Improving computer-aided diagnosis of interstitial disease in chest radiographs by combining one-class and two-class classifiers.

Yulia Arzhaeva[a], David Tax[b], and Bram van Ginneken[c]

[a,c]Image Sciences Institute, Utrecht University, Utrecht, Netherlands;
[b]Information and Communication Theory Group, Delft University of Technology,
Delft, Netherlands

ABSTRACT

In this paper we compare and combine two distinct pattern classification approaches to the automated detection of regions with interstitial abnormalities in frontal chest radiographs. Standard two-class classifiers and recently developed one-class classifiers are considered. The one-class problem is to find the best model of the normal class and reject all objects that don't fit the model of normality. This one-class methodology was developed to deal with poorly balanced classes, and it uses only objects from a well-sampled class for training. This may be an advantageous approach in medical applications, where normal examples are easier to obtain than abnormal cases. We used receiver operating characteristic (ROC) analysis to evaluate classification performance by the different methods as a function of the number of abnormal cases available for training. Various two-class classifiers performed excellently in case that enough abnormal examples were available (area under ROC curve $A_z = 0.985$ for a linear discriminant classifier). The one-class approach gave worse result when used stand-alone ($A_z = 0.88$ for Gaussian data description) but the combination of both approaches, using a mean combining classifier resulted in better performance when only few abnormal samples were available (average $A_z = 0.94$ for the combination and $A_z = 0.91$ for the stand-alone linear discriminant in the same set-up). This indicates that computer-aided diagnosis schemes may benefit from using a combination of two-class and one-class approaches when only few abnormal samples are available.

Keywords: Computer-aided diagnosis, interstitial lung disease, one-class classification, classifier combining

1. INTRODUCTION

The purpose of this work is the investigation of different discriminative approaches to the automated classification of given small regions of interest (ROIs) within the lung fields in frontal chest radiographs on presence or absence of interstitial abnormalities. Detection of interstitial lung disease (ILD) in chest radiographs is one of the most difficult areas in radiology, for which computer-aided diagnosis (CAD) systems may provide valuable assistance.

ILD is the common term for more than 200 types of disorders, which may cause significant morbidity and mortality.[1] The interstitium of the lung is the tissue between the air sacs, and when it is damaged the textural appearance of the lung is changed on radiological images. Whereas a large variation of abnormal patterns can represent one type of ILD, radiographs of patients with different types of ILD may look alike. Moreover, the difference between normal and abnormal texture patterns is ambiguous even for human experts, which is revealed by high inter-observer variability.[1, 2] Figure 1 shows an example of normal and diseased lungs on a radiograph. Recently high-resolution computed tomography has become a modality of choice for the diagnostic of ILD.[3] The role of chest radiographs remains in initial detection of abnormalities and providing a preliminary diagnosis and a recommendation for the following computed tomography examination.

Reliable classification of ROIs is an important part of an efficient CAD system for detection of ILD. The majority of works in this field over the last two decades focused on the classification of a complete radiograph being normal or abnormal.[4–9] In those classification schemes multiple ROIs were manually or automatically selected within the lung fields and texture measurements, or features, were computed from them. Then the classification of ROIs was performed using rule-based or pattern recognition methods, and classification opinions (class labels or probabilities to be normal/abnormal) about each ROI were obtained. Finally probabilities over

Medical Imaging 2006: Image Processing, edited by Joseph M. Reinhardt, Josien P. W. Pluim,
Proc. of SPIE Vol. 6144, 614458, (2006) · 0277-786X/06/$15 · doi: 10.1117/12.652208

