Prediction of Malignant and Benign of Lung Tumor using a Quantitative Radiomic Method

Jun Wang, Xia Liu*, Di Dong, Jiangdian Song, Min Xu, Yali Zang*, Jie Tian*, *Fellow, IEEE*

Abstract—**Lung cancer is the leading cause of cancer mortality around the world, the early diagnosis of lung cancer plays a very important role in therapeutic regimen selection. However, lung cancers are spatially and temporally heterogeneous; this limits the use of invasive biopsy. But radiomics which refers to the comprehensive quantification of tumour phenotypes by applying a large number of quantitative image features has the ability to capture intra-tumoural heterogeneity in a non-invasive way. Here we carry out a radiomic analysis of 150 features quantifying lung tumour image intensity, shape and texture. These features are extracted from 593 patients computed tomography (CT) data on Lung Image Database Consortium Image Database Resource Initiative (LIDC-IDRI) dataset. By using support vector machine, we find that a large number of quantitative radiomic features have diagnosis power. The accuracy of prediction of malignant of lung tumor is 86% in training set and 76.1% in testing set. As CT imaging of lung tumor is widely used in routine clinical practice, our radiomic classifier will be a valuable tool which can help clinical doctor diagnose the lung cancer.**

I. Introduction

Lung cancer is the leading cause of cancer mortality around the world. By 2030, up to 10 million patients in the world will die of lung cancer in terms of the report from the World Health Organization [1]. In medicine and therapy, we can get a lot of information from medical images which are often used to help diagnosis and therapy. Lung cancers are spatially and temporally heterogeneous that limits the genomic and proteomic based technologies which require biopsies or invasive surgeries to extract and analyses small portions of tumor tissue [2], [3]. But medical imaging has the ability to capture intra-tumoral heterogeneity in a non-invasive way [4]. The most widely used imaging modality in oncology is x-ray CT. Lung cancer CT images presented a strong contrast, reflects the differences in tumor gray value intensity, tumor texture and tumor shape. Early diagnosis of lung cancer has an impact on survival benefit improvements [5]. However, in clinical practice, there is a lack of a non-invasive method for the diagnosis of lung cancer with relatively high accuracy.

Radiomics is an emerging field, the high-throughput extraction of large amounts of features from radiographic images, which converts imaging into mineable data with high fidelity and high throughput [6]. We hypothesize that these imaging features capture distinct phenotypic differences of tumors may have prognostic power [7]. The radiomics can be divided into four distinct processes: (1) image acquisition; (2) image segmentation and rendering; (3) feature extraction and selection; (4) informatics analyses.

II. Materials and Methods

A. Patients and Data Selection

In this retrospective study, CT data of 593 patients from LIDC-IDRI dataset were analyzed [8]. Training set used to identify a predictor including a set of 400 patients. This predictor was then validated in an independent cohort containing 193 patients. According to the information about malignant degree of nodule provided by LIDC-IDRI, we divided the nodules into two categories (benign or malignant). Patients may have more than one nodule, but we just use the most malignant one. To maximize our ability to predict status of nodules, we deliberately designed the training set to contain equal numbers of patients with benign and malignant nodules (200 patients with benign nodules vs 200 patients with malignant nodules). The testing set includes 71 patients with benign nodules and 122 patients with malignant nodules.

B. Image Segmentation

Segmentation of lung cancer CT images is a crucial step for subsequent informatic analyses. Manual segmentation by expert radiologists is often treated as golden standard. However, this method has high inter-reader variability and is very time consuming; thus not feasible for radiomic analysis requiring big data sets.

We use preliminary laboratory research “toboggan based growing automatic segmentation approach” for images segmentation [9]. The algorithm automatically initializes seed point without any human interaction and dice coefficient value of the algorithm on 819 lesions from the LIDC-IDRI dataset is 81.57% which can minimize the bad segmentation (Fig. 1).

*Research supported by the National Natural Science Foundation of China under Grant No. and 81227901, 81501549, 61231004, 81501616, 81301346, the National Basic Research Program of China under Grant No. 61301002 and No. 61302025 and the Natural Science Foundation of Heilongjiang Province of China under Grant No. F201311.

