J Digit Imaging (2017) 30:761–771
DOI 10.1007/s10278-017-9957-6

An Ensemble Method for Classifying Regional Disease Patterns of Diffuse Interstitial Lung Disease Using HRCT Images from Different Vendors

Sanghoon Jun[1] · **Namkug Kim**[1,2] · **Joon Beom Seo**[2] · **Young Kyung Lee**[3] · **David A. Lynch**[4]

Published online: 21 February 2017

Abstract We propose the use of ensemble classifiers to overcome inter-scanner variations in the differentiation of regional disease patterns in high-resolution computed tomography (HRCT) images of diffuse interstitial lung disease patients obtained from different scanners. A total of 600 rectangular 20 × 20-pixel regions of interest (ROIs) on HRCT images obtained from two different scanners (GE and Siemens) and the whole lung area of 92 HRCT images were classified as one of six regional pulmonary disease patterns by two expert radiologists. Textual and shape features were extracted from each ROI and the whole lung parenchyma. For automatic classification, individual and ensemble classifiers were trained and tested with the ROI dataset. We designed the following three experimental sets: an intra-scanner study in which the training and test sets were from the same scanner, an integrated scanner study in which the data from the two scanners were merged, and an inter-scanner study in which the training and test sets were acquired from different scanners. In the ROI-based classification, the ensemble classifiers showed better ($p < 0.001$) accuracy (89.73%, SD = 0.43) than the individual classifiers (88.38%, SD = 0.31) in the integrated scanner test. The ensemble classifiers also showed partial improvements in the intra- and inter-scanner tests. In the whole lung classification experiment, the quantification accuracies of the ensemble classifiers with integrated training (49.57%) were higher ($p < 0.001$) than the individual classifiers (48.19%). Furthermore, the ensemble classifiers also showed better performance in both the intra- and inter-scanner experiments. We concluded that the ensemble classifiers provide better performance when using integrated scanner images.

Keywords Interstitial lung disease (ILD) · Ensemble learning · Support vector machine (SVM) · Inter-scanner variation · Multi-center trial

✉ Namkug Kim
namkugkim@gmail.com

[1] Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-Ro 43-Gil, Songpa-Gu, Seoul, South Korea

[2] Department of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-Ro 43-Gil, Songpa-Gu, Seoul, South Korea

[3] Department of Laboratory Medicine, Hallym University College of Medicine, Anyang, South Korea

[4] Department of Radiology, National Jewish Medical and Research Center, Denver, CO, USA

Introduction

Diffused interstitial lung disease (DILD) is a group of disease showing disorders in the interstitium, which is a collection of tissues within the lung including the alveolar epithelium, pulmonary capillary endothelium, basement membrane, and perivascular and perilymphatic tissues. DILD shows complex disorders in the lung whose cause should be removed with proper therapy due to its critical effect to respiratory failure [1]. Studies on the DILD have been generally driven by the pathologic reclassification and known to be associated with different clinical outcomes. Five major categories of the DILD had been defined, and different approaches to therapy were studied for each category. Recently, the importance of revealing its correlation to high-resolution computed tomography (HRCT) has been increased as the imaging technology advances.

HRCT is a popular diagnostic tool for detecting and characterizing numerous disorders of the lung parenchyma and

airways [2, 3]. To automatically classify lung diseases using HRCT images, texture features as well as density histograms have been used [4–6]. Furthermore, to avoid variations caused by patient breath-hold variations and to detect early-stage diseases, shape features were added for accurate classification [7, 8]. The usefulness of automatic classification systems using both texture and shape features such as run length, co-occurrence matrix, cluster analysis, and top-hat transforms has been verified [9–12]. Furthermore, studies for classification systems for DILD have been discovering new features and their combining methods for more accurate classification. Table 1 shows the comparisons of representative studies on DILD classification using HRCT images [13–17]. As a multidimensional CT (MDCT) coming into wide use, various studies have proposed the automatic classification and quantification system for DILD using three-dimensional images and volume of interest (VOIs) [18–20]. Since the HRCT image has been used for long periods in corresponding studies including clinical trials and pathology and used to be a groundwork for MDCT imaging, studies on characterizing HRCT images still have an important role in DILD studies. In this paper, we focus on implementing an accurate and robust classification and quantification of HRCT images, which has potentials to be extended to further studies.

General approaches to lung classification use HRCT images from a single CT machine with well-controlled variations and parameters. However, to consider both lung changes and disease progress or treatment, the analysis and comparison of large numbers of patient images from large multicenter cohorts are required. In this case, the use of images from various kinds of scanners with distinct settings and mechanical differences is unavoidable. In [21], the effect of radiation dose and scanner type on lung volume, mass, mean density, and the extent of emphysema was investigated using paired statistical testing. In [22], authors observed that the different reconstruction parameters and other variations such as the kernel can significantly influence the feature extraction results of the images and, thus, investigated several classifiers to reduce the inter-scanner variations in the DILD classification [23].

In the present study, we focused on the inter-scanner variability in the HRCT image classification of regional lung disease patterns of DILD. To overcome the difficulties in the use of an integrated dataset obtained from different scanners, we apply the ensemble classification method which is known to be robust in data having high variability. Firstly, we collect typical regional disease patterns of DILD and implement descriptors extracting textural and shape features from HRCT images. We apply various single and ensemble classification method for classifying regional disease patterns. The classification methods are evaluated and compared in terms of classification accuracy using data from each scanner, inter-scanner, and integrated data scanner. While evaluating the classification methods, we apply them in two different purposes, which are ROI-based classification which classifies square and small ROI images and the whole lung quantification which accesses the whole lung parenchyma and classifies areas locally. In following sections, we describe the details of materials and methods and discuss with the results.

