Classification of Lung Cancer Nodules using SVM Kernels

S.Shaik Parveen
Research Scholar,
Bharathiar University,
Coimbatore

C.Kavitha, Ph.D
Assistant Professor/Computer Science,
Thiruvalluvar Government Arts College,
Rasipuram

ABSTRACT
Support Vector Machines (SVM) is a machine learning method used for classifying the system. It analyses and identifies the categories using the trained data. It is widely used in medical field for diagnosing the disease. The proposed method consists of four phases. They are lung extraction, segmentation of lung region, feature extraction and finally classification of normal, benign and malignancy in the lung. Threat pixel identification with region growing method is used for segmentation of focal areas in the lung. For feature extraction gray level co- occurrence Matrix (GLCM) is been used. Extracted features are classified using different kernels of Support Vector Machine (SVM). The experimentation is performed with the help of real time computer tomography images.

Keywords
Computer Tomography, lung nodules, Classification, SVM kernels

1. 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. The most recent estimates according to the latest statistics provided by world health organization indicates that around 7.6 million deaths worldwide each year because of this type of cancer. Furthermore, mortality from cancer are expected to continue rising, to become around 17 million worldwide in 2030. Early detection of lung cancer is valuable. The 5-year-survival- rate of lung cancer has stagnated in the last 30 years and is now at approximately just 15%. Lung cancer takes more victims than breast cancer, prostate cancer and colon cancer together. This is due to the asymptomatic growth of this cancer. In the majority of cases it is too late for a successful therapy if the patient develops first symptoms (e.g. chronic croakiness or hemoptysis). But if the lung cancer is detected early (mostly by chance), there is a survival rate at 47% according to the American Cancer Society.

The Chest computed tomography (CT) images are difficult in diagnostic imaging modality for the detection of lung cancer and the resolution of any equivocal abnormalities detected on chest radiographs [1]. In particular, the expanding volume of thoracic CT studies along with the increase of image data, bring in focus the need for CAD algorithms to assist the radiologists [2].

A variety of computer assisted detection techniques have been proposed. In the development of CAD system, it involves two main categories such as Computer Aided Detection (CADe) and Computer Aided Diagnosis (CADi). The construction of CAD system would increase the mortality rate and reduce the unnecessary biopsies in patients with benign case and thus prevent physical and mental depression of patients. Thus CAD acts as a second reader and assists radiologist for accurate and efficient detection of cancer cells in the earlier stages. Thus the radiologist use CAD scheme to improve the detection accuracy.

In this paper previous works in this field are discussed in section 2. Section 3 explains the methodology of the proposed system. Section 4 gives the results of the implementation. Finally, conclusions are drawn in section 5.

2. PREVIOUS RESEARCH
There are many CAD systems are proposed earlier for the detection of lung cancer nodules and for the diagnosis. Some are discussed here.

Kenji Suzuki et al. in [3] proposed a different methodology using Massive Training Artificial Neural Networks (MTANN). A novel idea and solution to the problem that the lung nodules overlap with the ribs or clavicles in chest radiographs which gives complication to radiologists as well as computer-aided diagnostic (CAD) systems to detect these nodules. An MTANN is a non-linear filter that can be trained by use of input chest radiographs and the equivalent "teaching" images. They used a linear-output back-propagation (BP) algorithm that was derived for the linear-output multilayer ANN model in order to train the MTANN. The dual-energy subtraction is a technique used for separating bones from soft tissues in chest radiographs by using the energy dependence of the x-ray attenuation by different materials.

Using fuzzy rules, [4] proposed a Template-matching technique using genetic algorithms (GA) template matching (GATM) for detection of nodules in lung region. In their work, GA was used to determine the target position in the input image efficiently and to select an appropriate template image from several reference patterns for quick template matching.

[5] proposed a three step segmentation process for the analysis of lung image. In their approach, if the area in the CT image occupied by GGO is large, then it is easy for a medical doctor to extract the features. However, the possibility to overlook the light gray shadow becomes higher when GGO exists as a small area. In the first step of their model, extracting ROI is performed to segment the lung

area. To achieve better segmentation accuracy, preprocessing the CT slices is carried out by employing binarization, labeling, shrinking and expansion. In the second step, calculation of characteristics of GGO shadows such as mean value, standard deviation, and semi interquartile range have been carried out. In the final step, the GGO shadow's regions were extracted by linear discriminant function. Suspicious shadows are extracted by Variable N-Quoit (VNQ) filter from GGO. The suspicious shadows are classified into a certain number of classes using feature values calculated from the suspicious shadows.

