Physics in Medicine & Biology

PAPER

Computer-aided diagnosis of lung cancer: the effect of training data sets on classification accuracy of lung nodules

To cite this article: Jing Gong *et al* 2018 *Phys. Med. Biol.* **63** 035036

IOP Publishing Phys. Med. Biol. 63 (2018) 035036 (11pp) https://doi.org/10.1088/1361-6560/aaa610

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

Computer-aided diagnosis of lung cancer: the effect of training data sets on classification accuracy of lung nodules

Jing Gong[1], Ji-Yu Liu[2], Xi-Wen Sun[2], Bin Zheng[3] and Sheng-Dong Nie[1,4]

[1] University of Shanghai for Science and Technology, School of Medical Instrument and Food Engineering, 516 Jun Gong Road, Shanghai 200093, People's Republic of China

[2] Radiology Department, Shanghai Pulmonary Hospital, 507 Zheng Min Road, Shanghai 200433, People's Republic of China

[3] University of Oklahoma, School of Electrical and Computer Engineering, Norman, OK 73019, United States of America

[4] Author to whom any correspondence should be addressed.

E-mail: nsd4647@163.com

Keywords: computer-aided diagnosis, CADx, lung cancer, early stage, advanced stage

RECEIVED
24 October 2017

REVISED
12 December 2017

ACCEPTED FOR PUBLICATION
8 January 2018

PUBLISHED
5 February 2018

Abstract

This study aims to develop a computer-aided diagnosis (CADx) scheme for classification between malignant and benign lung nodules, and also assess whether CADx performance changes in detecting nodules associated with early and advanced stage lung cancer.

The study involves 243 biopsy-confirmed pulmonary nodules. Among them, 76 are benign, 81 are stage I and 86 are stage III malignant nodules. The cases are separated into three data sets involving: (1) all nodules, (2) benign and stage I malignant nodules, and (3) benign and stage III malignant nodules. A CADx scheme is applied to segment lung nodules depicted on computed tomography images and we initially computed 66 3D image features. Then, three machine learning models namely, a support vector machine, naïve Bayes classifier and linear discriminant analysis, are separately trained and tested by using three data sets and a leave-one-case-out cross-validation method embedded with a Relief-F feature selection algorithm.

When separately using three data sets to train and test three classifiers, the average areas under receiver operating characteristic curves (AUC) are 0.94, 0.90 and 0.99, respectively. When using the classifiers trained using data sets with all nodules, average AUC values are 0.88 and 0.99 for detecting early and advanced stage nodules, respectively. AUC values computed from three classifiers trained using the same data set are consistent without statistically significant difference ($p > 0.05$).

This study demonstrates (1) the feasibility of applying a CADx scheme to accurately distinguish between benign and malignant lung nodules, and (2) a positive trend between CADx performance and cancer progression stage. Thus, in order to increase CADx performance in detecting subtle and early cancer, training data sets should include more diverse early stage cancer cases.

1. Introduction

Lung cancer is the leading cause of cancer-related death worldwide, with a 5 year survival rate of less than 16% (Stewart and Wild 2015, Siegel *et al* 2016). However, the prognosis for lung cancer patients is strongly associated with the stage of cancer at the time of diagnosis (Torre *et al* 2016). Thus, early detection and diagnosis of lung cancer is important in order to increase the patients' survival rate. Currently, computed tomography (CT) is increasingly being used as an effective imaging tool for detecting pulmonary nodules at the early stage. However, detecting pulmonary nodules and determining which ones are highly likely to be malignant is a difficult and tedious task for radiologists in reading and interpreting a large number of thin-slice CT images, which results in inaccurate detection, misinterpretation and large inter-reader variability. Thus, developing computer-aided detection and/or diagnosis (CAD) schemes to assist radiologists in reading and interpreting lung CT images has been attracting extensive research interest in the past two decades (van Ginneken *et al* 2011).

As a result, many computer-aided detection (CADe) schemes of CT images for detecting suspicious lung nodules have been developed and proposed to be used as 'a second reader' to assist radiologists (Gong *et al* 2016, Liang *et al* 2016). In the clinical practice of reading lung CT images, accurately determining the likelihood of a

nodule being malignant is also a challenging task for radiologists. Thus, developing high-performance computer-aided diagnosis schemes (CADx) to distinguish between benign and malignant pulmonary nodules can also have high clinical significance to detect more early cancers, while eliminating many unnecessary biopsies or surgery. For this purpose, researchers have been working to develop the CADx scheme using different nodule segmentation and multi-feature fusion-based machine learning methods. However, developing CADx schemes still faces a number of challenges (Amir and Lehmann 2016).

Recently, a number of studies have used a publicly available LIDC/IDRI (Lung Image Database Consortium/Image Database Resource Initiative) database to develop the proposed CADx schemes trained using either conventional machine learning models (Han *et al* 2015, Dhara *et al* 2016, Jacobs *et al* 2016, Wei *et al* 2018) or deep convolutional neural network methods (Hua *et al* 2015, Shen *et al* 2017, Sun *et al* 2017, Wang *et al* 2017). However, due to the lack of histopathology-based 'ground truth', these schemes only classified nodules based on the suspiciousness scores as malignant rated by radiologists (i.e. grouping the cases rated 1–3 as 'benign' cases and 4–5 as 'malignant' cases). As a result, the performance of these CADx schemes to classify real benign and malignant nodules has not been tested. Although several other studies used the image data set collected from their cooperative hospitals (Suzuki *et al* 2005, Sun *et al* 2013), the distinction between benign and malignant nodules was also determined based on the diagnostic rating scores of radiologists. Due to the difficulty to collect a large number of biopsy and histopathology confirmed benign and malignant pulmonary nodules, several studies trained the CADx scheme with the LIDC/IDRI database and later validated the scheme performance using an in-house collected small data set with histopathology confirmed results (Way *et al* 2006, Chen *et al* 2010).