Materials and Methods

The overall scheme of the proposed system and its evaluation is illustrated in Fig. 1. First, HRCT images of DILD were acquired and assessed. To generate a training dataset of the classifiers, radiologists selected representative ROIs for each regional disease pattern and a normal lung. The proposed system extracted a textural and shape feature set and selected the near-optimal feature set by using a forward feature selection method. The quantification system used the classifiers trained by the feature dataset. In this experiment, we evaluated the performances of the classifiers using both ROIs and whole lung comparisons. The ROI classification evaluated the classifiers by calculating the match between the predicted and radiologist-marked ROIs using a cross-validation method. The whole lung quantification compared the accuracy of the regional lung disease pattern in whole lung parenchyma drawn by the automatic quantification system with that drawn by the expert radiologists. In this experiment, the accuracies of

Table 1 Comparisons of DILD classification systems using HRCT images

Authors	Features	Classifier	Number of cases	Number of classes	Overall accuracy
Delorme et al. [13]	Texture features	Multivariate discrimination analysis	1022 ROIs	6	70.7%
Uppaluri et al. [14]	Texture features	Adaptive multiple feature method	72 subjects	6	51.7%
Gangeh et al. [15]	Texton features	SVM-RBF	168 ROIs	3	96.4%
Sorensen et al. [16]	LBP intensity histogram	*k*-nearest neighbor	168 ROIs	3	95.2%
Vo et al. [17]	Gaussian derivative filter Wavelet and contourlet transform-based features	Multi-class multiple kernel learning	38 subjects	4	94.2%

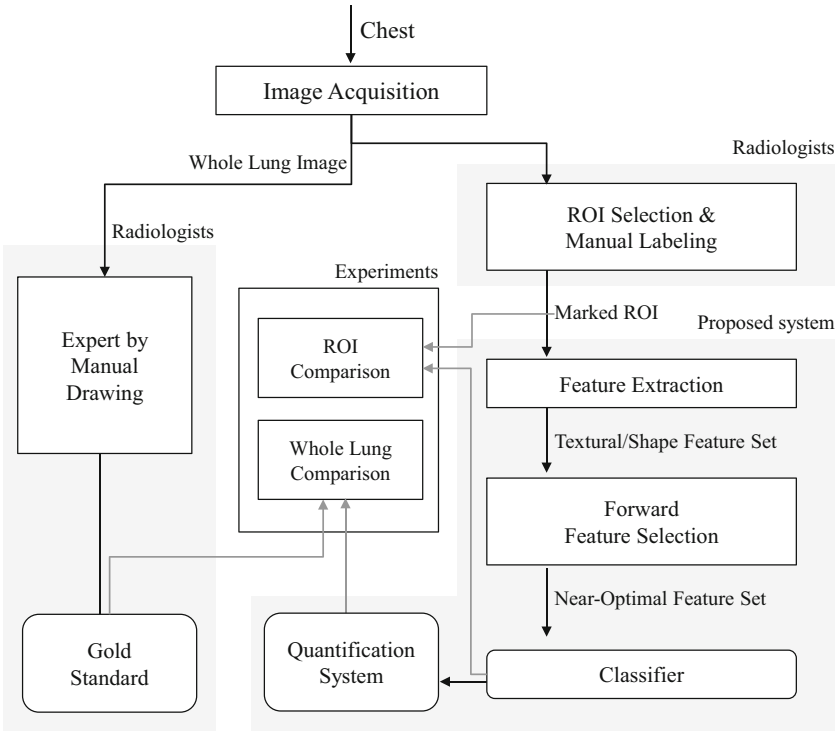

Fig. 1 Overall scheme of the proposed system, data construction, and experiment

the classifiers were measured by comparison to the gold standard that was manually drawn by the expert radiologists. More details on each step are provided in the following sections.

Subjects

The institutional review board for human investigation of Asan Medical Center, Seoul, South Korea, approved the study protocol; removed all patient identifiers; and, due to the retrospective design of this study, waived informed consent requirements. From 14 healthy individuals and 92 patients with lung disease, CT images were obtained using a Siemens CT scanner (Sensation 16; Siemens Medical Solutions, Forchheim, Germany) at Asan Medical Center or a GE CT scanner (GE Lightspeed 16; GE Healthcare, Milwaukee, WI) at the National Jewish Health Center, Denver, CO, between November 2000 and July 2005. Typical HRCT protocol parameters were used when obtaining the HRCT images, 220 mAs and 120–140 kVp with the patients at full inspiration. Image reconstruction used a 1-mm slice thickness and 10-mm slice interval with a reconstruction kernel B70f in the Siemens scanner and the sharp kernel in the GE scanner.

In this study, six classes—normal lung and five regional lung disease patterns—were defined, which are normal, ground-glass opacity, consolidation, reticular opacity, emphysema, and honeycombing. Representative examples of each class are shown in Fig. 2. Ground-glass opacity is an abnormally hazy focus in the lungs that is not associated with obscured underlying vessels. The latter is defined as consolidation. Increased reticular lung opacity is the product of a thickening of the interstitial fiber network of the lung by fluid, fibrous tissue, or cellular infiltration. In emphysema, there are focal areas of very low attenuation that can be easily contrasted with the surrounding higher-attenuation normal parenchyma at sufficiently low window levels (≤−600 HU). Emphysema can usually be distinguished from honeycombing because areas of emphysematous destruction lack a visible wall, whereas honeycomb cysts have thick walls of fibrous tissue. Honeycombing is additionally characterized by extensive fibrosis, with lung destruction and a resulting cystic, reticular appearance.

