Detection of Lung Cancer at Early Stage using Neural Network Techniques for Preventing Health Care

Megha Bhatnagar¹ Bhawana Malik² Shashi Bhushan Tyagi³ Prashant Naresh⁴

¹,²,³,⁴Department of Computer Science and Engineering
¹,³UTU, Uttarakhand, India ²,⁴UPTU, Lucknow, India

Abstract— Nowadays cancer has become huge threat in human life. There are different kinds of cancer. Lung cancer is common type of cancer causing very high ephemerality rate. Lung cancer is a serious illness which can be cured if it is diagnosed at prior stages. In this paper, we address the problem of extraction and segmentation the sputum cells based on the analysis of sputum color image with the aim to attain a high specificity rate and reduce the time consumed to analyze such sputum samples. A lung cancer risk prediction system is proposed here which will detect lung cancer at an early stage using CT scan images of DICOM format. One of the key challenges is to remove white Gaussian noise from the CT scan image, which is done using non local mean filter and to segment the lung Otsu’s thresholding is used. The textual and structural features are extracted from the processed image to form feature vector. Image Processing plays significant role in cancer detection. Various techniques used in Image Processing for information retrieval are Image acquisition, Noise Removal, Segmentation, and Morphological operations etc. In this approach, three classifiers are applied for the detection of lung cancer to find the severity of disease. It has been found from results that SVM classifier achieves higher accuracy of 95.12% while ANN classifier achieves 92.68% accuracy on the given data set and k-NN classifier shows least accuracy of 85.37%.

Key words: Computer aided diagnosis, SVM, ANN, k-NN, CT-Scan images, Feature Extraction, Segmentation

I. INTRODUCTION

Cancer has become a big threat to humans globally among various diseases, as per Indian population census data. According to GLOBOCAN 2012, an estimated 14.1 million new cancer cases and 8.2 million cancer-related deaths occurred in 2012, compared with 12.7 million and 7.6 million, respectively, in 2008. The most commonly diagnosed cancers worldwide were those of the lung (1.8 million, 13.0% of the total) [1]. Lung cancer is considered to be the main cause of cancer death, and in its prior stages it is hard to detect because only in the advanced stage symptoms appear causing the ephemerality rate to be the highest among all other types of cancer [2]. If lung nodules can be recognised accurately at a prior stage, the patient’s survival rate can be increased by a significant percentage.

Lung cancer is caused by uncontrollable irregular growth of cells in lung tissue. These lung tissue abnormalities are often called Lung nodules. They are small and roughly spherical masses of tissue, usually about 5 millimeters to 30 millimeters in size. Researchers are becoming more and more concerned with the elaboration of automated CAD systems for lung cancer. Many publications proposed different automated nodule recognition systems using image processing, and including, different techniques for segmentation, feature extraction and classification. In the modern era of computerized fully automated course of living, the field of automated diagnostic systems plays a vital role. Automated diagnostic system designs in Medical Image Processing is one such field where numerous systems are proposed and still many more under conceptual design due explosive growth of the technology today [3]. Data mining tool has proved to be successful in disease diagnosis [4].

Image processing is a method to convert an image into digital format and perform various operations on it to extract useful information from it. Extraction of certain features that characterize the nodule, but excludes the insignificant attributes is the way of describing nodule. Image segmentation is the process of assigning a label to every pixel in an image in such a way that pixels with the same label share certain visual characteristics. After features are extracted, algorithms are used for classifying the data into the categories.

This paper is organized into six sections. In section 2 related work carried out in this field is described. In section 3 proposed methodology for early detection of cancer is explained. In section 4 system architecture of nodule predictor is discussed. In section 5 experimental result and discussion is explained followed by conclusion in section 6.

II. RELATED WORK

In this section, some of the works on prediction of lung cancer, pre-processing, segmentation and classification techniques have been discussed:

A two stage scheme for automatic lung nodules detection in Multi-Slice Computed Tomography (MSCT) [5] scans with multiple SVMs to reduce number of false positive with accuracy of 87.82% is presented. An automatic CAD system [6] of 80% accuracy is developed for early detection of lung nodule by analysing LUNG CT images using several steps. First, several image processing techniques are applied to extract the lung regions from the CT image. After extraction, the extracted lung regions are segmented using region growing segmentation algorithm. Then rule based technique is applied to classify the cancer nodules.

In the image processing procedures of [7], processes such as image pre-processing, segmentation and feature extraction are discussed in detail. An objective and quantitative assessment of centrosomal numeral and morphological abnormalities in [8] is given and the magnitude of these differences is shown. Regions of interest were selected to include one cell and its centrosomes. After segmentation, feature abstraction, and optimization, six non
redundant features were used for statistical analysis and classification.

Optimal thresholding [9] is applied to the denoised image to segregate lung regions. Region growing method is used to segment Lung nodules which are of relatively high density found within the lung regions. An efficient lung nodule detection scheme of accuracy 80.36% [10] is developed by performing nodule segmentation through weighted fuzzy probabilistic [11] based clustering is carried out for lung cancer images. An automatic Computer-Aided Detection (CAD) scheme in [12] is presented having accuracy 95% that can identify the pulmonary nodule at an early stage from CT images. A database is used to focus on denoising task in order to determine the benefits and drawbacks for each algorithm [13].