Asterisk indicates corresponding author.

J. Wang is with the Measurement-Control Technology and Communications Engineering School, Harbin University of Science and Technology, Harbin, 150080, China. And the Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangjun.542@163.com).

*X. Liu is with the Harbin University of Science and Technology, Harbin, 150080, China (phone: +8613804516003, e-mail: liuxia@hrbust.edu.cn).

D. Dong and M. Xu is with the Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

*Y. Zang is with the Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (phone: +86-10-82618465, e-mail: yali.zang@ia.ac.cn).

*J. Tian is with the Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (phone: +86-10-82618465, e-mail: jie.tian@ia.ac.cn).

J. Song is with the Sino-Dutch Biomedical and Information Engineering School, Northeastern University, Shenyang, 110819, China.

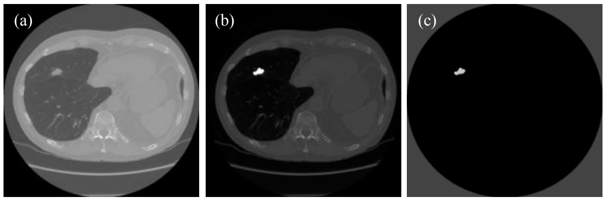

Figure 1. Segmentation result of lung lesion. (a) Original CT image of lung lesion. (b) Segment lung lesion from original CT image. (c) Segmented lung lesion area.

C. Feature Extraction and Selection

We extracted 150 quantitative image features from lung lesions describing tumor phenotype characteristics. These features can be divided into four groups: (I) lung lesion image intensity, (II) lung tumor shape and size, (III) lung tumor texture, and (IV) wavelet features. In the first group, we used first-order statistics to quantify tumor intensity characteristics by calculating the histogram of all tumor voxel intensity values. Group 2 describes the shape and size of the lung lesion area. Group 3 consists of textual features that have ability to quantify intratumor heterogeneity differences in the texture. Group 4 calculates the intensity and texture features from wavelet decompositions of original image, reflecting features on different frequency ranges within the tumor volume.

Correlation and redundancy between features may reduce the accuracy of the classification, while medical imaging research usually belong to small sample learning, too many features will increase the complexity of classifier resulting in over-fitting and decreasing the generalization of the classifier. Therefore we need to select and optimize feature set [10].

To get an optimal feature subset for latter training, a random forest classifier was used to select features. Random forest is a notion of general technique of random decision forests that are an ensemble learning method for feature extraction and other tasks.

Random forest algorithm is a kind of robust and efficient algorithm; it can be used to select features by calculating the variable importance of features [11]. The goal of the algorithm is to find features which highly associated with classification results. And the selected feature subset is relative small but able to adequately predict the outcome with a high accuracy.

Briefly, the random forest algorithm we used including two main steps. The first step includes four processes: (1) calculate the feature variables importance and sort in descending order; (2) select a certain proportion of top features and generate a new feature set; (3) generate new random forest repeat the process one using the new feature set; (4) repeat the above processes until feature set including m features. In the second step, we select the feature set which has the minimum out of bag error.

D. Definition of radiomic classifier

First, random forest algorithm is used to select features on the entire training set. Second, the selected features are token as input of SVM to train a prediction model. Finally, we validate the diagnostic accuracy on testing set (Fig. 2). Support vector machine (SVM) is a machine learning algorithm which based on structural risk minimization principle, in order to control the generalization ability, we need to dominate empirical risk and confidence range [12]. SVM takes minimizing the confidence range as the optimization objective under the constraint of training error. Eventually, it turns into solving a convex quadratic programming problem, so the solution of SVM is unique and global optimum.

