Physics in Medicine & Biology

PAPER

Malignant-benign classification of pulmonary nodules based on random forest aided by clustering analysis

To cite this article: Wenhao Wu *et al* 2019 *Phys. Med. Biol.* **64** 035017

IOP Publishing Phys. Med. Biol. 64 (2019) 035017 (12pp) https://doi.org/10.1088/1361-6560/aafab0

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

Malignant-benign classification of pulmonary nodules based on random forest aided by clustering analysis

RECEIVED
12 October 2018

REVISED
7 December 2018

ACCEPTED FOR PUBLICATION
21 December 2018

PUBLISHED
31 January 2019

Wenhao Wu[1,3], Huihui Hu[1,3], Jing Gong[1], Xiaobing Li[1], Gang Huang[2] and Shengdong Nie[1,4]

[1] School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai 200093, People's Republic of China
[2] Shanghai University of Medicine & Health Science, Shanghai 201318, People's Republic of China
[3] These authors contributed equally to this work and should be considered co-first authors.
[4] Author to whom any correspondence should be addressed.

E-mail: nsd4647@163.com (S Nie)

Keywords: CT images, pulmonary nodule, computer-aided diagnosis, random forest, class decomposition

Abstract

To help the radiologists better differentiate the benign from malignant pulmonary nodules on CT images, a novel classification scheme was proposed to improve the performance of benign and malignant classifier of pulmonary nodules.

First, the pulmonary nodules were segmented with the references to the results from four radiologists. Then, some basic features of the segmented nodules such as the shape, gray and texture are given by calculation. Finally, malignant-benign classification of pulmonary nodules is performed by using random forest (RF) with the aid of clustering analysis.

The data with a set of 952 nodules have been collected from lung image database consortium (LIDC). The effect of proposed classification scheme was verified by three experiments, in which the variant composite rank of malignancy were got from four radiologists (experiment 1: rank of malignancy '1', '2' as benign and '4', '5' as malignant; experiment 2: rank of malignancy '1', '2', '3' as benign and '4', '5' as malignant; experiment 3: rank of malignancy '1', '2' as benign and '3', '4', '5' as malignant) and the corresponding (A_z) (area under the receiver operating characteristic curve) are 0.9702, 0.9190 and 0.8662, respectively.

It can be drawn that the method in this work can greatly improve the accuracy of the classification of benign and malignant pulmonary nodules based on CT images.

1. Introduction

Lung cancer is the second commonest cancers for women and the most cause of death for men in the world. The high mortality rate is usually associated with the late diagnosis, and the early diagnosis of lung cancer can significantly increase the five-year survival rate from 4% to 55% compared to the advanced-stage diagnosis (Society 2010). In lung parenchyma, the pulmonary nodules, blob-like structures with the size of 3 to 30 mm in diameter manifest the lung cancer and indicate the stage of the disease (Gould *et al* 2013). With the development of CT scanning technology, more and more human anatomical information can be obtained from CT images, which make the lung lesions be recognized more easily (Menezes *et al* 2010). It's promising to substantially reduce the mortality rate of lung cancer by raising the diagnosis rate of pulmonary nodules at an earlier and more curable stage. However, a mass of CT images limits the number of daily diagnosis for radiologist because the diagnosis still relies on visual inspection now, which is very difficult and prone to fatigue. It has been reported that the misdiagnosis rate will rise to 30% with the extension of working hours (Venjakob *et al* 2012). So the computer-aided diagnosis (CADx) system is imperative for radiologists to improve the diagnosis accuracy as a second reader.

The traditional method of computer-aided classification for malignant-benign pulmonary nodules include three steps: the segmentation of pulmonary nodules, the extraction and selection of nodule features, and classification (Firmino *et al* 2014). Correspond, there are three major factors impacting on the classification: (a)

the accuracy of pulmonary nodule segmentation; (b) the representativeness of nodule characterization; (c) the performance of the classifier. Over the years, many efforts have been made to optimize the performance of factors in every step. Way *et al* (2009) provide some new nodule surface features to characterize the lung nodule surface smoothness and shape irregularity, and the reported A_z increased up to 0.857 from 0.821 when the linear discriminant analysis (LDA) were applied to a data set of 256 nodules. The 2D shape-based, 3D shape-based, 3D margin-based, 2D texture-based and 3D texture-based features were combined by Dhara *et al* (2016a) to improve the classification performance of pulmonary nodules. A data set of 891 nodules from LIDC (Armato *et al* 2010) was used and the reported A_z were 0.9505, 0.8822 and 0.8488 using SVM (support vector machine) on three configurations. Suzuki *et al* (2005) developed a computer-aided diagnostic (CAD) scheme for distinction between benign and malignant nodules by using a massive training artificial neural network (MTANN) and the reported A_z was 0.882 on a data set of 76 malignant and 413 benign nodules from LIDC. Lee *et al* (Lagerkvist *et al* 2010) proposed a CADx using a two-step supervised learning system combining GA (genetic algorithm) with the RSM (random subspace method). The reported A_z was 0.889 on a data set of 125 pulmonary nodules (63 benign; 62 malignant) and higher than that in the RSM and the GA-LDA. Tartar *et al* (2014) proposed a CADx using a two-step supervised learning system combining GA (genetic algorithm) with the RSM (random subspace method). The reported A_z was 0.889 on a data set of 125 pulmonary nodules (63 benign; 62 malignant) and higher than that in the RSM and the GA-LDA. Tartar *et al* (Shen *et al* 2017) presented a multi-crop convolutional neural network (MC-CNN) to automatically extract nodule salient information by employing a novel multi-crop pooling strategy which cropped different regions from convolutional feature maps and then applied max-pooling different times, avoiding relying on cautious segmentation of nodules and time-consuming feature extraction.

Based on the review of related literature, it can be found that most of studies focused on improving the accuracy of nodule segmentation and representativeness of nodule characterization, but few works concerned about designing or modifying classifiers to improve the performance of CADx schemes. The previous classification schemes, whether the single classifier or the ensemble classifier, always directly divided nodules into two categories (benign and malignant) without considering the internal structure of samples within a class, making the classification performance of existing classifiers almost reach bottlenecks in classification of malignant-benign pulmonary nodules. Classifier design is the key step to determine overall performance (Chen *et al* 2012). However, there are three critical limitations in the classifier design process for malignant-benign nodules classification: (1) the restricted training data with exacting class labels due to cost, time, and availability to patient clinical information; (2) the enormous variance in the appearance of nodules. It is difficult to describe and analyze such variance with a single classifier; (3) the limited performance of existing classifiers in classification of malignant-benign nodules.

