RESEARCH ARTICLE OPEN ACCESS

Early Detection of Lung Cancer Using Neural Network Techniques

Prashant Naresh*, Dr. Rajashree Shettar**
*Dept. of CSE, RVCE, Bangalore
**Professor and Assoc.Dean (PG-CSE), Dept. of CSE, RVCE, Bangalore

ABSTRACT
Effective identification of lung cancer at an initial stage is an important and crucial aspect of image processing. Several data mining methods have been used to detect lung cancer at early stage. In this paper, an approach has been presented which will diagnose lung cancer at an initial stage using CT scan images which are in Dicom (DCM) 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 textural and structural features are extracted from the processed image to form feature vector. In this paper, three classifiers namely SVM, ANN, and k-NN are applied for the detection of lung cancer to find the severity of disease (stage I or stage II) and comparison is made with ANN, and k-NN classifier with respect to different quality attributes such as accuracy, sensitivity(recall), precision and specificity. It has been found from results that SVM achieves higher accuracy of 95.12% while ANN achieves 92.68% accuracy on the given data set and k-NN shows least accuracy of 85.37%. SVM algorithm which achieves 95.12% accuracy helps patients to take remedial action on time and reduces mortality rate from this deadly disease.

Keywords: Dicom, SVM, ANN, k-NN, Accuracy, Sensitivity, Precision and Specificity.

I. INTRODUCTION

Cancer is the most disastrous and life threatening disease to human beings globally among various diseases. Cancer is the second largest disease in India which is responsible for maximum mortality about 0.3 million deaths per year [1]. According to GLOBOCAN 2012 statistics, it was found that around 14.1 million new cases were diagnosed and around 8.2 million deaths occurred in 2012, due to cancer which is quiet high when compared to statistics of 2008 which was 12.7 million new cases and 7.6 million deaths due to cancer. According to study [2] it has been found that out of all the cancers, Lung Cancer is the main cause of mortality worldwide amongst all types of cancers. Main reason behind high rate of mortality due to lung cancer is that it is not easily detected in the initial stage and it is very difficult to overcome this disease at later stages of cancer [3]. If lung nodules can be identified accurately at an early stage, the patient's survival rate can be increased by a significant percentage. In today's era, the field of automated diagnostic systems plays crucial role in the diagnosis of any disease. Image Processing is one such field where automated diagnostic system designed especially for medical diagnosis leads to solution which will help in decreasing the mortality rate and these medical diagnostics systems helps in detecting the disease at initial filed which is very remarkable in the field of bioinformatics [5].

Data mining provides the methodology and technology to analyze the useful information from data for decision making. Extracting meaningful knowledge from the voluminous database is an important aspect of data mining. As the size of data increases exponentially some techniques are needed which will proven to be helpful for the extraction of relevant data and predicting the outcome of the disease is one of the most interesting and challenging tasks of data mining [6]. Some tools developed using data mining approaches proved to play a significant role in medical diagnosis [7]. Some applications where Data mining used in the diagnosis of cancer are: cancer lesion detection [8], pulmonary nodule detection [9], and classification of cancer stage from tree-text histology report [10], breathe biomarker detection [11] and so on. According to classification criteria data mining tasks are categorized in two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the already stored data. Predictive mining tasks derive conclusion on the basis of current data. Broadly used Data Mining Learning Approaches for Data mining algorithms are classified as: supervised, unsupervised, or semi-supervised. In supervised learning, the algorithm works with a set of examples whose class labels are known. These labels are nominal values if the task if of classification type and if the task is regression task these labels are numerical values. In unsupervised learning, in contrast, the labels of the examples in the dataset are

unknown, and the algorithm itself aims at grouping some data according to the similarity score of their attribute values, which is commonly known as clustering task. Finally, semi-supervised learning is a combination of above discussed two approaches in which small subset of labeled examples is available together with small subset of unlabelled examples. The classification task can be seen as a supervised technique where each instance belongs to a set of some labeled class [12].