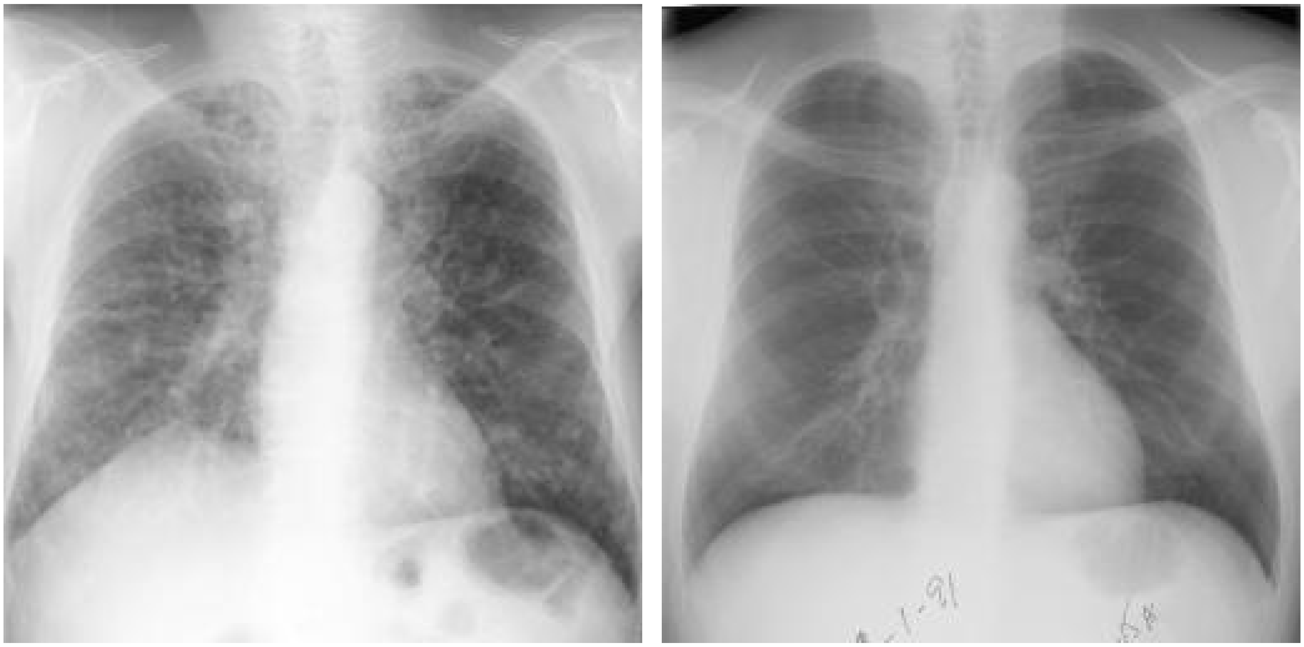

Figure 1. A radiograph of a patient with interstitial lung disease (left) and a radiograph of a healthy person (right). The left image clearly shows widespread shadowing with specific underlying patterns. In other cases the signs of abnormality can be much more subtle.

regions were fused to get a conclusion for a complete image, whether it contained any interstitial abnormalities or not. Besides its contribution to the classification of a complete image, good region classification can provide a radiologist with a credible map showing how a disease is spread over the lung fields. In this work we focus only on a region classification task. Two approaches are considered: standard two-class classifiers and recently developed one-class classifiers. The two-class technique can be thought of as directly distinguishing between normal and abnormal, while the one-class technique is a form of outlier detection: it evaluates if a region deviates from its normal appearance. Our goal is to compare the performances of two-class and one-class classifiers, in particular to study the effect of reducing the amount of abnormal samples in a training set, and to combine both approaches.

For a long time no quantitative analysis was provided for region classification accuracy. Performances of the automated classification systems developed in[4–8] were evaluated on image level only, though region classification was an essential step in there. In the work of Katsuragawa *et al.*[4] it was visually shown that texture measures extracted from ROIs were different for normal ROIs and ROIs with different types of abnormal tissue. Van Ginneken *et al.*[9] first estimated the region classification performance. In their work the lung fields were subdivided into overlapping regions of various sizes. A probability measure for a region to be abnormal was calculated using a two-class k-nearest neighbor classifier and a separate training set for each location. It was shown that the region classification performance was the higher the more abnormal examples were in a training set and the less superimposed structures (vessels, rib crossings) were present in that location. Further in that work the region probabilities were combined to receive a verdict over an image, which resulted in the image classification performance much higher than the classification performance for the majority of regions.

