IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Automatic benign and malignant classification of pulmonary nodules in thoracic computed tomography based on RF algorithm

ISSN 1751-9659
Received on 8th December 2016
Revised 11th November 2017
Accepted on 10th March 2018
E-First on 11th April 2018
doi: 10.1049/iet-ipr.2016.1014
www.ietdl.org

Xiang-Xia Li[1], Bin Li[1] ✉, Lian-Fang Tian[1], Li Zhang[1]

[1]*School of Automation Science and Engineering, South China University of Technology, Guangzhou, Guangdong, People's Republic of China*

✉ *E-mail: binlee@scut.edu.cn*

Abstract: Classification of *benign* and *malignant* pulmonary nodules can provide useful indicators for estimating the risk of lung cancer. In this study, an improved random forest (RF) algorithm is proposed for classification of benign and malignant pulmonary nodules in thoracic computed tomography images. First, an improved random walk algorithm is proposed to automatically segment pulmonary nodules. Then, intensity, geometric and texture features based on the grey-level co-occurrence matrix, rotation invariant uniform local binary pattern and Gabor filter methods are combined to generate an effective and discriminative feature vector. Mutual information is employed to reduce the dimensionality. Finally, an improved RF classifier is trained to classify benign and malignant nodules. An appropriate feature subset is selected by the bootstrap method and an effective combination method is introduced to predict a class label. The proposed classification method on the lung images dataset consortium dataset achieves a sensitivity of 0.92 and the area under the receiver-operating-characteristic curve of 0.95. An additional evaluation is performed on another dataset coming from General Hospital of Guangzhou Military Command. A mean sensitivity and a mean specificity of the proposed method are 0.85 and 0.82, respectively. Experimental results demonstrate that the proposed method achieves the satisfactory classification performance.

1 Introduction

According to the American Cancer Society, about 1,688,780 new cancer cases would have been diagnosed and about 589,430 cancer deaths. About 13% of all newly diagnosed cancers were lung cancer and lung cancer deaths account for about 27% of all cancer deaths in 2015 in the USA [1]. Lung cancer has become the leading cause of malignancy-related deaths worldwide. One of the major causes of the high mortality rate of lung cancer is that lung cancer is unlikely to produce the apparent symptoms until the disease is advanced. About 85% of lung cancers are diagnosed after metastatic disease progression, and the follow-up treatment is not available. Lung cancer is curable if it is timely diagnosed and appropriately treated. Therefore, early diagnosis is crucial. It can improve 5 year survival rate from only 15 to 18% [1]. Lung cancer potentially manifests itself as pulmonary nodules at an early stage [2]. Computed tomography (CT) is one of the most prominent imaging modalities for inspecting and diagnosing lung cancer due to its non-invasiveness, high spatial resolution and low expense. Although most pulmonary nodules are benign, the consequences of producing malignant nodules are extremely severe. Thereby, the demands for early-stage classifications of benign and malignant pulmonary nodules have increased as well.

In clinical practise, the classification of benign and malignant pulmonary nodules is still judged by radiologists; however, it is extremely time-consuming and subjective. Therefore, automatic classification of benign and malignant pulmonary nodules has an urgent need for monitoring the disease progression, planning the treatments and predicting the patient outcomes. Unfortunately, it has received considerably less attention. The classification of benign and malignant pulmonary nodules is a very challenging problem due to low contrast, variable sizes, irregular shapes and random positions. Despite some unsupervised or supervised classification methods for classifying benign and malignant pulmonary nodules have been extensively researched over the past decade, there is still a lack of proper classification method with the high performance, which is the focus of this paper.

In this paper, an automatic segmentation and classification framework based on an improved random walker (RW) and random forest (RF) is proposed for classifying benign and malignant pulmonary nodules. The main contributions of this paper are summarised as follows:

(i) An improved RW method is proposed to automatically and accurately segment pulmonary nodules for subsequent feature extraction process. First, an automatic seed acquisition method is introduced. Then, the intensity, texture features and shape index are incorporated to construct the weight function. Finally, a new energy function is defined to obtain the final segmentation results.
(ii) In the feature extraction process, intensity, texture and geometric features are extracted from the segmented nodules. The grey-level co-occurrence matrix (GLCM)-based, rotation invariant uniform local binary pattern (LBP)-based and Gabor filter-based methods are combined to generate a discriminative texture feature vector. In addition, mutual information (MI) is employed to reduce the dimensionality of GLCM texture features, because a high-dimensional texture feature vector can increase the computational complexity.
(iii) An improved RF is applied to classify benign and malignant pulmonary nodules. An RF classifier is trained by the lung image dataset consortium (LIDC) datasets. A rational and effective composition method of all decision trees is presented to predict the class label.

The manuscript of this paper is organised as follows: in Section 2, some relevant state-of-the-art literatures are reviewed. In Section 3, we implement a detailed description of the proposed method. In Section 4, massive experimental results are shown, which include overall performance results of the improved RW and RF, performance comparisons with traditional RW and other previously published classification methods. Discussion is presented in Section 5. Finally, conclusions and future works are provided in Section 6.

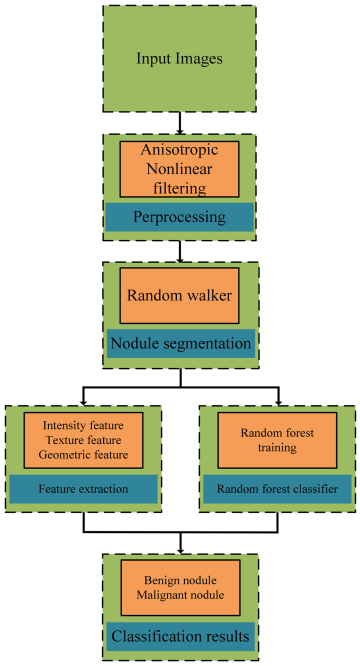


Fig. 1 *Flowchart of the proposed method*

2 Related works

In this section, we concisely review machine-learning-based methods for classifying benign and malignant pulmonary nodules in Section 2.1. We also will give a review of random walks and RFs in Section 2.2, which are closely related to our work.