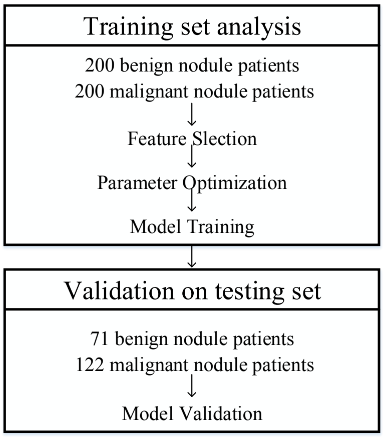
Training set analysis
200 benign nodule patients
200 malignant nodule patients
Feature Slection
Parameter Optimization
Model Training
Validation on testing set
71 benign nodule patients
122 malignant nodule patients
Model Validation

Figure 2. Generation and validation of radiomic classifier.

To obtain higher prediction accuracy, the parameters of SVM are optimized. Before optimizing the parameters, the training set and the validation set data should be normalized. Equation (1) is the “Min-Max Normalization”:

$$x' = \frac{x - \max(x)}{\max(x) - \min(x)} \quad (1)$$

The classification performance of SVM is influenced by many factors, among which the following two factors are the key factors: (1) The error penalty parameter C; (2) kernel and kernel parameter g. The error penalty parameter C is used to adjust the proportion of confidence range and the experience risk in determined data subspace to maximize the generalization ability of the learning machine [13]. In determined data subspace, the small C means the penalty of empirical error is little, the complexity of learning machine is low and empirical risk value is large, resulting “under-fitting”. On the contrary, if C is too large, the data will be over fitted. When C exceeds a certain value, the complexity of SVM reaches the maximum, and then the empirical risk and generalization will hardly change. For radial basis kernel function is nonlinear, has little parameters and can map data into high dimension, we choose it for SVM.

For the above reasons, genetic algorithm (GA) was used to optimize C and g. GA exclusively rely on repeated evaluations of the objective function, and the subsequent search direction after each evaluation follows certain heuristic guidelines [14].

Random initialization of population, evaluation of fitness function and generation of new population are the three main steps for GA (Fig. 3). In the random initialization of population step, all the parameters of the optimization problem are encoded to fix-length binary bit strings, called chromosomes or individuals. In the evaluation of fitness function step, all the individuals are evaluated by means of a fitness function. Then, the fitness function is used in next step to create a genetic pool. After the fitness of all individuals is calculated, a new population is created. The creation of a new generation basically needs three stages, reproduction, crossover and mutation. The overall objective of this step is to receive a new population with individuals which have high fitness values.

In the stage of reproduction, we select the individuals which have high fitness values among the population. The population after reproduction stages is called mating pool. In the step of crossover, crossover operator is applied to the mating pool to generate new individuals. In the step of mutation, the mutation process introduces further changes to a bit string. It is required that if the population does not contain all the encoded information required to solve a specific problem, no amount of gene mixing can provide a satisfactory solution. It is possible to produce new chromosomes by applying the mutation operator. We use the most common technique to change a randomly chosen bit in the bit string of the individual to be mutated. Thus certain bit is changed from 1 into 0 or from 0 into 1 [15].

Here we consider the prediction accuracy of SVM in cross validation as the fitness function, and select individuals who have high fitness value which mean high prediction accuracy. When the change of fitness value less than an invariant constant, or the generation over a certain iteration number, or the prediction of the individual is high enough, the algorithm is exited and we can get the optimal parameters of SVM. All parameters of the algorithm are set before the step of population initialization.

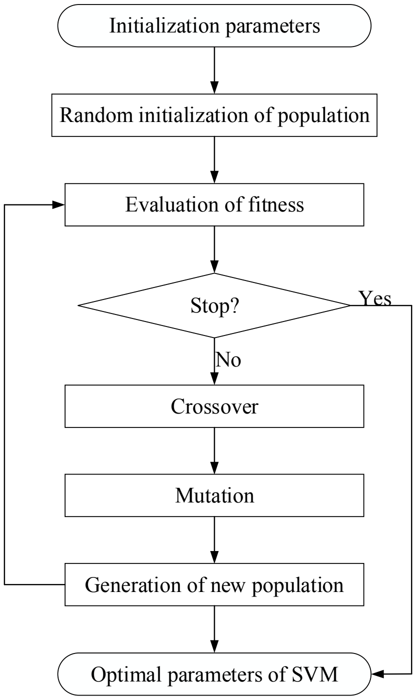
Initialization parameters
Random initialization of population
Evaluation of fitness
Stop?
Yes
No
Crossover
Mutation
Generation of new population
Optimal parameters of SVM

Figure 3. Flow of genetic algorithm for SVM parameters optimization.

