RESEARCH ARTICLE

Prediction Models for Solitary Pulmonary Nodules Based on Curvelet Textural Features and Clinical Parameters

Jing-Jing Wang[1,2], Hai-Feng Wu[1,2], Tao Sun[1,2], Xia Li[1,3], Wei Wang[1,2,4], Li-Xin Tao[1,2], Da Huo[1,2], Ping-Xin Lv[5], Wen He[6], Xiu-Hua Guo[1,2]*

Abstract

Lung cancer, one of the leading causes of cancer-related deaths, usually appears as solitary pulmonary nodules (SPNs) which are hard to diagnose using the naked eye. In this paper, curvelet-based textural features and clinical parameters are used with three prediction models [a multilevel model, a least absolute shrinkage and selection operator (LASSO) regression method, and a support vector machine (SVM)] to improve the diagnosis of benign and malignant SPNs. Dimensionality reduction of the original curvelet-based textural features was achieved using principal component analysis. In addition, non-conditional logistical regression was used to find clinical predictors among demographic parameters and morphological features. The results showed that, combined with 11 clinical predictors, the accuracy rates using 12 principal components were higher than those using the original curvelet-based textural features. To evaluate the models, 10-fold cross validation and back substitution were applied. The results obtained, respectively, were 0.8549 and 0.9221 for the LASSO method, 0.9443 and 0.9831 for SVM, and 0.8722 and 0.9722 for the multilevel model. All in all, it was found that using curvelet-based textural features after dimensionality reduction and using clinical predictors, the highest accuracy rate was achieved with SVM. The method may be used as an auxiliary tool to differentiate between benign and malignant SPNs in CT images.

Keywords: Solitary pulmonary nodule - curvelet - texture extraction - support vector machine

Asian Pac J Cancer Prev, **14 (10)**, 6019-6023

Introduction

Lung cancer has been the most common form of cancer in the world since 1985. It is the leading cause of cancer death, especially in males (Hart, 2011; Ma et al., 2012; Soerjomataram et al., 2012). If patients with lung cancer can be diagnosed at an early stage and treated timely, their 5-year survival rate could be improved from 15 to 70% (Beadsmoore and Screaton, 2003). Thus, it is very important to explore effective methods of diagnosing lung cancer in its early stages. Unfortunately, early stage lung cancer and benign lesions have similar appearances in images — both appearing as solitary pulmonary nodules (SPNs) in computerized tomographic (CT) slices, which cannot be distinguished by the naked eye.

By definition, an SPN is a single, spherical, well-circumscribed, radiographically opaque object that measures up to 3 cm in diameter and is completely surrounded by aerated lung tissue (Sun et al., 2013b). Wavelet package (Acharya et al., 2012; Dua et al., 2012; Khorasani and Daliri, 2013) and gray-level co-occurrence matrix (GLCM) methods (Arebey et al., 2012; Wu et al., 2013) have been used to extract the texture of SPNs. It was found that the texture at the edge of the nodules is critical in distinguishing malignant from benign nodules (Wang et al., 2010). Curvelet transformation (Guo et al., 2012), which is ideally suited to the analysis of two-dimensional (2D) images, has proven to be particularly effective at detecting image activity along curves instead of radial directions when compared with other transforms (Ko et al., 2012). It has been used in the analysis of medical images (Dettori and Semler, 2007; Eltoukhy et al., 2010; Meselhy Eltoukhy et al., 2010), such as CT scans, endoscope images, X-rays, etc.

Several studies have used clinical models, such as the logistic (Swensen et al., 1997; Herder et al., 2005; Gould et al., 2007) and neural network models (Henschke et al., 1997; Nakamura et al., 2000; Matsuki et al., 2002), to estimate the pretest probability of lung cancer in patients with SPNs. However, the most classical models may be the Mayo Clinic model (Swensen et al., 1997) and the Department of Veterans Affairs (VA) model (Gould et

*[1]Department of Epidemiology and Health Statistics, School of Public Health, Capital Medical University, [2]Beijing Municipal Key Laboratory of Clinical Epidemiology, [5]Department of Radiology, Beijing Chest Hospital, Capital Medical University, [6]Department of Radiology, Friendship Hospital, Capital Medical University, Beijing, China, [3]Department of Epidemiology and Public Health, University College Cork, Cork, Ireland, [4]School of Medical Sciences, Edith Cowan University, Perth, Australia *For correspondence: statguo@ccmu.edu.cn*

al., 2007). Li et al. (2011) established a clinical prediction model using clinical and radiological information and found it to be more accurate than the Mayo Clinic and VA models. However, textural features (TA), a vital component of computer-assisted diagnosis (CAD), were not included. Way et al. (2009) used a fully automated system to extract image features to differentiate between malignant and benign lung nodules in CT scans. In addition, Wang et al. (2010) used fourteen textural features obtained using GLCM and demographic features to establish a multilevel model with a sensitivity of 90.6%.