In this work small ROIs are extracted manually from the middle periphery of both lung fields, where the lung texture is less obscured by other structures. In this way we purposefully simplify the classification task because we aim at comparing performances of different methods depending on parameters other than the 'difficulty' of a region. Texture features are computed from each ROI, using the moments of responses to a multi-scale Gaussian filter bank. Several two-class classifiers are trained on subsets of data with the varying amount of abnormal samples to study how classification performance depends on unequal representation of classes in training data. In clinical practice normal cases are more often encountered and therefore are easier to collect than abnormal ones. That motivates studying two-class classification performance on unbalanced data sets, and application of one-class classification methodology[10] in particular. Moreover, interstitial lung disease reveals such a large variety of abnormal patterns that collection of a sufficient amount of abnormal representatives for training is an effortful and time-consuming task. One-class classifiers only need normal samples for training to estimate parameters of a distribution model to fit normal data. We investigate the performance of several one-class classifiers on our data. Furthermore, we use the mean combing rule to combine posterior probabilities yielded by a one-class classifier and a two-class classifier trained on a highly unbalanced training set, and compare combining classification performance with performances that each method is able to achieve alone.

The remainder of this paper is structured as follows. In Section 2 the data is described. In Section 3 classification methods are explained. In Section 4 the results are presented. We conclude the paper in Section 5.

2. DATA

We conducted our experiments on a database of digitized chest radiographs, consisting of 100 healthy images and 100 images containing areas with ILD collected at the University of Chicago Hospitals. This data was also used and described in.[4] All normal cases were selected based on consensus of a panel of experienced radiologists. Abnormal cases were selected based on radiological findings, CT, clinical data and follow-up radiographs, by consensus of the same radiologists. The radiographs were digitized to 2000 by 2000 pixels with 0.175 mm pixel size and 10 bits intensity.

Regions of interest were manually extracted from the middle periphery of both lung fields, 4 ROIs per image (see Figure 2). The size of regions was chosen such that a region roughly covered two ribs and a space between them. In this way region sizes differed from region to region, but regions had similar anatomical structure. ROIs were classified by an experienced chest radiologist into one of four possible categories–'normal', 'definitely abnormal', 'possibly abnormal' and 'containing other abnormalities'. Normal ROIs for experiments were taken only from healthy images, 399 normal ROIs in total. We obtained 228 definitely and possibly abnormal ROIs from 78 abnormal images. Normal ROIs from abnormal images and ROIs containing other, not ILD-related abnormalities were excluded from the study.

2.1. Features

Each ROI was described by a set of texture features. Left lung fields were mirrored before extracting features from ROIs placed there. Each ROI was filtered with Gaussian derivatives of the order $0, 1$ and 2 at scales $\sigma = 1, 2, 4, 8, 16$. Then the mean, standard deviation, skewness and kurtosis over ROIs were calculated from filtered images, as well as from an initial image, and used as features. Features were normalized to have zero mean and unit variance. The total amount of 124 features was reduced by means of principal component analysis (PCA) retaining 99% of the variance. During classification a probability is determined for a region to be abnormal based on these texture features.

3. CLASSIFICATION

3.1. Two-class classification

Two-class classification is a supervised classification method, which means that a classifier is first trained on labelled samples from both normal and abnormal classes. A decision boundary between two classes or class distributions are learned via training. In order to obtain probabilities for a test sample to belong to one or the other class, its feature vector is passed to a trained classifier that defines class' posterior probabilities for this sample.

For two class classification experiments we have chosen the linear discriminant (LDC), the nearest mean, and the nearest neighbor (1-NN) classifiers.[11] The domain of supervised classification has a major division into parametric and nonparametric methods. Our selection of classifiers represents both types of methods, though we selected the simplest representatives. LDC is a parametric classifier. It assumes Gaussian distribution with equal covariance matrix for both, normal c_0 and abnormal c_1 classes. Distribution parameters are estimated during training. During classification posterior probabilities $p(x|c_0)$ and $p(x|c_1)$ of a test sample x are determined according to estimated distributions. Both the nearest mean and the nearest neighbor classifiers are nonparametric

