World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:4, No:11, 2010

Support Vector Machine Prediction Model of Early-stage Lung Cancer Based on Curvelet Transform to Extract Texture Features of CT Image

Guo Xiuhua[1], Sun Tao[1], Wu Haifeng[1], He Wen[2], Liang Zhigang[3], Zhang Mengxia[4],Guo Aimin[1], Wang Wei[1]*

Abstract*—*Purpose: To explore the use of Curvelet transform to extract texture features of pulmonary nodules in CT image and support vector machine to establish prediction model of small solitary pulmonary nodules in order to promote the ratio of detection and diagnosis of early-stage lung cancer. *Methods:* 2461 benign or malignant small solitary pulmonary nodules in CT image from 129 patients were collected. Fourteen Curvelet transform textural features were as parameters to establish support vector machine prediction model. *Results:* Compared with other methods, using 252 texture features as parameters to establish prediction model is more proper. And the classification consistency, sensitivity and specificity for the model are 81.5%, 93.8% and 38.0% respectively. *Conclusion:* Based on texture features extracted from Curvelet transform, support vector machine prediction model is sensitive to lung cancer, which can promote the rate of diagnosis for early-stage lung cancer to some extent.

***Keywords*—**CT image, Curvelet transform, Small pulmonary nodules, Support vector machines, Texture extraction.

I. INTRODUCTION

IN recent years, lung cancer is today the most frequent cause of cancer deaths all over the world and carries a severe prognosis with 5-year survival rates. On a global basis it is estimated that 1.2 million people are diagnosed with this disease every year (12.3% of the total number of cancer diagnosed), and about 1.1 million people are dying of this disease yearly (17.8% of the total cancer death)[1]. Early detection of lung cancer is essential in reducing life fatalities[2]. However, achieving this early detection of lung cancer is not an easy task. More than 80% patients are already in middle or advanced stage when diagnosed and they miss the timing for the surgery. The 5-year survival rate is only 14%[3], which can reach more than 70% if lung cancer can be diagnosed in an earlier stage. Because of the difficulty of diagnosis at the early period, it has been the heat research area all over the world.

Although histology diagnosis is the most accurate detection method in the medical environment, it is an aggressive invasive procedure that involves some risks, patient discomfort and some trauma, which restricts it to be used in the clinical practice. Digital CT (Computed Tomography), overcoming shortages of histology diagnosis, has gradually become the best imaging diagnosis method of lung cancer[4]. But pulmonary nodules (referring to the lesion of lung field ≤ 3 cm in diameter) of lung cancer in CT images share similarity with benign cases to some extent, such as tuberculosis, inflammatory pseudotumor, hamartoma, and aspergillosis [5], which makes it difficult to distinguish, especially for the doctors who are not rich in clinical experience. With technique of computer rising, the computer-aided diagnosis (CAD) has become an auxiliary diagnosis tool[6,7], especially in diseases that can not be diagnosed efficiently. To improve the accuracy and efficiency of CT screening programs for the detection of early-stage lung cancer, a number of research projects, such as texture analysis[8,9] and image segmentation[10,11], have been done to assist radiologists in diagnosing lung cancer. The purpose of our research is to establish support vector machine prediction model for small pulmonary nodules based on Curvelet transform to extract texture features of CT image. In accordance with the prediction model we can predict the characteristics of pulmonary nodules and aid to diagnose early-stage lung cancer effectively.

II. MATERIALS AND METHODS

A. Materials

Please 2461 CT images used in this study are extracted from 129 patients with small solitary pulmonary nodules, including 537 CT images (25 benign cases) related to benign nodule and 1924 CT images (104 malignant cases) to malignant tumors.

1. School of Public Health and Family Medicine, Capital Medical University, Beijing, 100069, P.R.China

2. Department of Radiology, Friendship Hospital, Capital Medical University, Beijing 100050, China

3. Department of Radiology, Xuan Wu Hospital, Capital Medical University, Beijing 100053, China

4. College of Business and Administration, Capital University of Economics and Business, Beijing 100070, China

* Corresponding author. Tel.: +86 1083911504; fax: +86 1083911501.E-mail address: guoxiuh@ccmu.edu.cn (W. Wang)..