In addition, CAD performance may also heavily depend on the difficulty and diverse levels of the training and testing data set (Zheng *et al* 2010). How to build optimal training data sets in developing CADx schemes of lung nodules, that have not been previously investigated, are reported in the literature. Thus, the objective of this study is to develop a new CADx scheme for distinguishing between malignant and benign lung nodules, and also to assess whether and how the performance of the CADx scheme changes in detecting and classifying nodules associated with early and advanced stage lung cancer. For this purpose, we first retrospectively assemble a diverse data set, in which all targeted suspicious nodules are biopsied and have histopathological confirmed results. Second, we assess the possible trend between classification performance of the CADx schemes and the cancer stage. The details of our image data set, new CADx scheme, experimental procedures and data analysis results are presented in the following sections of this article.

2. Materials and methods

2.1. Image data set

We retrospectively collect study subjects of lung CT images from two sources. First, we collect CT images and histopathological test data of 89 patients who underwent lung cancer diagnosis at Shanghai Pulmonary Hospital, Shanghai, China. Among these 89 patients (34 males; 55 females), 98 nodules are detected and biopsied. Among these 89 patients, 22 are diagnosed and confirmed with stage I lung cancer, which includes 22 malignant nodules, while 67 are cancer-free. These cancer-free cases include 76 biopsy-proven benign nodules. CT imaging scans are performed using a multi-slice CT scanner (manufacturer: Siemens or Philips) with a tube voltage of 120 kVp and a tube current of 100–300 mA. The reconstruction slice thickness is 1 or 2 mm, and the pixel spacing is 0.68 or 0.77 mm depending on patient size. Each reconstructed CT section has an image matrix size of 512 $\times$ 512 pixels.

The second part of the data set is obtained from the Lung 1 data set in the NSCLC-Radiomics database (https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI) (Aerts *et al* 2014). All CT scans of the 145 malignant nodules selected from this cohort study contain the correct manual delineations made by the radiation oncologists. Among them, 86 malignant nodules are confirmed at clinical stage III (29 stage IIIa; 57 stage IIIb) and the remaining 59 malignant nodules are at clinical stage I. The pixel spacing of the CT scan is 0.98 mm and the slice thickness is 3 mm. Data are also reconstructed with a 512 $\times$ 512 matrix.

In summary, by combining the data collected from these two data sources, the total data set includes 243 nodules (76 benign and 167 malignant) detected from 234 patients (130 males and 104 females). All benign and malignant nodules are biopsy-proven. Of the 167 malignant nodules, 81 (48.5%) are determined at stage I (early stage group) and 86 (51.5%) are determined at stage III (advanced stage group). In order to develop a new CADx scheme and investigate the effect of training data sets on the CADx performance, we separate 243 nodules into three data groups namely, (1) overall data group of 243 nodules (including all benign and malignant nodules); (2) early stage data group of 157 nodules (all benign nodules and stage I malignant nodules); and (3) advanced stage data group of 162 nodules (all benign nodules and stage III malignant nodules).

2.2. CADx scheme

Figure 1 shows a block diagram of the proposed CADx scheme. We first apply and implement a hybrid 3D semi-automatic segmentation method in our CADx scheme to segment the lung nodules depicted on CT images.

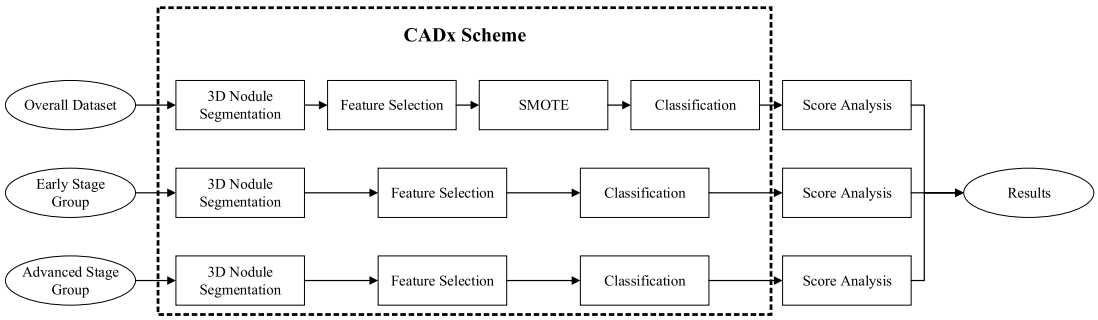

Figure 1. Flow diagram of the proposed CADx scheme.

In brief, we use the nodule center positions marked by the radiologists as the initial segmentation seeds, and then apply different methods or algorithms to three different types of nodules namely, the solitary nodule, juxta-pleural nodule and vascularized nodule, respectively. As an example, figure 2 shows the regions of interest (ROIs) for these three types of nodules. Unlike CADe schemes that search for and detect suspicious lung nodules and their locations, a CADx scheme is applied to the nodules that have been detected by the radiologists. In order to more accurately classify the detected nodules into benign and malignant classes, accurate segmentation of nodules is important in order to compute accurate and reliable image features. Due to heterogeneity or variation of lung nodules depicted on the CT images, three methods are applied to segment the three different types of nodules in our CADx scheme.

(1) ***Solitary nodule***: First, a 3D region growing algorithm with the criterion shown in equation (1) is applied to segment the nodule from the background (Gong *et al* 2014). A threshold is empirically selected with a value of 100. Second, a binary dilation operation with a ball-structuring element is applied on the 3D mask. Last, a flood-fill operation is applied to fill the holes in the mask. Figures 2(a-1)–(a-4) show an example of the process to segment a solitary nodule.

$$|I(x, y, z) - I_{\text{seed}}| \leqslant \text{Threshold} \tag{1}$$

where $I(x, y, z)$ is the CT value of the position (x, y, z). I_{seed} is defined as the CT value of the initial seed point.

(2) ***Juxta-pleural nodule***: First, the same 3D region growing method as applied for solitary nodule segmentation is used to conduct initial lung nodule segmentation. Second, a chain code algorithm proposed in the previous study (Gong *et al* 2016) is used to repair the initial mask. Last, a binary dilation operation and a flood-fill operation are applied to remove the holes in the 3D mask. An example of a segmentation result of the juxta-pleural nodule is shown in figures 2(b-3) and (b-4).

