LUNG CANCER DIAGNOSIS BASED ON CLUSTERING ALGORITHM AND ARTIFICIAL NEURAL NETWORK

Jayalakshmi S.[1], Manikandan T.[2]

[1,2]Department of Electronics and Communication Engineering, Rajalakshmi Engineering College, Chennai
Email: [1]contactjayaselvaraj@gmail.com, [2]mani_stuff@yahoo.co.in

Abstract

The most familiar cancer that occurs usually for men and women is lung cancer. The survival rate for the cancer patient can be increased by detecting the occurrence of cancer in earlier stages. But, the early detection of lung cancer is a challenging problem due to the structure of the cancerous cells. This paper presents segmentation of the suspected lung nodules from the input Computed Tomography (CT) image by K-means clustering algorithm and classification by Artificial Neural Network (ANN). Initially the input CT image is preprocessed to remove noise. Then, the suspected lung nodules are segmented from the input CT image using K-means clustering algorithm. For the suspected nodules, the features are extracted and given as input to the ANN, which classifies whether the suspected nodules are benign (normal) or malignant (cancerous) in early stages.

Index Terms Image Segmentation, Image Classification, Nodule, K-means clustering, Artificial Neural Network (ANN), Computed Tomography (CT).

I. INTRODUCTION

As per the *United States Cancer Statistics: 2007 Incidence and Mortality Web-based report (USCS)*, the three most common cancers among men include Prostate cancer, Lung cancer and Colorectal cancer. The three most common cancers among women include Breast cancer, Lung cancer and Colorectal cancer. The leading cause of cancer death among men and women is Lung cancer. Lung cancer can be broadly classified into two main types based on the cancer's appearance under a microscope: Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC). NSCLC accounts for 80% of lung cancers, while SCLC accounts for the remaining 20%. SCLC is characterized by small cells that multiply quickly and form large tumors that travel throughout the body. Almost all cases of SCLC are due to smoking. NSCLC can be further divided into four different types: Squamous cell carcinoma or epidermoid carcinoma, Adenocarcinoma, Bronchioalveolar carcinoma, Large-cell undifferentiated carcinoma. Lung Cancer is caused by smoking, passive smoking, carcinogens, genes, viruses, particulate matter, asbestos fibers, radon gas, and air pollution. Cancer symptoms are quite varied and depend on where the cancer is located, where it has spread, and how big the tumor is. Lung cancer symptoms may take years before appearing, usually after the disease is in an advanced stage. Many symptoms of lung cancer affect the chest and air passages. These include persistent or intense coughing, pain in the chest shoulder, or back from coughing, changes in color of the mucus that is coughed up from the lower airways (sputum), difficulty breathing and swallowing, hoarseness of the voice, harsh sounds while breathing (stridor), chronic bronchitis or pneumonia, coughing up blood, or blood in the sputum. Lung cancer is usually found in older persons because it develops over a long period of time. The average five-year survival rate after lung cancer diagnosis is about 15%. If lung cancer is detected at its earliest stage, the five-year survival rate can reach 70%.

II. RELATED WORK

In the literature, there are number of proposals for detecting the cancerous nodules from lung images. To assess the benefit of Artificial Neural Network to diagnosing lung cancer the systematic review was conducted by N.Ganesan et. al [1]. The CAD system to diagnose the lung cancer based on ANN to assist radiologists in distinguishing malignant from pulmonary nodules was proposed by Yongjun et. al [2]. In neural network based classifier, the difficult task is designing of ANN structure. To provide the solution for the above problem Genetic algorithm (GA)-ANN hybrid intelligence was described by Fazil Ahmad et. al [3]. In their approach GA was used to select significant features simultaneously as input to ANN and optimal number of

hidden node is determined automatically. A feature extraction model described by Vivekanandan et. al [4] for assessing the growth of lung cancer in computer aided diagnosis makes use of snake algorithm for segmenting the suspected lung nodules and nearest neighbor (NN) for classification. The Lung cancer diagnosis system described by Fatma Teher et. al [5] deals the fuzzy clustering technique for segmentation and proved that fuzzy clustering is better than the Hopefield neural network. Fuzzy logic is a rule based system it uses if-then rules. Fuzzy logic has been found applications in many areas from control theory to artificial intelligence. The Fuzzy logic is proved to be a potential tool for decision making systems such as pattern recognition and expert systems. Nguyen Hoang Phuong et.al [6] described the Fuzzy set theory and Fuzzy logic can be used in medicine to develop the knowledge based system. An automated method to detect the lung nodules based on Wavelet transform, Bi-histogram equalization, Morphology filter and Fuzzy logic was proposed by C. Clifford Samuel et. al [7], however in their work the bronchovascular details are also detected as nodules. M.A. Saleem Durai et. al [8] described the technique to determine disease name, stage and diagnostic treatment of the cancer patient using Fuzzy rules but the diagnosis were complex because of the analysis of all the information gathered about the symptoms.