Program of Funds: The program of Natural Science Fund of China (Serial Number: 30972550); the program of Natural Science Fund of Beijing (Serial Number: 7092010); the program of Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality(Serial Number: PHR201007112)

The final diagnosis of malignant cases was determined by either an operation or biopsy. The diagnosis of benign cases was confirmed by an operation, CT diagnosis or follow-up. The original format is DICOM, and diameter of the images is between 0.3 cm and 3 cm. 129 cases were provided by four hospitals, and details are as follows: Beijing Xuanwu Hospital of Capital Medical University (26 malignant cases, 11 benign cases), Beijing Friendship Hospital affiliated to Capital Medical University (35 malignant cases,6 benign cases), Chaoyang Hospital affiliated to Capital Medical University (20 malignant cases,7 benign cases) and Fuxing Hospital affiliated to Capital Medical University (23 malignant cases,1 benign cases)

CT scans were obtained with an 64-slice helical CT scanner using a tube voltage of 120 kV and current of 200mA. Reconstruction thickness and reconstruction intervals for routine scanning were 0.625mm, Kernel is B31f/B70. Window width,1,500 HU and window level,-500 HU. Data were reconstructed with a matrix of 512×512.

B. Basic Theories

1. Support vector machines

SVM is a popular classifier based on structural risk minimization principle[12], whose object is to minimize the generalization error of the classifier. Recently, SVM has gained much attention as a useful tool for image recognition. Youngjoo Lee[13] investigated the performance of Bayesian classifier, ANN (artificial neural net) and SVM (support vector machine) for differentiating obstructive lung diseases using texture analysis. Results showed that SVM showed the best performance for classification. Michael E. Mavroforakis[14] used fractal analysis to extract texture features of breast mass in mammograms. Then A wide range of linear and non-linear classification architectures was employed, including linear discriminant analysis (LDA), least-squares minimum distance(LSMD), K-nearest-neighbors (K-nn), radial basis function (RBF) and multi-layer perceptron (MLP) artificial neural network (ANN), as well as support vector machine(SVM) classifiers to classify mammographic breast tumors. Results showed that Non-linear classifiers, especially SVM, have been proven superior to any linear equivalent.

In contrast to other classifiers, such as Artificial Neural Networks, SVM searches for the hyperplane that maximizes the distance from the hyperplane to the nearest examples in each class. An attractive feature of SVM is that it can map linearly inseparable data into higher dimensional space where they can be linearly separated. This makes it useful for the supervised non-parametric classification of hyerspectral images. There are two types of SVM, linear and non-linear. The training data of linear SVM may be analyzed as either linearly separable or linearly non-separable.

We could use K for training even if we do not know function φ. The problem is still a linear separation problem but in a different space. It is easy to find kernel functions such that the training algorithm and solution are independent of the dimension of H. Such kernels must hold Mercer´s condition[15] which tells us whether or not a perspective kernel is a dot product in some space. The polynomial, radial, anova kernels are now often seen choices in SVM-based CAD applications.

The use of SVM, like any other machine learning technique, involves two basic steps namely training and testing. Training an SVM involves feeding known data to the SVM along with previously known decision values, thus forming a finite training set. It is from the training set that an SVM gets its intelligence to classify unknown data.

2. Curvelet transform

During the past two decades, wavelets theory has been widely used, because wavelets provide a powerful tool for multi-resolution analysis of images. Many studies have investigated wavelet-based features in various applications such as image denoising, image compression and tumor recognition. The success of wavelets lies in its good performance for piecewise smooth functions in one dimension[16], however wavelet is not suitable to capture more directional features in an image. In 2000, in an attempt to overcome the weakness of traditional multiscale representations using wavelets, Candes and Donoho[17] developed Curvelet for providing efficient representation of smooth objects with discontinuities along curves. The basic idea of Curvelet transform is to represent a curve as a superposition of functions of various lengths and widths obeying a specific scaling law. Regarding 2D images, it can be done first by decomposing an image into wavelet sub-bands, i.e., separating the object into a series of disjoint scales. Each sub-image of a given scale is then analyzed with a local ridgelet transform, another kind of new multi-resolution analysis tool.