(3) ***Vascularized nodule***: First, a multi-scale 3D dot filtering algorithm is used to enhance the nodule region, which has a ball-like structure (Gong *et al* 2016). Second, the 3D region growing method is applied on the enhanced images to extract the nodule mask. Third, the vessel structures on the initial mask are removed based on the geodesic distance histogram. Last, the morphological operators are utilized to fill the holes and repair the boundaries. Figures 2(c-1)–(c-4) show an example of segmenting a vascularized nodule.

Next, our CADx scheme initially computes a set of 66 3D image features to represent the CT value distribution (heterogeneity), shape and texture of the segmented nodule (as shown in table 1). Thus, in this study, three initial image feature pools are created. The first is an overall feature pool, which includes all 66 features computed from all 243 nodules (76 benign and 167 malignant nodules). The second is named an early stage cancer feature pool, which includes features computed from 157 nodules (76 benign and 81 stage I malignant nodules), and the third is an advanced stage cancer feature pool containing features computed from 162 nodules (76 benign and 86 stage III malignant nodules).

Then, we build and test several multi-feature fusion-based machine learning classifiers, which are implemented into our CADx scheme to generate a classification score ranging from 0–1 to represent the likelihood of a queried nodule being malignant. Although many machine learning methods or models can be used to build classifiers for this purpose, in this study we select and use three relatively simple classifiers, which have relatively high generalization capability when they are trained using small-feature data sets. These are a support vector machine classifier (SVM), Naïve Bayes classifier (NBC) and linear discriminant analysis classifier (LDA). In this step, each

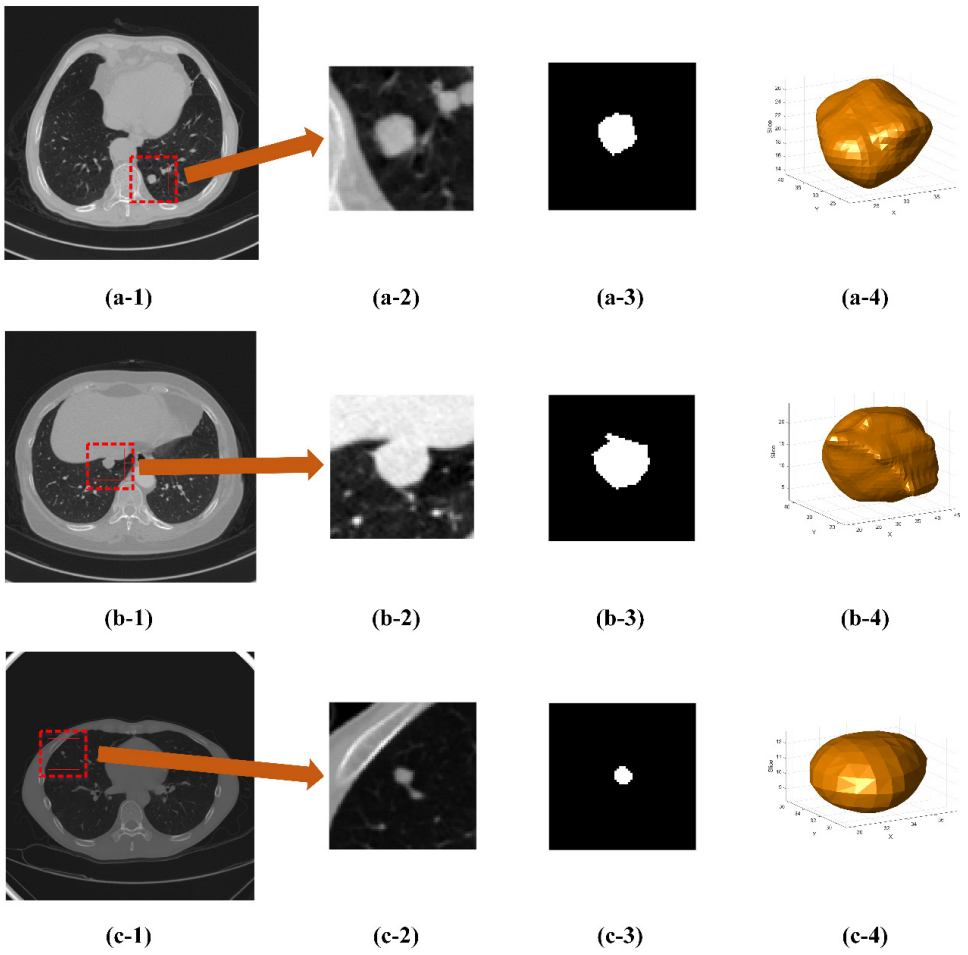

Figure 2. Example of nodule segmentation results. (a-1)–(c-1) Case with solitary nodule, case with juxta-pleural nodule and case with vascularized nodule. (a-2) and (c-2) ROIs of the corresponding nodules. (a-3)–(c-3) Masks of the corresponding nodules. (a-4)–(c-4) 3D volume of the corresponding nodules.

Table 1. Description of quantitative 66 image features.

Number	Feature group	Feature	Description
F1–F44	Texture feature	GLCM[a]	Contrast, correlation, energy, homogeneity 1, homogeneity 2, entropy, autocorrelation, cluster prominence, cluster shade, cluster tendency, dissimilarity, inverse variance, difference entropy, maximum probability, sum average, sum entropy, variance
		GLRLM[b]	Short run emphasis, long run emphasis, gray level non-uniformity, run length non-uniformity, run percentage, low gray level run emphasis, high gray level run emphasis, short run low gray level emphasis, short run high gray level emphasis, long run low gray level emphasis, long run high gray level emphasis
		GLSZM[c]	Small area emphasis, large area emphasis, intensity variability, size zone variability, zone percentage, low intensity emphasis, high intensity emphasis, low intensity small area emphasis, high intensity large area emphasis, high intensity small area emphasis, low intensity large area emphasis
		NGTDM[d]	Coarseness, contrast, busyness, complexity, texture strength
F45–F58	Histogram feature		Minimum, maximum, mean, median, standard deviation, skewness, kurtosis, entropy, uniformity, energy, range, variance, root mean square, mean absolute deviation
F59–F66	Shape feature		Surface area, volume, surface to volume ratio, compactness 1, compactness 2, spherical disproportion, sphericity, maximum 3D diameter

[a] GLCM: gray-level co-occurrence matrix.
[b] GLRLM: gray-level run length matrix.
[c] GLSZM: gray level size zone matrix.
[d] NGTDM: neighborhood gray-tone difference matrix.

classifier is trained using one of the three feature pools with (1) overall 243 nodules, (2) 157 nodules (including 81 stage I malignant nodules), and (3) 162 nodules (including 86 stage III malignant nodules). In order to eliminate or minimize bias in both feature selection and case partition during the classifier training and testing, we embed both feature selection and cross-validation methods into the process of building and validating the classifiers. The detailed steps are described as follows.