To improve the accuracy and efficiency of the diagnosis of lung cancer via SPNs, in this study, dimensional reduction of textural features extracted using curvelets is performed by principal component analysis. Also, clinical predictors among the demographic parameters and morphological features are found using non-conditional logistic regression. Three classifiers, a least absolute shrinkage and selection operator (LASSO) regression method, a support vector machine (SVM), and a multilevel model are established.

Materials and Methods

Subjects

This study was performed with ethics approval (Ethics Committee of Xuanwu Hospital, Capital Medical University; Approval Document No. [2011] 01). Written consent was given by all patients involved and all were provided by five hospitals between November 1, 2006 and November 1, 2011. All the cases were identified by 8 thoracic radiologists with experience ranging between 4 and 20 years. The final diagnosis of malignant and benign status was determined by either operation or biopsy. Conflicts in the final interpretation of the CT images were resolved by consensus discussion.

A total of 4,742 regions of interest (ROIs) were acquired from 502 patients (275 men, 227 women). There were 1,343 benign ROIs in 152 patients and 3,399 malignant ROIs in 350 patients, as presented in Table 1. The ages of the patients ranged from 18 to 89 years (the mean age is 58.8 years). Another 283 ROIs from 18 patients (10 men, 8 women) were used as a validation sample. These comprised of 187 benign (13 patients) and 96 malignant CT images (5 patients). In addition, 20 variables related to demographic and clinical information were collected.

Table 1. Description of the Training Data Set

Diagnosis	Cases (%)	ROIs (%)
Benign	152(100.00)	1,343(100.00)
Tuberculosis	49 (32.24)	405 (30.16)
Inflammatory pseudotumor	32 (21.05)	277 (20.63)
Hamartoma	40 (26.32)	372 (27.70)
Pulmonary interstitial edema	3 (1.97)	21 (1.56)
Sclerosing hemangioma	16 (10.53)	143 (10.65)
Clear cell tumor	4 (2.63)	24 (1.79)
Chondroma	8 (5.26)	101 (7.52)
Malignant	350(100.00)	3,399(100.00)
Adenocarcinoma	270 (77.14)	2,642 (77.73)
Squamous carcinoma	53 (15.14)	484 (14.24)
Adenosquamous carcinoma	20 (5.71)	204 (6.00)
Malignant carcinoid tumor	7 (2.00)	69 (2.03)

Texture features extraction

All the CT scans were obtained using a 64-slice helical CT scanner (GE/Light Speed Ultra System CT99, USA) using a tube voltage of 120 kV and a current of 200 mA. The reconstruction thickness and intervals for routine scanning were 0.625 mm. The data were reconstructed using a 512 × 512 matrix. CT images were supplied in standard DICOM format and the nodule sizes were 0.3–3.0 cm. All of the SPNs in the CT images were segmented manually to obtain ROIs and the textural features were extracted ROI by ROI. The region growing algorithm (Zhu et al., 2013), a popular tool for image segmentation, was used to remove any background pixels. Curvelet transformation was used to extract textural features (Wang et al., 2010), including entropy, mean, correlation, energy, homogeneity, standard deviation, maximum probability, inverse difference moment, cluster tendency, inertia, sum-mean, difference-mean, sum-entropy, and difference-entropy. Overall, 140 textures were extracted from the ROIs.

Data analysis

Dimensional reduction of the textural features extracted by curvelet transformation was performed using principal component analysis. Furthermore, clinical predictors among the demographic parameters and morphological features were found by non-conditional logistic regression. LASSO, SVM, and multilevel models were used as prediction models, while back substitution and 10-fold cross validation were used to assess them. The 10-fold cross validation method randomly divides the data set into 10 parts — 9 of them are used as training sets and the other one is used, in turn, as the test set. In contrast, the back substitution method uses all the data as a training set and then selects one part to use as the test set.