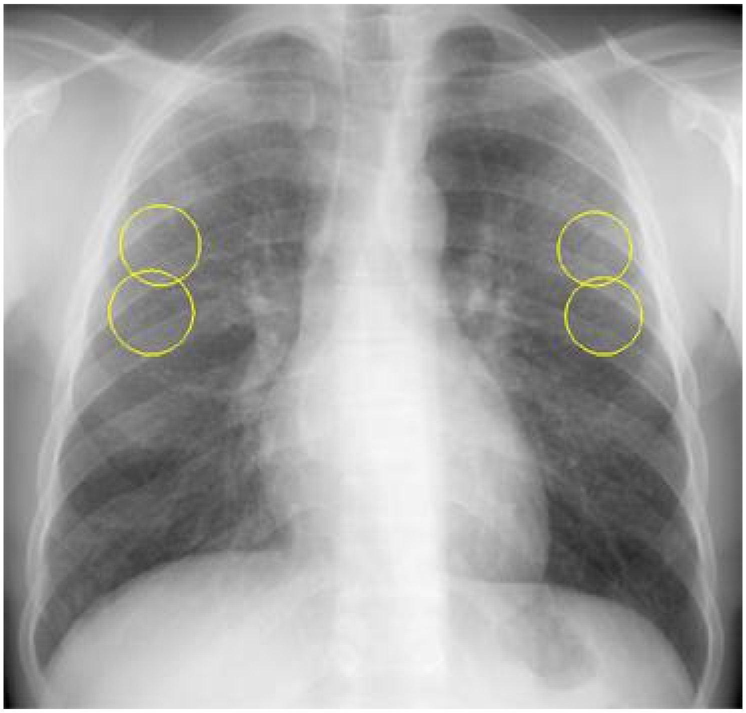

Figure 2. An example of regions of interest within the lung fields.

classifiers. The nearest mean classifier assigns a new sample x to a class which estimated mean is closer to x in terms of Euclidean distance. The 1-NN rule is very similar to the nearest neighbor rule: it assigns x to the same class as its first nearest neighbor in a training set belongs to. In both rules posterior probabilities instead of class labels can be approached using relative distances to the nearest neighbors (or means) from both classes.

In experiments we took care that ROIs from the same image were either all in a training set or all in a test set. A classifier performance was evaluated using a modification of a leave-one-out method. Each time a different image from our data set was used to provide test ROIs (i.e. from 1 to 4 ROIs from an image). The standard leave-one-out technique implies that a classifier is retrained each time with ROIs from the rest of images. In our modification we restrained the number of abnormal images to be used for training. Instead of using all abnormal images (excluding a test one if it was abnormal) to train a classifier, a randomly selected subset of M abnormal images was included in a training set. For M equal to the number of all abnormal images this modification converted to the standard leave-one-out technique. By changing M training sets with a variable degree of disbalance are obtained.

3.2. One-class classification

One-class classification is an unsupervised classification strategy because it assumes that only information of one of the classes, the target class, is available for training. In our case the target class is the normal class. One-class classification tries to describe the normal class of samples and learns nothing about the abnormal class during training. Later it distinguishes abnormal samples by their dissimilarity to the normal class. This strategy fits well into medical image analysis where poorly sampled classes are frequently encountered.

To describe the normal class of our data the Gaussian model, the Parzen density estimator, and the nearest neighbor data description were used.[10] The first method tries to fit the normal or Gaussian distribution model to the data. The Parzen density estimator is an extension of this: the estimated density is a mixture of Gaussian kernels centered on individual training samples. The nearest neighbor method does not estimate density explicitly but uses distances to the first nearest neighbor. A new sample is accepted as normal when its resemblance (distance, probability) to the modelled normal class is above a given threshold. Otherwise the sample is rejected as being abnormal. In our experiments a threshold was taken such that 5% of normal samples from a training set would be rejected.

A leave-one-out technique adapted to one-class methodology was used to conduct experiments. ROIs from every normal image were classified separately by a classifier retrained with ROIs from the remaining normal images. ROIs from abnormal images were classified by a classifier trained with ROIs from the whole set of normal images.