(1) Since among the different cross-validation methods, a leave-one-case-out (LOCO) method has the advantages of providing the maximum training samples and avoiding the possible case partition bias (Li *et al* 2006), each classifier is trained and tested using the LOCO cross-validation method. In each training/testing iteration cycle, one case is removed from the training data set and the remaining cases are used to train the classifier. The trained classifier is then applied to one case that is excluded from the training process to generate a classification score for this test case. This iteration process is repeated until all cases in the data set are independently tested and have received classification scores.
(2) Since each initial feature pool includes 66 image features, which can contain a large fraction of low-performance and/or redundant features, in order to build a high-performance and robust classifier, applying an optimal feature selection process to reduce the dimensionality of feature space is important. In order to avoid bias, the test case also needs to be excluded from the feature selection process. Thus, the feature selection process is embedded inside the LOCO cross-validation iteration cycle. Specifically, in each iteration cycle, we apply a Relief-F feature selection method (Awai *et al* 2004, Hawkins *et al* 2014) to reduce the dimensionality of the feature space by selecting the effective features and removing the redundant ones from the initial image feature pool. The top ten features in the performance sorting list determined by the feature selection algorithm are selected and applied to build the classifier.
(3) Due to the unbalanced number of nodules in two classes of the overall data set involving 243 nodules in which the ratio between malignant and benign nodules is 2.20 (167/76), we use a synthetic minority over-sampling technique (SMOTE) (Chawla *et al* 2002) to generate a balanced data set by creating synthetic instances of 'benign' nodules with a 100% oversampling rate (Sun *et al* 2013). As a result, we expand the number of nodules from 243 to 319 (167 malignant versus 152 'benign' nodules), which increases the balance ratio between nodules in two classes. The SMOTE method is also embedded inside a LOCO cross-validation-based iteration cycle. The added synthetic 'benign' nodules are only used in the training process. They are not involved in the testing process. Thus, upon the completion of a LOCO cross-validation-based training and testing process, we only obtain 243 classification scores computed from the original 243 nodules in our data set. A similar SMOTE method has been applied and reported in previous studies to develop quantitative imaging feature classification schemes for predicting or assessing cancer prognosis (Aghaei *et al* 2016, Yan *et al* 2016).

2.3. Performance evaluation

After completing the training and cross-validation of each classifier using one of the three feature pools, we evaluate the performance of each classifier. First, we apply a receiver operating characteristic (ROC)-based data analysis method to compute the area under a ROC curve (AUC) and use it as a performance assessment index to evaluate and compare performance of the classifiers. Second, since in the clinical practice, radiologists need to make a binary decision (i.e. whether to recommend a biopsy or not), we apply several operation thresholds to stratify the cases into two groups of malignant and benign cases. We count the number of true-positive (TP), false-positive (FP), true-negative (TN) and false-negative (FN) nodules in the two stratified groups. We then compute and compare three additional evaluation indices: overall accuracy ($ACC = \frac{TP+TN}{TP+FP+TN+FN}$), sensitivity ($SE = \frac{TP}{TP+FP}$), and specificity ($SP = \frac{TN}{TN+FN}$) (Awai *et al* 2004, Amir and Lehmann 2016). Using these evaluation indices, we compare and analyze the trends of the performances of the classifiers trained with three different feature data sets.

All of the above CADx computation processes and data analysis processes are implemented on the MATLAB R2016b by using a computer with Intel Core i5-6600 CPU 3.3 GHz $\times$ 2, 8 GB RAM. For building all three machine learning classifiers (SVM, NBC, and LDA), we use the default parameters set by the MATLAB package. Thus, the training and testing process is straightforward and can be easily applied and/or validated in future studies.

3. Results

Since the feature selection is embedded inside each LOCO cross-validation-based training and testing iteration cycle to train and optimize each classifier, the top ten features selected by the Relief-F feature selection method

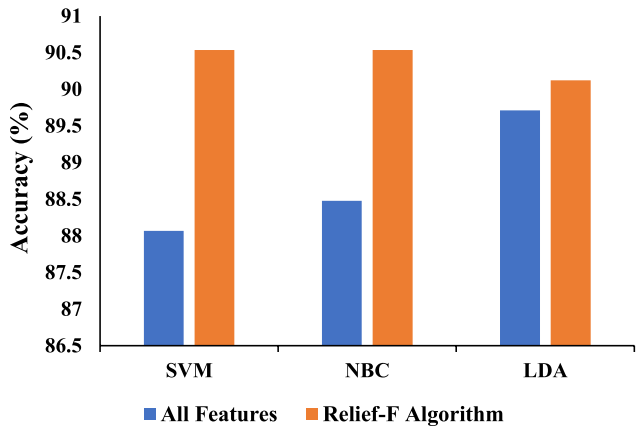

Figure 3. Comparison of accuracy generated by using all features and Relief-F algorithm selected features with three different classifiers.

Table 2. Frequently selected features in each feature pool, which were selected at least 50% times in the LOCO cross-validation training/testing process cycles.