Dougherty et al. [6] presented a temporal registration of CT scans of the lung. Their approach is based on an optical flow method and assumes a certain measure of intensity correspondence that scans containing pathology do not exhibit.

Penedo et al., [7] described a computer-aided diagnosis scheme which has two-level artificial neural network (ANN) architecture. In first level artificial neural network identifies suspicious regions in a low-resolution image. The curvature peaks calculated for all pixels in each suspicious region is given as the input to the second artificial neural network . The small size tumors are identified by the signature in curvature-peak feature space, where curvature is the local curvature of the image data when sighted as a relief map. The result gives a true positive identification in this network is threshold at a particular level of importance.

Okada et al [8] proposed a multiscale joint segmentation and model fitting solution which extends the robust mean shift-based analysis to the linear scale-space theory. In their paper, an ellipsoidal (anisotropic) geometrical structure of pulmonary nodules in the multislice X-ray computed tomography (CT) images was used for target's center location, ellipsoidal boundary approximation, volume, maximum/average diameters.

Hu et al.[9] used conventional region-based methods that implemented the concept of thresholding, regiongrowing and component labeling, and morphological processing. The automatic segmentation works on 3-D CT volumes and is tested on data sets of 8 normal subjects that were scanned three times at biweekly intervals. They used this template to find the structures with similar properties of nodules. Ingrid Sluimer et. al. [10] proposed a refined segmentation-by-registration scheme in which an atlas based segmentation of the pathological lungs is refined by applying voxel classification to the border volume of the transformed probabilistic atlas approach.

Yang Song et al. [11] presented a new method to automatically detect both tumors and abnormal lymph nodes based on the low-level intensity and neighborhood features and high-level contrast-type features, with a two-level SVM classification. One level of conditional random field (CRF) is based on unary level contextual and spatial features and pair wise-level spatial features. Other level is based by relabeling the detected tumors as positive or mediastinum by filtering the high-uptake myocardium areas.
Xujiong Ye et al. in [12] presented a new computer tomography (CT) lung nodule computer-aided detection (CAD) method. The method handles both solid nodules and ground-glass opacity (GGO) nodules. It uses fuzzy thresholding technique to segment the lung region from the CT data using a. The next step is of the volumetric shape index map and the ―dot map calculation. First map is based on local Gaussian and mean curvatures, and the next is based on the Eigen values of a Hessian matrix. The main advantages is high detection rate, fast computation, and applicability to different imaging conditions and nodule types make the method more reliable for clinical applications.

By analyzing the materials [13] we proposed automatic region growing method for segmentation [15]. For feature extraction GLCM is used and SVM kernels for classification to diagnose the occurrence of lung cancer.

3. METHODOLOGY

The proposed system consists of four steps for the classification of lung cancer nodules. They are as follows:

- Data collection.
- Extraction of lung region from Computer Tomography (CT) images with different preprocessing techniques.
- Segmentation of lung region using threat pixel identification with region growing method.
- Feature extraction using Gray level co- occurrence Matrix (GLCM).
- Classification to identify normal, benign and malignant cancer of the lung using different Support Vector Machine (SVM) kernels.

The data used here was got from reputed hospital. It contains CT images of 11 patients which has 3278 images for experiment.

After performing the segmentation [15], the features have to be extracted for detecting the cancer in the lung region correctly. This step concerns with two feature extraction such as Gray level co- occurrence Matrix (GLCM). Texture feature is used in identifying normal and abnormal pattern. Texture is an alteration and variation of surface of the image. Texture is characterized as the space distribution of gray levels in neighborhood. A gray level co-occurrence matrix (GLCM) contains information about the positions of pixels having similar gray level values. Texture descriptors derived from GLCM are contrast, Energy, Homogeneity and Correlation.

$$\text{Contrast} = \sum_i \sum_j \frac{P_d[i,j]}{1+|i-j|} \quad (1)$$

$$\text{Homogeneity} = \sum_{i=1}^{m} \sum_{j=1}^{n} |C(i,j)| \quad (2)$$

$$\text{Energy} = \sum_{i=1}^{m} \sum_{j=1}^{n} |C(i,j)| \quad (3)$$

$$\text{Correlation} = \frac{\sum_i \sum_j [ijP_d[i,j]] - \mu_i \mu_j}{\sigma_i \sigma_j} \quad (4)$$

Mean

$$\mu_i = \sum iP_d[i, j], \qquad (5)$$

Variance

$$\sigma_i^2 = \sum i^2 P_d[i, j] - \mu_i^2 \qquad (6)$$

Where P(i,j) Element i, j of the normalized symmetrical GLCM.
N is total number of gray levels in the image.