2.1 Machine-learning-based methods for classification of benign and malignant pulmonary nodules

Machine-learning-based methods have been attracting more and more attentions of researchers for classifying benign and malignant pulmonary nodules. For example, Iwano *et al.* [3] calculated circularity and second central moment for characterising pulmonary nodules and the discriminant analysis was performed by using two thresholds to differentiate malignant pulmonary nodules from benign pulmonary nodules. They obtained a sensitivity of 76.9% and a specificity of 80%. Chen *et al.* [4] proposed a neural network (NN) ensemble-based computer-aided diagnosis (CAD) scheme for classifying benign and malignant pulmonary nodules, which achieved areas under the receiver-operating-characteristic (ROC) curve of 0.79. Chen *et al.* [5] also performed a comparison between artificial NNs (ANNs) and multivariable logistic regression (LR) for differentiating malignant nodules from benign nodules. ANNs had a higher classification performance than LR, which obtained an area under the ROC curve (AUROC) of 0.955, an accuracy rate of 90.0%. Lin *et al.* [6] proposed a fractal-based feature set derived from the fractional Brownian motion model for differentiating malignant nodules from benign solitary pulmonary nodules. The support vector machine (SVM) classifier was employed to distinguish malignant nodules from benign nodules. They obtained an AUROC of 0.8437, an accuracy of 83.11%, a sensitivity of 90.92%, a specificity of 71.70%, a positive predictive value of 80.05% and a negative predictive value of 87.52%. Gabor and LBP-based texture features were employed to construct three-dimensional (3D) Haralick features by Han *et al.* [7] for classifying benign and malignant nodules. Considering pulmonary nodules with a composite rank of malignancy '1' and '2' as benign and '4' and '5' as malignant, this paper achieved an AUROC of 0.94 by using an SVM classifier. Cheng *et al.* [8] proposed a deep learning-based computer-aided diagnosis (CADx) for classifying benign and malignant nodules. The stacked denoising auto-encoder was used to automatically extract features. Dhara *et al.* [9] applied an SVM for classification of benign and malignant pulmonary nodules. The shape-based, texture-based and margin-based features were extracted from semi-automated segmented nodules. They achieved an AUROC of 0.9505 when configuration 1 (composite rank of malignancies '1' and '2' as benign and '4' and '5' as malignant) was used. Liu *et al.* [10] proposed a machine-learning method to predict cancer status, which scored 24 radiological traits and built a linear classifier to classify pulmonary nodules. They had obtained an AUROC of 0.88, an accuracy of 81%, a sensitivity of 76.2% and a specificity of 91.7%. Tajbakhsh and Suzuki [11] performed a comparison between massive-training ANNs (MTANNs) and convolutional NNs (CNNs) for classification of benign and malignant nodules in CT images. MTANNs obtained an AUROC of 0.8806, which was greater than the CNNs with an AUC of 0.7755.

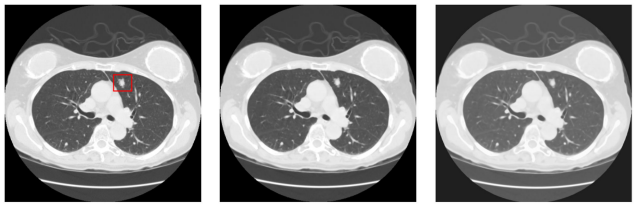
2.2 RW and RF for segmentation and classification

Recently, RW has gained much attention in the image segmentation domains due to its simplicity and effectiveness. Grady [12] first introduced RW for interactive image segmentation. In essence, RW regards the segmentation problem as a labelling assignment problem on a weighted graph. The user first marks some seeds indicating image regions belonging to $\boldsymbol{K}$ objects. Then, the user assigns a $\boldsymbol{K}$-tuple vector to each node that specifies the probability that an RW starting from each unlabelled node will first reach each of the adjacent labelled nodes. A final segmentation result can be derived from these $\boldsymbol{K}$-tuples by selecting each node the most probable seed destination. RW and its variants have been developed in computer vision applications. More closely related to our work, RWs have been successfully applied in the field of medical image segmentation [13–21]. The theoretical investigations and practical studies have demonstrated that RW is very versatile for image segmentation. Since the optimisation problem is solved by solving a sparse linear system in the Laplacian matrix of a graph, RWs perform simply and quickly. In addition, RWs have the capability of dealing with the obscure boundaries and the outliers. RWs capture the spatial information by constructing the neighbouring system. Therefore, noise limitation can be alleviated.

RF, proposed by Breiman [22] in 2001, has been widely applied as a powerful classification method, which yields the competitive classification accuracy. RF is a tree-structure classifier used both for regression and classification [23, 24], which consists of an ensemble of binary decision trees and a combination function of all decision trees. RF has many appealing advantages, making it very well-suited for classification. During the training stage, each tree is constructed by using a different bootstrap aggregation sample selected from the input data. This scheme makes RF robust against noise and immune to overfitting. It is insensitive to a large number of irrelevant features because it performs embedded feature selection. Decision trees take into account high interaction depth between variables, resulting in reducing bias. In each tree, each node is split by using the best predictor randomly chosen at the node. This scheme turns out that the RF is capable of producing high prediction accuracy. The final prediction result for a given data is the average score achieved by all decision trees or the majority vote. RF can be easily parallelised and robust to outliers and noise in the training and testing processing. RF also can deal with the high-dimensional non-linear input data. However, the drawbacks of RF are obvious. As the number of decision trees increases, the errors always converge even without pruning the trees. To address this issue, we take into account the probability density function at each feature vector of each tree.