3.3. Combining classifiers

Many experimental studies have shown that combining classifiers can improve classification accuracy. It was shown in[12] that when classifiers are applied on identical data representations, classifiers outputs should be averaged to suppress errors of individual classifiers caused by the same noise in data. In our experiments after the initial classification of a test sample, we averaged the posterior probabilities obtained with a one-class classifier and a two-class classifier.

4. RESULTS

In all our experiments the area under a receiver operating characteristic (ROC) curve, indicated as A_z, was used as classification performance measure[13] on test data. The ROC curve plots the true positive fraction as a function of the false positive fraction. Points on the ROC curve can be obtained by varying a threshold of the posterior probabilities that defines abnormality of a ROI. A_z indicates how reliable classification can be performed. A value of $A_z = 1$ represents a perfect test, $A_z = 0.5$ is equivalent to guessing.

For three two-class classifiers an ROC curve was obtained for each number M, $2 \leq M \leq 78$, of abnormal images in a training set, and A_z was calculated. Experiments were repeated ten times. Average A_z values for the classifiers are plotted versus M in Figure 3. In is shown in this Figure that the best performing classifier is LDC. It yielded a high performance even when the abnormal class was very ill-sampled. The ultimate results for LDC were: $A_z = 0.985$ for $M = 78$, and $A_z = 0.909(\pm 0.013)$ for $M = 2$. All three classifiers dropped their steady performances at $M \leq 15$, i.e. when there were 6 times more normal images than abnormal in the training set.

In Table 1 the performances of one-class methods are shown. The best performing one-class classifier was Gaussian, $A_z = 0.882$. After applying the mean combination rule to the posterior probabilities resulted from LDC with $M = 2$ (that is on average 6 abnormal ROIs) and the Gaussian model, the classification performance improved to $A_z = 0.941(\pm 0.006)$ (see Figure 5). For $M \geq 4$ the combining classifier showed the same performance as the stand-alone LDC. For the 1-NN classifier and the nearest mean classifier, their combination with the Gaussian one-class classifier improves their performances for any amount of abnormal images in a training set (see Figure 4). When comparing Figure 3 and Figure 4 it is clear that the classification performance of combining classifiers becomes less dependent on the abnormal fraction than the performance of stand-alone two-class classifiers.

Table 1. Classification performance of one-class classifiers in term of the area under the ROC curve, A_z.

One-class classifier	$\mathbf{A_z}$
Gaussian model	0.882
Parzen data description	0.870
Nearest neighbor data description	0.750

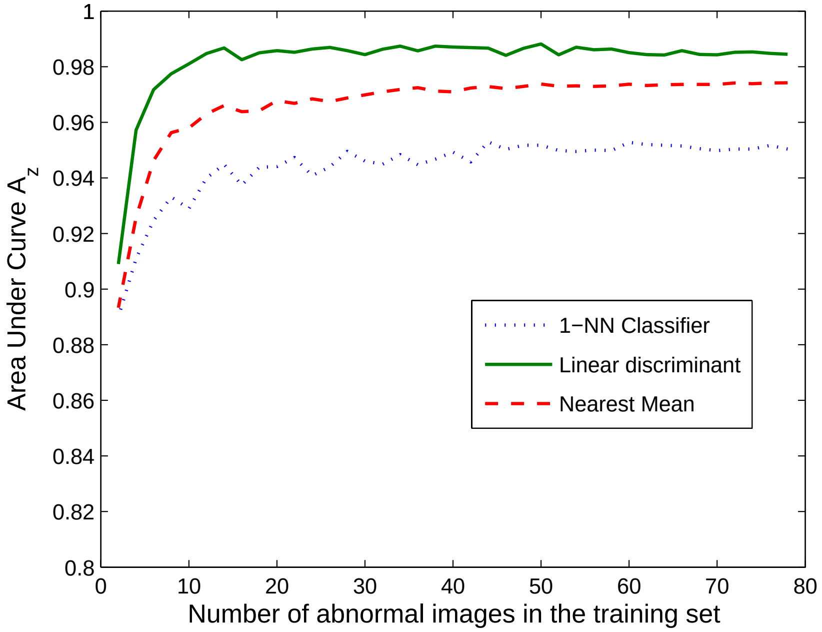

Figure 3. The dependency of different two-class classifiers performance on the number of abnormal images in the training set in terms of the area under ROC curve A_z.