III. PROPOSED WORK

The proposed work is carried in two stages. In the first stage the suspected lung nodule is segmented. In the second stage the ANN classifier makes the diagnosis from the extracted features from the suspected nodules. The flow diagram of the proposed method is shown in fig.1.

The proposed method consists of four steps. They are Preprocessing, Lung Nodule Segmentation and Feature Extraction and ANN Classification.

A. Preprocessing

In image analysis, before the image is processed it needs to be preprocessed. The preprocessing aims to reduce the noise in the input CT lung images. The preprocessing is achieved by either increasing the contrast of the image or suppressing the noise. The noise distribution in CT images follows the Gaussian distribution therefore Wiener filter is employed to remove the noise. The Wiener filter is an optimum filter

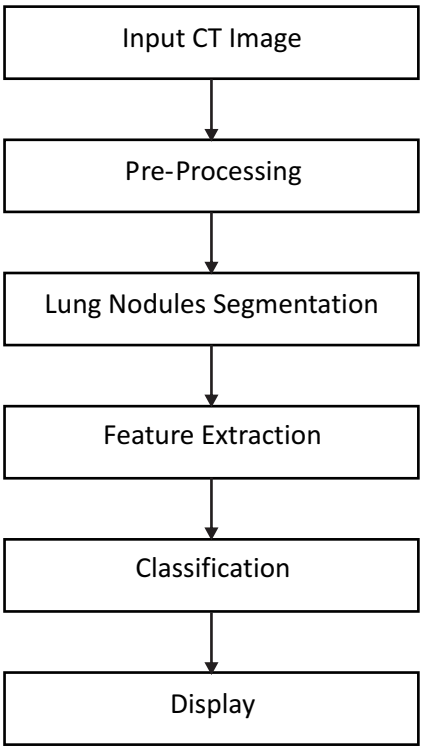

Fig. 1. Flow diagram of proposed method

which minimizes the means square error hence it maximizes the image quality. In order to balance between the noise removal and over blurring the size of the Wiener filter is selected as 3×3. The metrics used to evaluate the performance of filters are Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR). The PSNR is a measure of the peak error whereas the MSE is the cumulative squared error between the filtered and the original image. The mathematical formulae for the above two are,

$$MSE = 1/MxN\{ g(x, y) - g'(x, y)]^2 \} \quad (1)$$

$$PSNR = 20 \times \log_{10} [255/sqrt(MSE)] \quad (2)$$

where, $g(x, y)$ is the original image, $g'(x, y)$ is the filtered image and M,N are the dimensions of the image.

2. Segmentation

Segmentation refers to the process of partitioning digital image into multiple segments or sets of pixels. The goal of segmentation is to simplify or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries in

images. Image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image. Each of the pixels in a region is similar with respect to some characteristic or computed property, such as colour intensity, or texture. Adjacent regions are significantly different with respect to the same characteristics. Clustering partitions a data set into several groups such that similarity within a group is larger than that among groups. Clustering algorithms are used extensively not only to organize and categorize data but are also useful for data compression and model construction. Clustering techniques are validated on the basis of two assumptions. One assumption is similar inputs to the target system to be modeled should produce similar outputs. Another assumption is the similar input-output pairs are bundled into clusters in the training data set.