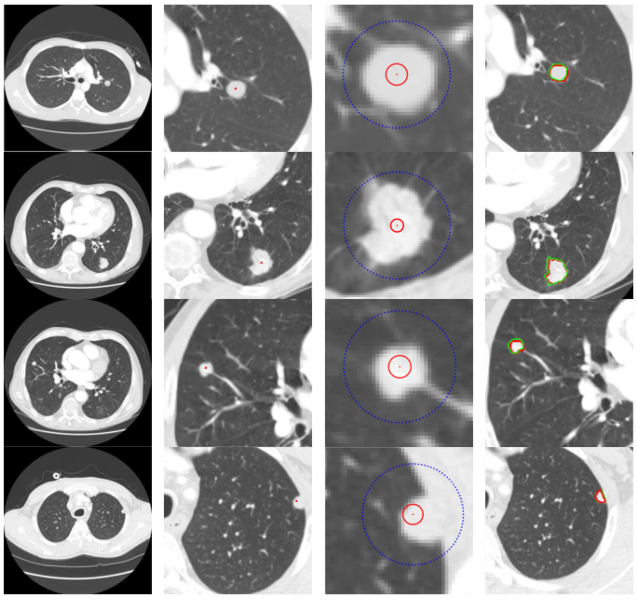
3 Method

In this paper, a novel segmentation and classification method based on an improved RW and RF is proposed for classifying benign and malignant pulmonary nodules. The flowchart of the proposed method is shown in Fig. 1. The proposed method consists of four major steps: preprocessing, pulmonary nodule segmentation, pulmonary nodule feature extraction and RF classifier, which are described in more details in the following sections.



a b c

Fig. 2 *Illustration of visual results in the preprocessing step*
(a) Original CT image, ***(b)*** Filtered result obtained by the median filter, ***(c)*** Final filtered result obtained by anisotropic non-linear diffusion filter of (b)



a b c d

Fig. 3 *Examples of pulmonary nodule segmentation obtained by the improved RW*
(a) Original CT images, ***(b)*** Nodule centres obtained by geodesic distance, ***(c)*** Nodule and background seeds. The red and blue points indicate nodule and background seeds, respectively, ***(d)*** Close-ups of the improved RW segmentation results (red) and ground truths (green)

3.1 Preprocessing

The preprocessing process is a key step in doing further the segmentation of pulmonary nodules. In this paper, an anisotropic non-linear diffusion filter is employed to suppress the image noise while preserving nodule boundaries very well. Fig. 2 shows a visual result of CT image preprocessing. As shown in Fig. 2*c*, we can see that the anisotropic non-linear diffusion filter can remove the CT image noise without blurring the nodule boundaries.

3.2 Pulmonary nodule segmentation

The accurate segmentation of pulmonary nodules is an essential step for subsequent feature extraction step. Although massive pulmonary nodule segmentation methods [25–35] have been developed over the past decade, how to precisely segment pulmonary nodules still remains an open challenge. RW can achieve the high segmentation quality in the images, whose foreground and background intensities are well separable, but it often fails to segment the images whose foreground and background share the similar intensity distributions. RW requires the user to provide some labelled pixels as parts of foreground and background regions. Thus, it is time-consuming and subjective. In addition, RW is sensitive to the locations and quantities of seeds. To address the aforementioned drawbacks, we propose an improved random walk method for pulmonary nodule segmentation. Considering the fact that the improved RW method is automatic in pulmonary nodule segmentation step, an automatic seed acquisition method is introduced. All training images are cropped from the annotations of radiologists to obtain the binary images of pulmonary nodules. Then, geodesic distance [36] is used to locate the nodules centres. Finally, nodule seeds are sampled from a circle with a radius R centred on it. The background seeds are sampled from a circle with a radius $4R$ centred on it. Nodule and background seeds are sampled adaptively according to the input image information. LBPs texture features are scale invariant and also invariant to intensity distortions. Therefore, LBP texture descriptor is calculated [37] in this paper. However, it captures too many trivial variations of images. So, the maximum response (MR) filter is employed before computing LBP histogram. The MR includes isotropic and anisotropic filters at multiple scales and orientations and records the angle of maximum response, which makes it possible to discriminate textures that appear to be very similar. To enhance the discriminative power between a pair of adjacent nodes, the intensity, texture and shape index features [38] are incorporated to construct a new weight function. The affinity entry w_{ij} from a node i to any neighbouring node j is calculated in the equation below: (see (1)) where $I(i)$, $I(j)$ and $T(i)$, $T(j)$ are intensity and texture values at nodes i and j, respectively. $SI(i)$ and $SI(j)$ are shape index values. $\mathcal{N}(i)$ is a neighbourhood system of a node i. $\| \, . \, \|_2$ denotes L_2 norm, which calculates the distance between a pair of neighbouring nodes on each feature space. To obtain the desired probabilities vector $\boldsymbol{\mathcal{F}} = [\mathcal{F}(i)]_{N\times 1}$, a new energy function is defined in equation below:

$$E(\boldsymbol{\mathcal{F}}) = \frac{1}{2}\sum_{e_{ij}\in E} w_{ij}(\boldsymbol{\mathcal{F}}(i) - \boldsymbol{\mathcal{F}}(j))^2 + \frac{1}{2}\lambda\left(\sum_{i=1}^{|V_M|}(1-\delta(\boldsymbol{\mathcal{F}}(i), b(i)))(\boldsymbol{\mathcal{F}}(i) - b(i))^2\right) \tag{2}$$

where $\mathcal{F}(i)$ and $\mathcal{F}(j)$ are probabilities of nodes i and j. $\delta(\,.\,)$ is the Kronecker delta function with $\delta(\mathcal{F}(i), b(i)) = 1$ for $\mathcal{F}(i) = b(i)$ and $\delta(\mathcal{F}(i), b(i)) = 0$ for $\mathcal{F}(i) \neq b(i)$. λ is set to ten empirically for all the experiments. $b(i)$ is a preassigned label of the node i, which is defined in the equation below:

$$b(i) = \begin{cases} -1 & \text{if } i \in V_M^B \\ 1 & \text{if } i \in V_M^F \end{cases} \tag{3}$$