RESULTS

First, we segmented lung tumor CT images of 593 patients from LIDC-IDRI dataset. Then, four kinds of quantitative radiomic features are extracted from each segmented image, totally 150 features including intensity features, shape features, texture features and wavelet features. The random forest defined a feature set including 15 features: Group 1 (Energy, Entropy); Group 2 (Compactness, Spherical Disproportion, SurfacetoVolumeRatio, Volume), 'Spherical Disproportion' quantifying nodules spherical disproportion, 'SurfacetoVolumeRatio' describing surface to volume ratio; Group 3 (Run Percentage, Maximum Probability, Variance, Autocorrelation, Cluster Prominence, Cluster Shade, Cluster Tendency); Group 4 (GMTRmean, GPTRentropy), 'GMTRmean' based on Gabor and describing the feature of magnitude-based texture representation that make a averaging, ' GPTRentropy' describing the feature of Gabor phase-based texture representation that make an operation of entropy. The selected feature subset is normalized by "Min-Max Normalization". When we separately used normalized feature subset and original feature subset for optimizing, we found that using the normalized subset to solve the optimization problem is much faster than the non-normalized subset. In order to increase the accuracy of SVM classifier, GA was used to optimize the SVM parameters. To approximate the optimal solution we use four decimal places (optimal C=1.1167, g=6.3334), the fitness curve is shown in Fig. 4. Taking normalized feature subset as input of SVM to train a predictive model. The diagnostic accuracy of the prediction model for status classification (benign vs malignant) in the training set was 86% (344 of 400), with a sensitivity of 82.5% (165 of 200), specificity of 89.5% (179 of 200), positive predictive value (PPV) of 88.7% (165 of 186) and negative predictive (NPV) of 83.6% (179 of 214). We next validated the predictive ability of the prediction model in the testing set, which received a diagnostic accuracy of 76.1% (147 of 193), with a sensitivity of 74.6% (91 of 122), specificity of 78.9% (56 of 71), PPV of 85.8% (91 of 106) and NPV of 64.37% (58 of 87). Overall, the diagnostic accuracy of the prediction model across the entire study set in predicting benign and malignant lung tumor was 82.7% (491 of 593), with a sensitivity of 79.5% (256 of 322) and a specificity of 77.6% (235 of 271). All prediction results are detailed in Table I.

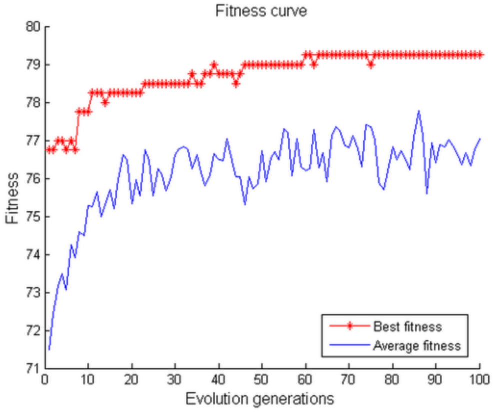

Figure 4. GA fitness curve.

TABLE I. DIAGNOSTIC ACCURACY

Subset	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)
Training set	82.5(165/200)	89.5(179/200)	88.7(165/186)	83.6(179/214)	86.0(344/400)
Testing set	74.6%(91/122)	78.9(56/71)	85.8(91/106)	66.7(58/87)	76.1(147/193)
Overall	79.5(256/322)	86.7(235/271)	87.7(256/292)	78.7(237/301)	82.7(491/593)

III. CONCLUSION

We develop and validate a radiomic classifier that predict the status (benign or malignant) of lung tumor from CT image data. The final feature set selected by random forest algorithm including 15 quantitative radiomic features which have prognostic value. The results reveals that our radiomic prediction classifier the diagnostic accuracy in training set is 86% and in test set is 76.1%. The goal of radiomic classifier is not to replace molecular testing but to enable radiologists to better understand the CT images with malignant lung tumors and to translate this understanding into clinical practice. There two benefits of radiomic classifier：(1) We can get the status of lung tumor in a non-invasive way; (2) It is very convenient to calculate if we get lung CT imaging which is routinely used in treatment. Therefore, the radiomic classifier can be applied in clinical practice.