Based on Curvelet transform, we extracted fourteen texture features of pulmonary nodules of CT images, including Entropy, Mean ,Correlation, Energy, Homogeneity, StdDev, MP, IDM, ClustTend, Inertia, SumMean, DiffMean, SumEntr, DiffEntr. The meanings of fourteen texture features are as follows.

Energy is defined to measure the number of repeated pairs, which is expected to be high if the occurrence of repeated pixel pairs is high. In statistical mechanics, entropy is defined as a factor or quantity that is a function of the physical state of a mechanical system and is equal to the logarithm of the probability of the occurrence of the particular molecular arrangement in that state. Inverse Difference Moment tells us about the smoothness of the image, like homogeneity. The IDM is expected to be high if the gray levels of the pixel pairs are similar. Inertia reflects the roughness of texture, which is expected to be low if the more elements are near to diagonal line of matrix when texture is rougher. Correlation is expected to measure the relevance on the gray of pixel. Sun–mean (mean) and Difference-mean provide the mean of the gray levels in the image. The sum–mean is expected to be large if the sum of the gray levels of the image is high, so is Difference-mean.Standard deviation tells us how spread out the distribution of gray levels is. The variance is expected to be large if the gray levels of the image are spread out greatly.

Results in the pixel pair which is most predominant in the image. The Maximum probability (MP) is expected to be high if the occurrence of the most predominant pixel pair is high. the mean of the gray reflects the central tendency of the gray. Cluster tendency measures the grouping of pixels that have similar gray level values. Homogeneity measures the local homogeneity of a pixel pair. The homogeneity is expected to be large if the gray levels of each pixel are similar.

Curvelet transform is a new image representation approach that codes image edges more efficiently than wavelet transform. Curvelet will be better than wavelet in following cases: [18]

(1) Optimally sparse representation of objects with edges.

(2) Optimal image reconstruction in severely ill-posed problems.

(3) Optimal sparse representation of wave propagators.

Some studies have been done using Curvelet transform in image processing. Dettori and Semler [19] presented a comparative study between Wavelet, Ridgelet and Curvelet transform on some computed tomography (CT) scans. The comparative study indicated that Curvelet yields better results than Wavelet or Ridgelet.

C. Methods

Programs to extract texture features of CT image and establish prediction model of SVM were performed respectively using MATLAB, version 7.0, software (The MathWorks, Inc.) and Microsoft Visual C++ 6.0, software(The Microsoft, Inc.)

Curvelet transform was used as a multiscale level decomposition to represent pulmonary nodules of CT images as a pre-process for classification. Then we chose 252 texture features to establish prediction model of SVM.

III. RESULTS

A. Basic Situation of Case

1. Age Distribution

The youngest patient is 22 years old. The oldest patient is 86 years old. The average age is 63.2 years old.

2. Differences Comparison of Gender Distribution of 129 Cases Between Benign and Malignant Cases

There are 25 benign cases (11 males, 14 females) and 104 malignant cases (62 males, 42 females) .Results are showed in TABLE I.

TABLE I
GENDER DISTRIBUTION OF 129 CASES BETWEEN BENIGN AND MALIGNANT CASES

Pathological Diagnosis	Gender		Total
	Male	Female	
Benign	11	14	25
Malignant	62	42	104
Total	73	56	129

We performed chi-square test on the gender distribution of 129 cases between benign and malignant cases (Pearson χ^2 =2.001, P =0.157). Gender distribution in cases showed no statistically significant differences, indicating that the cases of the sample distribute evenly in gender between benign and malignant cases.