The first term denotes a cost for assigning two neighbouring nodes to different probabilities, which encourages the consistence of local probabilities. The second term reflects the information of seeds and enforces the desired probability, which penalises the discrepancies between the obtained probability and the preassigned probability. Fig. 3 shows some examples of various types of pulmonary nodule segmentation obtained by the improved RW algorithm. We utilise the regions of pulmonary nodules marked by specialists, and geodesic distance is employed to obtain the nodule centres, which are shown in Fig. 3*b*. Nodule seeds and the background seeds are sampled, which are shown in Fig. 3*c*. The red and blue points indicate the nodule and background seeds, respectively. The final segmentation results are shown in Fig. 3*d*. The red and green outlines indicate the results by the improved RW and ground truths, respectively. As observed in Fig. 3*d*, the segmentation results of the proposed RW algorithm are visually close to ground truths. The improved RW algorithm is capable of segmenting various types of pulmonary nodules and achieves a desirable segmentation quality.

$$w_{ij} = \begin{cases} \exp(-\| I(i) - I(j) \|_2 - \| T(i) - T(j) \|_2 - \| SI(i) - SI(j) \|_2) & \text{if } j \in \mathcal{N}(i) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Table 1 Intensity features and their equations

Feature	Equation description
Max_Intensity	$\max\,[I(x,\,y)]$
skewness	$\frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y}\left(\frac{I(x,\,y)-u}{\sigma}\right)^3$
kurtosis	$\frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y}\left(\frac{I(x,\,y)-u}{\sigma}\right)^4-3$
$\Delta I(x,\,y)$	$\sqrt{G_x^2+G_y^2}$
RDSm	$\text{mean}\left(\frac{\sqrt{(x_n-x_c)^2+(y_n-y_c)^2}}{\max\sqrt{(x_i-x_c)^2+(y_i-y_c)^2}}\right)$
RDSv	$\text{std}\left(\frac{\sqrt{(x_n-x_c)^2+(y_n-y_c)^2}}{\max\sqrt{(x_i-x_c)^2+(y_i-y_c)^2}}\right)$

3.3 Pulmonary nodule feature extraction

After all training and testing, CT lung images are segmented completely by using the improved RW, intensity, texture and geometric features are extracted from the segmented pulmonary nodules for training an RF classifier. Three feature categories are described in more details in the following sections.

3.3.1 Intensity features: In previous nodule classification tasks, intensity feature is often used as the primary source of image information in thoracic CT images. Intensity features and their equation descriptors are described in more details in Table 1.

3.3.2 Texture features: Texture descriptor can be a characterised property of nodule surfaces such as contrast, regularity, coarseness and structural arrangement. The extraction of texture features has various methods such as GLCMs [39–43], LBP [44–46], Gabor filter [47] and fractal dimension [48]. In this paper, a hybrid method is proposed for extracting texture features. GLCMs, LBP and Gabor filter are integrated to improve the discriminating power of texture features. The GLCM is a second-order statistical method for the extraction of texture features. First, the image is translated into an l-grey-level image and GLCM is generated by counting occurrences of intensity pairs between the current and neighbour pixels for each scale and orientation. We construct a feature vector by using the average of matrices for each scale and all orientations. The normalised GLCM is calculated in the equation below:

$$G(i,\,j)=\frac{N(i,\,j)}{\sum_{m=0}^{l-1}\sum_{n=0}^{l-1}N(m,\,n)} \tag{4}$$

where i and j are grey values in the l-grey-level image. $N(i,\,j)$ is the co-occurrence relative frequency matrix defined by the equation below:

$$\begin{aligned}N(i,\,j)=\text{num}(\{[(x_1,\,y_1),\,(x_2,\,y_2)]\}\,|\,&x_2-x_1=d\cos\theta,\\&y_2-y_1=d\sin\theta,\,I(x_1,\,y_1)=i,\,I(x_2,\,y_2)=j)\end{aligned} \tag{5}$$

where $(x_1,\,y_1)$ and $(x_2,\,y_2)$ are pixel positions, and $I(\cdot)$ is the grey level of the pixel. $\text{num}(\cdot)$ denotes the number of the pixel pairs that satisfy the corresponding conditions. In this paper, 15 texture features are calculated from the corresponding GLCM. Texture features and their equations in l-GLCMs are described in more detail in Table 2. About 15 texture features from the GLCM are selected in this paper.

LBPs have brought a revolution as one of the most prominent texture descriptors due to simplicity and efficiency, which have been successfully applied to the texture extraction of focal lesions such as breast cancer [45, 49], sub-solid pulmonary nodules [50] and renal lesions [51]. LBP features were extracted by using circular neighbourhood pixels uniformly at the radius of R to the centre pixel, denoted as $\text{LBP}_{P,\,R}$. The formula is defined in the equation below:

Table 2 Texture features in CMs and their equation descriptions

Feature	Equation description
autocorrelation	$\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}(i\cdot j)P(i,\,j)$
contrast	$\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}\lvert i-j\rvert^2P(i,\,j)$
cluster prominence	$\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}(i+j-u_x-u_y)^4P(i,\,j)$
dissimilarity	$\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}\lvert i-j\rvert P(i,\,j)$
energy	$\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}P(i,\,j)^2$
entropy	$-\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}P(i,\,j)\log(P(i,\,j))$
maximum probability	$\max_{i,\,j}P(i,\,j)$
sum of squares	$\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}(i-j)^2P(i,\,j)$
sum average	$\sum_{i=0}^{l-1}i^2P_{x+y}(i)$
sum variance	$\sum_{i=0}^{l-1}\left(i-\sum_{i=0}^{l-1}i^2P_{x+y}(i)\right)^2P_{x+y}(i)$
sum entropy	$-\sum_{i=2}^{2l}P_{x+y}(i)\log\left[P_{x+y}(i)\right]$
difference variance	$\sum_{i=0}^{l-1}i^2P_{x-y}(i)$
difference entropy	$-\sum_{i=0}^{l-1}P_{x-y}(i)\log\left[P_{x-y}(i)\right]$
information measure of correlation *I*	$\frac{HXY-HXY1}{\max(HX,\,HY)}$
inverse difference normalised (INN)	$\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}\frac{P(i,\,j)}{1+\lvert i-j\rvert^2/l}$

$$\text{LBP}_{P,\,R}(c)=\sum_{i=0}^{P-1}s(g_i-g_c)\,2^i\quad s(x)=\begin{cases}1 & x\geq 0\\0 & x<0\end{cases} \tag{6}$$

where g_i and g_c denote intensity values of the neighbour pixel i and centred pixel c on a circle of radius R. $s(\cdot)$ is a sign function to ensure that LBP code is invariant against any monotonic transformation of image intensities. P is the number of sampled pixels at the radius R. If the intensity value of a neighbour pixel i is higher than or equal to that of the central pixel c, the value of $s(x)$ is set to 1; otherwise, to 0. In this paper, we obtain 2^N (e.g. $2^8=256$) distinct binary pattern in a neighbourhood.

