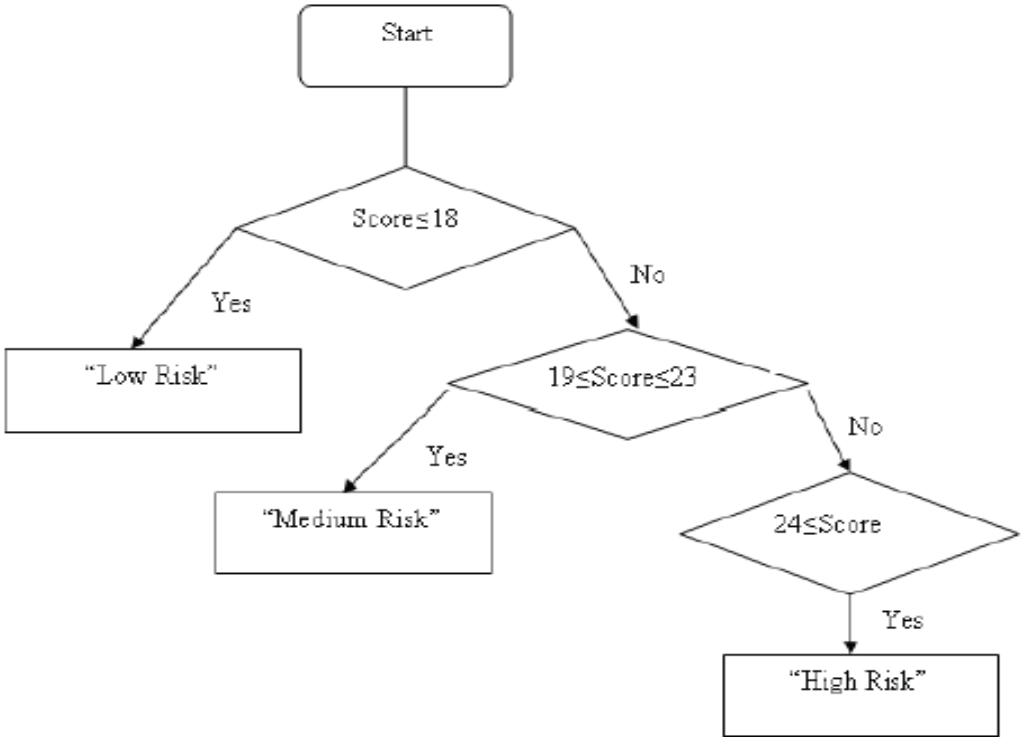
- Creating feed-forward back propagation network.
- Training neural network with the already available samples and the group defined for it.
- The input image mined PCA consistent data as the test samples, fires the neural network to check whether the particular selected input sample has cancer or not.
- From the results which are obtained from the neural network and the samples trained in network classification rate is calculated using some mathematical formulas.

7. Diagnosis Result

After completion of all the processes in the last stage i.e. in the diagnosis stage or in diagnosis result the proposed system show whether the image is in normal or in abnormal state.

8. Prediction process

There is no remedy for cancer after completely affected. Death is inevitable. So the ability to predict Lung cancer plays an important role in the diagnosis process. In this paper we have proposed an effective Lung cancer prediction system based on data mining. This lung cancer risk prediction system should prove helpful in detection of a person's predisposition for lung cancer**.** We will be considering various risk factors which includes-age, gender, hereditary, previous health examination, use of anti-hypersensitive drugs, smoking, food habit, physical activity, obesity, tobacco, genetic Risk, environment, mental trauma, uptake of red meat, balance diet, hypertension, heart disease, excessive alcohol, radiation therapy and chronic lung diseases. Various algorithms will be used such as decision tree algorithm for the prediction process. Figure below illustrates the working of decision tree algorithm. After its processing, system will predict the chances of patient to get affected by lung cancer in the years to come.



[5]**Figure2:** Flow diagram of decision tree

CONCLUSION

Thus, we can conclude that using the combination of neural network classifier along with binarization [1] and GLCM will increase the accuracy of lung cancer detection process. This system will also decrease the cost and time required for cancer detection. Also if the patient is not detected with the lung cancer the system will proceed further for the prediction process. As this system will be available online, patients from remote areas can also avail its benefits. So this system is beneficial for huge number of people all over the world. Also tests required for cancer detection is required.

REFERENCES

1. T. Sowmiya, M. Gopi, M. New Begin L.Thomas Robinson "*Optimization of Lung Cancer using Modern data mining techniques."* International Journal of Engineering Research ISSN:2319-6890)(online),2347-5013(print)VolumeNo.3,Issue No.5, pp : 309-3149(2014)
2. Ada[1], Rajneet Kaur[2] "*Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier*", (IJAIEM)Volume 2, Issue 6, June 2013
3. Dasu Vaman Ravi Prasad,*"Lung cancer detection using image processing techniques",* International journal of latest trends in engineering and technology.(2013)
4. S Vishukumar K. Patela and Pavan Shrivastavab, *"Lung A Cancer Classification Using Image Processing",* International Journal of Engineering and Innovative Technology Volume 2, Issue 3, September 2012.

5. Fatma Taher1,*, Naoufel Werghi1, Hussain Al-Ahmad1, Rachid Sammouda2, "*Lung Cancer Detection Using Artificial Neural Network and Fuzzy Clustering Methods*," American Journal of Biomedical Engineering 2012, 2(3): 136-142
6. Morphological Operators, CS/BIOEN 4640: "*Image Processing Basics*", February 23, 2012.
7. 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 :2011.
8. American Cancer Society, "*Cancer facts & figures2010*" http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-026238.pdf (2010).
9. "*Multileve-l Thresholding Based on Histogram Difference*," in 17th International Conference on Systems, Signals and Image Processing. 2010.
10. Nunes, É.d.O. and M.G. Pérez., Nunes, É.d.O. and M.G. Pérez., *"Medical Image Segmentation by Multilevel Thresholding Based on Histogram Difference,"* in17th International Conference on Systems, Signals and Image Processing. 2010.
11. S.Shah, "*Automatic Cell Images segmentation using a Shape-Classification Model*", Proceedings of IAPR Conference on Machine vision Applications, pp. 428-432,Tokyo, Japan,(2007)