Meanwhile, Pulmonary nodules contains multiple latent types, such as solid, part-solid and non-solid nodules, it is not easy for a classifier to directly divide them into benign and malignant nodules. The complex distribution degrades the discrimination of malignant-benign nodules and may lead to high misdiagnosis rate or over-fitting problem. The decomposition methodology (Apte *et al* 1998) can break down a complex problem into several manageable sub-problems to achieve a better result with less running time, including roughly following categories: functional modularity, domain modularity, class decomposition and state decomposition, according to different partition strategies (Guan and Zhu 2005).Thus, it is helpful for classification tasks to combine the class decomposition with typical classifiers by analyzing internal latent distribution of the dataset (Rahman and Verma 2013, Elyan and Gaber 2016).

To overcome the above issues, this study developed CADx scheme by designing or modifying the classifier. Therefore, we propose a novel classification scheme which combines clustering with RF by applying class decomposition and tuning weights to distinguish benign nodules from the malignant ones. We first segment pulmonary nodules by consulting the segmented results drawn by four experienced radiologists. Then, some based-shape, based-gray, and based-texture features are computed to represent pulmonary nodules. Finally, in the process of classification, *k*-means clustering algorithm is employed to form multiple sub-clusters within a class. Then we can train RF using the new data, and form a mapping relationship between decomposed subclasses and the real classes. For a test sample, the preliminary class is decided according to the maximum weighted category distribution value, and the weights are reciprocals of the Euclidean distances between the test sample and centers of multiple clusters. The final predicted class is obtained according to the established mapping relationship.

In the present work, the proposed method is discussed in section 2 to improve the classification performance of malignant-benign pulmonary nodules by RF. Then thorough experiments were presented in section 3 to verify the validity of the proposed method. Finally, the discussion and conclusion are given in section 4.

2. Methods

We referenced the segmentation results drawn by LIDC's four radiologists to segment the lung nodules, avoiding the effects of inaccurate segmentation. A total of 50 features were extracted. And the feature selection

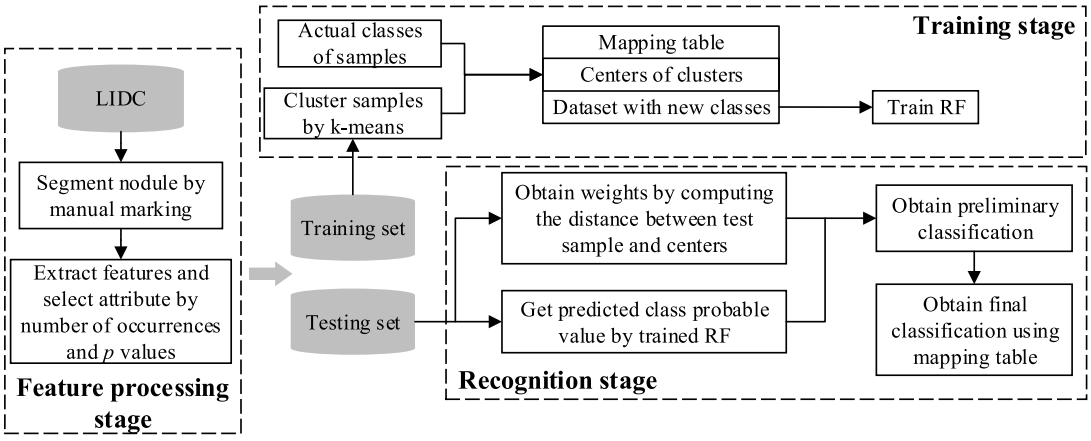

Figure 1. The specific flow chart of the three stages in the classification scheme and the connection between them.

was implemented according to the criterion that the number of occurrences is greater than or equal to 3 in the statistical significance analysis and p is less than 0.05 in the correlation-based analysis. Ultimately, we selected 18 features from the feature candidates.

In this paper, we propose a classification scheme which combines the clustering with RF by applying class decomposition and tuning weights for malignant-benign classification of pulmonary nodules. Figure 1 provides an overview of the proposed classification scheme. As shown in figure 1, this scheme largely consists of two parts: (1) feature extraction, (2) classification including training stage and recognition stage. Each of these steps will be described in more detail in the following sections.

2.1. Feature extraction

It is important for a classifier to get more relevant and discriminatory input data. In the process of malignant-benign classification of pulmonary nodules, nodule features as the input data is computed to represent segmented nodules. Therefore the feature processing is the first step for malignant-benign classification of pulmonary nodules, including the segmentation of pulmonary nodules, feature extraction and selection of segmented nodules.

Several works on segmentation of nodule (e.g. threshold method, region growing and classification of segmentation (Ahuja 2002, Dhara *et al* 2016b), etc) are reported in the literature. But no matter what method is designed, their reference standards are the manual markings from experienced radiologists. In this paper, we segment pulmonary nodules using manual marking by experienced radiologists, partly reducing the influence of segmentation errors. The nodules collected from LIDC are marked by four experienced radiologists and we obtain segmented pulmonary nodules by consulting the segmented results from four radiologists with Probability Map, then the p-map is segmented with a threshold of 0.5 to obtain the final segmentation result (Meyer *et al* 2006), as shown in figure 2.

Feature extraction is the conversion process of image attributes from visual information to digitization. The surface of malignant nodules is uneven due to the uncontrollable growth, whereas benign nodules have smooth surface. So, a total of 50 features about gray, shape and texture are computed from segmented nodules in this work. We calculated the area under the curve (A_z) of 50 features and the confidence interval (CI) of 95%, and selected 18 features with A_z values greater than 0.6 which have higher performance in lung nodule classification. The results are shown in table 1. And Relevant attribute selection is also necessary to reduce the irrelevant and redundant features, improving the chances of avoiding over-fitting and reducing the running time. The relevant features are selected using statistical significance analysis and correlation-based analysis (Ferreira and Oliveira 2017). The statistical significance analysis is performed with p values using 2-tailed Student's t test by SPSS 22.0. Correlation-based feature selection is used to find a small subset of features that are highly correlated with the class while having low inter-correlation. This is implemented using a stratified 10-fold cross-validation by the MATLAB software v9.1.0, and the number of occurrences within ten times represents the relevance on distinguishing malignant and benign pulmonary nodules.