B. Texture Features

Based on Curvelet transform, we extracted fourteen texture features of pulmonary nodules of CT images. Every image could be decomposed into 18 sub-images. Also the 18 sub-images could be classified into three parts: inner layer, middle layer and outer layer. So 252 texture features were extracted from every image. Among those texture features, 158 texture features showed statistically significant differences between benign and malignant cases through two independent samples tests of nonparametric test or two independent samples t-test.

C. Establishment of Prediction Model of Support Vector Machines

By 80% and 20%, we divided the 2461 images into two parts: one was a training sample (80%) and the other was a test sample (20%). The training sample was to establish the database and the test sample was to evaluate the validity of prediction model of SVM. (TABLE II)

TABLE II
BENIGN AND MALIGNANT CASES DISTRIBUTION

Samples	Benign	Malignant	Total
Training sample	429	1539	1968
Test sample	108	385	493
Total	537	1924	2461

Based on Curvelet transform, 252 texture features we extracted were as parameters to establish prediction model for small pulmonary nodules (TABLE III)

The validity of prediction model of SVM is evaluated by the following three indexes: sensitivity (93.8%), specificity (38.0%) and consistency (81.5%). The high sensitivity (93.8%) can reduce the false negative rate of early-stage lung cancer effectively.

TABLE III
PREDICTION RESULTS OF PULMONARY NODULES BASED ON SVM

SVM	Pathological Diagnosis		Total
	Benign	Malignant	
Benign	41	24	65
Malignant	67	361	428
Total	108	385	493

There are other methods used in published researches to select texture features. Leandro Lu’ s Galdino Oliveira[20] used wavelet transform to extract the chest radiography, and used Energy as the only parameter to establish the prediction model. The same attempt had been done in other research projects[21,22]. Lucia Dettori[19] selected Mean, StaDev, Energy and Entropy to establish the prediction model. Principal component analysis，a very useful tool to deal with colinearity, has various applications in texture extraction and tumor recognition[23,24,25]. Mohamed Meselhy Eltoukhy[26,27] used Curvelet transform to decompose mammogram images into 4 levels, then selected the largest 100 texture features as parameters.

In order to select texture features which are more accurate to reflect characteristics of pulmonary nodules, we have made many attempts. Results are followed (TABLE IV).

TABLE IV
PREDICTION RESULTS OF PULMONARY NODULES USING OTHER METHODS

	Sensitivity	specificity	consistency
Using Energy As The Only Parameter	93.2%	29.6%	79.3%
Using Texture Features of Inner Layer As Parameters	96.4%	31.5%	82.2%
Using Texture Features of Middle Layer As Parameters	94.8%	25.0%	79.5%
Using Texture Features of Outer Layer As Parameters	100.0%	0.0%	78.1%
Using Mean, StaDev, Energy and Entropy As Parameters	94.8%	29.6%	80.5%
Using Principal Component Analysis	100.0%	0.0%	78.1%
Using 158 Texture Features As Parameters	94.5%	34.3%	81.3%%
The Largest 100 Texture Features As Parameters	93.8%	28.7%	79.5%

In order to promote sensitivity and specificity, we had made some attempts to select proper texture features. Compared with other methods, using 252 texture features as parameters to establish prediction model is more satisfying.

IV. DISCUSSION

Based on published reports, characteristics of pulmonary nodules can been detected by texture features[9]. However, 2D images are irregular when decomposed, so Curvelet transform is more suitable than the wavelet transform to extract texture features. The methods to establish prediction model are variable, such as multiple linear regression, logistic regression, discriminant analysis, artificial neural networks, but the result of support vector machine is better[28,29].In this research, we establish support vector machine prediction model for small pulmonary nodules using Curvelet transform to extract texture features of CT image, which has not been reported to our knowledge.