To collect representative examples of the five regional disease and normal lung patterns from CT images, two expert radiologists (J.B.S. and Y.K.L.) with more than 15-year experience were asked to independently select ROIs on CT images. Based on the study on DILD classification using textural features, ROIs with a 20 × 20-pixel size were used [24]. To prevent the unintended selection bias, no ROIs were selected more than once in the same lobe of CT images. For the categorization of the selected ROIs, the radiologists selected 100 ROIs per regional disease and normal lung pattern by mutual consent. By repeating the process on each scanner, 1200 ROIs for the different regional disease patterns were collected from the different scanners. In addition, 92 HRCT images were collected to investigate the classification of the whole lung parenchyma. The expert radiologists were asked to draw area maps of the five regional diseases and normal lung in each 92

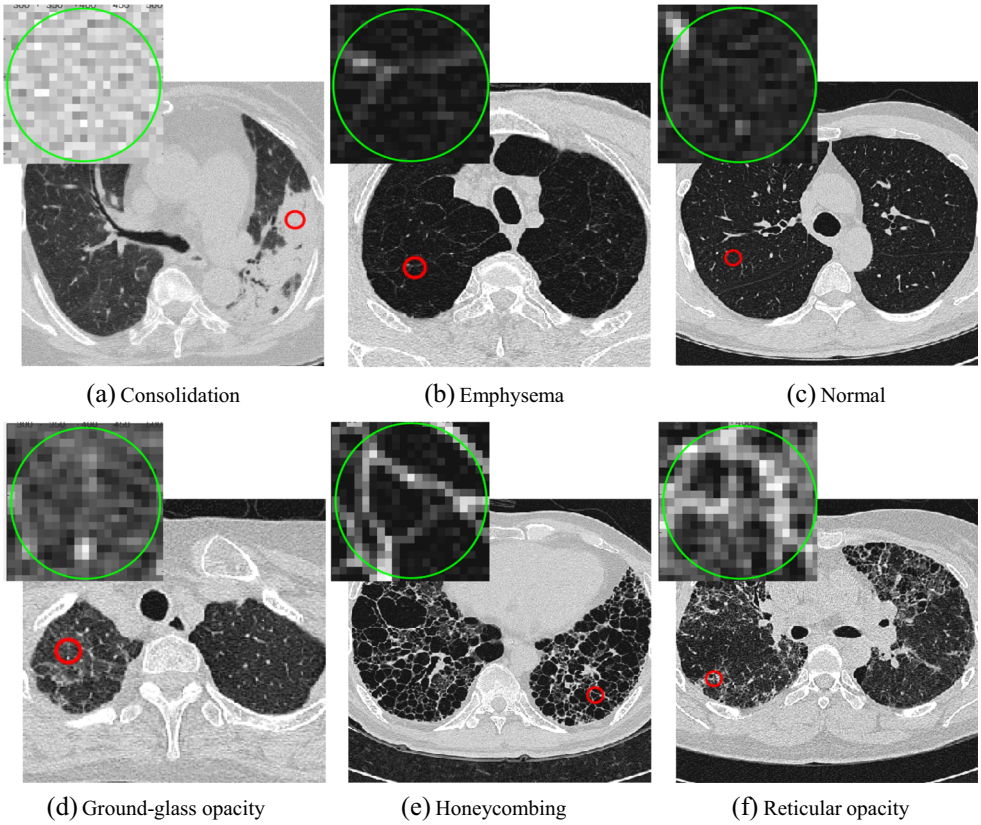

Fig. 2 Representative examples of the five regional lung disease and normal lung patterns

HRCT image. As a result, 92 pairs of the gold standard dataset of the whole lung were obtained.

Image Pre-Processing and Representation

To characterize the patterns of the six categories (five regional lung disease and normal lung patterns), 28 textural and shape features were extracted from each ROI [6, 9]. The extracted textural and shape patterns could be divided into six types of descriptor by their calculation methods, which are histogram, gradient, run-length matrix, co-occurrence matrix, cluster analysis, and top-hat transform. The histogram of an image is the distribution of gray level values of all pixels in the image. The gradient represents the variations in the gray level between black and white. The run-length matrices calculate the gray tone run length within the image. These matrices can help to distinguish coarse and fine textures based on the length of the run [25].

To describe spatial dependency characteristics, the other features were computed from a co-occurrence matrix containing spatial gray tone relationships [26]. Binning of the gray level into a smaller number is necessary to calculate the run-length and co-occurrence matrices. In this paper, the linear bin sizes of the run-length and co-occurrence matrices were optimized to 196 and 32, respectively [27]. Two additional descriptors, cluster analysis and top-hat transform, were additionally computed to describe not only the textural patterns but also the shape of the regional patterns. By thresholding the image under −950 HU, low-attenuation areas regarded as emphysema were obtained [28]. Subsequently, cluster analysis and top-hat transform were used [29]. As a result, 28 dimensions of the extracted feature vectors were extracted. The corresponding descriptors are listed in Table 2. The feature extractions are implemented and imported from a publicly available tool written in C++ (ITK 4.7) [30]. The extracted feature vectors were standardized with the zero mean and uniform variation before applying to the classifiers.

Research Design

Comparison of Individual and Ensemble Classifiers

To investigate the effect of the ensemble classifier in classifying regional lung disease patterns and normal lung based on both the ROI and whole lung, we used a number of individual and ensemble classifiers. For the individual classifier, a support vector machine (SVM) and naïve Bayes' classifier were used. For the ensemble classifiers, we applied four ensemble methods, which are random forest, bagging, voting, and stacking. To implement these classifiers, a publicly available tool written in Python language (scikit-learn) was used [31].