Contrast	Measures the local variations in the gray-level co-occurrence matrix.
Homogeneity	Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.
Energy	Provides the sum of squared elements in the GLCM. Also known as uniformity or the angular second moment.
Correlation	Measures the joint probability occurrence of the specified pixel pairs.

These extracted features can be passed to SVM classifier in order to detect the cancer nodules.

3.1 Classification

SVM [14] is usually used for classification task which is introduced by Vapnik. For binary classification SVM is used to determine an Optimal Separating Hyper plane (OSH) which produces a maximum margin between two categories of data. A transform that nonlinearly maps the data into a higher-dimensional space allows a linear separation of classes that cannot be linearly separated in the original space.

Theory of SVM is defined as:
Consider training set $D=\{(xj, yi)\}_{i=1}^{L}$ 1 with every input n i $x\in$ Rn we have, $f(x,\{w,b\}) = \text{sign}(w.x + b)$. Also, an associated output yi$\in\{-1, +1\}$. Every input x is initially mapped into a higher dimension feature space F, by $z=\phi(x)$ through a nonlinear mapping ϕ: Rn$\rightarrow$F. If the data are linearly non-separable in F, then a vector w$\in$ F and a scalar b will exist which describe the separating hyper plane as:

$$y_i = (w'.z_i \square\square b) \square\geq 1-\square\xi_i, \forall i \qquad (7)$$

where $\xi(\geq 0)$ are knows as lack variable. The hyper plane that optimally splits the data in F is one that

$$\text{minimize} \quad \|w\|^2 + C\sum_{i=1}^{m} \xi_i \qquad (8)$$

$$\text{subject to } y_i\left(w'.z_i + \beta\right) \geq -\xi_i, \qquad \xi_i \geq 0, \forall I \qquad (9)$$

where C is knows as regularization parameter that finds the tradeoff between maximum margin and minimum classification error.
Maximise

$$L(\alpha)=\sum_{i=1}^{n} \alpha i - 1/2\sum \alpha i\ \alpha j\ yi\ yjk\ (xj, xj) \qquad (10)$$

$$\text{subject } \sum_{i=1}^{L} y_i\alpha_i = 0, 0 \leq \alpha_i \leq C, \forall\ i=1,2,\ldots,n \qquad (11)$$

where $\alpha1,\ldots..,\alpha L$ represents the nonnegative Lagrangian multipliers. The data points xi that corresponding to $\alpha i>0$ are considered as Support Vectors.

In this study, the following three kernel functions have been applied to build SVM classifiers:

Linear kernel function, K(x,z) =x, z;

Polynomial kernel function K(x, z) =(x, z +1) ^{d} is the degree of polynomial;

Radial basis function $K(x,x') = \exp(-\|x-x'\| /2\ \sigma^2)$ Where σ is positive real number and the width of the function.

After completing the above process by using various constraints, the proposed method can able to detect normal, benign and malignant class in the lung region automatically.

4. RESULTS AND DISCUSSION

Experiments are carried out from real time datasets obtained from the reputed hospitals. Totally 11 patients are considered with 1 normal case, 2 benign cases and 8 malignant cases about 3278 sectional images. 50% of dataset are considered for training and 50% as testing phases. The performance results of the three kernels are measured using the parameters sensitivity and specificity. Results for sensitivity and specificity are given in Table 1 and Figure 1.

5. CONCLUSION

In this study automatic CAD system for detecting lung cancer nodules is proposed. It guarantees early detection of lung cancer nodules using computer tomography images. In the preprocessing step lung region is been extracted. In the next step segmentation is performed using threat pixel identification and region growing method. In the third step features are extracted using GLCM. Finally, Support Vector Machines (SVM) of three kernels is used for classification of normal, benign and malignant classes in the lungs. Experimental results show that the proposed system shows that RBF kernel works better than other kernels of SVM.