This paper is organized into five sections. In section 2 related work carried out in this field is described. In section 3 proposed model for early detection of cancer is explained. In section 4 experimental results are discussed followed by conclusion in section 5.

II. RELATED WORK

Automatic lung nodule detection scheme in [4] is presented in Multi-Slice Computed Tomography (MSCT) scans using SVM. An automatic CAD system [13] is developed for early detection of lung nodule by analyzing LUNG CT images which achieves 80% result accuracy. Kernel RX-algorithm [14] which is a nonlinear anomaly detector is applied to CT images for malignant nodule detection. An automated computerized system for the detection of lung cancer in CT scan images consists of two stages [15]: a) lung segmentation and enhancement, b) feature selection and classification. Sensitivity of 95% in [16] is achieved. In [17] the efficiency of the diagnosis system for lung cancer is improved through a region growing segmentation method applied to segment CT scan lung images. In [18] assessment of centrosomal numeral and morphological abnormalities is presented and the magnitude of these differences is shown. Linear Discriminant Analysis (LDA) [19, 20] and support vector machines (SVM) with 10-fold cross validation used for classification and gets an accuracy of 85%. An automated system using hybrid approach which includes image processing and data mining techniques for the prediction of lung tumor from Computed Tomography (CT) images is presented in the paper [21] through image processing techniques coupled with neural network classification as either benign or malignant is presented. For the segregation of lung regions an approach known as Optimal thresholding, is applied to the de noised images in [22]. In [23] lung nodules are detected for the dataset taken from Lung Image Database Consortium (LIDC) [12] and techniques like acquisition of image from the database, background removal and detection of nodule for lung nodule detection have been applied. Computer aided lung nodule detection scheme is presented in [24] which works upon the analysis of enhanced voxel in three dimensional (3D) CT image and evaluated the performance of the proposed scheme on two CT data sets. An efficient lung nodule detection scheme with accuracy of 80.36% in [25] is developed by performing nodule segmentation through weighted fuzzy probabilistic method In [26] clustering is carried out for lung cancer images. Another hybrid approach is presented in the paper [27] which is a combination of image processing and data mining techniques. In this paper lung segmentation is done using Genetic Algorithm (GA) [28] and morphological image processing techniques. GA is applied on the normalized histogram for the determination of threshold values for the separation of background and object. Susan thinning algorithm is used in [29] to reduce the borders to the width of one pixel. An automatic Computer-Aided Detection (CAD) scheme in [30] is presented which achieves accuracy of 95% for the prediction of pulmonary nodule at an initial stage from CT images. Bayesian classification and Hopfield Neural Network algorithm [31,32] for extracting and segmenting the sputum cells is presented for the purpose of lung cancer early diagnosis. A new classification method has been proposed [33] known as iterative linear discriminant analysis which is used in addition to fuzzy c-means clustering for reducing the false positive values to enhance accuracy of the classifier.

III. PROPOSED METHOD

In the proposed method, CT scan images of lung cancer patients are taken as input. During preprocessing Gaussian white noise is removed. Figure 1 shows the lung CT image and its different regions. Segmentation is done to segment the lung part in an image. The preprocessed image is then fed to feature extraction phase in which textual and structural features of nodule are extracted and fed to the classifier. The classifier is trained and tested, it is then used to predict the severity of the cancer patient (stage-I or stage-II).

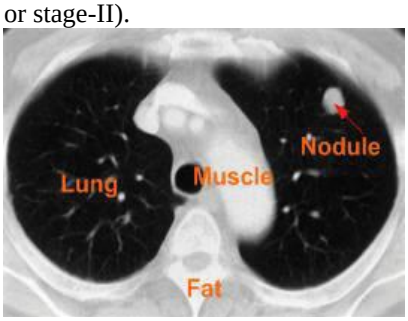


Fig.1: Lung CT Scan Image

Post processing enhancement is done to get clear image for detection of nodule (tumor). In nodule's feature extraction module, output of post processing is given as input to extract feature of nodule and classifier is trained and tested on the basis of those features to provide final output i.e. severity of the disease. Section A and B describes the lung CT

image segmentation and post processing enhancements of the image.