Table 2 Feature ($N = 28$) vectors for characterizing regional lung disease patterns

Descriptor	Dimension	Description
Histogram (4)	Mean	- Number of pixels in each ROI with a given gray level value
	SD	
	Skewness	
	Kurtosis	
Gradient (2)	Mean	- Variation in the gray level from black to white, with a high gradient value defined as an abrupt change in gray level from black to white
	SD	
		- Horizontal and vertical direction
Run length (2)	Short primitive emphasis (SPE)	- Run of gray tones, with coarse and fine textures defined as large and small numbers in the run, respectively, in a constant gray tone run
	Long primitive emphasis (LPE)	
		- Horizontal, vertical, 45°, and 135° directions
		- 196 nonlinear binning
Co-occurrence matrix (12)	Angular second moment mean and SD	- Spatial gray tone relationships in textural patterns and computed by co-occurrences at each ROI
	Contrast mean and SD	
	Correlation mean and SD	- Horizontal, vertical, 45°, and 135° directions
	Inverse difference moment mean and SD	- 32 nonlinear binning
	Entropy mean and SD	
	Inertia mean and SD	
Cluster analysis (4)	LAA	- Low-attenuation area (LAA) regarded as emphysema below a threshold of −950 HU
	Number of LAAs	
	Area mean	
	Area SD	
Top-hat transform (4)	White top-hat mean	- Morphological filter extracts small elements and details by calculating differences between images and its structuring element
	White top-hat SD	
	Black top-hat mean	
	Black top-hat SD	

Table 3 shows the list of individual and ensemble classifiers used in this study.

SVM is a supervised learning model that is widely used in various areas such as pattern recognition and data analysis [32]. While training the model, SVM defines the hyperplane, which is a boundary separating different classes, by finding support vectors. To find support vectors generating the optimal hyperplane, the Lagrange multiplier and Karush-Kuhn-Tucker condition were used to solve the optimal problem with the boundary margin constraints. In this study, we used SVM with a radial-based function (RBF) kernel and the optimal parameter such as gamma and cost are selected by grid searching algorithm. RBF kernel is the kernel function widely using in various classification problem with SVM classifier. With RBF kernel, feature space is reformed by calculating given two data point x and x' as shown in the following equation:

$$K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right) \quad (1)$$

where C_k is the class variables and x is the feature vector.

The naïve Bayes' classifier defines classification criterion based on the maximum a posteriori probability and with Bayes' theorem. According to the Bayes' theorem, the prior probability multiplied by the likelihood is proportional to the posteriori as the following equation:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (2)$$

Table 3 Individual and ensemble classifiers

Type	Classifier
Individual classifier	Support vector machine
	Naïve Bayes' classifier
Ensemble	Random forest
	Bagging
	Voting
	Stacking

The naïve property applied to Bayes' classifier assumes that each feature is independent to the other features. The naïve Bayes' classifier uses maximum a posteriori (MAP) as a decision rule which choose the most probable hypothesis class among the posteriors of the classes. Following equation shows how the function assigns the hypothesis class C_k when feature vector x is given:

$$y = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}}\, p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \tag{3}$$

The ensemble method combines a set of individual classifiers and integrates their decisions by taking a vote of their predictions [33]. In this paper, we applied four ensemble methods to the classifications, which are voting, bagging, random forest, and stacking. The voting and bagging methods are the most straightforward ways of combining individual classifiers. The voting method simply involves taking a vote of the predictions of individual classifiers. In the bagging method, training datasets are randomly sampled and generated from the original training dataset. The generated training datasets are called bootstrap replicates, and the prediction is computed by taking a vote of the predictions of classifiers trained by each replicate. The random forest method applies the bootstrapping approach to decision trees [34]. These ensemble methods apply the bootstrapping method to both the training dataset and feature space. For the classification, the prediction is selected by taking a vote on the result predicted by multiple decision trees. The stacking method (stacked generalization) is similar to the voting ensemble method but uses meta-classification instead of voting on the predictions of individual classifiers [35]. In other words, the meta-classifier is trained and makes a final decision using the dataset obtained by the predictions of each individual classifier.

Cross-Vendor Study

To investigate the effects of ensemble classifiers on the classification of the regional disease patterns obtained using different scanners, we designed three types of experiments, which are intra-scanner, inter-scanner, and integrated scanner. In the intra-scanner experiment, classifiers used the training and testing datasets from the same scanner. In the inter-scanner experiment, the training and testing datasets for the classification were obtained from different scanners. In the integrated scanner experiment, the training and testing datasets were equally integrated using the datasets from the two different scanners. Table 4 describes these experimental designs.

In our experimental design, the traditional cross-validation scheme was inapplicable to the intra-scanner experiment. Thus, we modified the cross-validation method for selecting testing and training subsets. In the intra-scanner experiment, two datasets from different scanners were both shuffled and divided into five exclusive subsets. After selecting a testing subset from a scanner, four training subsets were selected from the other scanner. By sequentially rotating the testing and training subsets, subsets were selected exclusively from each dataset as if in the traditional cross-validation scheme.

Evaluation and Statistics

In each experiment, individual and ensemble classifiers were compared in two aspects, which are ROI-based classification and whole lung quantification. For the ROI-based classification, sequential forward selection and fivefold cross validation were used to estimate the prediction accuracies of the five regional disease patterns and normal lung. We used the sequential forward selection algorithm, which selects and adds features to gradually increase classification accuracy until maximum accuracy is reached. For the whole lung quantification, trained classifiers predicted the regional disease patterns in the whole lung images and the predictions were compared with the quantification results of the expert radiologists. In the fivefold cross validation, the dataset was shuffled and divided into given subsets. One of the subsets was selected as the testing dataset, while the remaining subsets were selected as the training set. Classifiers were trained and the prediction accuracies were calculated. By rotating the testing and training roles of the five subsets, the average accuracy was calculated as the cross-validation accuracy. To measure reliable estimations in terms of statistics, the fivefold cross validations were repeated 20 times by applying variations in shuffling. As a result, the averages and standard deviations of the classification accuracy were calculated and statistically evaluated (paired t test).