While these results are promising, our radiomic classifier require more data for further training and validation to improve the diagnostic accuracy. Further investigation is required since we just focus on the classification of begin and malignant of lung tumor. As our method is easily reproducible, it could be explored in other cancer types for which lung tumor imaging is widely available.

REFERENCES

[1] W. H. Organization. Description of the global burden of NCDs, their risk factors and determinants. *Geneva, Switzerland: World Health Organization*, 2011.

[2] Lambin P1, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, and Aerts HJ, “Radiomics: Extracting more information from medical images using advanced feature analysis,” *European journal of cancer*, vol.48, issue 4, pp. 441-446, 2012.

[3] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar,Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nat Commun*, vol.5, no.4006, 2014.

[4] Brana, and L. L. Siu. “Locally advanced head and neck squamous cell cancer: treatment choice based on risk factors and optimizing drug prescription,” *Annals of Oncology*, vol.23, pp. 178-185, 2012.

[5] Chintan Parmar, Ralph T. H. Leijenaar, Patrick Grossmann, Emmanuel Rios Velazquez, Johan Bussink, Derek Rietveld, Michelle M. Rietbergen, Benjamin Haibe-Kains, Philippe Lambin and Hugo J.W.L. Aerts, “Radiomic feature clusters and Prognostic Signatures specific for Lung and Head & Neck cancer,” *Scientific Reports*, vol.5, no.11044, 2015.

[6] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Schabath, Kenneth Forster, Hugo J.W.L. Aerts, Andre Dekker, David Fenstermacher, Dmitry B Goldgof, Lawrence O Hall, Philippe Lambin,Yoganand Balagurunathan, Robert A Gatenby, and Robert J Gillies, “QIN “Radiomics: The Process and the Challenges”,” *Magn Reson Imaging*, vol.30, issue 9, pp. 1234-1248, 2012.

[7] Robert J. Gillies, Paul E. Kinahan, Hedvig Hricak, “Radiomics: Images Are More than Pictures, They Are Data,” “Radiological Society of North America”, vol.278, issue 2, pp. 563-577, 2015.

[8] S. G. Armato, G. McLennan, D. Hawkins, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. Van Beeke, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batrah, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Vande Casteele, S. Gupte, M. Sallamm, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke, “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans,” *Med. Phys*., vol. 38, no. 2, pp. 915–931, 2011.

[9] Jiangdian Song, Caiyun Yang, Li Fan, Kun Wang, Feng Yang, Shiyuan Liu, and Jie Tian, “Lung lesion extraction using a toboggan based growing automatic segmentation approach,” *Medical Imaging, IEEE Transactions on,* vol.35, issue 1, pp. 337-353, 2015.

[10] Edward R Dougherty, Jianping Hua and Chao Sima, “Performance of Feature Selection Methods,” *Current Genomics*, vol. 10, issue 6, pp. 365-374, 2009.

[11] Ramón Díaz-Uriarte, and Sara Alvarez de Andrés, “Gene selection and classification of microarray data using random forest,” *BMC Bioinformatics*, vol. 7, issue 1, pp. 1-13, 2006.

[12] Vapnik, V. N., “An overview of statistical learning theory,” *Neural Networks, IEEE Transactions on*, vol. 10 issue 5 pp. 988-999, 1999

[13] Chih-Chung Chang, and Chih-Jen Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2 issue 3 pp. 1-27, 2011.

[14] D Whitley, T Starkweather, and C Bogart, “Genetic algorithms and neural networks: optimizing connections and connectivity,” *Parallel Computing*, vol. 14, issue 3 pp. 347-361, 1990.

[15] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *Evolutionary Computation, IEEE Transactions on*, vol. 6, issue 2, pp. 182-197, 2002.