2.2. Classification

At the classification stage, the main used classifier is RF. It sets up a forest in a random way. There are many decision trees in the forest, and there is no correlation between each decision tree in the random forest.

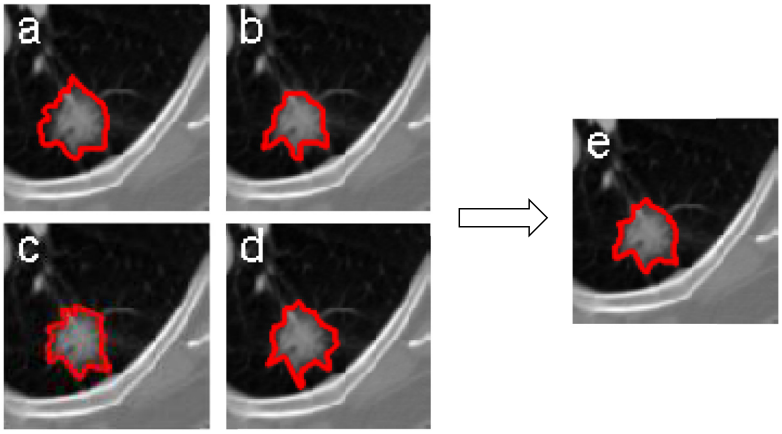

Figure 2. An example of a nodule with four segmented regions (a)–(d) from radiologists. Regions (e) represents the final segmentation of the nodule.

Table 1. The 18 features selected from candidate features.

Features	A_z	95% CI
Intensity feature-variance	0.83 ± 0.01	[0.82, 0.85]
Shape feature-lobulation (2D)	0.53 ± 0.01	[0.50, 0.56]
Shape feature-compactness (2D)	0.87 ± 0.01	[0.85, 0.88]
Shape feature-Fourier descriptor2 (2D)	0.87 ± 0.01	[0.86, 0.89]
Shape feature-sphericity (3D)	0.60 ± 0.01	[0.57, 0.62]
Shape feature-compactness (3D)	0.93 ± 0.01	[0.92, 0.95]
GLCM texture feature-entropy (middle layer, orientations = 0°)	0.94 ± 0.01	[0.93, 0.96]
GLCM texture feature-maximum probability (middle layer, orientations = 90°)	0.77 ± 0.01	[0.75, 0.79]
GLCM texture feature-energy (MIP, orientations = 0°)	0.86 ± 0.01	[0.84, 0.87]
GLCM texture feature-entropy (MIP, orientations = 45°)	0.93 ± 0.01	[0.92, 0.94]
GLCM texture feature-energy (MIP, orientations = 90°)	0.86 ± 0.01	[0.84, 0.87]
GLCM texture feature-sum entropy (MinIP, orientations = 0°)	0.86 ± 0.01	[0.84, 0.88]
GLCM texture feature-sum of squares (MinIP, orientations = 0°)	0.68 ± 0.01	[0.65, 0.70]
GLCM texture feature-energy (MinIP, orientations = 90°)	0.85 ± 0.01	[0.83, 0.87]
GLCM texture feature-energy (AIP, orientations = 0°)	0.93 ± 0.01	[0.92, 0.94]
GLCM texture feature-sum entropy (AIP, orientations = 0°)	0.94 ± 0.01	[0.93, 0.95]
GLCM texture feature-entropy (AIP, orientations = 45°)	0.93 ± 0.01	[0.92, 0.95]
GLCM texture feature-maximum probability (AIP, orientations = 135°)	0.85 ± 0.01	[0.83, 0.87]

After getting the forest, each decision tree in the forest will make a separate judgment to see which class the sample should belong to, then the mostly selected type and sample will be predicted.

Considering the internal structure of nodule samples within a class, the overlapping of multiple sub-classes may cause the difficulty for a classifier to directly divide them into benign and malignant nodules, and may result in the over-fitting problem. To solve this problem, we consider not only the spatial distribution of nodule training samples but also characteristics of testing samples. Hence, we utilize the class decomposition to fully explore the internal distribution of nodule data, and use weights to reflect the similarity degree between testing samples and multiple found clusters. By considering the two aspects above, a novel classification approach of RF aided by clustering analysis is proposed to improve the classification performance of malignant-benign pulmonary nodules.

As shown in figure 1, the classification component of our scheme includes two stages: the training and the recognition. In the training stage, we train RF using the new class decomposed data, and form a mapping relationship between decomposed subclasses and the real classes. In the recognition stage, for a querying sample, the preliminary class is obtained by choosing one whose weighted category distribution value is the maximum and the final predicted class is obtained according to the established mapping relationship. The details of two stages are given in the following sections.

2.2.1. The training procedure

First, class decomposition was applied to the training set to form a new training data. The heterogeneous clustering with patterns from different classes and the homogeneous clustering with patterns within a class are two different approaches for class decomposition, and the homogeneous clustering plays a more important part in improving classification performance (Verma and Rahman 2012). Thus k-means clustering algorithm is employed to achieve homogeneous class decomposition in this paper. Further description is shown as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \cdots & x_{22} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & \cdots & \cdots & x_{mn} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \cdots \\ \cdots \\ y_m \end{bmatrix} \tag{1}$$

where X represents the feature vector of m number of samples and each sample is represented by an n number of features as $x_i = (x_{i1}, x_{i2}, \cdots, x_{in})$. $y_i \in Y$ represents the true class of ith sample. If X is a two-category samples, Y can also be expressed as in equation (2). Y' is the converted form of Y after class decomposition, where c and d are numbers of clusters within each class in X. It should be noted here that both c and d are varying due to the unpredictability of internal structure of the dataset, which will be discussed in the experimental section. Meanwhile, we can form a mapping relationship between decomposed subclasses and the real classes

$$Y = \begin{bmatrix} a_0 \\ a_0 \\ \cdots \\ a_1 \\ a_1 \\ \cdots \end{bmatrix} \Rightarrow \quad Y' = \begin{bmatrix} a_{01} \\ \cdots \\ a_{0c} \\ a_{11} \\ \cdots \\ a_{1d} \end{bmatrix}. \tag{2}$$

Then we can train a classification model with the new training set. With the development of the classification study, the ensemble classifiers can achieve more satisfactory effect than single classifiers. RF emerged as an accurate classifier which can perform better than other ensemble classifiers (Breiman 2001). And RF has been used widely in the computer-aided medical fields due to simple decision-making mechanism and good classification performance (Azar *et al* 2014). Mainly the two effective settings of RF are *ntrees*—the number of trees in the ensemble, and *mfeatures*—the number of features to be assessed for goodness at each split point of any tree. Hence, we choose RF to establish the scheme for malignant-benign classification of pulmonary nodules.