C. K-means Clustering Algorithm

K-means algorithm is an unsupervised clustering algorithm that classifies the input data points into multiple classes based on their inherent distance from each other. The algorithm assumes that the data features form a vector space and tries to find natural clustering in them. The points are clustered around centroids $\mu_j A_j = 1 \ldots \ldots k = l::k$ which are obtained by minimizing the objective where there are k clusters $s_{ij} = 1, 2, \ldots k$ and μ_i is the centroid or mean point of all the points x_j * s_i. As a part of this project, an iterative version of the algorithm was implemented. The algorithm takes a 2 dimensional image as input. Various steps in the algorithm are as follows:

1. Compute the intensity distribution (also called the histogram) of the intensities.
2. Initialize the centroids with k random intensities.
3. Repeat the following steps until the cluster labels of the image do not change anymore.
4. Cluster the points based on distance of their intensities from the centroid intensities

$$c^{(i)} = argmin\, ||x^{(i)} - \mu_j||^2 \quad (3)$$

5. Compute the new centroid for each of the clusters.

$$\frac{\Sigma_{t=0}^{m} 1\{c(i) = j\}\, x^{(t)}}{\Sigma_{i=0}^{m} 1\{c_{(i)} = j\}} \quad (4)$$

where k is a parameter of the algorithm (the number of clusters to be found), i iterates over the all the intensities, j iterates over all the centroids and μ_i are the centroid intensities.

$$\mu_i = \frac{\Sigma_{i=0}^{m} 1\{c(i) = j\}\, x^{(i)}}{\Sigma_{j=0}^{m} 1\{c_{(i)} = j\}} \quad (5)$$

IV. FEATURE EXTRACTION AND FORMULATION OF DIAGNOSTIC RULES

After the segmentation is performed on lung region, the features can be obtained from it and the diagnosis rule can be designed to exactly detect the cancer nodules in the lungs. This diagnosis rules can eliminate the false detection of cancer nodules resulted in segmentation and provides better diagnosis.

A. Feature Extraction

The features that are used in this paper in order to generate diagnosis rules are:

- Area of the candidate region
- The Maximum Drawable Circle (MDC) inside the candidate region
- Mean intensity value of the candidate region

(i) Area of the candidate region

This feature can be used here in order to

- Eliminate isolated pixels.
- Eliminate very small candidate object.

With the help of this feature, the detected regions that do not have the chance to form cancer nodule are detected and can be eliminated. This helps in reducing the processing in further steps and also reduces the time taken by further steps.

(ii) The Maximum Drawable Circle (MDC)

This feature is used to indicate the candidate regions with its maximum drawable circle (MDC). All the pixels inside the candidate region is considered as center point for drawing the circle. The obtained circle within the region is taken for consideration. Initially radius of the circle is chosen as one pixel and then

the radius is incremented by one pixel every time until no circle can be drawn with that radius. Maximum drawable circle helps in the diagnostic procedure to remove more and more false positive cancerous candidates.

(iii) Mean intensity value of the candidate region

In this feature, the mean intensity value for the candidate region is calculated which helps in rejecting the further regions which does not indicate cancer nodule. The mean intensity value indicates the average intensity value of all the pixels that belong to the same region and is calculated using the formula:

$$Mean\,(j) = \frac{\Sigma_{i=0}^{n}\,\text{intensity}\,(i)}{n} \tag{6}$$

where j characterizes the region index and ranges from 1 to the total number of candidate regions in the whole image. *intensity(i)* indicates the CT intensity value of pixel i, and i ranges from 1 to n, where n is the total number of pixels belonging to region j.

B. Formulation of Diagnostic Rules

After the necessary features are extracted, the following diagnosis rules can be applied to detect the occurrence of cancer nodule. There are three rules which are involved are as follows:

Rule 1: Initially the threshold value T1 is set for area of region. If the area of candidate region exceeds the threshold value, then it is eliminated for further consideration. This rule will helps in reducing the steps and time necessary for the upcoming steps.

Rule 2: In this rule maximum drawable circle (MDC) is considered. The threshold T2 is defined for value of maximum drawable circle (MDC). If the radius of the drawable circle for the candidate region is less than the threshold T2, then that is region is considered as non-cancerous nodule and is eliminated for further consideration. Applying this rule has the effect of rejecting large number of vessels, which in general have a thin oblong, or line shape.

Rule 3: In this, the rage of value T3 and T4 are set as threshold for the mean intensity value of candidate region. Then the mean intensity values for the candidate regions are calculated. If the mean intensity value of candidate region goes below minimum threshold or goes beyond maximum threshold, then that region is assumed as non-cancerous region. By implementing all the above rules, the maximum of regions which does not considered as cancerous nodules are eliminated. The remaining candidate regions are considered as cancerous regions. This CAD system helps in neglecting all the false positive cancer regions and helps in detecting the cancer regions more accurately. These rules can be passed to the Extreme learning machine (ELM) in order to detect the cancer nodules for the supplied lung image.