For the whole lung quantification, individual and ensemble classifiers trained by ROIs were used. After manually segmenting the lung parenchyma from the whole lung image, the moving ROI function traveled and made predictions for each pixel with the trained classifiers. Since the boundary areas of the parenchyma usually cause miscalculation of both textural and shape features, they are eroded before the feature extraction. After the classification, the eroded areas are filled with the nearest neighbor classes. As a result, each pixel of the whole lung parenchyma was represented by the corresponding class among the five regional diseases or the normal lung. This whole lung quantification process was applied to the 92 collected HRCT lung images. To evaluate the accuracy of the whole lung quantification, we defined two criteria for computing the agreement, which are regions and areas. The region agreement was calculated by the number of matched pixels between the classifiers and the two radiologists. The area agreement was calculated by the area proportion of each regional pattern comprising the whole lung parenchyma. The mean square error (MSE) was used to represent the

Table 4 Study designs for the classification of the five diseases or the normal region

Study design		Training set (*N*)	Test set (*N*)
Intra-scanner study	GE	GE (600)	GE (600)
	Siemens	Siemens (600)	Siemens (600)
Integrated-scanner study	Integrated set	GE + Siemens (1200)	GE + Siemens (1200)
Inter-scanner study	Train GE and test Siemens (GE → Siemens)	GE (600)	Siemens (600)
	Train Siemens and test GE (Siemens > GE)	Siemens (600)	GE (600)

differences. The MSEs between each pair of classifiers and the radiologist read were calculated by the following equation:

$$\mathrm{MSE}_{i,k} = \frac{1}{6}\sum_{j=1}^{6}\left(R_{i,j} - C_{i,j,k}\right)^2 \quad (4)$$

where $R_{i,j}$ and $C_{i,j,k}$ are the proportions of the regional lung disease pattern j to the whole lung parenchyma calculated from the radiologists and the classifier k for the HRCT image i, respectively. A lower MSE indicated a higher quantification agreement between the classifier and radiologists. Using the MSE, we could define the area accuracy by the following equation:

$$\text{Area accuracy}_{i,k} = 1 - \mathrm{MSE}_{i,k} \quad (5)$$

All statistical evaluations were performed by SPSS v20 (20.0.0) with a 0.05 significance level.

Results

ROI-Based Classification

As shown in the experimental design in Table 4, each individual and ensemble classifier was applied to the comparison of the classification accuracy. For each classifier, optimal feature sets giving the highest classification accuracy were selected. Table 5 shows the results of the comparison. Among the individual classifiers, the SVM (RBF kernel; parameter optimized by grid search algorithm) was better than the other individual classifier ($p < 0.001$). In most experiments, the ensemble classifiers showed better performance than the individual classifiers. In the intra-scanner experiment using the GE dataset, the stacking ensemble classifier had higher classification accuracy than the SVM classifier ($p < 0.01$). In contrast, for the Siemens dataset in the intra-scanner experiment, the ensemble classifiers were not different from the individual classifiers ($p = 0.916$). In the inter-scanner experiment, the random forest (with 200 estimators) classifier showed better classification accuracy than the SVM when training with the Siemens dataset and testing with the GE dataset ($p < 0.001$). In the inter-scanner experiment using GE training and the Siemens testing dataset, the individual and ensemble classifiers did not show significant differences in classification accuracy ($p = 0.797$). In the experiment using the integrated dataset, the voting ensemble classifier performed better than the SVM classifier ($p < 0.001$).

Whole Lung Quantification by Regions

To more clearly investigate the effects of the classification methods, the individual and ensemble classifiers were applied to the whole lung area with moving ROIs. For the evaluations, we used similar study designs as those applied to the ROI-based classification (Table 4), which are intra-scanner, inter-scanner, and integrated scanner. Based on the study designs, we compared the voxel-by-voxel quantification agreements between the classified results and the visual assessments of the two expert radiologists. Table 6 shows the quantification agreement between the classifiers and radiologists using various training datasets and classification methods.

As shown in Table 6, the ensemble classifiers, including bagging, stacking, and random forest, performed better quantification agreements in all study designs. In the integrated dataset, the bagging ensemble classifier had higher quantification agreement with the two radiologists than the SVM classifier ($p = 0.001$). When the classifiers were trained by the Siemens dataset, the quantification agreements were higher than those of the other study designs because the 92 whole lung images were obtained from the same scanner (comparison to the integrated experiment t test, $p < 0.001$). In this design, the stacking ensemble method also shows significantly higher agreement with both radiologists than with the individual SVM classifiers ($p = 0.022$ and 0.034, respectively). When the classifiers were trained by the GE dataset, quantification of the whole lung HRCT images obtained from the Siemens scanner showed the lowest agreement among the study designs. In this study design, the random forest ensemble method showed higher agreement than the other individual classifiers ($p = 0.004$ and 0.000). Figure 3 illustrates several examples of the quantification results obtained using the integrated training dataset with a comparison of the results of the classifiers and radiologists using colored overlays.

Table 5 Classification accuracy comparisons using individual and ensemble classifiers for ROI images of regional lung disease patterns

Study design	Training dataset	Testing set	Accuracy (%)		*p*
			Individual classifiers	Ensemble classifiers	
Intra-scanner	GE	GE	91.56 ± 0.62 (SVM)	92.02 ± 0.56 (stacking)	0.002
	Siemens	Siemens	89.91 ± 0.39 (SVM)	89.92 ± 0.36 (bagging)	0.916
Inter-scanner	GE	Siemens	68.31 ± 0.52 (SVM)	68.27 ± 0.69 (bagging)	0.797
	Siemens	GE	65.88 ± 0.39 (SVM)	69.73 ± 0.61 (random forest)	<0.001
Integrated	GE + Siemens	GE + Siemens	88.38 ± 0.31 (SVM)	89.73 ± 0.43 (stacking)	<0.001

Whole Lung Quantification by Areas

For the lung quantification, the amount of each regional disease pattern composing the lung parenchyma is important information in the clinical approach. In this experiment, the ratios of each regional disease pattern to the whole lung area were calculated from the classifier and radiologist reads. The MSE was computed between the classifiers and the radiologists to compare the individual and ensemble methods with three different training datasets, which are the GE, Siemens, and integrated datasets. Table 6 shows the area accuracy of the individual and ensemble classifiers.