Classifier	Feature pool 1[a]	Feature pool 2[b]	Feature pool 3[c]
SVM	F7, F15, F22, F23, F24, F45, F61, F62, F64, F66	F15, F22, F23, F24, F28, F45, F61, F62, F64, F66	F7, F10, F15, F22, F23, F24, F35, F61, F62, F66
NBC	F15, F22, F23, F24, F28, F45, F61, F62, F64, F66	F15, F22, F23, F24, F28, F45, F61, F62, F64, F66	F7, F10, F15, F22, F23, F24, F35, F61, F62, F66
LDA	F15, F22, F23, F24, F28, F45, F61, F62, F64, F66	F15, F22, F23, F24, F28, F45, F61, F62, F64, F66	F7, F10, F15, F22, F23, F24, F35, F61, F62, F66

[a] Feature Pool 1: overall feature pool.
[b] Feature Pool 2: early stage cancer feature pool.
[c] Feature Pool 3: advanced stage cancer feature pool.

may vary in different LOCO iteration cycles. Figure 3 shows and compares the accuracy of three classifiers using either all 66 features or ten optimal features selected by the Relief-F algorithm from the overall data group. For three classifiers, the accuracy generated using the optimal features selected by the Relief-F algorithm are higher than those generated using all features. The results show that using the Relief-F feature selection method is effective to improve the classifier's performance.

Table 2 illustrates nine sets of image features that are frequently selected by three classifiers trained using each of three feature pools. Each feature listed in the table is selected at least 50% times in the LOCO training/testing cycles. From this table, we can make the following observations. (1) For each of the three classifiers (SVM, NBC, and LDA) trained using the same feature pool, the most frequently selected features are quite consistent, which indicates that identifying and computing effective image features plays a dominant and important rule in the classification of lung nodules. (2) For using three different training data sets with 243, 157 and 162 nodules, respectively, there is a bigger difference in feature selection between the overall feature pool and the early stage cancer feature pool than between the overall feature pool and the advanced stage cancer feature pool, which indicates that the nodules associated with the early and advanced stages of lung cancer have increased differences in the image feature characteristics. (3) The following six image features are among the most frequently selected features of all nine classifiers trained using three feature pools or data sets. These are F15 (GLCM-sum average), F22 (GLRLM-run percentage), F23 (GLRLM-low gray level run emphasis), F24 (GLRLM-high gray level run emphasis), F62 (Shape feature-compactness1), and F66 (Shape feature-maximum 3D diameter). Selecting these features clearly indicates the importance of using tumor morphologic and texture features in developing CADx schemes of lung nodule classification. (4) Overall, table 2 indicates 13 frequently selected features by one or more classifiers. By conducting the ROC analysis on each of these features, the results indicate that all features yield an AUC value greater than 0.7 (as shown in table 3), which indicates that each of these features has a relatively higher discriminatory power to help distinguish between benign and malignant nodules individually.

Figure 4 shows and compares three sets of ROC curves corresponding to the classification results generated using SVM, NBC and LDA classifiers, respectively. Each set also includes and compares three ROC curves trained and tested using three different data groups namely, the overall data group, early stage data group, and advanced stage data group. The detailed AUC values of these three sets of classifiers trained using these three different data sets or feature pools are summarized in table 4. The results from these three sets of ROC curves and AUC values

Table 3. Summary of the AUC values using individual frequently selected image features.

Feature	Description	AUC	95% CI
F7	GLCM-autocorrelation	0.91 ± 0.02	[0.87, 0.95]
F10	GLCM-cluster tendency	0.91 ± 0.02	[0.87, 0.94]
F15	GLCM-sum average	0.91 ± 0.02	[0.87, 0.95]
F22	GLRLM-run percentage	0.89 ± 0.03	[0.83, 0.93]
F23	GLRLM-low gray level run emphasis	0.93 ± 0.02	[0.89, 0.96]
F24	GLRLM-high gray level run emphasis	0.90 ± 0.02	[0.85, 0.94]
F28	GLRLM-long run high gray level emphasis	0.83 ± 0.03	[0.76, 0.88]
F35	GLSZM-high intensity emphasis	0.91 ± 0.02	[0.86, 0.95]
F45	Histogram feature-minimum	0.81 ± 0.03	[0.75, 0.86]
F61	Shape feature-surface to volume ratio	0.94 ± 0.02	[0.90, 0.96]
F62	Shape feature-compactness1	0.94 ± 0.01	[0.91, 0.97]
F64	Shape feature-spherical disproportion	0.73 ± 0.03	[0.67, 0.79]
F66	Shape feature-maximum 3D diameter	0.93 ± 0.02	[0.89, 0.96]

