Developing a Radiomics Framework for Classifying Non-Small Cell Lung Carcinoma Subtypes

Dongdong Yu[a], Yali Zang[a], Di Dong*[a], Mu Zhou[b], Olivier Gevaert[b], Jingyun Shi*[c], and Jie Tian*[a]

[a]The Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[b]The Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University
[c]Department of Radiology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China

ABSTRACT

Patient-targeted treatment of non-small cell lung carcinoma (NSCLC) has been well documented according to the histologic subtypes over the past decade. In parallel, recent development of quantitative image biomarkers has recently been highlighted as important diagnostic tools to facilitate histological subtype classification. In this study, we present a radiomics analysis that classifies the adenocarcinoma (ADC) and squamous cell carcinoma (SqCC). We extract 52-dimensional, CT-based features (7 statistical features and 45 image texture features) to represent each nodule. We evaluate our approach on a clinical dataset including 324 ADCs and 110 SqCCs patients with CT image scans. Classification of these features is performed with four different machine-learning classifiers including Support Vector Machines with Radial Basis Function kernel (RBF-SVM), Random forest (RF), K-nearest neighbor (KNN), and RUSBoost algorithms. To improve the classifiers' performance, optimal feature subset is selected from the original feature set by using an iterative forward inclusion and backward eliminating algorithm. Extensive experimental results demonstrate that radiomics features achieve encouraging classification results on both complete feature set (AUC=0.89) and optimal feature subset (AUC=0.91).

Keywords: Non-Small Cell Lung Carcinoma, Histologic subtype classification, Radiomics analysis, Computed tomography, Computed-aided diagnosis

1. INTRODUCTION

Non-small cell lung carcinoma (NSCLC) is a lethal disease accounting for about 85% of all lung cancers with a dismal 5-year survival rate of 15.9% .[1] NSCLC can be primarily sub-divided into adenocarcinoma (ADC) and squamous cell carcinoma (SqCC) on the basis of where the lung cancer cell starts from. Overall, patients of ADC and SqCC account for 65% to 70% of all the lung cancer population.

NSCLC patients' treatment has been well documented over the past decade to indicate targeted therapy according to the different histologic subtypes.[2,3] The histologic subtypes are discerned by histopathological examination which is a golden standard in the lung nodule classification. However, the histologic examination can be inefficient if the availability tissue samples are inadequately provided. On the other hand, Computed Tomography (CT) has been a major imaging modality for early cancer detection in NSCLC. The recent emergence of Radiomics studies,[4,5] focusing on extracting large amounts of computational image features, presents new opportnutnities in cancer predictive analysis. A challenging yet important task is to infer the diagnostic value from growing volumes of CT images, allowing identification of discriminative radiomics features that are able to predict the histologic subtypes in NSCLC.

In this study, we focus on developing computational CT image features for predicting non-small cell lung carcinoma subtypes of ADC and SqCC. In particular, we introduce a computational framework utilizing the radiomics feature analysis for the histopathologic prediction. The presented method includes image feature extraction, feature selection and histopathologic prediction. More specifically, 52-dimensional feature including

Table 1. CT-based Texture feature list. Full notations are: **GLRLM features**: Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray-Level Nonuniformity (GLN), Run-Length Nonuniformity (RLN), Run Percentage (RP), Low Gray-Level Run Emphasis (LGRE), High Gray-Level Run Emphasis (HGRE), Short Run Low Gray-Level Emphasis (SRLGE), Short Run High Gray-Level Emphasis (SRHGE), Long Run Low Gray-Level Emphasis (LRLGE), Long Run High Gray-Level Emphasis (LRHGE), Gray-Level Variance (GLV), Run-Length Variance (RLV). **GLSZM features**: Small Zone Emphasis (SZE), Large Zone Emphasis (LZE), Gray-Level Nonuniformity (GLN), Zone-Size Nonuniformity (ZSN), Zone Percentage (ZP), Gray-Level Variance (GLV), Zone-Size Variance (ZSV), Low Gray-Level Zone Emphasis (LGZE), High Gray-Level Zone Emphasis (HGZE), Small Zone Low Gray-Level Emphasis (SZLGE), Small Zone High Gray-Level Emphasis (SZHGE), Large Zone Low Gray-Level Emphasis (LZLGE), Large Zone High Gray-Level Emphasis (LZHGE).

Texture Type	Texture Subtype	Feature Abbreviation
First Order	Global feature (#5)	Variance, Skewness, Kurtosis, Entropy, Uniformity
Second Order	GLCM texture feature (#9)	Contrast, Energy, Variance, Average, Correlation, Homogeneity, Entropy, Dissimilarity, IDM
High Order	GLRLM texture feature (#13)	SRE, LRE, GLN, RLN, RP, LGRE, HGRE, SRLGE, SRHGE, LRLGE, LRHGE, GLV,RLV
	GLSZM texture feature (#13)	SZE, LZE, GLN, ZSN, ZP, GLV, SZV, LGZE, HGZE, SZLGE, SZHGE, LZLGE, LZHGE
	NGTDM texture feature (#5)	Coarseness, Contrast, Busyness, Complexity, Strength

statistical features and texture features are extracted to characterize lung nodules in CT imaging. Next, we use four different machine-learning classifiers, including Support Vector Machines with radial basis function kernel (RBF-SVM), random forest (RF), K-nearest neighbor (KNN), RUSBoost classifier to build up the prediction model. Finally, we use the iterative forward inclusion and backward eliminating algorithm to select the significant features and improve the prediction model's ability. The purpose of this study is to investigate the correlation between radiomics CT imaging features and histologic subtypes of NSCLC.