V. ARTIFICIAL NEURAL NETWORK CLASSIFIER

Artificial Neural Networks (ANN) are networks of interconnected computational units, usually called nodes. The input of a specific node is the weighted sum of the output of all the nodes to which it is connected. The output value of a node is, in general, a non-linear function (referred to as the activation function) of its input value. The multiplicative weighing factor between the input of node j and the output of node i is called the weight w_{ji}. An Artificial Neural Network is an adaptive, most often nonlinear system that learns to perform a function (an input/output map) from data. Adaptive means that the system parameters are changed during operation, normally called the Learning/Training phase. After the training phase the Artificial Neural Network parameters are fixed and the system is deployed to solve the problem at hand (The Recognition/Testing phase). Back-propagation ANN's used in this study consist of one input layer, one or two hidden layers, and one output layer. With back-propagation, the input data (Extracted Features) is repeatedly presented to the Artificial Neural Network, with each presentation the output of the neural network is compared to the desired output and an error is computed. This error is then fed back (back-propagated) to the Artificial Neural Network and used to adjust the weights such that the error decreases with each iteration and the neural model gets closer and closer to producing the desired output. This process is known as Training. The Training of these networks consists in finding a mapping between a set of input values and a set of output values. This mapping is accomplished by adjusting the value of the weights w_{ji} using a learning algorithm, the most popular of which is the generalized delta rule. After the weights are adjusted on the training set, their value is fixed and the ANN's are used to classify unknown input

images. The generalized delta rule involves minimizing an error term defined as

$$E_p = \frac{1}{2} \Sigma_j^3 (t_\pi - o_\pi)^2 \quad (7)$$

In this equation, the index p corresponds to one input vector, and the vectors t_p and o_p are the target and observed output vectors corresponding to the input vector p, respectively.

VII. RESULTS

The benefits of K-mean clustering are its simplicity and its speed which allow us to run on large datasets. The suspected lung nodules are shown in Fig. 2.