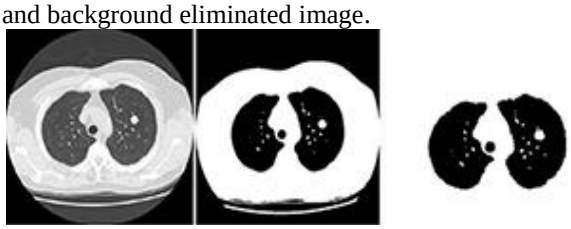
A. Lung CT Image Segmentation

Segmentation of an image involves the separation of lung nodule from other part of the CT scan images and then enhancement of the resultant image to get details. This process includes series of steps which are listed below:

1) The input image is converted to gray image and Non Local Mean filter is applied to remove Gaussian white noise.

2) Otsu's threshold is used to do segmentation of lung part from lung CT image.

Figure 2 shows the original image, segmented image and background eliminated image.



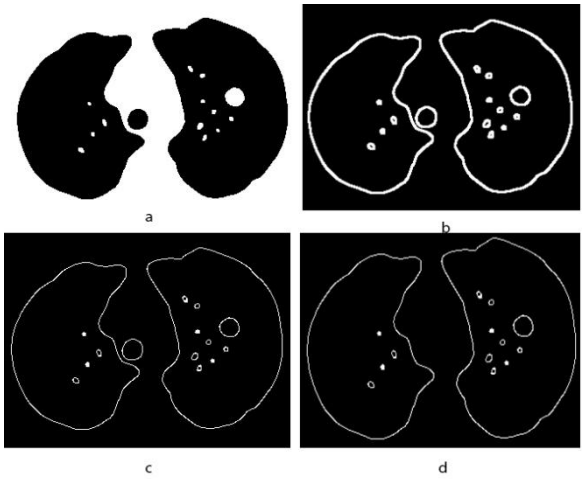
a b c

Fig 2: Segmentation: a) Original Image, b)Background removed Image, c)Threshold Image

B. Post processing Enhancement

The lung image will have much clarity with post processing enhancement in order to detect nodules. The series of steps evolved in enhancement after segmentation are listed below:

1) Small objects are eliminated by morphological opening present inside and outside the lungs in segmented image.
2) After that borders enhancement and the gaps in the border is filled by morphological closing.
3) After Morphological operation canny edge detection is used to detect boundary of the enhanced image.
4) Morphological thinning is then applied on the boundary extracted image.
5) To get the final post-processed image morphological filling is applied to remove extra muscle part from an image except the lungs. Figure 3 shows the post-processing enhancement process in detail.



a b c d

Fig. 3: Post-processing Enhancement a) Morphological Operations, b) Border Detected Image c) Border Thinned Image d) Filled Image

C. Lung Nodule Feature Extraction and Classification

The aim of Feature Extraction is to capture the necessary characteristics of the nodule, and it is usually accepted that this is one of the most challenging problems of nodule prediction. Extraction of certain features that characterize the nodule, but excludes the insignificant attributes is the way of describing nodule. Hence varying features for the lung nodule detection are considered and the feature vector thus formulated is $FV = \{F1, F2, F3, F4, F5, F6\}$. These features are described from section 2.3.1 to 2.3.2

3.3.1 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: It is a 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 (2.1).

$$\text{Equiv diameter} = \sqrt{\frac{4*\text{area}}{\sqrt{\pi}}} \tag{2.1}$$

SOLIDITY: It is the proportion of the pixels in the convex hull that are also in the Region Of Interest as defined in (2.2).

$$solidity = \frac{area}{convexarea} \tag{2.2}$$

3.3.2 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 (2.3).

$$energy(j) = \sum_{\kappa} Intensity(\kappa)^2 \quad (2.3)$$

MEAN: The mean intensity value indicates the average intensity value of all the pixels that belong to the same region, calculated using equation (2.4).