2. MATERIALS AND METHODS

2.1 Dataset and Radiomics feature extraction

We collect a clinical dataset including 324 ADCs and 110 SqCCs patients with CT image scans. All the lung nodules are segmented by using an automatic segmentation method.[6] In order to character the lung nodules, we extract 52 CT-based features. There are 7 statistical features: IntensityMax, IntensityMin, IntensityAve, IntensityStd, Volume, Solidity, and Eccentricity. The first 4 features describe CT intensity variations. Volume stands for the nodule size. Solidity indicates the ratio of the number of voxels in the nodule to the number of voxels in the 3-D convex hull of the nodule. Eccentricity denotes the ellipsoid which best fit the nodule. We additionally extract 45 texture features including the first order statistics, second order statistics and higher order statistics. A full list of texture feature is shown in Table 1.

2.2 Feature selection

In this study, we use the iterative forward including and backward elimination to find the optimal feature set that characterizes differences between ADC and SqCC. Starting from an empty feature set, iterative forward inclusion and backward elimination [7] are employed to include and eliminate feature attribute in the current feature set to increase the cost function. The cost function is defined as below: $cost = \frac{1}{2} * (\frac{TP}{TP+FN} + \frac{TN}{FP+TN})$. True Positive (TP) stands that ADC patients correctly identified as ADC, False Positive (FP) denotes that SqCC patients incorrectly identified as ADC, True Negative (TN) refers that SqCC patients correctly identified as SqCC, and False Negative (FN) indicates that ADC patients incorrectly identified as SqCC.

2.3 Histopathologic prediction

For the classification model development, we adopt four different algorithms: Support Vector Machines with Radial Basis Function kernel (RBF-SVM), random forest (RF), K-nearest neighbor (KNN), and RUSBoost algorithms. In addition, we also build up the classification model using the optimal feature set selected by the

iterative forward inclusion and backward elimination to improve the best classifier's performance. The original image feature set and the optimal feature set are both trained with the 10-fold cross validation. We report the area under the receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and geometric mean (GM) as metrics to assess the average classification results from 10-fold cross validation. Following the definition of TP, TN, FP, and FN in Sec. 2.2, evaluation metrics are given as $Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$, $Sensitivity = \frac{TP}{TP+FN}$, $Specificity = \frac{TN}{FP+TN}$, $PPV = \frac{TP}{TP+FP}$, $NPV = \frac{TN}{TN+FN}$, $GM = \sqrt{Sensitivity * Specificity}$.

3. RESULTS AND DISCUSSION

To build up a robust histologic classification model, we report average results from 100 times 10-fold cross validation on the patient cohort (324 ADCs and 110 SqCCs) with four different classifiers and two different feature sets. We use the AUC and GM as the main rules to measure the performance of classification. AUC is a generalized indicator of the classifier that is independent of the sample class distribution, and the GM maximizes the accuracy on each of the two classes while keeping these accuracies balanced.[8]