As shown in Table 7, ensemble classifiers performed better in the integrated dataset for the lung quantification by area ($p < 0.001$). With the Siemens training dataset, ensemble classifiers showed the lowest MSE, although it was not always significantly lower than the individual classifier ($p = 0.021$ and 0.054, respectively). When ensemble and individual classifiers were trained by the GE dataset, the ensemble classifiers showed a lower MSE than the individual classifier, although the difference was not always significant ($p = 0.665$ and 0.098, respectively).

Discussion

In our present study, we evaluated the potential use of ensemble classifiers to overcome inter-scanner variations in the differentiation of the regional lung disease patterns of DILD on HRCT images with texture and shape features. In [23], authors compared the differences in the image features obtained from CT scanners from two different vendors. Some extracted features, such as run-length and co-occurrence matrices, showed significant differences due to the different mechanisms of the two scanners. We believe that these factors could have a negative effect on stable lung classification among different scanners with only individual classifiers.

In multicenter trials using scanners from different vendors, there are fundamental limitations to equal classifications. Different scanners have different mechanisms and reconstruction kernels that cannot be controlled. Both limitations prevent a consistent lung classification. The experimental results of this study imply that the use of an ensemble classifier can ease this problem and increase the classification accuracies.

Ensembles of classifiers perform better at classifying robust data than individual classifiers [33]. Thus, we assumed that use of the ensemble method to classify regional disease patterns in lung HRCT images would show better performance in the integrated scanner dataset. In our experiments, the ensemble classifier showed better performance than the individual classifier in the ROI-based classifications when the integrated dataset was used. Furthermore, in the whole lung quantification, performance significantly increased in the integrated dataset in both regions and area. In all study designs, the increase in accuracy was greater than that of the other dataset

Table 6 Average region accuracy between classifiers and radiologists for the 92 HRCT whole lung images

Training dataset	Radiologist	Accuracy (%)						Maximum difference comparison *p*
		SVM	NB	RF	Bagging	Voting	Stacking	
Integrated	SJB	48.13**	44.67***	44.81***	49.49	48.53*	44.67***	0.001 (SVM-bagging)
	LYK	48.25**	42.71***	43.47***	49.64	47.79***	42.71***	0.001 (SVM-bagging)
Siemens	SJB	54.26*	49.31***	51.81***	53.51**	54.40	55.13	0.022 (SVM-stacking)
	LYK	50.82*	42.85***	49.13**	48.23***	48.18***	51.78	0.034 (SVM-stacking)
GE	SJB	24.75**	25.75**	27.35	24.49**	23.83***	22.57***	0.005 (NB-RF)
	LYK	23.20***	25.18**	26.89	23.15***	22.49***	21.29***	0.003 (NB-RF)