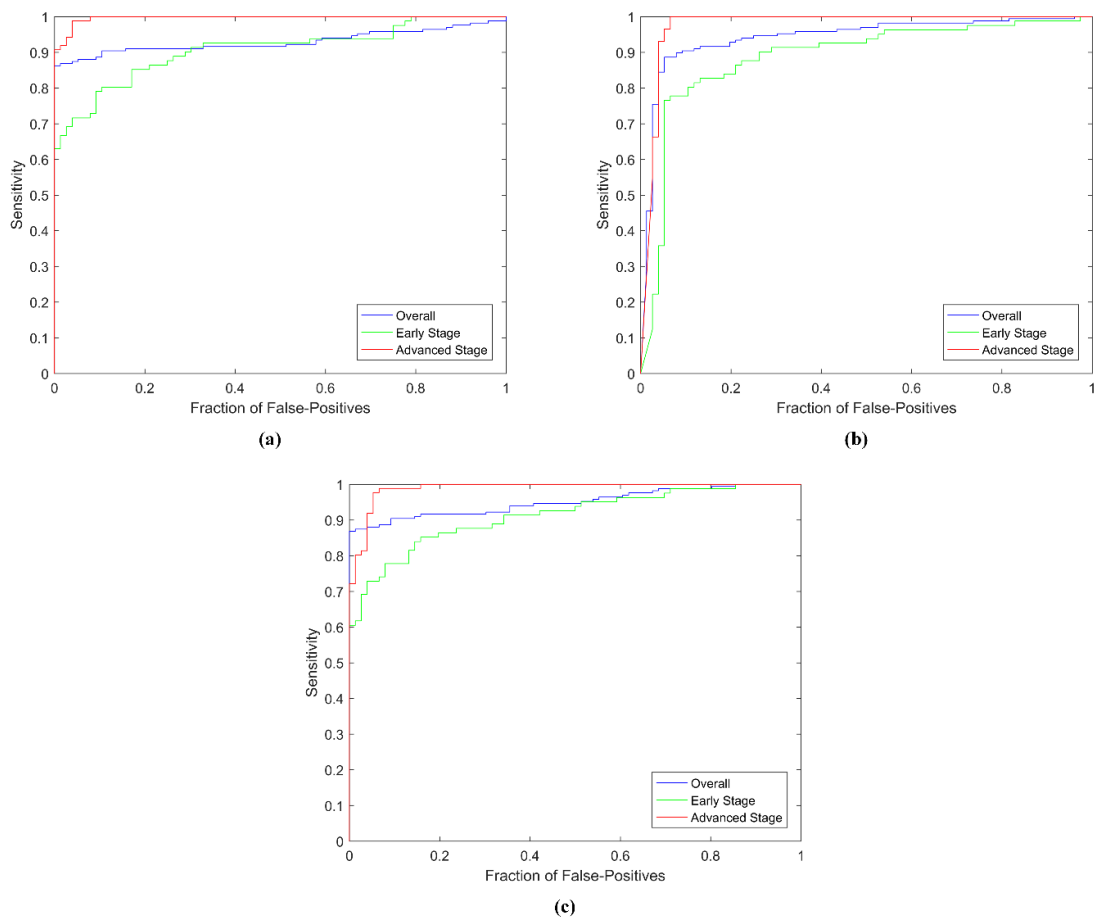

Figure 4. Comparison of ROC curves generated by using different classifiers with three data groups. (a)–(c) Correspond to the SVM, NBC, LDA classifier, respectively.

show a consistent trend. For all three classifiers, classification performance levels increase based on the following order namely, the classifier has the lowest performance when it is trained using the second data set with early stage cancer cases, followed by that trained using the first data set involving all cases. The classifier trained using the third data set with the advanced stage cancer cases yields the highest performance level. In addition, when three classifiers are trained using the first data set of all 243 nodules, they all yield higher performance in classifying the nodules associated with advanced stage cancer than in classifying the nodules associated with early stage cancer. For example, AUC = 0.99 ± 0.01 versus AUC = 0.86 ± 0.03 for classifying advanced and early stage cancer using the same SVM classifier, respectively (as shown in table 4).

Table 4. AUC values of the three classifiers built with different data groups.

Data set	Subset	SVM	NBC	LDA
Overall data group	Overall	0.93 ± 0.02	0.93 ± 0.01	0.95 ± 0.01
	Group I[a]	0.86 ± 0.03	0.89 ± 0.03	0.89 ± 0.03
	Group II[b]	0.99 ± 0.01	0.98 ± 0.02	0.99 ± 0.01
Early stage data group	—	0.91 ± 0.02	0.89 ± 0.02	0.91 ± 0.02
Advanced stage data group	—	0.99 ± 0.01	0.98 ± 0.02	0.99 ± 0.01

[a] Group I: benign and early stage malignant nodules in overall data group.
[b] Group II: benign and advanced stage malignant nodules in overall data group.

Table 5. The SE and ACC scores of the CAD scheme built with three data groups under SP = 80%, 90%.

		SP = 80.26% (80%)			SP = 89.47% (90%)		
		Group 1[a]	Group 2[b]	Group 3[c]	Group 1[a]	Group 2[b]	Group 3[c]
SVM	SE (%)	91.02	85.19	100	90.42	80.25	100
	ACC (%)	87.65	82.80	90.74	90.12	84.71	95.06
NBC	SE (%)	92.81	83.95	100	90.42	80.25	100
	ACC (%)	88.89	82.17	90.74	90.12	84.71	95.06
LDA	SE (%)	91.62	86.42	100	90.42	77.78	98.84
	ACC (%)	88.07	83.44	90.74	90.12	83.44	94.44

[a] Group 1: overall data group.
[b] Group 2: early stage data group.
[c] Group 3: advanced stage data group.

Table 5 summarizes and compares the overall classification accuracy and sensitivity scores generated by all nine classifiers under two specificity values of 80% and 90%. The results show that under both specificity levels, each classifier trained and tested using the third data set including the nodules associated with the advanced lung cancer yield the highest overall accuracy and classification sensitivity, while the classifier trained and tested using the second data set involving the nodules associated with early stage lung cancer yield the lowest performance. Table 6 summarizes the correlation coefficients and p values of the related AUC values for testing the statistical significance of the differences between these three classifiers. For the same training data group, the differences between the three classifiers are not statistically significant ($p > 0.05$).

4. Discussion

In this study, we developed and tested a new CADx scheme of lung CT images to distinguish between malignant and benign pulmonary nodules. The study has a number of unique characteristics. First, we assembled a diverse lung CT data set, in which all nodules have the confirmed biopsy and histopathology test results. Thus, our CADx scheme enables us to distinguish between malignant and benign lung nodules, not only to predict the risk or likelihood of the nodule being malignant as reported in many of the previous studies (i.e. Sun *et al* (2017) and Wang *et al* (2017)). In order to increase the robustness of the CADx scheme developed by using a limited data set, we only used several simple image processing and feature analysis methods or processes. The study results demonstrated the feasibility of extracting image features or a feature set and building the multi-feature fusion-based machine learning classifiers with high discriminatory power to distinguish between malignant and benign nodules. Since many previous studies have tested and demonstrated the feasibility of applying a CADx scheme of lung CT images as 'a second reader' to assist radiologists detecting suspicious lung nodules and improving the diagnosis performance (Awai *et al* 2004, 2006, Li *et al* 2004a, 2004b, van Ginneken *et al* 2011, Amir and Lehmann 2016, Liang *et al* 2016), if the new scheme is robust in the future tests using larger and more diverse data sets, the study will have a high clinical impact in assisting radiologists in their decision making in the detection and diagnosis of lung cancer in future clinical practice.