Considering that the prediction results of this study are binary (viz. the case is either malignant or benign), and that there are three levels to be considered [hospitals (level-3), patients (level-2), and their CT slices (level-1)], a three-level binomial logistic model was selected. A radial basis function kernel was chosen for the kernel of the SVM as a trial. Sensitivity (Sen), specificity (Spe), accuracy (Acc), and Youden index (Yi) were used to evaluate the results made by these models in the external validation set. These are defined as:

$$Sen = \frac{\text{the number of malignant images correctly predicted}}{\text{the total number of malignant images}}, \quad (1)$$

$$Spe = \frac{\text{the number of benign images correctly predicted}}{\text{the total number of benign images}}, \quad (2)$$

$$Acc = \frac{\text{the number of images correctly predicted}}{\text{the total number of images}}, \quad (3)$$

$$Yi = Sen - (1 - Spe), \quad (4)$$

The program for implementing the prediction models was performed using R software (version 2.14.2) and SAS (version 9.2).

Table 2. The Eleven Clinical Predictors Selected by Logistic Regression

Variables (control)	B	SE	*P*	OR	95%CI	
Age (<55 years)	1.412	0.288	<0.001	4.104	2.334	7.215
Nodule diameter (<2cm)	2.138	0.705	0.002	8.482	2.131	33.766
History of turmor (No)	1.256	0.514	0.015	3.510	1.281	9.620
Lymphadenectasis (No)	0.810	0.354	0.022	2.249	1.124	4.498
Ground-Glass Opacity (No)	2.924	0.747	<0.001	18.618	4.306	80.489
Boundary (Smooth)	1.248	0.305	<0.001	3.483	1.917	6.329
Lobulation (No)	1.263	0.313	<0.001	3.537	1.916	6.530
Vacuole sign (No)	1.374	0.447	0.002	3.953	1.647	9.488
Calcification (No)	-1.672	0.394	<0.001	0.188	0.087	0.407
Cavitations (No)	-1.089	0.533	0.041	0.336	0.118	0.957
Pleural indentation (No)	0.620	0.310	0.045	1.860	1.013	3.412

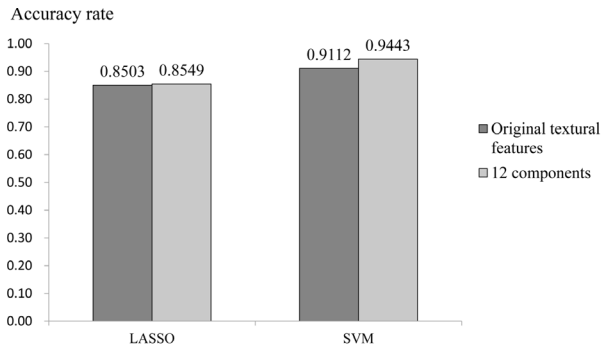

Figure 1. The 10-fold Cross Validation of the LASSO and SVM Methods using 12 Principal Components vs. the Original Curvelet-based Textural Features. Abbreviations: LASSO, least absolute shrinkage and selection operator; SVM, support vector machine

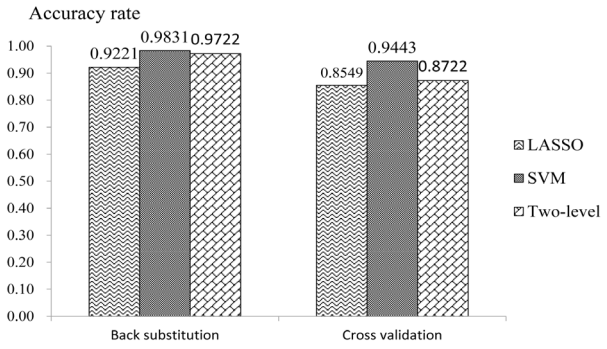

Figure 2. Cross Validation and Back Substitution in the LASSO, SVM, and Two-level models. Abbreviations: LASSO, least absolute shrinkage and selection operator; SVM, support vector machine; Two-level, two-level logistic regression model

Results

Non-conditional logistic regression

Non-conditional logistic regression was used to find the clinical predictors among the demographic parameters and morphological features which were used as independent variables in the analysis. The SPNs formed the dependent variables: 1 if malignant, and 0 if benign. Finally, 11 variables were selected (Table 2).

Principal component analysis

The results show that the first principal component can account for 66.33% of the total variance, while the top 12 can account for 90.68%. In light of the accuracy rates obtained using 10-fold cross validation in the LASSO and SVM methods (using 12 principal components vs. the original curvelet-based textural features, see Figure 1), the top 12 components were used in the models in place of the original textural features.