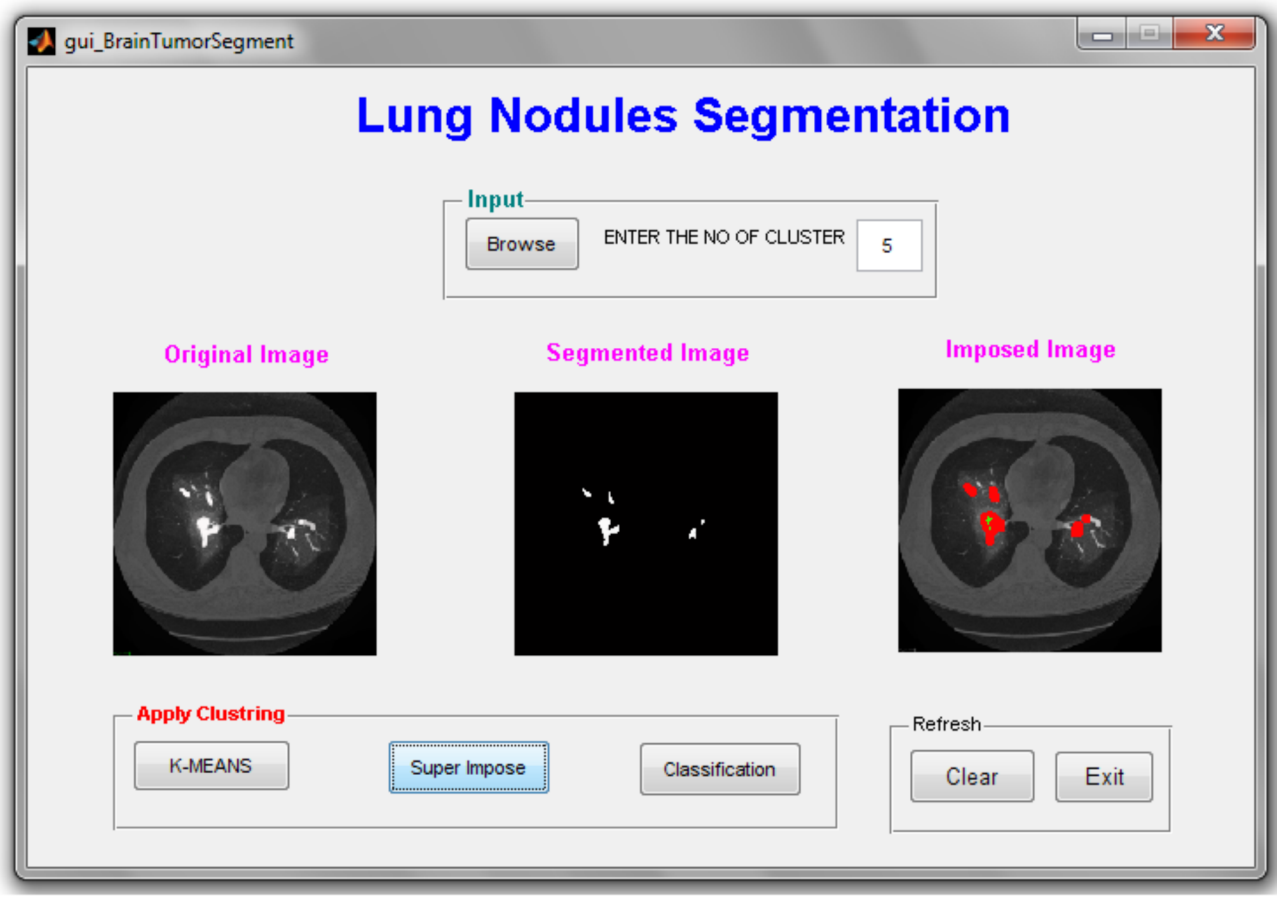

Fig. 2 Lung Nodules Segmentation

Based on the extracted features, the suspected lung nodules are classified as benign or malignant. The above results show that the system works efficiently for Detection and Classification of Lung cancer.

VI. CONCLUSION

In this paper, the lung cancer diagnosis based on clustering algorithm and Artificial Neural Network is proposed. The proposed system consists of preprocessing, segmentation of the suspected lung nodules, feature extraction and classification to detect the lung cancer. Future work includes analyzing other clustering algorithms like Fuzzy C-means, Modifies Fuzzy C-means for image segmentation and Extreme Learning Machine for image classification.

REFERENCES

[1] Ganesan. N, Venkatesh. K and Rama: Application of Neural Networks in diagnosing cancer diseases using Demographic data" International journal of computer applications, Volume 1, 2010.

[2] Yongjun WU, Na Wang, Hongshezhang: Application of Artificial Neural Networks in the Diagnosis of Lung Cancer by Computed Tomography, China, Sixth International Conference on Natural Computation (ICNC2010).

[3] Fadzil Ahmad, Nor Ashidi Mat-Isa and Rozan Boudville: Genetic algorithm-Artificial Neural network Hybird intelligence for cancer diagnosis, Second international conference on Computational Intelligence, Communication Systems and Networks, 2010.

[4] Vivekanandan.D, Sunil Retmin Raj: A future extraction model for assessing the growth of lung cancer in computer aided diagnosis, IEEE International conference on Recent trends in Information Technology, 2011.

[5] Fatma Taher, Rachid sammouda: Lung cancer detection by using Artificial Neural Network and Fuzzy clustering methods, IEEE GCC conference and exhibition, 2011.

[6] Nguyen Hoang Phuong and Vladik Kreinovich: Fuzzy logic and its application in medicine, IEEE transactions on Medical imaging, Vol. 10, No. 6, December 2005.

[7] Clifford Samuel.C, Saravanan.V, VimalaDevi.M.R: Lung nodule diagnosis from CT images using fuzzy logic, International Conference on Computational Intelligence and Multimedia Applications 2007.

[8] Saleem Durai.M.A and N.Ch.S.N.Iyengar: Effective analysis and diagnosis of lung cancer using Fuzzy rules, International journal of Engineering Science and Technology, Vol 2, 2010.

[9] Gonzalez, Digital Image Processing and its applications, Pearson education 2010.

[10] Kenji Suzuki, Junji Shairaishi: False-Positive reduction in computer-aided diagnostic scheme for detecting in chest radiographs by means of massive training artificial neural network, Elsevier, Volume 12, issue 2, pp. 191-205, 2005.

[11] Ayman El-Baz, Georgy Gimel'farb, Robert Falk: Promising results for early diagnosis of lung cancer, IEEE transactions on Medical imaging, 2008.