Second, it is well known that in order to build an optimal and robust machine learning classifier, it requires a large and diverse image data set, as well as a small set of image features to yield a higher or acceptable ratio between the training samples and number of image features. However, the size of the cancer imaging data set is typically limited, including in this study, while a large number of image features can often be initially computed. For example, we initially compute 66 image features that cover or represent density heterogeneity, shape and texture features of the nodules in this study. Thus, we add and integrate three methods or steps namely, a Relief-F feature selection algorithm, a SMOTE algorithm, and a LOCO cross-validation method into the process of

Table 6. AUC correlation coefficients and *p*-values for testing the statistical significance of the differences between the three classifiers by using same data group.

	Data group 1[a]	Data group 2[b]	Data group 3[c]
SVM versus NBC	0.45 (*p*-value: 0.65)	−0.77 (*p*-value: 0.44)	−1.84 (*p*-value: 0.07)
SVM versus LDA	1.35 (*p*-value: 0.18)	0.07 (*p*-value: 0.95)	−1.63 (*p*-value: 0.10)
NBC versus LDA	0.16 (*p*-value: 0.87)	0.94 (*p*-value: 0.35)	1.44 (*p*-value: 0.15)

[a] Data Group 1: overall data group.
[b] Data Group 2: early stage data group.
[c] Data Group 3: advanced stage data group.

training and testing the classifiers. Specifically, we apply the Relief-F feature selection algorithm to reduce the dimensionality of the feature space. We apply the SMOTE algorithm to create a more balanced training data set between the benign and malignant cases. Then, both the Relief-F and SMOTE algorithms are embedded into the LOCO cross-validation process, which produces the largest number of training samples, and avoids or minimizes the potential bias in both feature selection and case partition. The study results demonstrate that using this process with optimally selected small features improves the performance of the classifier (as shown in figure 3). In addition, after analyzing the most frequently selected features in the LOCO iteration cycles (i.e. >50%), we find that these features have higher discriminatory power (i.e. AUC value as shown in table 3) to distinguish between benign and malignant nodules. The results indicate that applying and embedding a feature selection and possible SMOTE algorithms within a LOCO cross-validation process is effective in building high-performance and robust machine learning classifiers using the relatively small image data sets.

Third, since CAD performance depends on the difficulty and diversity of the training and testing data set (Zheng *et al* 2010), we divided the malignant nodules or cases into two groups of early and advanced stage cancer, and assembled three data sets to investigate performance changes of the CADx scheme in detecting or classifying lung nodules associated with different stage lung cancers. A number of classifiers were separately trained and tested using these three data sets. The data analysis results showed a consistent trend. The more advanced stage cancer cases that are included in the training and testing data set, the higher the classification performance (i.e. AUC value) that can be achieved. This indicates that as the cancer grows or migrates to a more advanced stage, the image feature computed from the nodule also shows increased difference compared to the benign nodules. Thus, accurate classification between benign nodules and the malignant nodules associated with early stage lung cancer is more difficult or challenging. Since the previous studies do not report the distribution of case difficulty levels and/or cancer stages, whether the CADx schemes are optimally trained and whether the reported performance levels are comparable are unknown. The results of this study suggest that in order to increase the performance of applying CAD schemes to accurately detect more subtle and/or early stage lung cancer with a lower FP rate, the researchers need to assemble an optimal training data set dominated with more nodules associated with early stage lung cancer.

Fourth, in order to preliminarily test the robustness of using image features computed and/or selected in this study to develop a CADx scheme without another independent test data set, we build and test three different types of machine learning classifiers (including SVM, NBC and LDA). These classifiers use different statistical or machine learning concepts or approaches, but they have been frequently used in the imaging-based CADx schemes with relatively high generalization capability when they are trained using relatively small image data sets. The data analysis results showed a consistent trend of using any of these three classifiers to detect or classify early and advanced stage lung cancer cases. The result provides new evidence to support the robustness of our approach to develop a new CADx scheme in this study.

Despite the promising results, this study also has several limitations. First, our data set is small, with 243 cases and unbalanced with 76 benign nodules. Thus, the diversity of this data set may be limited and cannot sufficiently represent the general lung cancer population in clinical practice. The reproducibility and generalization of the reported results need to be validated with large diverse data sets in future studies. Second, although in order to improve nodule segmentation accuracy, an adaptive nodule segmentation method with three algorithms is applied to segment three types of solitary, juxta-pleural and vascularized nodules, respectively, how to evaluate nodule segmentation accuracy remains a challenge and is subjective due to the lack of 'ground-truth' for the actual boundary contour of the nodules. Thus, developing more accurate and reliable nodule segmentation and evaluation methods is still one of the research focuses to be explored. Third, we only computed 66 image features that focus on nodule density heterogeneity and texture variation. Some of the other potentially useful image features, such as nodule speculation and lobulation have not yet been explored. Fourth, although the deep CNN transfer learning methods have been developed and tested to develop CADx schemes to predict the malignant risk of the lung nodules (i.e. Sun *et al* (2017) and Wang *et al* (2017)), due to the limited data set of this study, we only use a conventional machine learning approach or classifiers. Whether using the deep learning methods can

produce significantly higher and more robust classification accuracy than well-trained conventional machine learning classifiers needs to be further investigated in future studies with an increase of the training data set size. Last, this is only a primary technology development study to evaluate the stand-alone performance of a CADx scheme. Whether and how the CADx performance translates to human (radiologist) performance is unclear. Thus, clinical application of the CADx scheme needs to be investigated in future studies before the CADx scheme and technology can be accepted in clinical practice.

5. Conclusion

In this study, we developed a new CADx scheme to distinguish between malignant and benign pulmonary nodules, and investigated the effects of different image training data sets on CADx scheme performance. The experimental and data analysis results demonstrated a consistently positive trend between the CADx performance and cancer progression stage. Thus, in order to increase CADx performance in detecting subtle or early stage lung cancer cases, the training data set should include more diversely subtle and/or early stage lung cancer cases. In addition, embedding an effective feature selection method within the LOCO cross-validation method to train and build a machine learning classifier is also an optimal approach to develop a CADx scheme using a limited image data set.