This study present an overview of different algorithm for classification and image processing used in the field of lung cancer prediction. Summary of various segmentation and classification techniques with their classification accuracy and sensitivity of nodule detection has been presented, based on above s it has been found that not much work has been carried out for early detection of lung cancer. Hence it has been taken up here.

III. METHODOLOGY
Methodology is composed of two phases.

1) In first phase, the CT scan images are pre-processed to remove Gaussian white noise by using non-local mean filter. As the accuracy of the segmentation algorithm depends on the quality of images so, the images are cleansed and segmented using Otsu’s thresholding [14] and then, the textural and structural features are extracted from the segmented image by the application of feature extraction techniques.

2) In the second phase, the SVM classifier is implemented and then it is trained and tested on sample data for the prediction of lung cancer. The output of the classifier is checked for accuracy which is a measure of how accurately the classifier predicts the status of patient.

IV. SYSTEM ARCHITECTURE OF NODULE PREDICTOR
Lung Nodule prediction aims to automatically predict the information of nodule presented in lung’s medical images by using dedicated computer systems to provide ‘second opinions’. CAD is a relatively young interdisciplinary technology combining elements of artificial intelligence and digital image processing with radiological image processing. CAD systems and technology show that CAD can help to improve diagnostic accuracy of radiologists lighten the burden of increasing workload; reduce cancer missed due to fatigue [15].

CAD system takes CT scan images of as input and provides status of patient as output on the basis of classifiers. In lung CT image segmentation process, Gaussian noise is removed from CT scan image which is most common type of noise present in medical images. After that, segmentation is done using Otsu’s thresholding to segment the lung part in an image. post processing enhancement is done to get clear image for detection of nodule (tumor) by detecting boundary in image using canny edge detection. Then, two largest regions are filled to remove extra muscle part from an image except lungs. In nodule’s feature extraction module, output of post processing is given as input to extract textural and structural feature of nodule after that SVM classifier is trained and tested on the basis of those features to provide final output i.e severity of the disease. The figure.1 depicts the system architecture of CAD system.

A. Pre-Processing
Pre-processing is the method to correct different kind of errors in images, done before processing. Pre-processing stage of the images is necessary to improve the quality of the images and make the feature extraction phase more reliable and make it available for other phase. Pre-processing stage is important of a Computer Aided Diagnosis system for prediction of lung cancer lies in its ability to remedy some of the problems that may occur due to some factors.

B. Image Segmentation
The term image segmentation refers to the partition of an image into a set of regions. It refers to the process of partitioning the pre-processed CT image into multiple regions to separate the pixels or voxels corresponding to lung tissue from the surrounding anatomy. Image segmentation is usually used to locate objects and boundaries (lines, curves, etc.) in images. The goal in many tasks is to find regions that represent meaningful area of an image. Thresholding, clustering, comparison based, histogram based, edge detection and region growing are several general-purpose techniques have been developed for image segmentation.
C. Post Processing

Post processing is done so that the lung image will become clearer in order to detect nodules. The post processing means filling and thinning. Thinning brings down the width of the line. While filling gets rid of small breaks and holes in the boundary, separate extra part from image and make image more clear for nodule detection.

D. Feature Extraction

Feature Extraction captures the main characteristics of the nodule, and it is usually accepted that this is one of the most challenging problems of nodule prediction. We can characterize the nodule by extraction of certain features, to define the nodule. It excludes the insignificant attributes. There are two main types of feature descriptors, namely textural features and structural features.

1) Textural Feature:
Computes the structural features value of nodule i.e. Energy, Mean, and Standard Deviation.
- ENERGY: is used to describe measure of information in an image, represented in equation (2.1).
  \[
  \text{energy}(j) = \sum_{\kappa} \text{Intensity}(\kappa)^2 
  \]
  (2.1)
- MEAN: The mean intensity value indicates the average intensity value of all the pixels that belong to the same region, calculated using equation (2.2).
  \[
  \text{Mean}(g) = \frac{1}{N} \sum_{\kappa=1}^{N} \text{Intensity}(\kappa) 
  \]
  (2.2)
- STANDARD DEVIATION: is a measure of how much that gray levels differ from mean, defined by equation (2.3).
  \[
  \text{std}(g) = \frac{1}{N} \sum_{\kappa=1}^{N} (\text{Mean}(\kappa) - \text{Intensity}(\kappa))^2 
  \]
  (2.3)

2) Structural Feature:
Computes the structural features value of nodule i.e. Area, Convex Hull Area, Equiv Diameter and Solidity.
- AREA: It is a scalar value that gives the actual number of pixels in the Region Of Interest (ROI).
- CONVEX AREA: Scalar value that gives the number of pixels in convex image of the Region Of Interest which is a binary image with all pixels within the hull filled in.
- EQUIV DIAMETER: It is the diameter of a circle with the same area as the Region Of Interest, defined in (2.4).
  \[
  \text{Equiv diameter} = \frac{4*\text{area}}{\sqrt{\pi}} 
  \]
  (2.4)
- SOLIDITY: It is the proportion of the pixels in the convex hull that are also in the Region Of Interest as defined in (2.5).
  \[
  \text{solidity} = \frac{\text{area}}{\text{convex area}} 
  \]
  (2.5)

The extracted features are then used by the classifier either for training the system or for classifying the nodule.