Table 3. Prediction Results in the External Validation Set. Abbreviations Used: LASSO, Least Absolute Shrinkage and Selection Operator; SVM, Support Vector Machine; and Two-level, Two-level Logistic Regression Model

Model	AUC	Sensitivity	Specificity	Accuracy
LASSO	0.908	0.979	0.775	0.845
SVM	0.855	0.802	0.808	0.958
Two-level	0.868	0.99	0.647	0.763

Comparison of the models

Based on the number of levels involved [hospitals (level-3), patients (level-2), and their CT slices (level-1)], a three-level binomial logistic model was intended to be established. However, the level-3 variance was found to lack statistical significance in the zero model ($\chi^2 = 0.866$, $P = 0.352$). On the other hand, level-2 did ($\chi^2 = 51.600$, $P < 0.001$). Therefore, a two-level binomial logistic model was established instead.

The accuracy rates using 10-fold cross validation and back substitution in the SVM, LASSO, and multilevel models using 12 principal components and 11 clinical predictors are shown in Figure 2. In addition, results from the external validation set are given in Table 3.

Discussion

The diagnosis of pulmonary nodules is still an important clinical topic. In this study, curvelet transformation was used to extract textural features and principal component analysis was used to simplify the textual features. The results show that using 12 of the principal components of the textural features to establish SVM or LASSO models is better than using the original set of textual features. This amply illustrates the superiority of using principal component analysis on the data. In addition, back substitution and 10-fold cross validation was used to assess the prediction models internally. The accuracy rate is relatively high (0.9443 and 0.9831 using SVM), and is better than that obtained in an earlier study using the original textural features and three demographic

parameters (Sun et al., 2013a). Furthermore, another 283 ROIs from 18 patients were used as an external validation set. The least area under the curve (AUC) was 0.855, significantly higher than 0.5 ($P < 0.001$). This illustrates that all these models have a certain diagnostic accuracy. More importantly, the highest sensitivity obtained (0.990) was higher than in other comparable research (Herder et al., 2005; Schultz et al., 2008; Li et al., 2011).

While some demographic parameters and morphological features taken into the data sets had similar results with the Mayo Clinic model and other research (Khan et al., 1991; Gurney, 1993; Erasmus et al., 2000; Alzubi et al., 2011), smoking history (i.e. whether the patient smoked or not) did not reflect a difference between malignant and benign nodules. The reason for this may be that most of the malignant cases involved adenocarcinoma (77.14%), which does not have as clear a relationship with smoke as squamous carcinoma does (Li et al., 2011). Furthermore, more specific indices are needed to accurately reflect the true situation relating to smoke, such as smoking index, smoking cessation, etc.

Other limitations also need addressing. As all of the cases were obtained from hospitals, it was hard to balance benign and malignant cases. Unfortunately, the models used are known to have weaknesses with imbalanced training sets (Borrajo et al., 2011). Besides this, the cases used as external verification samples were selected as they were diagnosed latter, which is not random. This reflects a selection bias that needs to be acknowledged. Therefore, further studies are required.

In conclusion, after using a dimensional reduction method, the data set with highest accuracy rate combined textural features extracted by curvelet transformation and personal features and involved using a SVM method. It is clear that this method may therefore be used as a useful auxiliary tool to differentiate between benign and malignant SPNs in CT images.

Acknowledgements

The authors thank the doctors from the radiology departments of Beijing Xuanwu Hospital, Friendship Hospital, Chaoyang Hospital, Beijing Chest Hospital, and Fuxing Hospital, for their assistance with data acquisition. Acknowledgment is also gratefully made to the Natural Science Fund of China (serial nos.: 81172772 and 30972550) and the Natural Science Fund of Beijing (serial nos.: 4112015 and 7131002) for their generous financial support.

References

Acharya UR, Faust O, Sree SV, et al (2012). ThyroScreen system: high resolution ultrasound thyroid image characterization into benign and malignant classes using novel combination of texture and discrete wavelet transform. *Comput Methods Programs Biomed*, **107**, 233-41.

Alzubi S, Islam N, Abbod M (2011). Multiresolution analysis using wavelet, ridgelet, and curvelet transforms for medical image segmentation. *Int J Biomed Imaging*, **2011**, 136034.

Arebey M, Hannan MA, Begum RA, et al (2012). Solid waste bin level detection using gray level co-occurrence matrix feature extraction approach. *J Environ Manage*, 104, 9-18.

Beadsmoore CJ, Screaton NJ (2003). Classification, staging and prognosis of lung cancer. *Eur J Radiol*, **45**, 8-17.

Borrajo L, Romero R, Iglesias EL, et al (2011). Improving imbalanced scientific text classification using sampling strategies and dictionaries. *J Integr Bioinform*, **8**, 176.