Let us assume that we have applied RF to the new training set obtained by combining X with Y', and the trained classification model (Q) is obtained. Supposing continuously that we input testing samples into Q, we can obtain a confusion matrix (h) shown in equation (3). We can evaluate the performance of Q in term of accuracy denoted by equation (4)

$$h = \begin{bmatrix} & a_{01} & \cdots & a_{0c} & a_{11} & \cdots & a_{1d} \\ a_{01} & p & \cdots & q & r & \cdots & s \\ \cdots & & & & & & \\ a_{0c} & . & \cdots & . & . & \cdots & . \\ a_{11} & . & \cdots & . & . & \cdots & . \\ \cdots & & & & & & . \\ a_{1d} & . & \cdots & . & . & \cdots & . \end{bmatrix} \tag{3}$$

$$acc = \frac{\sum_{i=a_{01}}^{a_{0c}} h_i + \sum_{j=a_{01}}^{a_{0d}} h_j}{u} \tag{4}$$

where u is the number of instances in the testing set, and we can see from equation (4) that we not only sum all the diagonal elements in h, but also count all elements within the same clusters even if they are not on the diagonal of the matrix, which are also considered as correct classifications. The fault-tolerance of the classification model is enhanced and so the accuracy is improved accordingly.

In general, at this stage, the classification mode can be obtained by applying the RF to the decomposed data with new labels obtained by implementing the *Homogeneous clustering* algorithm. The pseudocode of *Homogeneous clustering* is shown in figure 3 ($KValue = [c, d]$, a two-element vector, is considered to represent the cluster numbers of each class in the two-category nodule dataset). Then the trained RF, centers of multiple clusters, the mapping relationship between the true classes of samples and the decomposed ones are got in this stage.

Pseudocode of *Homogeneous Clustering*

Input: sample matrix X; class label Y; vector of cluster numbers $KValue$

Output: class decomposed data set D

Method :

1. $Name_{class} = unique(Y)$.
2. $for \quad c = 1 : length(Name_{class})$.

   $[x_{ij}{}^{c}] =$ examples in X whose class equal to $Name_{class}(c)$.

   partition $[x_{ij}{}^{c}]$ into $KValue(c)$ clusters by $k-means\ clustering$.

   form new class label $[L_i{}^{c}]$ using $Name_{class}(c)$ and clusting cluster.
3. end
4. $D = [X, L]$.

Figure 3. The pseudocode of homogeneous clustering for partitioning data within a class into multiple clusters.

2.2.2. The recognition procedure

In the recognition phase, given an input pattern, the trained classifier (RF_D) output a vector of predicted category distribution values. To further improve classification accuracy, we recommend weighing the predicted class distribution values to adjust accuracy.

Formally, let T be an input testing sample with a feature vector, K be the number of all sub-clusters obtained in the process of class decomposition, $P(C_i/T)$ $(i = 1, \cdots K)$ be the predicted category distribution values outputted by RF_D for representing the membership degree that T belongs to ith class. First, we determine the weights by computing the reciprocal of the distance between the testing sample and cluster centers, as follows:

$$w_i = \frac{1/dist\,(T, Center_i)}{\sum_{i=1}^{K} {}^{1}\!/_{dist(T,Center_i)}} \quad i = 1, \cdots K \tag{5}$$

$$class_{prel} = \underset{i \in \{1,2\cdots,K\}}{\arg\max}\ (w_i * P(C_i/T)) \tag{6}$$

where *dist* represents the distance between the input pattern and multiple clusters centers. Obviously, the smaller this *dist* is, the higher the similarity degree representing that T belongs to one cluster is. And this can help RF improve classification accuracy, to some extent. Then, the weights are assigned to $P(C_i/T)$ for deciding the preliminary class of the input testing sample. And the preliminary class can be determined according to equation (6) which means choosing the class whose weighted category distribution value is the maximum. Finally, we obtain the final classification decision according to the mapping relationship established before.

The specific classification algorithm of our classification method is presented in figure 4. In Algorithm *Classification*, the meaning of *KValue* is the same as mentioned above and the *bestKValue* represents the best value of *KValue*, which means the optimal cluster numbers for all classes in the nodule dataset. In addition, considering that a bigger number of clusters within a class may against the training of model, we vary the number of clusters subject to max_C which is set as the maximum value of clusters within a class.

3. Experiments and results

3.1. Experimental setup

To validate the effectiveness of the proposed scheme, we used 952 pulmonary nodules marked by all four radiologists in LIDC database to conduct the experiments. The rank of malignancy for each nodule is a comprehensive result from four radiologists' annotations. We designed three different experiments (see table 2) for the evaluation of our classification scheme by categorizing the pulmonary nodules into benign and malignant based on the composite rank of malignancy (Dhara *et al* 2016a). To verify the accuracy of the classification for benign and malignant lung nodules, an experiment named experiment 1 was carried out in the LIDC with radiologists, which labels benign and malignant lung nodules clearly. In Experiments 2 and 3, the data of 'unknown' is placed in a benign or malignant group. As the uncertainty of these data, they were used to verify the enhancement of the fault tolerance of the classification scheme.