$$Mean(g) = \frac{1}{N}\sum_{\kappa=1}^{N} Intensity(\kappa) \quad (2.4)$$

STANDARD DEVIATION: is a measure of how much that gray levels differ from mean, defined by equation (2.5).

$$std(g) = \frac{1}{N}\sum_{\kappa=1}^{N} (Mean(\kappa) - Intensity(\kappa))^2 \quad (2.5)$$

For classification purpose, the feature vector is given as input to classifier. SVM Classifier has 3 functions to perform classification. Select data from database to train classifier for 2 classes. Selected feature input data is transformed into a high dimensional space using nonlinear mapping, then next step searches for linear separating hyper plane in the new space. Using following steps, SVM classifier is trained for 2 classes. This classifier is then used for predicting the lung cancer at early stage or predicts the status of patient.

IV. EXPERIMENTAL RESULTS

The dataset of the lung images considered are CT scan images of National Lung Screening Trial (NLST) data/images of stage I and stage II. The total number of sample images taken for experimentation is 111 for stage-I and 73 samples for stage-II type of lung cancer. Out of this four-fifth of the data is used for training and the remaining one-fifth is taken for testing the classifiers.

The Confusion Matrix for the SVM, ANN and KNN classification are shown respectively from table 1 to table 3. The tabulations are shown with respect to total number of images.

24 images of stage I and 17 images of stage II are in test dataset. Confusion matrix is shown in table 1. TP is 24, means 24 images of stage I are predicted as stage I, FP is 2, means 2 images of stage II are predicted as stage I, FN is 0, means no image of stage I are predicted as stage II. TN is 15, means 15 images of stage II are predicted as stage II.

Table 1: Confusion Matrix for SVM Classification

		Actual	
		Positive	Negative
Predicted	Positive	24	2
	Negative	0	15

24 images of stage I and 17 images of stage II are in test dataset. Confusion matrix is shown in table 2. TP is 17, means 17 images of stage I are predicted as stage I. FP is 0, means no images of stage II are predicted as stage I. FN is 3, means 3 images of stage I are predicted as stage II. TN is 17, means 17 images of stage II are predicted as stage II.

Table 2: Confusion Matrix for ANN Classification

		Actual	
		Positive	Negative
Predicted	Positive	21	0
	Negative	3	17

24 images of stage I and 17 images of stage II are in test dataset. confusion matrix is shown in table 3. TP is 22, means 22 images of stage I are predicted as stage I. FP is 4, means 4 images of stage II are predicted as stage I. FN is 2, means 2 images of stage I are predicted as stage II. TN is 13, means 13 images of stage II are predicted as stage II.

Table 3: Confusion Matrix for KNN Classification

		Actual	
		Positive	Negative
Predicted	Positive	22	4
	Negative	2	13

The various Performance Metrics (Accuracy, Precision, Recall, and Specificity) for the test data are shown in Table 4. The tabulations are shown in percentage, each column indicates the Classifier used and the rows indicate the Metric value.

Table 4: Performance Metrics in percentage for test data

		Classifier		
		SVM	ANN	KNN
Metrics	Accuracy(%)	95.12	92.68	85.37
	Precision(%)	92.31	87.50	84.62
	Recall(%)	100.00	100.00	91.67
	Specificity(%)	88.24	100.00	76.47

From table 4 it is shown that accuracy of SVM is 95.12% which is better that ANN (92.68%) and k-NN (85.37%). SVM predicts images of stage I and stage II more accurately.

V. CONCLUSION

The field of Disease Diagnosis is a continuously evolving and very active field of research. The intention of the current study was to predict the status of patient for early detection of lung cancer. A novel approach for predicting Lung cancer nodule at early stage using SVM Classifier has been proposed here. The Structural and Textural Features have been used for describing the nodule. The results got are very encouraging, data was tested on SVM Classifier with RBF kernel obtained an accuracy of 95.12%. A comparison of classification accuracy for ANN, KNN and SVM Classifiers was made on lung CT scan images of stage I and stage II. The classification rates obtained for the SVM, ANN and k-NN Classifier were 95.12%, 92.68% and 85.37% for the test images.