SVM support vector machine, *NB* naïve Bayes, *RF* random forest

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

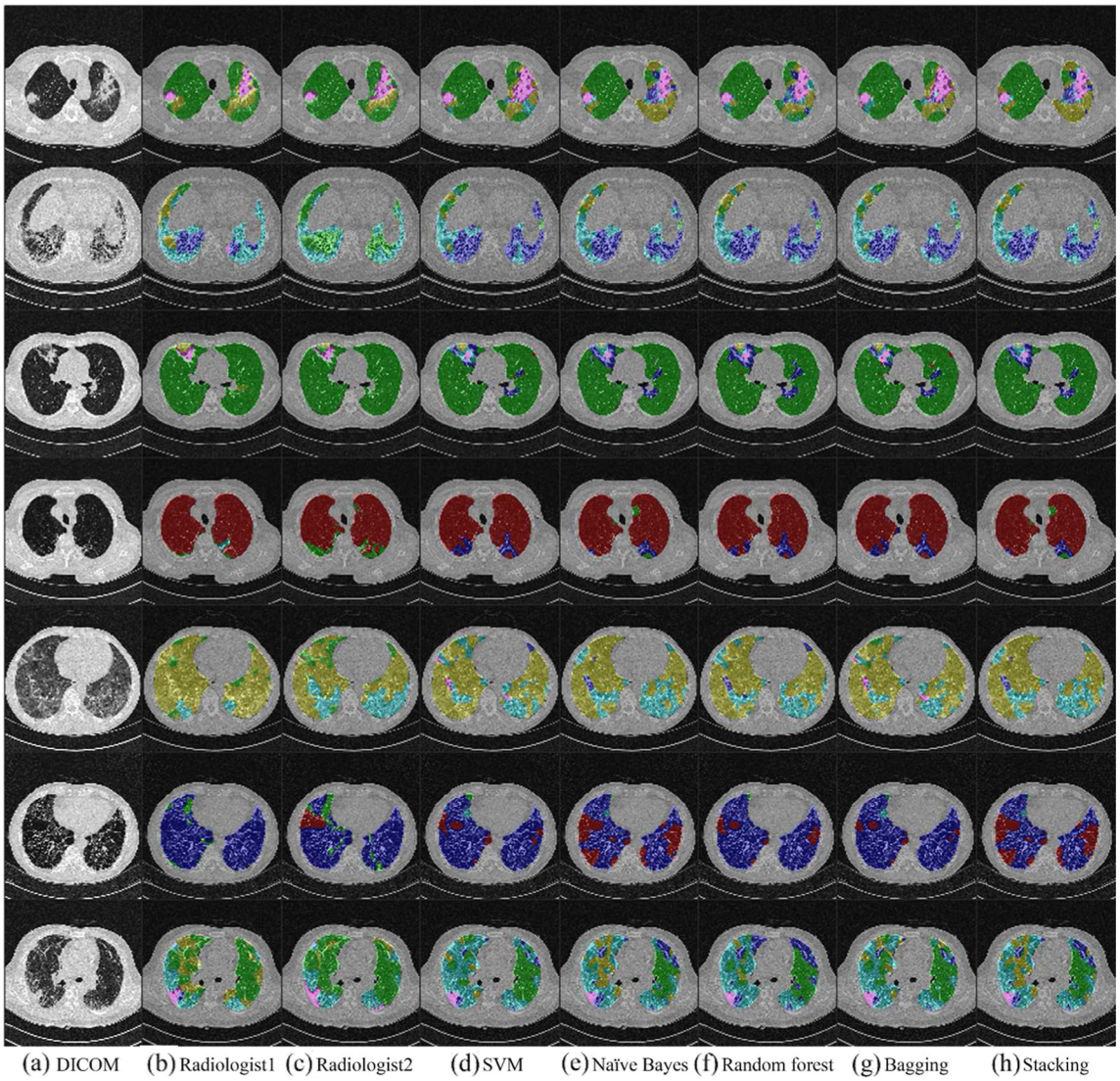

Fig. 3 Examples of the quantification results of radiologists and classifiers trained by integrated dataset using colored overlays

when the ensemble classifier and the integrated dataset were used for training. Additionally, we observed that use of the ensemble classifier in the intra-scanner dataset showed better performance in whole lung quantification, even though it did not show a significant difference in the ROI classification. On the other hand, the inter-scanner dataset led to poor

Table 7 Average area accuracy between classifiers and radiologists for the 92 HRCT whole lung images

Training dataset	Radiologist	Accuracy (%)						Maximum difference comparison *p*
		SVM	NB	RF	Bagging	Voting	Stacking	
Integrated	SJB	0.51***	0.51***	0.51*	0.54	0.52*	0.51*	<0.001 (SVM-bagging)
	LYK	0.44***	0.40***	0.41***	0.47	0.44***	0.40***	<0.001 (SVM-bagging)
Siemens	SJB	0.62*	0.54***	0.58***	0.61**	0.63	0.64	0.021 (SVM-stacking)
	LYK	0.49	0.38***	0.46***	0.45***	0.46**	0.50	0.054 (SVM-stacking)
GE	SJB	0.25	0.23**	0.26	0.25	0.23**	0.20***	0.665 (SVM-RF)
	LYK	0.17	0.16***	0.20	0.18	0.16***	0.14***	0.098 (SVM-RF)

SVM support vector machine, *NB* naïve Bayes, *RF* random forest

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

classification accuracy in the ROI classification, which was also reflected in the performance of the whole lung quantification.

In our current analysis, we found a limitation in pairwise comparisons between the ROI-based classification and the whole lung quantification using the same classifier. In the ROI-based classification, the stacking ensemble showed the highest accuracy when the integrated dataset was used. Nevertheless, the stacking ensemble was not the best performer and even showed lower accuracy than the individual SVM classifier in the whole lung quantification. This result is not directly comparable because the whole lung data consists of Siemens data only, unlike the ROI dataset.

There was an inevitable difference between the radiologists and the classifiers in recognizing the regional disease patterns. While radiologists assess the patterns by area, classifiers analyze the patterns by pixel. In other words, radiologists might ignore pixel-wise patterns in their assessment.

Our present findings show that there is still room for improvement. We observed that airways or large vessels tend to be easily misclassified into disorder patterns. In a future study, the 3D volumetric images need to be analyzed, not only to leave out unnecessary patterns but also so that more precise features can be extracted and used for the classification with more sophisticated segmentation algorithms.

Conclusion

In multicenter trials, the HRCT images of DILD obtained from scanners from different vendors cause variations in classification accuracy. In the present study, we investigated the effects of ensemble classifiers on regional pattern classification and whole lung quantification using HRCT images from different scanners. The main contribution of our present study is that an ensemble classifier that combines individual classifiers can improve classification performance in analyses that uses the same scanner and those that use different scanners. When using an integrated dataset containing HRCT images from different scanners, ensemble classifiers such as bagging and stacking performed better than the individual classifiers in both the ROI-based classification and whole lung quantification. In intra- and inter-scanner studies, ensemble classifiers generally showed better performance than individual classifiers. This implies that ensemble classifiers could differentiate regional interstitial lung disease patterns with better accuracy in multicenter trials.

Acknowledgements This work was partially supported by the Technological Innovation R&D Program (S2259881) funded by the Small and Medium Business Administration (SMBA, Korea); and by Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korean Government (MSIP) (No. R6910-15-1023).