The control groups of the comparison experiment were taken as follows: (1) the typical RF, SVM, NB (Naive Bayesian), and LDA using the same dataset obtained in this study, (2) prior works about classification of benign and malignant pulmonary nodules (Suzuki *et al* 2005, Lagerkvist *et al* 2010, Nascimento *et al* 2012, Madero *et al* 2015, Dhara *et al* 2016a). The RF was initialized with the following parameter $ntrees = 60\% * m$ and $mFeature = [\log_2(n) + 1]$. m is the number of instances in the training set and n is the number of features.

Algorithm *Classification*
Input: sample set ***X***; class labels ***Y***; the max of cluster number within a class **max_*C***
Output: trained classifier model ***f***; the vector of optimal cluster numbers ***bestKValue***
Method:
1. Divide ***X*** into training set and testing set, ***trainX*** and ***testX***.
2. Obtain the cluster numbers for all classes ***KValue*** by grid search strategy.
3. Get new dataset *D* and multiple sub-cluster centers by applying *Homogenous clustering* to ***trainX***.
4. Train RF with *D*.
5. for the testing set, ***testX***
   Computer *w* = the reciprocal of the Euclidean distance between ***testX*** and centers.
   Obtain category distribution values, *classprobs*, by the trained RF.
   Weigh *classprobs* by *w* and obtain the initial classes = labels whose weighted category distribution values are the maximum.
   Obtain the final classes according to the relationship between decomposed classes and true ones.
   Compute accuracy and save.
6. Repeat steps 1-5 until the ***max_C*** is achieved.
7. Obtain the vector of optimal cluster numbers ***bestKValue*** and model ***f*** whose accuracy is the maximum.

Figure 4. Algorithm of the proposed method.

Table 2. The description of pulmonary nodule in three experiments.

Description	Experiment	Benign	Malignant
'1', '2' as benign and '4', '5' as malignant	1	294	320
'1', '2' as benign and '3', '4', '5' as malignant	2	294	658
'1', '2', '3' as benign and '4', '5' as malignant	3	632	320

And all these classifiers were configured with the default parameters. In addition, we set the maximum number of clusters within a class, $\max\_C = 6$, which was obtained experimentally for many time—that increasing the value beyond this number does not improve the performance of our method with the data of pulmonary nodules. Training and testing sets were constructed by using 10-fold cross-validation approach. To ensure the accuracy of the experiment, all experiments were implemented on the MATLAB platform and each of these approaches was performed over 20 times for each experiment.

The evaluation criteria of classification model were listed as follows: (1) accuracy (Acc), the proportion of correctly classified examples to the number of all examples; (2) true positives rate or sensitivity (Sen), the proportion of actual positives which are correctly identified; (3) true negative rate or specificity (Spe), the proportion of negatives which are correctly identified. If a positive example can be recognized correctly by the algorithm, we call it 'true positive' (TP); otherwise we call it 'false negative' (FN). The means of 'true negative' (TN) and 'false positive' (FP) are defined similarly

$$Sen = TP/(TP + TN) \quad 0 \leqslant Sen \leqslant 1 \tag{7}$$

$$Spe = TN/(TN + FP) \quad 0 \leqslant Spe \leqslant 1 \tag{8}$$

$$Acc = (TP + TN)/(TP + TN + FP + FN) \quad 0 \leqslant Acc \leqslant 1. \tag{9}$$

Area under the receiver operating characteristic curve (A_z) was also applied to evaluate and analyze the performance of our new scheme (van Erkel and Pattynama 1998).

3.2. Results and analysis

The training samples of one sub-class will be too few if the number of clusters within a class increases limitlessly. And in that case, the trained model may be inadequate. For the purpose of finding the optimal value of sub-clusters within each class, we vary the number of clusters subject to $\max\_C$. As shown in table 3, the vector representing the best numbers of sub-clusters for all classes in nodule dataset, *bestKValue* was obtained for three experiments through multiple iterations by using the grid search strategy (Bergstra and Bengio 2012). It can be seen from the heat map of figure 5 that the higher accuracy can be obtained when the cluster number is smaller, which means there are fewer potential types in the data in experiment 1. For experiments 2 and 3, there are more potential types in benign or malignant sample data, so we need to increase the number of clusters to improve

Table 3. The vector of optimal cluster numbers for experiment 1, experiment 2, and experiment 3.

Accuracy (%)	*bestKValue*	Our method	Classical RF
Experiment 1	[1, 2]	92.37 ± 1.25	90.54 ± 1.50
Experiment 2	[6, 1]	85.62 ± 2.01	83.49 ± 2.31
Experiment 3	[3, 4]	82.05 ± 2.15	81.39 ± 2.56