Since edge orientation is an important malignant, rotation invariant LBP [52] is employed to extract texture features in a rotated image, which is introduced by the equations below:

$$\text{RLBP}_{P,\,R}^{ri}=\left\{\text{ROR}(\text{LBP}_{P,\,R},\,\alpha)\,|\,\alpha=\parallel\theta\cdot\frac{P}{2\pi}\parallel\right\} \tag{7}$$

$$\theta=\tan^{-1}\frac{y_m-y_c}{x_m-x_c} \tag{8}$$

$$U(\text{LBP}_{P,\,R})=\sum_{\alpha=1}^{P}\left|s(g_{\text{mod}(\alpha,\,P)}-g_c)-s(g_{\alpha-1}-g_c)\right| \tag{9}$$

$$\text{LBP}_{P,\,R}^{\text{riu}}=\begin{cases}\sum_{\alpha=0}^{P-1}s(g_\alpha-g_c) & \text{if } U(\text{LBP}_{P,\,R})\leq 2\\P+1 & \text{otherwise}\end{cases} \tag{10}$$

where $\text{ROR}(\text{LBP}_{P,\,R},\,s)$ denotes a circular s-step bit-wise right shift on the P-bit pattern binary string $\text{LBP}_{P,\,R}$, s times. $(x_m,\,y_m)$ and $(x_c,\,y_c)$ are coordinates of neighbour pixels and the centre pixel of

the nodule. This yields a lower-dimensional texture feature descriptor around the centre pixel of the nodule. In practise, $\mathrm{LBP}^{\mathrm{riu}}_{P,R}$ has $P+2$ distinct output pattern values. Therefore, we have a total of 70 texture features from the LBP.

The Gabor filter is rotationally invariant and appropriate to highlight the texture features. A Gabor filter is a multiplication of a Gauss distribution by a harmonic, which is formulated in the equations below:

$$\psi(x,\ y) = \exp\left(\frac{-x^2+\gamma y^2}{2\sigma^2}\right)\cos\left(2x\frac{x'}{\lambda}+\varphi\right) \tag{11}$$

$$x' = x\cos\theta + y\sin\theta \tag{12}$$

$$y' = -xv\sin\theta + y\cos\theta \tag{13}$$

where x and y are x and y axes pixel coordinates, respectively. The parameter σ denotes the standard deviation of 2D Gaussian envelope. λ and θ are wavelength and orientation, respectively. φ and γ are phase shift and spatial aspect ratio, respectively. In this paper, eight orientations $[\theta = 0, (\pi/8), \ldots, (7\pi/8)]$, two wavelengths ($\lambda = 1, 2$) and two standard deviations ($\sigma = 1, 2$) are applied to extract Gabor texture features. The corresponding maximum amplitude is defined in the equation below:

$$I'(x,\ y) = \max_k I^{\mathrm{G}}_k(x,\ y) \tag{14}$$

where $I^{\mathrm{G}}_k(x,\ y)$ is the filtered image by a set of Gabor filters. k is the number of Gabor filters with a different orientation θ. The eight orientations $[\theta = 0, (\pi/8), \ldots, (7\pi/8)]$, two wavelengths ($\lambda = 1, 2$) and two standard deviations ($\sigma = 1, 2$) are applied to extract Gabor texture features. Therefore, we have a total of 64 texture features from the Gabor filter. MI [43] measures the level of statistical dependence among different variables, which is defined in the equation below:

$$\mathrm{MI}(\boldsymbol{v};\ \boldsymbol{v}') = \sum_{i,\ j} p(\boldsymbol{v}_i,\ \boldsymbol{v}'_j)\log\frac{p(\boldsymbol{v}_i,\ \boldsymbol{v}'_j)}{p(\boldsymbol{v}_i)p(\boldsymbol{v}'_j)} \tag{15}$$

where $\boldsymbol{v}_i$ and $\boldsymbol{v}'_j$ are two random feature vectors and $p(\boldsymbol{v}_i,\ \boldsymbol{v}'_j)$ is the joint probability density function. $p(\boldsymbol{v}_i)$ and $p(\boldsymbol{v}'_j)$ denote marginal probability density functions of feature vectors $\boldsymbol{v}_i$ and $\boldsymbol{v}'_j$, respectively. To reduce the redundancy, the selection of texture features is executed by minimising the MI, which is defined in the equation below:

$$\min\ R(\mathscr{H})R = \frac{1}{n}\sum_{\boldsymbol{v}_i,\ \boldsymbol{v}'_j \in \mathscr{H}} \mathrm{MI}(\boldsymbol{v}_i;\ \boldsymbol{v}'_j) \tag{16}$$

where n is the number of texture features in a feature space $\mathscr{H}$. The goal of texture feature selection in terms of MI is to rank the all feature sets in agreement to minimal-redundancy–maximal-relevance (mRMR) criterion.