Dettori L, Semler L (2007). A comparison of wavelet, ridgelet, and curvelet-based texture classification algorithms in computed tomography. *Comput Biol Med*, **37**, 486-98.

Dua S., Acharya UR, Chowriappa P, et al (2012). Wavelet-based energy features for glaucomatous image classification. *IEEE Trans Inf Technol Biomed*, **16**, 80-7.

Eltoukhy MM, Faye I, Samir BB (2010). Breast cancer diagnosis in digital mammogram using multiscale curvelet transform. *Comput Med Imaging Graph*, **34**, 269-76.

Erasmus JJ, Connolly JE., McAdams HP, et al (2000). Solitary pulmonary nodules: Part I. Morphologic evaluation for differentiation of benign and malignant lesions. *Radiographics*, **20**, 43-58.

Gould MK, Ananth L, Barnett PG (2007). A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. *Chest*, **131**, 383-8.

Guo L, Dai M, Zhu M (2012). Multifocus color image fusion based on quaternion curvelet transform. *Opt Express*, **20**, 18846-60.

Gurney JW (1993). Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. *Part I. Theory*. *Radiology*, **186**, 405-13.

Hart J (2011). Lung cancer in Oregon. Dose Response, 9, 410-5.

Henschke CI, Yankelevitz DF, Mateescu I, et al (1997). Neural networks for the analysis of small pulmonary nodules. *Clin Imaging* **21**, 390-9.

Herder GJ, van Tinteren H, Golding RP, et al (2005). Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. *Chest*, **128**, 2490-6.

Khan A, Herman PG, Vorwerk P, et al (1991). Solitary pulmonary nodules: comparison of classification with standard, thin-section, and reference phantom CT. *Radiology*, 179, 477-81.

Khorasani A, Daliri MR (2013). Estimation of neural firing rate: the wavelet density estimation approach. *Biomed Tech (Berl)*, **58**, 377-86.

Ko JP, Berman EJ, Kaur M, et al (2012). Pulmonary Nodules: growth rate assessment in patients by using serial CT and three-dimensional volumetry. *Radiology*, **262**, 662-71.

Li Y, Chen KZ, Wang J (2011). Development and validation of a clinical prediction model to estimate the probability of malignancy in solitary pulmonary nodules in Chinese people. *Clin Lung Cancer*, **12**, 313-9.

Ma X, Zhang Z, Hu Y, et al (2012). Combined surgical intervention treatments for lung cancer and coronary heart disease patients. *Zhongguo Fei Ai Za Zhi*, **15**, 602-5.

Matsuki Y, Nakamura K, Watanabe H, et al (2002). Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: evaluation with receiver operating characteristic analysis. *AJR Am J Roentgenol*, **178**, 657-63.

Meselhy Eltoukhy M, Faye I, Belhaouari Samir B (2010). A comparison of wavelet and curvelet for breast cancer diagnosis in digital mammogram. *Comput Biol Med*, **40**, 384-91.

Nakamura K., Yoshida H, Engelmann R, et al (2000). Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks. *Radiology*, **214**, 823-30.

Schultz EM, Sanders GD, Trotter PR, et al (2008). Validation

of two models to estimate the probability of malignancy in patients with solitary pulmonary nodules. *Thorax*, **63**, 335-41.

Soerjomataram I, Lortet-Tieulent J, Parkin DM, et al (2012). Global burden of cancer in 2008: a systematic analysis of disability-adjusted life-years in 12 world regions. *Lancet* **380**, 1840-50.

Sun T, Wang J, Li X, et al (2013a). Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set. Comput Methods Programs Biomed.

Sun T, Zhang R, Wang J, et al (2013b). Computer-aided diagnosis for early-stage lung cancer based on longitudinal and balanced data. *PLoS One*, **8**, e63559.

Swensen SJ, Silverstein MD, Ilstrup DM, et al (1997). The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch Intern Med*, **157**, 849-55.

Wang H, Guo XH, Jia ZW, et al (2010). Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image. *Eur J Radiol*, **74**, 124-9.

Way TW, Sahiner B, Chan HP, et al (2009). Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. *Med Phys*, **36**, 3086-98.

Wu H, Sun T, Wang J, et al (2013). Combination of radiological and gray level co-occurrence matrix textural features used to distinguish solitary pulmonary nodules by computed tomography. *J Digit Imaging*, **26**, 797-802.

Zhu L, Gao Y, Appia V, et al (2013). Automatic Delineation of the Myocardial Wall from CT Images via Shape Segmentation and Variational Region Growing. *IEEE Trans Biomed Eng*, **60**, 2887-95.