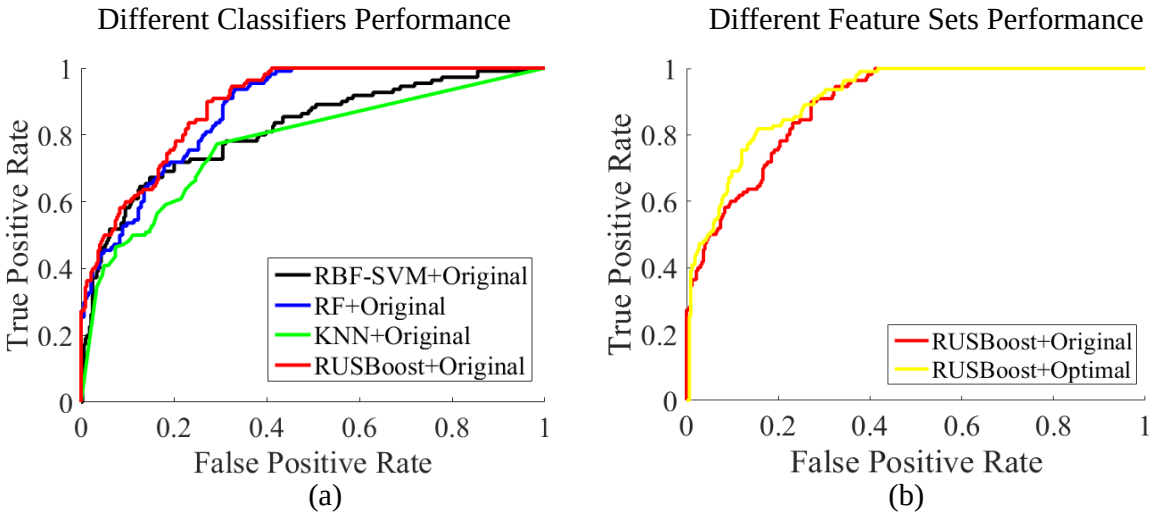

Figure 1. ROC curves of lung nodule histologic prediction using different types of classification algorithms based on original feature set (a): the black line, blue line, green line, and red line respectively indicate the ROC prediction curve by using RBF-SVM, RF, KNN, and RUSBoost. ROC curves of lung nodule histologic prediction using original feature set and optimal feature set based on RUSBoost classifier (b): the red line and the yellow line indicate the ROC prediction curves by using original feature set and the optimal feature set.

Table 2. Comparison of different classifiers for histologic prediction using different feature sets.

Classifiers	Feature Set	Accuracy	Sensitivity	Specificity	PPV	NPV	GM	AUC
RBF-SVM	Original	79.8%	**97.6%**	27.6%	79.9%	**79.5%**	0.52	0.82
RF	Original	81.0%	91.5%	50.1%	84.4%	66.7%	0.68	0.88
KNN	Original	77.2%	86.1%	51.1%	83.8%	55.5%	0.66	0.78
RUSBoost	Original	78.8%	80.2%	74.6%	90.3%	56.1%	0.77	0.89
RUSBoost	Optimal	**81.5%**	82.6%	**78.3%**	**91.8%**	60.5%	**0.80**	**0.91**

Tab. 2 shows the classification results with different classifiers with regards to two feature sets. In this study, we find that the classification model by using the RUSBoost classifier best predict the adenocarcinoma

(ADC) and squamous cell carcinoma (SqCC). RUSBoost is the best performing classifier which handles the dataset best among all the four different classifiers. Also, Tab. 2 shows the classification with the original feature set and the optimal feature set selected from the original feature set using iterative forward inclusion and backward elimination by using RUSBoost classifier. By using the optimal feature set with selected 20 features, the performance achieves an average accuracy of 81.5%, sensitivity of 82.6%, specificity of 78.3%, geometric mean of 0.80, and AUC of 0.91. With the RUSboost classification model, the GM and AUC of the model using optimal feature set outperform the original feature set by approximated 3.9% and 2.2%. We can see that the classification model built up by the optimal feature set outperformed the original feature set. This indicates that feature selection can be efficient to eliminate potential irrelevant features and improve prediction performance. Feature selection leads to a selected optimal feature set with 20-dimensional features. More specifically, the optimal feature set includes a statistical feature of IntensityMin and 19-dimensional texture features (3 first-order texture features, 2 second-order texture features, and 14 high-order texture features). We also report ROC curves in Fig. 1 to fully observe the classification outcomes.

4. CONCLUSION

In this paper, we investigate the association between CT imaging features and histologic subtypes for patients suffering from NSCLC. The proposed radiomics analytic framework presents encouraging results in predicting the adenocarcinoma and squamous cell carcinoma by extracting 52-dimensional radiomics feature. In particular, we achieve the highest classification results with AUC of 0.91 by applying the RUSBoost classifier with 20 selected radiomics features. This study based on radiomics analysis and histopathlogical characteristics supports the potential of computational CT-based analysis as an non-invasive means to facilitate NSCLC diagnosis. The proposed prediction model therefore holds promise to provide objective and reproducible diagnosis for non-small cell lung carcinoma.

5. ACKNOWLEDGMENT

This work is not being and has not been submitted for any publication or presentation elsewhere.

REFERENCES

[1] Kligerman, S. and White, C., "Epidemiology of lung cancer in women: risk factors, survival, and screening," *American Journal of Roentgenology* **196**(2), 287–295 (2011).

[2] Dubey, S. and Powell, C. A., "Update in lung cancer 2008," *American journal of respiratory and critical care medicine* **179**(10), 860–868 (2009).

[3] Sandler, A., Gray, R., Perry, M. C., Brahmer, J., Schiller, J. H., Dowlati, A., Lilenbaum, R., and Johnson, D. H., "Paclitaxel–carboplatin alone or with bevacizumab for non–small-cell lung cancer," *New England Journal of Medicine* **355**(24), 2542–2550 (2006).

[4] Gillies, R. J., Kinahan, P. E., and Hricak, H., "Radiomics: images are more than pictures, they are data," *Radiology* **278**(2), 563–577 (2015).

[5] Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., Zegers, C. M., Gillies, R., Boellard, R., Dekker, A., et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer* **48**(4), 441–446 (2012).

[6] Song, J., Yang, C., Fan, L., Wang, K., Yang, F., Liu, S., and Tian, J., "Lung lesion extraction using a toboggan based growing automatic segmentation approach," (2015).

[7] Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M., "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations," *Nature genetics* **44**(7), 825–830 (2012).

[8] Barandela, R., Sánchez, J. S., Garcıa, V., and Rangel, E., "Strategies for learning in class imbalance problems," *Pattern Recognition* **36**(3), 849–851 (2003).