3.3.3 Geometric features: Several 2D geometric features are computed from the segmented nodules. Major axis length and minor axis length are the major and minor axes lengths of the bounding ellipse of the nodule, respectively. The perimeter is the perimeter of a nodule multiplied by pixel resolution. Area is the number of nodule pixels encompassed by the nodule region at the maximum area slice multiplied by the pixel spacing. Eccentricity is defined in the equation below:

$$\mathrm{EC} = \sqrt{1-\frac{b^2}{a^2}} \tag{17}$$

where a and b are semi-major axis and semi-minor axis lengths of nodule region of interest, respectively. Curvature descriptor [53] is calculated in (18) with respect to intensity inside nodule region of interest, which depends on the intensity variation

$$\mathrm{CD} = \tan^{-1}\left(\frac{\sqrt{\lambda_1^2+\lambda_2^2}}{1+I(x,\ y)}\right) \tag{18}$$

where λ_1 and λ_2 ($\lambda_1 \le \lambda_2$) are two eigenvalues of Hessian matrix. A new structure descriptor indicates the intensity variation, which is calculated by using two eigenvalues of Hessian matrix in the equation below:

$$\mathrm{ST} = \mathrm{e}^{-\sqrt{|\lambda_1\lambda_2|}} \tag{19}$$

The compactness can represent characteristics of lobulation and circularity of a pulmonary nodule to a certain extent, which is calculated in the equation below:

$$\mathrm{comp} = \frac{R_{\mathrm{in}}}{R_{\mathrm{out}}} \tag{20}$$

where R_{in} and R_{out} denote the radius of the maximum inscribed tangent circle and the circumcircle of nodule contour. Circularity (CI) measures the degree to which a shape differs from a perfect circle. CI is calculated by the equation below:

$$\mathrm{CI} = \frac{\mathrm{area}}{\pi r^2} \tag{21}$$

where r is the maximum bounding box radius in the maximum area slice. Geometric moments are geometric invariant [54] for image translation, scaling and rotation. We employ a symmetric to computer geometric moments introduced by Wee *et al.* [55]. The circularity of the shape S(CI2) is calculated by the equation below:

$$\mathrm{CI2} = \frac{u_{2,\ 0}(S) + u_{0,\ 2}(S)}{u_{0,\ 0}(S)^2} \tag{22}$$

where $u_{p,\ q}$ is $(p,\ q)$ order-centralised moment of the shape. Spiculation of a nodule is calculated in (23) by Dhara *et al.* [56]

$$\mathrm{Sp} = \frac{\sum_{i=1}^{M} h_i\cos\varpi_i}{\sum_{i=1}^{M} h_i} \tag{23}$$

where ϖ_i is a solid angle subtended at the peak point of the ith spicule and h_i is the height of the spicule. M is the total number of the nodule spicules.

3.4 RF classifiers training

RF is a suitable classifier for differentiating benign and malignancy of pulmonary nodules. In this section, an RF is trained to predict the class labels. A subset of features is chosen at each node of the tree to find the best split, which is generated using the bootstrap method. Each split node of the tree is associated with a split function $f(\cdot)$ to send a node to the left or right child node. $f(\cdot)$ is defined in the equation below:

$$f(\boldsymbol{v},\ \theta_j): R^d \times \chi \to \{0,\ 1\} \tag{24}$$

where $\boldsymbol{v}$ is an input feature vector and θ_j is a feature parameter. χ denotes the space of all split parameters and R^d is the d-dimensional feature space. A threshold τ_j is selected to determine the node split, then all training images are split into left $\mathfrak{R}_{\mathrm{L}}(j,\ \tau_j) = \{[v_i;\ \theta_j] | \forall i,\ f(v_i;\ \theta_j) \le \tau_j\}$ and right $\mathfrak{R}_{\mathrm{R}}(j,\ \tau_j) = \{[v_i;\ \theta_j] | \forall i,\ f(v_i;\ \theta_j) > \tau_j\}$. In this paper, the probability density distribution at each feature vector of the tth tree is modelled as a multivariate in the equation below:

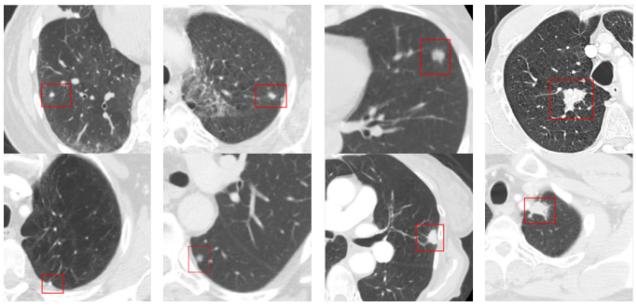


Fig. 4 *Examples of rated benign and malignant pulmonary nodules from the LIDC radiologist's marks. The first column: benign pulmonary nodules with the rank of malignancy '1'. The second column: benign pulmonary nodules with the rank of malignancy '2'. The third column: malignant pulmonary nodules with the rank of malignancy '4'. The fourth column: malignant pulmonary nodules with the rank of malignancy '5'*

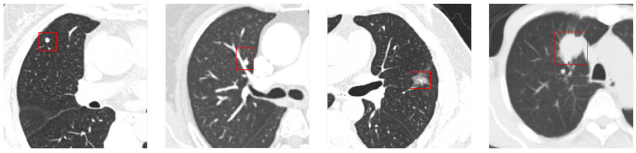


Fig. 5 *Examples of benign and malignant pulmonary nodules from the GHGMC dataset*

(a) Benign pulmonary nodule with the rank of malignancy '1', ***(b)*** Benign pulmonary nodules with the rank of malignancy '2', ***(c)*** Malignant pulmonary nodules with the rank of malignancy '4', ***(d)*** Malignant pulmonary nodules with the rank of malignancy '5'

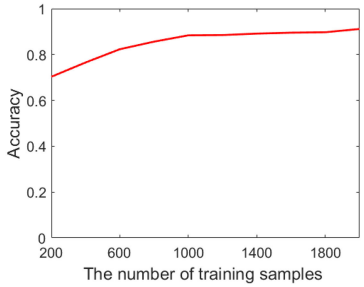


Fig. 6 *Quantitative results of the accuracy of ten different numbers of training samples on the LIDC datasets. The red curve indicates the accuracies on ten different numbers of training samples*

$$p_t(v|\mathscr{L}) = \frac{1}{(2\pi)^{d/2}\boldsymbol{\Lambda}} e^{-(1/2)(v-\bar{v})^{\mathrm{T}}\Lambda^{-1}(v-\bar{v})} \tag{25}$$

where d is the dimension of the feature vector. $\bar{v}$ and $\boldsymbol{\Lambda}$ are the mean value and covariance matrix of the feature vector. In Bayesian rule framework, a weight value is calculated based on Kullback–Leibler divergence in the equation below:

$$w_i' = \sum_j p_t(\boldsymbol{v}_i|\mathscr{L}) \ln \frac{p_t(\boldsymbol{v}_i|\mathscr{L})}{p_t(\boldsymbol{v}_j|\mathscr{L})} \tag{26}$$

where $\mathscr{L}$ is a binary class label indicating a malignant nodule $\mathscr{L} = 1$ or a benign nodule $\mathscr{L} = 0$. $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are input feature vectors at ith and jth pixels, respectively.