REFERENCES

[1] Imran Ali, Waseem A. Wani and Kishwar Saleem, "*Cancer Scenario in India with Future Perspectives*", Cancer Therapy, vol. 8, 2011, pp. 56-70.

[2] Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin D M, Forman D, Bray, F (2013). GLOBOCAN 2012 v1.0, *Cancer Incidence and Mortality World Wide: IARC Cancer Base No. 11, Lyon, France: International Agency for Research on Cancer.*

[3] S.Shaik Parveen, C.Kavitha, "*Detection of lung cancer nodules using automatic region growing meth*od", Proceedings of the 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013, pp. 201-206.

[4] Yang Liu, Jinzhu Yang, Dazhe Zhao, Jiren Liu, "*A Method of Pulmonary Nodule Detection utilizing multiple Support Vector Machines*", Proceedings of the International Conference on Computer Application and System Modeling (ICCASM 2010), 2010, pp. 118-121.

[5] Guruprasad Bhat, Vidyadevi G Biradar , H Sarojadevi Nalini, "*Artificial Neural Network based Cancer Cell Classification (ANN – C3)*", Computer Engineering and Intelligent Systems, vol. 3, (2), 2012, pp. 116-119.

[6] Juliet R Rajan1, Jefrin J Prakash, "*Early Diagnosis of Lung Cancer using a Mining Tool*", Proceedings of the National Conference on Architecture, Software systems and Green computing-2013, pp. 87-91.

[7] Ada, Rajneet Kaur, "*A Study of Detection of Lung Cancer Using Data Mining Classification Techniques*", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, (3), 2013, pp. 67-70.

[8] T. Jia , Y. Wei, D. Wu, "*A Lung Cancer Lesions Detection Scheme Based on CT Image*", Proceedings of the 2nd International Conference on Signal Processing Systems (ICSPS), 2012, pp. 45-50.

[9] L. Yang, Y. Jinzhu , Z. Dazhe, "*A Method of Pulmonary Nodule Detection utilizing multiple support Vector Machine*", Proceedings of the International Conference on Computer Application and System Modeling, 2010, pp. 203-207.

[10] M. Iain , M. Darren, F. Mary-Jane, "*Classification of Cancer Stage from Free-text Histology Reports*", Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006, pp.156-159.

[11] D. Siqi, H. Tianlin , S. Yang, L. Chun, H. Yuanqing, "*Detection of Lung Cancer with Breath Biomarkers Based on SVM Regression*", Proceedings of the Fifth International Conference on Natural Computation 2009, pp. 93-96.

[12] Sunita Beniwal, Jitender Arora, "*Classification and Feature Selection Techniques in Data Mining*", International Journal of Engineering Research & Technology (IJERT), vol. 1, (6), 2012, pp. 94-97.

[13] Disha Sharma, Gagandeep Jindal, "*Identifying Lung Cancer Using Image Processing Techniques*", Proceedings of the International Conference on Computational Techniques and Artificial Intelligence (ICCTAI), 2011 pp. 115-120.

[14] Aminmohammad Roozgard, Samuel Cheng, and Hong Liu, "*Malignant Nodule Detection on Lung CT Scan Images with Kernel RX – algorithm*", Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2012) Hong Kong and Shenzhen, China, 2012, pp. 499-502.

[15] Anam Tariq, M. Usman Akram and M. Younus Javed, "*Lung Nodule Detection in CT Images using Neuro Fuzzy Classifier*", Proceedings of the Fourth International Workshop on Computational Intelligence in Medical Imaging (CIMI), 2013, pp. 49-53.

[16] Atiyeh Hashemi, Abdol Hamid Pilevar, Reza Rafeh, "*Mass Detection in Lung CT Images Using Region Growing Segmentation and Decision Making Based on Fuzzy Inference System and Artificial*

Neural Network", I.J. Image, Graphics and Signal Processing, 2013, pp. 16-24.