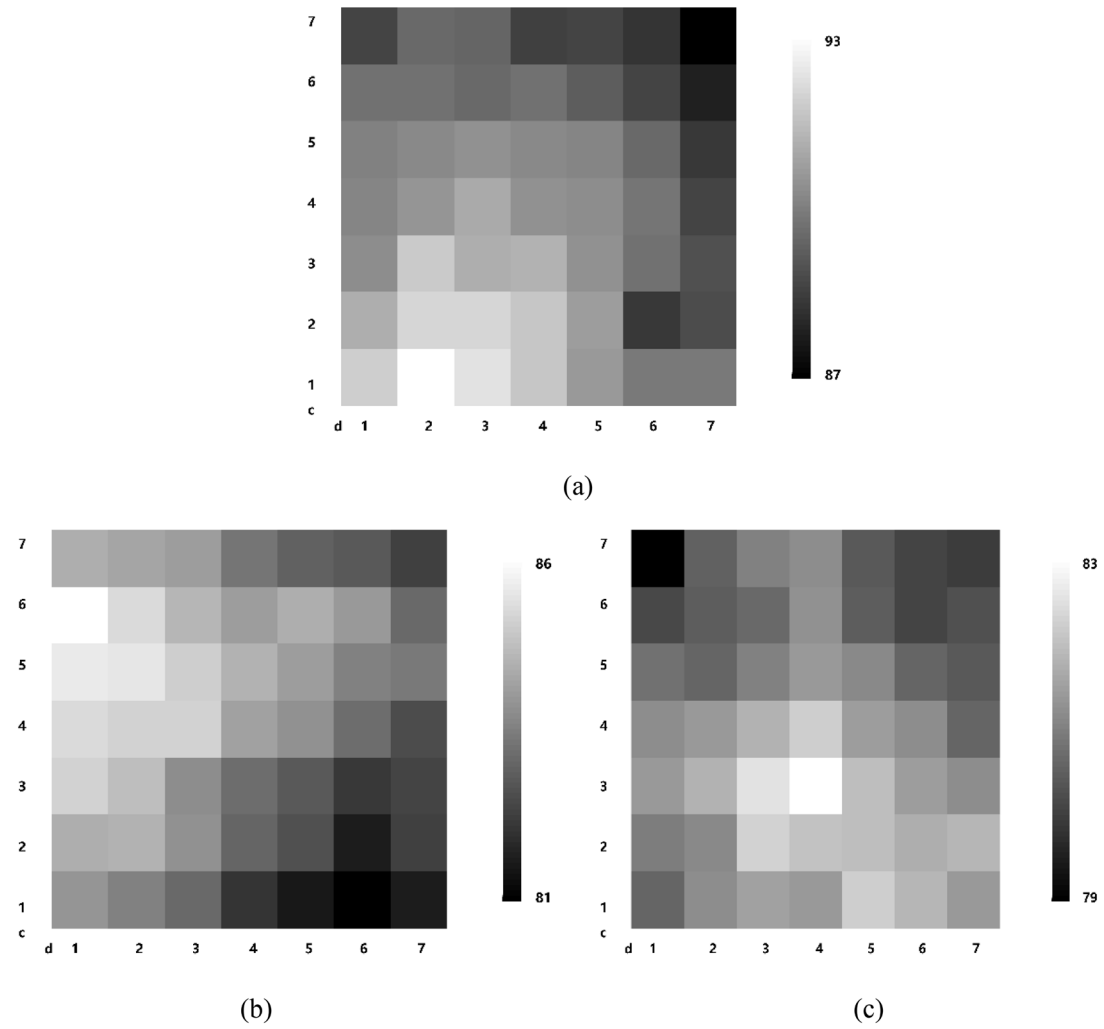

Figure 5. Figures (a)–(c) show the heat map between *bestKValue* and classification accuracy (%) for experiment 1, 2 and 3.

the classification accuracy. Table 3 also shows clearly that the accuracy of the proposed scheme is improved in lung nodules diagnosis, compared with the classical RF. And it shows us that the standard deviation of the three sets of experiments is less than 3%, which indicates that the proposed scheme is stable and reliable. One certain explanation is that there are multiple latent types in single benign or malignant nodules. When the distributions of multi-class samples are overlapping, it is difficult for a classifier to establish accurate decision boundaries. But after class decomposition, a classification model with high robustness is advantageously trained among some structured sub-regions. Another explanation is that tuning weights representing the closeness degree between testing samples and multiple found sub-clusters can play a role in the classification. And the following results proved these characteristics.

To vividly compare our method with typical RF classifiers, Let cw represent the method of class decomposition and weighting. RF_{cw} is denoted as the proposed classification method, while RF represents the typical random forest. Their ROC curves are plotted in figure 6. The red curve represents RF and the blue curve represents RF_{cw}. It can be seen that in the three sets of experiments, the RF_{cw} curve is above the RF curve. It indicates that the RF_{cw} is better in classifying malignant and benign pulmonary nodules.

To further evaluate the effectiveness of the proposed method, we compared our method with typical RF, SVM, NB, and LDA in the condition of same dataset. The results of each experiment with settings explained above are plotted in table 4. As mentioned above we let SVMcw mean combining cw with SVM, NBcw be the