Trees progressively grow the tree depth by adding new nodes. The number of points is fewer than a threshold or a maximum tree depth D and the growth stops. The output of random forests is predicted by weighting all individual tree predictions, which is defined in the equation below:

$$p(\mathscr{L}|v) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{\mathrm{NP}} w'_i p_t(\mathscr{L}|v_i) \tag{27}$$

where $\mathscr{L}$ is a binary class label indicating a malignant nodule $\mathscr{L} = 1$ or a benign nodule $\mathscr{L} = 0$. T is the number of trees and NP is the number of image pixels. $p_t(\mathscr{L}|v_i)$ is the probability density function given a multivariate feature vector of the pixel i in the tth tree. Given an unlabelled pulmonary nodule, feature subsets are sampled from the training set in the same manner previously. Each feature subset of test nodules is pushed through each tree starting at the root and the corresponding sequence of tests applied.

4 Experimental setup and results

4.1 Pulmonary nodule datasets

(i) *LIDC dataset:* The LIDC dataset of thoracic CT scans is applied to evaluate the performance of the RF for the classification of benign and malignant pulmonary nodules. The LIDC [57] is the largest library of thoracic CT scans publicly, which contains 1018 CT thoracic scans and associated XML files. In general, a pulmonary nodule appears in several slices of a CT scan. In the case of 2D slices, the slice with the greatest-sized pulmonary nodule is sampled for differentiating benign and malignancy pulmonary nodules. Nodule descriptions consist of lobulation, spiculation, malignancy and so on. The semantic rating used to test and train the classifier ranges from 1 to 5 by four experienced thoracic radiologists, which indicates an increasing degree of the manifestation of nodule characteristics. Malignancy denotes the malignant likelihood of a pulmonary nodule. In this paper, all training samples are classified into malignant and benign nodules. A malignancy score lower than 3 is labelled as a benign nodule and a malignancy score higher than 3 is labelled as a malignant nodule. The pulmonary nodules with a score of 3 in malignancy are removed to avoid the ambiguousness of nodule samples. About 1000 nodules are randomly selected per class to train an RF classifier, and the remaining 200 of 2200 CT images are used as a test set. Fig. 4 shows examples of rated benign and malignant pulmonary nodules with different ranks of malignancy from the LIDC dataset.

(ii) *General Hospital of Guangzhou Military Command (GHGMC) dataset:* The GHGMC real dataset is also employed to evaluate the classification performance of the proposed method. The GHGMC dataset was collected from the GHGMC, which is a large general and comprehensive hospital in Guangdong province, and one of the best hospital in Guangdong province of China. In the dataset, the certified radiologists annotated all visible pulmonary nodules, and recorded the nodule characteristic of each case. The dataset consists of 300 pulmonary nodules (180 benign nodules and 120 malignant nodules). Fig. 5 shows examples of benign and malignant pulmonary nodules with different ranks of malignancy from the GHGMC dataset.

4.2 Parameter setting

In RF classification, three important parameters are involved: the number of training samples, the number of trees T and the tree depth D. We perform sensitivity analyses of three parameters to estimate the effect on the accuracy of the results. Fig. 6 shows the accuracy of the proposed classification method with different numbers of training samples. The more training samples are selected, the larger accuracy is obtained. Therefore, more nodule samples are selected to improve the classification performance. Fig. 7 shows the classification accuracy of the proposed method with different numbers of decision trees T. As we can see, the optimal number of trees T is 45, and both smaller and larger numbers of trees lead to decreasing classification performances.

4.3 Experimental results

4.3.1 Segmentation results and performance: To understand the benefit of the proposed segmentation method, we conduct a comparative experiment on the proposed segmentation method and the ground truth. Fig. 8 shows segmentation results of different types of pulmonary nodules by using the proposed segmentation method. Figs. 8*a* and *c* show original CT lung images with

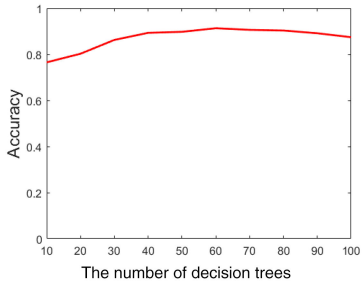


Fig. 7 *Accuracy of classification results with ten different numbers of decision trees by the proposed RF algorithm*

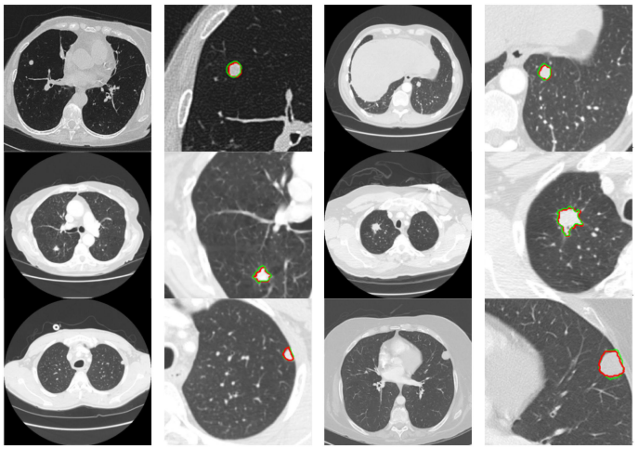


Fig. 8 *Examples of segmentation results by the proposed RW method in different types and sizes of pulmonary nodules*
(a), ***(c)*** Original CT images, ***(b)***, ***(d)*** Segmentation results (red) by the proposed RW method and ground truth (green)