In recent years, the incidence of lung cancer has been the top of cancers in the most countries. Because of the difficulty to diagnosis, more attention has been paid to lung cancer. Now the most accurate diagnosis method of lung cancer is histology diagnosis, but this method is traumatic, which restricts it to be used in clinical practice. In the decades, digital CT has been the main diagnosis tool of lung cancer for its convenience and safety, and widely used in clinical practice. However, it is difficult to distinguish between benign and malignant cases in the CT images of pulmonary nodules, especially for the doctors lack of experience. In this paper, using the computer-aided diagnosis, we extracted texture features of pulmonary nodules of CT images by curvelet transform. Through establishing the prediction model of SVM, we can predict the characteristics of pulmonary nodules. The result shows: sensitivity can reach 93.8%, the consistency rate is 81.5%, which are better than the same kind of research project[8].The prediction model is so sensitive that it can diagnose early-stage lung cancer effectively, reduces the difficulty of distinguishing characteristics of pulmonary nodules and improves accuracy rate of diagnosing early-stage lung cancer.

However, many shortcomings are available in this research, some of which may be the reasons why the specificity of the prediction model is low (38.0%). First the shortage of the quantity of benign cases makes the training sample lack of benign data. When we establish the prediction model of SVM, deficiency of data may be one reason that prediction model can not distinguish benign cases from malignant cases. Second using SVM as the classifier to distinguish characteristics of pulmonary nodules may be not suitable. Third we selected all the texture features to establish the prediction model. The methods which we selected texture features may not exclude texture features of pulmonary nodules which can not accurately reflect the differences between benign and malignant cases.

REFERENCES

[1] D.M. Parkin, Global cancer statistics in the year 2000. *Lancet Oncol.*,vol.2, 2001. pp.533-543.

[2] A. Motohiro, H. Ueda, H. Komatsu, N. Yanai, T. Mori, National Chest Hospital Study Group for Lung Cancer. Prognosis of non-surgically treated, clinical stage I lung cancer patients in Japan. *Lung Cancer*, vol.36, 2002. pp. 65–69.

[3] N.R. Wardwell, P.P Massion, Novel strategies for the early detention and prevention of lung cancer. *Seminars In Oncology*,vol.3 ,2005. pp. 259–268.

[4] C.I. Henschke, D F. Yankelevitz, D.M. Libby, M.W. Pasmantier, J.P. Smith, O.S. Miettinen, International Early Lung Cancer Action Program Investigators: survival of patients with stage I lung cancer detected on CT screening. *The New England Journal of Medicine*, vol.17, 2006. pp.1763–71.

[5] J.W. Chang, C.A. Yi, D.S. Son, N. Choic, J. Lee, H.K. Kim, Y.S. Choi, K.S. Lee, J. Kim, Prediction of lymph node metastasis using the combined criteria of helical CT and mRNA expression profiling for non-small cell lung cancer. *Lung Cancer*, in press, 2008.

[6] J. Jiang, B.Yao, A.M. Wason. A genetic algorithm design for micro calcification detection and classification in digital mammograms. *Computerized Medical Imaging and Graphics*, vol.1, 2007. pp.49–61.

[7] J. Staal, B. van Ginneken, M.A. Viergever, Automatic rib segmentation and labeling in computed tomography scans using a general framework for detection, recognition and segmentation of objects in volumetric data. *Medical Image Analysis*,vol.1,2007. pp.35–46.

[8] Y.N . Liu, H. Wang, X.H. Guo, ZG. Liang, Q. He, Application of artificial neural networks in prediction model of early-stage lung cancer. *Chinese journal of Medical Statistics,* vol.1,2008. pp. 30–33.

[9] H. Wang, X.H. Guo, Z.W. Jia, H.K. Li, Z.G. Liang, K.C. Li, Q. He. Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image. *European Journal of Radiology.* Online: doi:10.1016/j.ejrad.2009.01.024.

[10] J. Dehmeshki, X. Ye, X. Lin, M. Valdivieso, H. Amin, Automated detection of lung nodules in CT images using shape-based genetic algorithm. *Computerized Medical Imaging and Graphics*,vol.6,2007. pp.408–417.