E. Classification

After features are extracted, algorithms are used for classifying the data into the group. These are also known as classifiers. There are two kinds of classification: supervised classification and unsupervised classification. If the feature vectors are given with known labels (the corresponding correct outputs), then the training is called to be supervised. If the classifier categorizes the data automatically without any use of class labels, then it’s known to be unsupervised classification. Supervised classifiers are Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Artificial Neural Networks (ANN), and Classification Trees. Unsupervised classifiers are k-means clustering, mixture models, hierarchical clustering, Principal Component Analysis, Singular Value Decomposition etc [16].

SVM, ANN and k-NN are three supervised classification techniques discussed here.

1) SVM Classifier:
SVMs are a set of related supervised learning methods used for classification and regression. An SVM will construct a separating hyperplane for viewing input data as two sets of vectors in an n-dimensional space, we check the one which maximizes the margin between the two data sets. Then construct two parallel hyper-planes for calculating margin. Good separation is achieved by the hyper-plane to calculate the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier.

The MMH can be rewritten as the decision boundary [17] as described in equation (2.6).

\[
\begin{aligned}
  d(X_T) &= \sum_{i=1}^{l} y_i \alpha_i X_i X_T^T + b_0 \\
  \end{aligned}
\]

Where, \( y_i \) is the class label of support vector \( X_i \). \( X_T \) is a test tuple.

\( \alpha_i \) and \( b_0 \) are numerical parameters they were determined automatically by the optimization or SVM algorithm.

\( l \) is the no. of support vector.

Kernel function is used to make (implicit) nonlinear feature map e.g. Polynomial kernel, Radial basis function (RBF) kernel.

2) RBF Kernel:
The RBF kernel as in equation (2.7) is one of the most popular kernel functions. It adds a “bump” around each data point. The RBF kernel is a measure of similarity between two examples. The feature space is infinite-dimensional.

\[
K(x; x') = \exp(-\gamma \| x - x' \|^2) 
\]

(2.7)

3) ANN Classifier:
ANN usually called “neural network”, is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Neural networks are non-linear statistical data modeling tools. It can be modelled to define complex relationship.
4) **K-NN Classifier:**
The k-nearest neighbors algorithm (k-NN) is a method for classifying objects that is based on closest training examples in the feature space. Instance-based learning is same as K-NN, or lazy learning where the function is only approximated locally and all computation is deferred until classification.

V. EXPERIMENTAL RESULTS AND DISCUSSION
The various Performance Metrics (Accuracy, Precision, Recall, and Specificity) for the test data are shown in Table 1. The formulations are shown in percentage, each column indicates the neural networks Classifier used and the rows indicate the Metric value respectively.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>SVM</th>
<th>ANN</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>95.12</td>
<td>92.68</td>
<td>85.37</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>92.31</td>
<td>87.50</td>
<td>84.62</td>
</tr>
<tr>
<td>Recall (%)</td>
<td>100.00</td>
<td>100.00</td>
<td>91.67</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>88.24</td>
<td>100.00</td>
<td>76.47</td>
</tr>
</tbody>
</table>

Table 1: Performance Metrics In Percentage For Test Data

Fig. 2: Performance metrics of Classifiers
These results shown in figure 2 with 24 images of stage I and 17 images of stage II is used as test dataset. Value of accuracy, precision, recall and specificity is in percentage (%).

From experimentation of the technique, the CT images are obtained from a NIH/NCI Lung Image Database Consortium (LIDC) dataset. These images are progressed to this system. The diagnosis rules are then produced from these images and these rules are progressed to the classifier for the learning procedure. After learning operation a lung image is progressed to the proposed system. Then the proposed system executes its processing steps and finally it will detect whether the supplied lung image is with cancer or not. The presented CAD system is capable of detecting lung nodules with diameter ≥ 2.45 mm, which means that the system is capable of detecting lung nodules when they are in their initial stages. Thus early diagnosis of cancer will improve the patient’s survival rate.

VI. CONCLUSION
The field of Disease Diagnosis is a continuously evolving and very active field for research. The purpose of the current approach is to predict the status of patient for initial stage detection of lung cancer. A novel approach for predicting Lung cancer nodule at prior stage using SVM Classifier has been presented here. The Structural and Textural Features have been used for reporting the nodule. The results got are very stimulating, data was tested on Support Vector Machine Classifier with RBF kernel obtained an accuracy of 95.12%. The classification rates obtained for the SVM, ANN and k-NN Classifier were 95.12%, 92.68% and 85.37% respectively.

ACKNOWLEDGMENT
Initially, we would like to thank our almighty in the success of completing this work. We would extend our gratitude to all the experts for their critical comments.

REFERENCES