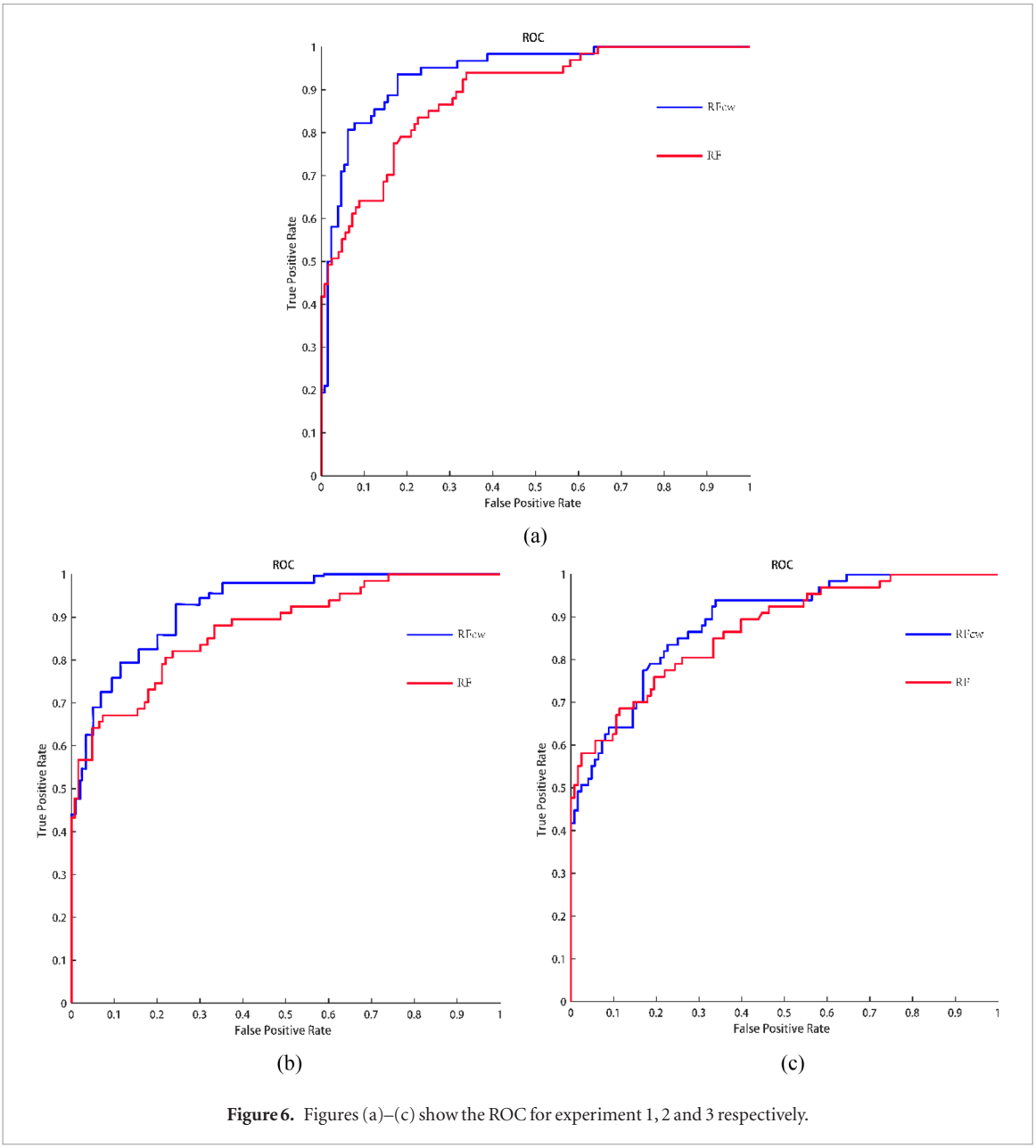

Figure 6. Figures (a)–(c) show the ROC for experiment 1, 2 and 3 respectively.

Table 4. Comparison of evaluation parameters between CLASSIFIER and CLASSIFIERcw.

Methods	Acc (%)	Sen (%)	Spe (%)	A_z
SVM	88.51	87.44	90.33	0.875
SVMcw	89.12	88.2	90.22	0.866
NB	86.32	80.24	95.8	0.902
NBcw	90.15	91.12	88.78	0.898
LDA	89.87	90.35	89.56	0.887
LDAcw	91.2	91.34	90.14	0.905

method of adding cw on NB, LDAcw mean combining cw with LDA. It is necessary to note that our aim of this step is to prove that cw is also available for other classifiers and we chose RF for establishing the classification model for malignant-benign pulmonary nodules, considering comprehensive performances for all samples.

In table 4, compared with typical classifier, the values of all evaluation criteria (except specificity) of converted classifiers are increasing. It indicates that the cw is effective. For the value of specificity, the difference between the typical classifier and its promoted modality is very small for RF, and is ignored for SVM nearly, and is badly huge for NB who has the best specificity and the worst sensitivity within these classifiers used in this paper. But the sensitivity of RF_{cw} is greatly increased compared with RF, resulting in a better classification performance. However, we obey the rule that improving the sensitivity as much as possible in the medical industry, because

Table 5. Comparison with other prior schemes for malignant-benign classification of pulmonary nodules.

Work	Method	Database	Sen (%)	Spe (%)	Acc (%)	A_z
Suzuki K *et al*	Multiple MTANNs using pixel values in a 9 × 9 sub-region	Thick-slice (10 mm) screening LDCT scans of 76 malignant and 413 benign nodules	100	48	—	0.882
Lee M C	LDA with stepwise feature selection using GA and a random subspace method	125 pulmonary nodules (63 benign; 62 malignant)	87	84	81	0.889
Nascimento L B *et al*	SVM based on texture features using diversity indexes of Shannon and Simpson	73 nodules (47 benign, 26 malignant) from LIDC	85.64	97.89	92.78	—
Madero H O *et al*	SVM using wavelet features	45 CT scans from ELCAP and LIDC	90.90	73.91	82.22	0.805
Dhara A K *et al*	SVM using a combination of 2D shape-based, 3D shape-based, 3D margin-based, 2D texture-based, and 3D texture-based features.	891 nodules from LIDC	89.73	86.36	88.58	0.9505
			82.89	80.73	82.21	0.8822
			76.14	74.91	75.16	0.8488
Our method	Several shape-based, gray-based, and texture-based features; class decomposition and RF	925 nodules from LIDC	94.28	90.27	92.37	0.9702
			87.23	82.54	85.62	0.9190
			93.56	56.59	82.05	0.8662

the costs of misclassification are different. Thus, it is worth substantially improving classification performance by adding the cw on basic classifiers at the expense of reducing a little specificity. The specificity, sensitivity and A_z are both increasing for LDA. It indicates that the proposed method cw can be applied to LDA classifier. Compared with other classification models, RF_{cw} can achieve the best performance for all samples. Thus it can be used to establish the model for malignant-benign classification of pulmonary nodules.

It was challenging to compare our scheme with other prior works reported in literature, because the experiment details (e.g. the ground truth of pulmonary nodules, the extracted features, and the parameter settings) were not explained clearly in any of the references. They usually only provided the information of which database and classifiers were used in their works. Thus, we were unable to perform a rigorous evaluation of our method with respect to other works. What we can do is to provide an overview (database, complexity of the methodology, overall performances, etc.) of the results known in the related works and our experiment. As shown in table 5, we chosen five papers reported in the last few years to compare with our work. And we can see that our scheme which improves accuracy by establishing a more appropriate classification model is promising.

For thorough comparison of these three experiments, the work by Dhara *et al* (2016a) whose experimental configurations are the same as ours is stated in table 5, representing the results of experiment 1, experiment 2, experiment 3 from top to bottom, respectively. Obviously, the classification performances (except specificity in experiment 3) of the proposed classification scheme are all better than the Dhara *et al*'s work, and the overall A_z in each configuration is still superior. It indicates that the proposed scheme is effective and promising. On the whole, overall classification performance of experiment 1 is better than other two experiments due to more specific malignancy levels of pulmonary nodules. The classification performance of experiment 2 and experiment 3 is lower slightly because uncertain nodules with rank of malignancy '3' are regarded as benign and malignant, respectively. The classification accuracy in experiment 2 is slightly higher than experiment 3, falling in line with the latent consensus that nodules with rank of malignancy '3' have much likelihood towards benign category (Dhara *et al* 2016a).

4. Discussion and conclusion

In this study, we design a novel classification scheme for classifying malignant-benign pulmonary nodules. The proposed method is evaluated on a dataset of 952 nodules collected from the LIDC. It is also compared with the existing classifiers and state-of-the-art classification schemes in classification of malignant-benign pulmonary nodules. Based on the comparative results, it can be concluded that our method has some merits which are described as follows.

- A great number of pulmonary nodules with higher determinacy marked by four radiologists are fully utilized to establish a better classifier with high reliability and applicability. The use of statistical significance analysis and correlation-based analysis in the process of feature selection is sufficient to assess the relations not only between attribute and class but also between the inter-attributes. It reduces the irrelevant and redundant

features simultaneously. It is beneficial for the process of training a better classification model to acquire more representative data.