[11] X.J. Sun, H.B. Zhang, H.C. Duan, 3D computerized segmentation of lung volume with computed tomography. Academic Radiology 2006;13(6):670–677.

[12] V. N. Vapnik, Statistical learning theory. New York: Wiley; 1998.

[13] Y. Lee, J.B. Seo, J.G. Lee, S.S. Kim, N. Kim, S.H. Kang, Performance testing of several classifiers for differentiating obstructive lung diseases based on texture analysis at high-resolution computerized tomography (HRCT). *Computer Methods And Programs In Biomedicine*,vol.93,2009. pp.206-215.

[14] M.E. Mavroforakis, H.V. Georgiou, N. Dimitropoulos, D. Cavouras, S. Theodoridis, Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers. *Artificial Intelligence in Medicine*, vol.37,2006. pp.145—162.

[15] V. Vapnik, Estimation of Dependencies based on Empirical Data, Springer Verlag, New York, 1982.

[16] M.N. Do, M. Vetterli, The finite ridgelet transform for image representation. *IEEE Transactions on Image Processing*,vol.12, 2003.pp. 16–28.

[17] E.J. Candes, D.L. Donoho, Curvelets, multi-resolution representation, and scaling laws, *Wavelet Applications in Signal and Image Processing VIII*, vol. 4119-01, SPIE, 2000.

[18] Candes EJ, Demanet L, Donoho DL, Ying L. Fast discrete curvelet transforms. *Multiscale Modelling and Simulation*,vol.5,2006. pp861–899.

[19] L. Dettori, L. Semler. A comparison of wavelet, ridgelet, and curvelet-based texture classification algorithms in computed tomography. *Computers in Biology and Medicine*,vol.37, 2007. pp.486-498.

[20] L.L. Oliveira, S.A. Silva, L.H. Ribeiro, R.M. de Oliveira, C.J. Coelho, A.L. S Andradea, Computer-aided diagnosis in chest radiography for detection of childhood pneumonia. *International journal of medical informatics*, vol.77,2008. pp.555–564.

[21] P.W. Huang, S.K. Da, Design of a two-stage content-based image retrieval system using texture similarity. *Information Processing & Management,* vol.1,2004. pp.81–96.

[22] M.K. Bashar, T. Matsumoto, N. Ohnishi, Wavelet transform-based locally orderless images for texture segmentation. *Pattern Recognition Letters,* vol.15, 2003. pp.2633-2650.

[23] R. Llobet, J.C. P´erez-Cort´es, A.H. Toselli, A. Juan, Computer-aided detection of prostate cancer. *International Journal of Medical Informatics*, vol.76,2007. pp.547–556.

[24] J. Zhang, L.Z. Tong, L. Wang, N. Li, Texture analysis of multiple sclerosis: a comparative study. *Magnetic Resonance Imaging,*vol.26, 2008. pp.1160–1166.

[25] I. Güler, A. Demirhan, R. Karakıs. Interpretation of MR images using self-organizing maps and knowledge-based expert systems. *Digital Signal Processing,*vol.19,2009. pp. 668–677.

[26] M.M. Eltoukhy, I Faye, B.B. Samir, A comparison of wavelet and curvelet for breast cancer diagnosis in digital mammogram. *Computers in Biology and Medicine*,vol.40,2010. pp.384–391.

[27] M.M. Eltoukhy, I Faye, B.B. Samir, Breast cancer diagnosis in digital mammogram using multiscale curvelet transform. *Computerized Medical Imaging and Graphics*,2009, in press.

[28] Z, Zheng, Y.X. Zhang, YX Hu, Investigation of eye gaze based on independent component analysis and support vectormachine. *Journal of Optoelectronics*,vol.7, 2007. pp. 491-494.

[29] D. Meye, F. Leisch, K. Hornik. The support vecor machine under test. *Neurocomputing*,2003.pp.169-186.