References

1. Hoffman EA, et al.: Characterization of the interstitial lung diseases via density-based and texture-based analysis of computed tomography images of lung structure and function 1. Academic Radiology 10:1104–1118, 2003
2. Scatarige JC, Diette GB, Haponik EF, Merriman B, Fishman EK: Utility of high-resolution CT for management of diffuse lung disease: results of a survey of US pulmonary physicians. Academic Radiology 10:167–175, 2003
3. Grenier P, Valeyre D, Cluzel P, Brauner MW, Lenoir S, Chastang C: Chronic diffuse interstitial lung-disease—diagnostic-value of chest radiography and high-resolution Ct. Radiology 179:123–132, 1991
4. Coxson HO, et al.: A quantification of the lung surface area in emphysema using computed tomography. American Journal of Respiratory and Critical Care Medicine 159:851–856, 1999
5. Kalender WA, Rienmuller R, Seissler W, Behr J, Welke M, Fichte H: Measurement of pulmonary parenchymal attenuation—use of Spirometric gating with quantitative Ct. Radiology 175:265–268, 1990
6. Kim N, Seo JB, Lee Y, Lee JG, Kim SS, Kang S-H: Development of an automatic classification system for differentiation of obstructive lung disease using HRCT. Journal of Digital Imaging 22:136–148, 2008
7. Xu Y, Sonka M, McLennan G, Guo JF, Hoffman EA: MDCT-based 3-D texture classification of emphysema and early smoking related lung pathologies. IEEE Transactions on Medical Imaging 25:464–475, 2006
8. Fujisaki T, et al.: Effects of density changes in the chest on lung stereotactic radiotherapy. Radiat Med 22:233–238, 2004
9. Chabat F, Yang GZ, Hansell DM: Obstructive lung diseases: texture classification for differentiation at CT. Radiology 228:871–877, 2003
10. Lim J, Kim N, Seo JB, Lee YK, Lee Y, Kang S-H: Regional context-sensitive support vector machine classifier to improve automated identification of regional patterns of diffuse interstitial lung disease. Journal of Digital Imaging 24:1133–1140, 2011
11. Moon JW, et al.: Perfusion-and pattern-based quantitative CT indexes using contrast-enhanced dual-energy computed tomography in diffuse interstitial lung disease: relationships with physiologic impairment and prediction of prognosis. European Radiology:1–10, 2015
12. Xu Y, van Beek EJ, Hwanjo Y, Guo J, McLennan G, Hoffman EA: Computer-aided classification of interstitial lung diseases via MDCT: 3D adaptive multiple feature method (3D AMFM). Academic Radiology 13:969–978, 2006
13. Delorme S, Keller-Reichenbecher M-A, Zuna I, Schlegel W, Van Kaick G: Usual interstitial pneumonia: quantitative assessment of high-resolution computed tomography findings by computer-assisted texture-based image analysis. Investigative Radiology 32:566–574, 1997
14. Uppaluri R, Hoffman EA, Sonka M, Hartley PG, Hunninghake GW, Mclennan G: Computer recognition of regional lung disease patterns. American Journal of Respiratory and Critical Care Medicine 160:648–654, 1999
15. Gangeh MJ, Sorensen L, Shaker SB, Kamel MS, Bruijne Md, Loog M: A texton-based approach for the classification of lung parenchyma in CT images. Proc. Proceedings of the 13th International Conference on Medical Image Computing and Computer-assisted Intervention: Part III
16. Sorensen L, Shaker SB, de Bruijne M: Quantitative analysis of pulmonary emphysema using local binary patterns. IEEE Transaction on Medical Imaging 29:559–569, 2010
17. Vo KT, Sowmya A: Multiple kernel learning for classification of diffuse lung disease using HRCT lung images. Proc. 2010 Annual

International Conference of the IEEE Engineering in Medicine and Biology
18. Xu R, Hirano Y, Tachibana R, Kido S: Classification of diffuse lung disease patterns on high-resolution computed tomography by a bag of words approach. Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention
19. Depeursinge A, et al.: Optimized steerable wavelets for texture analysis of lung tissue in 3-D CT: classification of usual interstitial pneumonia. Proc. 2015 I.E. 12th International Symposium on Biomedical Imaging (ISBI)
20. Zhao W, Xu R, Hirano Y, Tachibana R, Kido S: Classification of diffuse lung diseases patterns by a sparse representation based method on HRCT images. Proc. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)
21. Yuan R, et al.: The effects of radiation dose and CT manufacturer on measurements of lung densitometry. Chest Journal 132:617–623, 2007
22. Park SO, et al.: Feasibility of automated quantification of regional disease patterns depicted on high-resolution computed tomography in patients with various diffuse lung diseases. Korean Journal of Radiology 10:455, 2009
23. Chang Y, Lim J, Kim N, Seo JB, Lynch DA: A support vector machine classifier reduces interscanner variation in the HRCT classification of regional disease pattern in diffuse lung disease: comparison to a Bayesian classifier. Medical Physics 40:051912, 2013
24. Kim N, et al.: Effect of various binning methods and ROI sizes on the accuracy of the automatic classification system for differentiation between diffuse infiltrative lung diseases on the basis of texture features at HRCT. Proc. Medical Imaging, 2008
25. Haralick RM: Statistical and structural approaches to texture. Proceedings of the IEEE 67:786–804, 1979
26. Carr JR, de Miranda FP: The semivariogram in comparison to the co-occurrence matrix for classification of image texture. IEEE Transactions on Geoscience and Remote Sensing 36:1945–1952, 1998
27. Kim N, Seo JB, Lee YK, Kim SS, Kang SH: Optimal binning and ROI size of the automatic classification system for differentiation between obstructive lung diseases on the basis of texture features at HRCT. IEICE technical report 106:95–97, 2007
28. Gevenois PA, de Maertelaer V, De Vuyst P, Zanen J, Yernault JC: Comparison of computed density and macroscopic morphometry in pulmonary emphysema. American Journal of Respiratory and Critical Care Medicine 152:653–657, 1995
29. Sonka M, Hlavac V, Boyle R: Image processing, analysis, and machine vision, Pacific Grove, CA: PWS Pub., 1999
30. Yoo TS, et al.: Engineering and algorithm design for an image processing API: a technical report on ITK—the insight toolkit. Studies in Health Technology and Informatics: 586–592, 2002
31. Pedregosa F, et al.: Scikit-learn: machine learning in python. Journal of Machine Learning Research 12:2825–2830, 2011
32. Cortes C, Vapnik V: Support-vector networks. Machine Learning 20:273–297, 1995
33. Dietterich TG: Ensemble methods in machine learning: Springer, 2000
34. Breiman L: Random forests. Machine Learning 45:5–32, 2001
35. Wolpert DH: Stacked generalization. Neural Networks 5:241–259, 1992