[17] J. Quintanilla-Dominguez, B. Ojeda-Magaña, M. G. Cortina-Januchs, R. Ruelas, A. Vega-Corona, and D. Andina, "*Image segmentation by fuzzy and possibilistic clustering algorithms for the identification of microcalcifications,*" Sharif University of Technology Scientia Iranica, vol. 18, 2011, pp. 580–589.

[18] Dansheng Song, Tatyana A. Zhukov, Olga Markov, Wei Qian3, MelvynS. Tockman, "*Prognosis of stage i lung cancer patients through quantitative analysis of centrosomal features*", IEEE, 2012, pp. 1607-1610.

[19] Qiao Z, Zhou, L., Huang, J. Sparse, "*Linear discriminant analysis with application to high dimension low sample size data,*" IAENG International Journal of Applied Mathematics, vol. 39, 2009, pp. 48-60.

[20] Kumar K, Bhattacharya, S. "*Artificial neural network vs linear discriminant analysis in credit ratings forecast: A comparative study of prediction performances,*" Review of Accounting and Finance, vol. 5, 2006, pp. 216-227.

[21] S.K. Vijai Anand, "*Segmentation coupled Textural Feature Classification for Lung Tumor Prediction*", Proceedings of the International Conference on Computing, Communications and Networking Technologies ICCCCT, 2010, pp. 518-524.

[22] Shiy ingH u, EricA Huffman, and Jospe h M. Reinhard t, "*Automatic lung segementation for accurate quantitiation of volumetric X-Ray CT images*", IEEE Transactions on Med ical Imaging, vol. 20 , (6), June 2001, pp. 490 -498.

[23] S.L.A. Lee, A.Z. Kouzani, and E.J. Hu, "*A Random Forest for Lung Nodule Identification*", 2010, pp.56-60.

[24] Yang Liu, Jinzhu Yang, Dazhe Zhao, Jiren Liu, "*Computer Aided Detection of Lung Nodules Based on Voxel Analysis utilizing Support Vector Machines*", Proceedings of the International Conference on Future Biomedical Information Engineering, 2009, pp. 90-93.

[25] S.Sivakumar, Dr.C.Chandrasekar, "*Lung Nodule Detection Using Fuzzy Clustering and Support Vector Machines*", International Journal of Engineering and Technology (IJET), vol. 5, (1), Feb-Mar 2013, pp. 179-185.

[26] S.Sivakumar and C.Chandrasekar, "*Lung Nodule Segmentation through Unsupervised Clustering Models", Procedia Engineering*, vol. 38, pp. 3064-3073.

[27] M. Arfan Jaffar, Ayyaz Hussain, M. Nazir, Anwar M. Mirza and Asmatullah Chaudhry, "*GA and Morphology based automated Segmentation of Lungs from CT scan Images*", CIMCA, IAWTIC, and ISE, 2008, pp. 265-270.

[28] P. Kanungo, P. K. Nanda and U. C. Samal, "*Image Segmentation Using Thresholding and Genetic Algorithm*", 2008, pp.1-4.

[29] S.M .Smith and J.M. Brady. SUSAN, "*a new approach to low level image processing*", Int. Journal of Computer Vision, vol 23, (1), May 1997, pp. 45--78.

[30] JIA Tong, ZHAO Da-Zhe, YANG Jin-Zhu,WANG Xu, "*Automated Detection of Pulmonary Nodules in HRCT Images*", IEEE, 2007, pp. 38-41.

[31] Fatma Taher, Naoufel Werghi and Hussain Al-Ahmad, "*Bayesian Classification and Artificial Neural Network Methods for Lung Cancer Early Diagnosis*", IEEE, 2012, pp. 773-776.

[32] R. Duda, P. Hart, "*Pattern Classification*", Wiley-Interscience 2nd edition, October 2001.

[33] Negar Memarian, Javad Alirezaie, Paul Babyn, "*Computerized Detection of Lung Nodules with an Enhanced False Positive Reduction Scheme*", Proceedings of the International Conference on Image Processing ICIP, 2006, pp. 1921-1924.