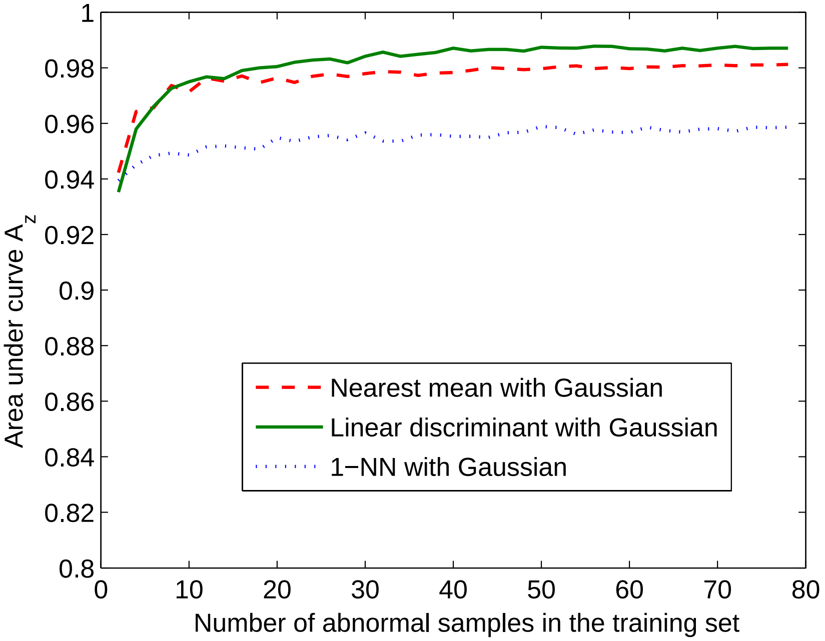

Figure 4. The classification performance of the mean combining classifiers in terms of the area under ROC curve depending on the amount of abnormal images in the training set.

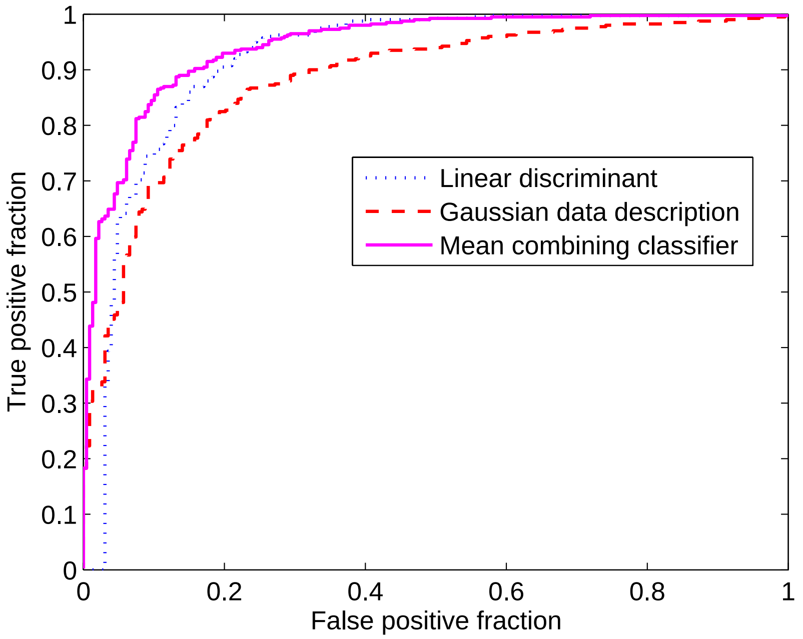

Figure 5. ROC curves for the two-class linear discriminant classifier ($A_z = 0.91$), the one-class Gaussian data description classifier ($A_z = 0.88$), and the mean combining classifier ($A_z = 0.94$).

5. CONCLUSIONS

This study has been the first one, to our knowledge, where one-class classification and the combination of two-class and one-class classifiers were applied to the automated detection of interstitial abnormalities on chest X-rays. It appears from this study that the standard two-class classification methods are very promising yet simple to be used as part of a CAD scheme.