Table 1. Results of Sensitivity and Specificity of SVM Kernels

Kernels	Sensitivity(%)	Specificity(%)
Linear	83.45	82.23
Polynomial	85.79	84.91
Radial Bias Function (RBF)	91.38	89.56

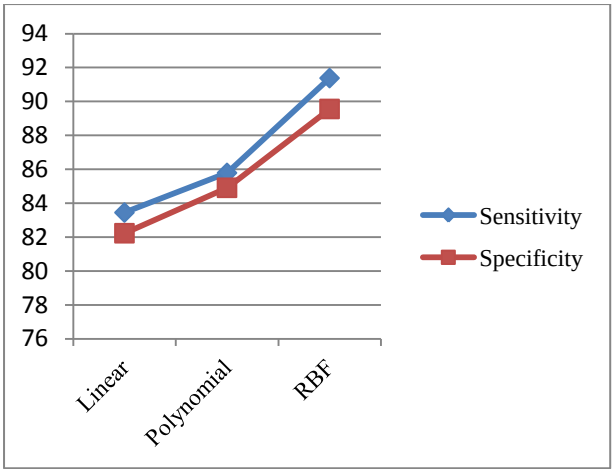

Fig 1: Results of Sensitivity and Specificity of SVM Kernels

6. ACKNOWLEDGMENT

Foremost, we would like to thank the almighty for the success in completing this work. We would like to extend our thanks to all the reviewers for their critical comments.

7. REFERENCES

[1] HoweMA, Gross BH. 1987 "CT evaluation of the equivocal pulmonary nodule", Computer Radiology, vol. 11, pp. 61–67

[2] H. Abe, H. MacMahon, J. Shiraishi, 2004 "Computer-aided diagnosis in chest radiology", Semin. Ultrasound CT MRI, Vol. 25, pp. 432-437.

[3] Kenji Suzuki, Hiroyuki Abe, Heber MacMahon, and Kunio Doi, 2006 "Image-Processing Technique for Suppressing Ribs in Chest Radiographs by Means of Massive Training Artificial Neural Network (MTANN)," IEEE Transactions on medical imaging, vol. 25, no. 4, pp. 406-416.

[4] Lee Y, Hara T, Fujita H, Itoh S, Ishigaki T, 2001 "Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique", IEEE Trans. Med. Imaging, Vol. 20, pp.595–604.

[5] Hyoungseop Kim, Seiji Mori, Yoshinori Itai, Seiji Ishikawa, Akiyoshi Yamamoto and Katsumi Nakamura, 2007" Automatic Detection of Ground-Glass Opacity Shadows by Three Characteristics on MDCT Images", World congress on medical physics and biomedical engineering, IFMBE Pro2 Vol. 14/4.

[6] L. Dougherty, J. C. Asmuth, and W. B. Gefter , 2003 "Alignment of CT lung volumes with an opticalflow method,"Acad. Radiol., vol. 10, no. 3, pp.249–254.

[7] Penedo, M.G., Carreira, M.J., Mosquera, A. and Cabello,D.,1998 "Computer-Aided Diagnosis: A Neural-Network-Based Approach to Lung Nodule Detection", IEEE Transactions on Medical Imaging, Pp: 872 – 880.

[8] Okada K, Comaniciu D, Krishnan, 2005 "A Robust Anisotropic Gaussian Fitting for Volumetric Characterization of Pulmonary Nodules in Multislice CT", IEEE Trans. Med. Imaging, Vol. 24, No. 3, pp. 409–423.

[9] S. Hu, E. A. Hoffman, and J. M. Reinhardt , 2001 "Automatic lung segmen-tation for accurate quantitation of volumetric X-ray CT images,"IEEE Trans. Med. Imag., vol. 20, no. 6, pp. 490–498.

[10] Ingrid Sluimer, Mathias Prokop, and Bram van Ginneken, 2005 " Toward Automated Segmentation of the Pathological Lung in CT", IEEE Transactions on Medical Imaging, vol. 24, no. 8.

[11] Yang Song, Weidong Cai, Jinman KimDavid Dagan Feng, 2012 " A Multistage Discriminative Model for Tumor and Lymph Node Detection in Thoracic Images", IEEE transactions on Medical Imaging, vol. 31, no. 5.

[12] Xujiong Ye, Xinyu Lin, Jamshid Dehmeshki, Greg Slabaugh, and Gareth Beddoe, 2009 "Shape-Based Computer-Aided Detection of Lung Nodules in Thoracic CT Images", IEEE Transactions on Biomedical Engineering, vol. 56, no. 7, pp. 1810-1820.

[13] S.Shaik Parveen, Dr.C.Kavitha, 2012 " A Review on Computer Aided Detection and Diagnosis of lung cancer nodules," International Journal of Computers & Technology, Volume 3 No. 3, Nov-Dec.

[14] Christopher J.C. Burges, 1998," A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery 2, 121-167.

[15] Parveen S. Shaik, Kavitha C, 2013, "Detection of lung cancer nodules using automatic region growing method", International Conference on Computing, Communications and Networking Technologies IEEE – ICCCNT Digital Object Identifier :10.1109/ICCCNT.2013.6726669.

.