- To overcome the limited performance of existing classifiers in malignant-benign nodules classification, our method settle the matter by establishing a classification model aided by class decomposition and tuning weights. Meanwhile, the fault-tolerance of the integral model is enhanced and the accuracy is improved accordingly. For the improvement of accuracy, it can be seen in tables 4 and 5, that the clustering-assisted lung nodule classification method has a certain improvement in accuracy compared with the traditional method, and we calculate the standard of the method. Poor to illustrate the stability of our method; for the enhancement of fault tolerance, we set up three sets of comparative experiments to verify this. The main difference between the three sets of experiments is that the processing of the tag 3 data, the tag 3 data is the data in the LIDC tag labeled 'Unknown' means that there is no evidence to prove its benign and malignant. We are not sure whether the data is benign, malignant or other types of nodules, so we set up experiments 2 and 3 to verify the reliability of the model. It can be proved that the method of clustering can handle fault data very well.
- Furthermore, compared with the prior works reported in literature, the proposed scheme is comparable in classification of malignant-benign pulmonary nodules, and this method can build more accurate decision boundaries by analyzing the inherent latent spatial distribution of samples, achieving an accuracy of 92.37%, sensitivity of 94.28%, and specificity of 90.27%.

The proposed method may be limited in the following aspects. First, the internal parameters of RF and other classifiers used in this paper are fixed with their default values. Second, the automaticity in the segmentation of pulmonary nodules is poor. We segment nodule by consulting the segmented results drawn by all four radiologists from the LIDC to reduce the influence of inaccurate segmentation. Thus more in-depth research is needed to take on segmentation of pulmonary nodules to perfect the classification scheme of benign and malignant pulmonary nodules in the future.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Number 81830052, 81530053) and the Natural Science Foundation of Shanghai (Grant Number 14ZR1427900).

References

Ahuja N 2002 Automated lung nodule segmentation using dynamic programming and EM-based classification *Proc. SPIE* **4684** 666–76

Apte C, Hong S J, Hosking J R M, Lepre J, Pednault E P D and Rosen B K 1998 Decomposition of heterogeneous classification problems *Intell. Data Anal.* **2** 81–96

Armato S, Mclennan G, Mcnitt-Gray M, Meyer C, Reeves A, Bidaut L, Zhao B, Croft B and Clarke L 2010 WE-B-201B-02: the lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed public database of CT scans for lung nodule analysis *Med. Phys.* **37** 3416–7

Azar A T, Elshazly H I, Hassanien A E and Elkorany A M 2014 A random forest classifier for lymph diseases *Comput. Methods Prog. Biomed.* **113** 465–73

Bergstra J and Bengio Y 2012 Random search for hyper-parameter optimization *J. Mach. Learn. Res.* **13** 281–305

Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

Chen H, Zhang J, Xu Y, Chen B and Zhang K 2012 Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans *Expert Syst. Appl.* **39** 11503–9

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

Dhara A K, Mukhopadhyay S, Gupta R D, Garg M and Khandelwal N 2016b Erratum to: a segmentation framework of pulmonary nodules in lung ct images *J. Digit. Imaging* **29** 148

Elyan E and Gaber M M 2016 Diversified random forests using random subspaces *Intelligent Data Engineering and Automated Learning–IDEAL 2014 Lecture Notes in Computer Science* vol 8669 (Berlin: Springer) pp 85–92

Ferreira J R Jr and Oliveira M C 2017 Selecting relevant 3D image features of margin sharpness and texture for lung nodule retrieval *Int. J. Comput. Assist. Radiol. Surg.* **12** 1–9

Firmino M, Morais A H, Mendoça R M, Dantas M R, Hekis H R and Valentim R 2014 Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects *Biomed. Eng.* **13** 41

Gould M K, Donington J, Lynch W R, Mazzone P J, Midthun D E, Naidich D P and Wiener R S 2013 Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd edn: American College of Chest Physicians evidence-based clinical practice guidelines *Chest* **143** e93S–120S

Guan S U and Zhu F 2005 A class decomposition approach for GA-based classifiers *Eng. Appl. Artif. Intell.* **18** 271–8

Lagerkvist B, Samuelsson B and Sjölin S 2010 Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction *Artif. Intell. Med.* **50** 43–53

Madero H O, Vergara O V, Cruz V S, Ochoa J D H and Nandayapa J A M 2015 Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine *Biomed. Eng.* **14** 1–20

Menezes R J *et al* 2010 Lung cancer screening using low-dose computed tomography in at-risk individuals: the Toronto experience *Lung Cancer* **67** 177–83

Meyer C R, Johnson T D, Mclennan G, Aberle D R, Kazerooni E A, Macmahon H, Mullan B F, Yankelevitz D F and van Beek E J 2006 Evaluation of lung MDCT nodule annotation across radiologists and methods *Acad. Radiol.* **13** 1254–65

Nascimento L B, Paiva A C D and Silva A C 2012 *Lung Nodules Classification in CT Images Using Shannon and Simpson Diversity Indices and SVM* (Berlin: Springer)

Rahman A and Verma B 2013 Ensemble classifier generation using non-uniform layered clustering and genetic algorithm *Knowl.-Based Syst.* **43** 30–42

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

Society A C 2010 Surveillance research *Technical Report* American Cancer Society, Inc. pp 216–7

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

Tartar A, Akan A and Kilic N 2014 A novel approach to malignant-benign classification of pulmonary nodules by using ensemble learning classifiers *36th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 4651–4

van Erkel A R and Pattynama P M 1998 Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology *Eur. J. Radiol.* **27** 88

Venjakob A, Marnitz T, Mahler J, Sechelmann S and Roetting M 2012 Radiologists' eye gaze when reading cranial CT images *Proc. SPIE* **8318** 8318OB

Verma B and Rahman A 2012 Cluster-oriented ensemble classifier: impact of multicluster characterization on ensemble classifier learning *IEEE Trans. Knowl. Data Eng.* **24** 605–18

Way T W, Sahiner B, Chan H P, Hadjiiski L, Cascade P N, Chughtai A, Bogot N and Kazerooni E 2009 Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features *Med. Phys.* **36** 3086–98