The two-class classifiers show good performance for the classification of regions that may contain interstitial lesions on conventional chest radiographs when enough abnormal samples were presented to the classifier. However, when only few abnormal samples are available, combined schemes of the two-class and one-class classifiers can achieve higher performance than either method individually. Another – slightly disappointing – result is that already for very unbalanced training sets the two-class classifiers clearly outperform the one-class methods.

The very good classification results that we got for the two-class methods could be partly explained by the 'easy' data. There are not so many subtle abnormal cases in the database, and regions in the middle periphery of the lung fields do not normally contain a lot of variable superimposed structures like vessels that might hinder classification. These could explain why few abnormal samples were sufficient for a classifier to generalize well about the abnormal class. It should be noted that we applied both one-class and two-class classifiers on the same feature sets, and we used PCA for feature space dimensionality reduction, which could be counter-productive for classification. A subspace with large variance is not necessarily one in which the normal class is well described. One-class classifiers might perform better after deliberate selection of features that capture characteristics of the normal class.

Future research direction is the classification of regions that cover other parts of the lung fields. We might construct different training sets for different locations because of the large differences in normal texture appearances, e.g. between regions from the middle periphery and those close to the hilum. Combining one-class and two-class methods could be practical for regions were ILD is less commonly encountered, for example in the lung tops.

ACKNOWLEDGMENTS

This research was supported by the Technology Foundation STW, Applied Science division of NWO and the Technology Programme of the Ministry of Economic Affairs of the Netherlands.

REFERENCES

1. Society, British Thoracic and Committee, Standards of Care, "The diagnosis, assessment and treatment of diffuse parenchymal lung disease in adults. introduction," *Thorax* **54(Suppl 1)**, pp. S1–S14, 1999.
2. S. Padley, D. M. Hansell, C. Flower, and P. Jennings, "Comparative accuracy of high resolution computed tomography and chest radiography in the diagnosis of chronic diffuse infiltrative lung disease," *Clin Radiol* **44**(4), pp. 222–226, 1991.
3. E. Kazerooni, "High-resolution CT of the lungs," *Am J Roentgenol* **177**(3), pp. 501–519, 2001.
4. S. Katsuragawa, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography: detection and characterization of interstitial lung disease in digital chest radiographs," *Med Phys* **15**(3), pp. 311–319, 1988.
5. S. Katsuragawa, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography: classification of normal and abnormal lungs with interstitial lung disease in chest radiographs," *Med Phys* **16**(1), pp. 38–44, 1989.
6. T. Ishida, S. Katsuragawa, T. Kobayashi, H. MacMahon, and K. Doi, "Computerized analysis of interstitial disease in chest radiographs: improvement of geometric-pattern feature analysis," *Med Phys* **24**(6), pp. 915–924, 1997.
7. T. Ishida, S. Katsuragawa, K. Ashizawa, H. MacMahon, and K. Doi, "Application of artificial neural network for quantitative analysis of image data in chest radiographs for detection of interstitial lung disease," *J Digit Imaging* **11**(4), pp. 182–192, 1998.
8. S. Kido, S. Tamura, N. Nakamura, and C. Kuroda, "Interstitial lung disease: evaluation of the performance of a computerized analysis systems versus observers," *Comput Med Imaging Graph* **23**(2), pp. 103–110, 1999.
9. B. van Ginneken, S. Katsuragawa, B. M. ter Haar Romeny, K. Doi, and M. A. Viergever, "Automatic detection of abnormalities in chest radiographs using local texture analysis," *IEEE Trans Med Imag* **21**(2), pp. 139–149, 2002.
10. D. Tax, *One-class classification; Concept-learning in the absence of counter-examples.* PhD thesis, Delft University of Technology, the Netherlands, June 2001.
11. F. van der Heiden, R. Duin, D. de Ridder, and D. Tax, *Classification, parameter estimation, state estimation: an engineering approach using MatLab*, Wiley, New York, 2004.
12. D. Tax, M. van Breukelen, R. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?," *Pattern Recognition* **33**, pp. 1475–1485, 2000.
13. C. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology* **21**(9), pp. 720–733, 1986.