Acknowledgments

This work was partially funded by the National Natural Science Foundation of China under Grant No. 60972122 and the Natural Science Foundation of Shanghai under Grant No. 14ZR1427900. The authors also acknowledge the support received from the Peggy and Charles Stephenson Cancer Center, University of Oklahoma, USA and Shanghai Pulmonary Hospital, Shanghai, China.

References

Aerts H J, Velazquez E R, Leijenaar R T, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B and Rietveld D 2014 Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach *Nat. Commun.* **5** 4006

Aghaei F, Tan M, Hollingsworth A B and Zheng B 2016 Applying a new quantitative global breast MRI feature analysis scheme to assess tumor response to chemotherapy *J. Magn. Reson. Imaging* **44** 1099–106

Amir G J and Lehmann H P 2016 After detection: the improved accuracy of lung cancer assessment using radiologic computer-aided diagnosis *Acad. Radiol.* **23** 186–91

Awai K, Murao K, Ozawa A, Komi M, Hayakawa H, Hori S and Nishimura Y 2004 Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance *Radiology* **230** 347–52

Awai K, Murao K, Ozawa A, Nakayama Y, Nakaura T, Liu D, Kawanaka K, Funama Y, Morishita S and Yamashita Y 2006 Pulmonary nodules: estimation of malignancy at thin-section helical CT—effect of computer-aided diagnosis on performance of radiologists *Radiology* **239** 276–84

Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique *J. Artif. Intell. Res.* **16** 321–57

Chen H, Xu Y, Ma Y and Ma B 2010 Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on CT images: clinical evaluation *Acad. Radiol.* **17** 595–602

Dhara A K, Mukhopadhyay S, Dutta A, Garg M and Khandelwal N 2016 A combination of shape and texture features for classification of pulmonary nodules in lung CT images *J. Digit. Imaging* **29** 466–75

Gong J, Gao T, Bu R-R, Wang X-F and Nie S-D 2014 An automatic pulmonary nodules detection method using 3D adaptive template matching *Life System Modeling and Simulation* (Berlin: Springer) pp 39–49

Gong J, Liu J-Y, Wang L-J, Zheng B and Nie S-D 2016 Computer-aided detection of pulmonary nodules using dynamic self-adaptive template matching and a FLDA classifier *Phys. Med.* **32** 1502–9

Han F, Wang H, Zhang G, Han H, Song B, Li L, Moore W, Lu H, Zhao H and Liang Z 2015 Texture feature analysis for computer-aided diagnosis on pulmonary nodules *J. Digit. Imaging* **28** 99–115

Hawkins S H, Korecki J N, Balagurunathan Y, Gu Y, Kumar V, Basu S, Hall L O, Goldgof D B, Gatenby R A and Gillies R J 2014 Predicting outcomes of nonsmall cell lung cancer using CT image features *IEEE Access* **2** 1418–26

Hua K-L, Hsu C-H, Hidayati S C, Cheng W-H and Chen Y-J 2015 Computer-aided classification of lung nodules on computed tomography images via deep learning technique *Oncotargets Ther.* **8** 2015–22

Jacobs C, Rikxoort E M, Murphy K, Prokop M, Schaefer-Prokop C M and Ginneken B 2016 Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database *Eur. Radiol.* **26** 2139–47

Li F, Aoyama M, Shiraishi J, Abe H, Li Q, Suzuki K, Engelmann R, Sone S, MacMahon H and Doi K 2004a Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy *Am. J. Roentgenol.* **183** 1209–15

Li F, Sone S, Abe H, MacMahon H and Doi K 2004b Malignant versus benign nodules at CT screening for lung cancer: comparison of thin-section CT findings *Radiology* **233** 793–8

Li Q and Doi K 2006 Reduction of bias and variance for evaluation of computer-aided diagnosis schemes *Med. Phys.* **33** 868–75

Liang M, Tang W, Xu D M, Jirapatnakul A C, Reeves A P, Henschke C I and Yankelevitz D 2016 Low-dose CT screening for lung cancer: computer-aided detection of missed lung cancers *Radiology* **81** 279–88

Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, Zang Y and Tian J 2017 Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification *Pattern Recognit.* **61** 663–73

Siegel R L, Miller K D and Jemal A 2016 Cancer statistics, 2016 *CA Cancer J. Clin.* **66** 7–30

Stewart B and Wild C P 2015 World cancer report 2014 *Lyon: International Agency for Research on Cancer, WHO; 2014.* (http://publications.iarc.fr/Non-Series-Publications/World-Cancer-Reports/World-Cancer-Report-2014)

Sun T, Zhang R, Wang J, Li X and Guo X 2013 Computer-aided diagnosis for early-stage lung cancer based on longitudinal and balanced data *PLoS One* **8** e63559

Sun W, Zheng B and Qian W 2017 Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis *Comput. Biol. Med.* **89** 530–9

Suzuki K, Li F, Sone S and Doi K 2005 Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network *IEEE Trans. Med. Imaging* **24** 1138–50

Torre L A, Siegel R L and Jemal A 2016 *Lung Cancer and Personalized Medicine* (Berlin: Springer) pp 1–19

van Ginneken B, Schaefer-Prokop C M and Prokop M 2011 Computer-aided diagnosis: how to move from the laboratory to the clinic *Radiology* **261** 719–32

Wang H *et al* 2017 A hybrid CNN feature model for pulmonary nodule malignancy risk differentiation *J. X-Ray Sci. Technol.* pp 1–17

Way T W, Hadjiiski L M, Sahiner B, Chan H-P, Cascade P N, Kazerooni E A, Bogot N and Zhou C 2006 Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours *Med. Phys.* **33** 2323–37

Wei G, Ma H, Qian W, Han F, Jiang H, Qi S and Qiu M 2018 Lung nodule classification using local kernel regression models with out-of-sample extension *Biomed. Signal Process. Control* **40** 1–9

Yan S, Qian W, Guan Y and Zheng B 2016 Improving lung cancer prognosis assessment by incorporating synthetic minority oversampling technique and score fusion method *Med. Phys.* **43** 2694–703

Zheng B, Wang X, Lederman D, Tan J and Gur D 2010 Computer-aided detection: the effect of training databases on detection of subtle breast masses *Acad. Radiol.* **17** 1401–8