Table 3 Overlap measures obtained by the improved RW method and traditional RW method

ID	Overlap measure (improved RW)	Overlap measure (conventional RW)
1	0.9435	0.8643
2	0.9034	0.8412
3	0.9427	0.8742
4	0.8741	0.8154
5	0.9243	0.8432
6	0.9374	0.9042
7	0.8942	0.7351
8	0.9514	0.8641
9	0.9812	0.8401
10	0.8714	0.9145
mean	0.9224	0.8496
variance	0.0357	0.0502

different types of nodules from LIDC dataset. Figs. 8*b* and *d* show the segmentation results by the improved RW method. To compare with ground truths and segmentation results, their segmentation contours of nodules and ground truths are plotted in Figs. 8*b* and *d*. The red and green curves indicate the results by the proposed segmentation method and the ground truths, respectively. As shown in Figs. 8*b* and *d*, we can see that the proposed RW method produces segmentation results closer to the ground truths.

The overlap measure is defined by the equation below:

$$\text{overlap} = \frac{|S_A \cap S_M|}{|S_A \cup S_M|} \tag{28}$$

where S_A and S_M denote the segmentation results by the proposed method and ground truth, respectively. $|S_A|$ and $|S_M|$ denote the

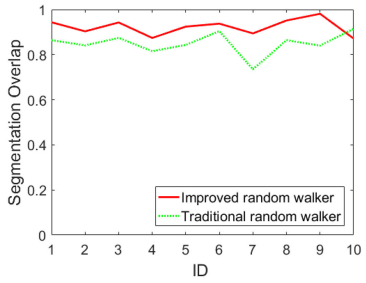


Fig. 9 *Curves of overlap measures on ten patient cases from the LIDC dataset. The red and green curves indicate the improved RW and traditional RW methods, respectively*

	Predicted Benign	Predicted Malignant	
Actual Benign	TN	FP	TN+FP
Actual Malignant	FN	TP	FN+TP
	TN+FN	FP+TP	

Fig. 10 *Confusion matrix*

numbers of pixels in S_A and S_M, respectively. $|S_A \cap S_M|$ is the number of pixels in both S_A and S_M and $|S_A \cup S_M|$ is the number of pixels in either S_A, S_M or both. Ideally, a good segmentation is expected to have a high overlap measure. We perform a comparison of the improved RW against traditional RW. We randomly select ten patient cases from the LIDC dataset. Table 3 lists the overlap measures on ten cases. The mean and standard variance of overlap measures is also calculated for intuitive observation. As observed from Table 3, the mean of the overlap measures obtained by the proposed RW method is 0.92, which is higher than that of the traditional RW method.

Fig. 9 shows the curves of overlap measures obtained by the improved RW method and traditional RW method. The red and green curves indicate the overlap measures obtained by the proposed method and the conventional RW method, respectively. As shown in Fig. 9, we can observe that the improved RW method obtains a higher overlap measure.

4.3.2 Classification results and performance: To verify the performance of the proposed classification method, the confusion matrix [6, 58, 59] is used as the metric. We have two classes: benign nodules and malignant nodules. Therefore, the confusion matrix with a size of 2×2 is defined in this paper. Fig. 10 shows the terminology the confusion matrix, and we also add the row and column totals.

True negative (TN) corresponds to the number of benign nodules correctly predicted by the proposed classification method. True positive (TP) corresponds to the number of malignant nodules correctly predicted by the proposed classification method. False positive (FP) corresponds to the number of benign nodules wrongly predicted as malignant nodules by the proposed classification method. False negative (FN) corresponds to the number of malignant nodules wrongly predicted as benign nodules by the proposed classification method. The number of correctly classified samples is the sum of diagonals and all others are incorrectly classified in the confusion matrix. The six terminologies are used when referring to the counts in a confusion matrix. The TP rate (TPR) or sensitivity is defined as the fraction of malignant nodules predicted correctly, which is calculated in the equation below:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{29}$$

Table 4 Results of confusion matrices of the proposed RF with different numbers of testing samples

Number	Number of test samples	TN(TN + FP)	TP(FN + TP)
no. 1	100	67(80)	17(20)
no. 2	200	108(130)	62(70)
no. 3	300	220(242)	52(58)
no. 4	400	248(264)	120(136)
no. 5	500	309(340)	147(160)
no. 6	600	360(400)	184(200)

Table 5 Sensitivity, specificity and accuracy of the proposed RF with different numbers of testing samples in confusion matrices

Number	Sensitivity	Specificity	Accuracy
no. 1	0.85	0.84	0.84
no. 2	0.89	0.83	0.85
no. 3	0.90	0.91	0.91
no. 4	0.88	0.94	0.92
no. 5	0.92	0.91	0.91
no. 6	0.92	0.90	0.91
mean	0.89	0.88	0.89

The TN rate is analogous to TPR, which is defined as the fraction of benign nodules predicted correctly in the equation below:

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{30}$$

FP rate (FPR) is the fraction of benign nodules predicted as malignant nodules, which is calculated in the equation below:

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{31}$$

The FN rate (FNR) is the fraction of malignant nodules predicted as benign nodules, which is calculated in the equation below:

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{32}$$

Recall, precision and accuracy are also widely used metrics employed in performance classification, which are defined in equations below:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{33}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{34}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{35}$$

To verify the proposed classification method on the LIDC dataset, we perform a comparison experiment with different numbers of test samples in terms of the confusion matrix. We select six different numbers of testing samples in this experiment and divide them into two cohorts: benign nodules and malignant nodules. The size of the testing samples is N, which ranges from 100 up to 600 at steps of 100. Table 4 summarises results of the confusion matrices with six different numbers of testing samples. Table 5 shows sensitivity and specificity and accuracy of the proposed RF with different numbers of testing samples in confusion matrices. As observed from Table 5, the proposed classification method obtains a high sensitivity and specificity. Fig. 11 shows confusion matrices on six different numbers of testing samples. When we select 600 testing samples from the LIDC dataset and divide them into benign nodules (TN + FP = 400) and malignant nodules (FN + TP = 200),

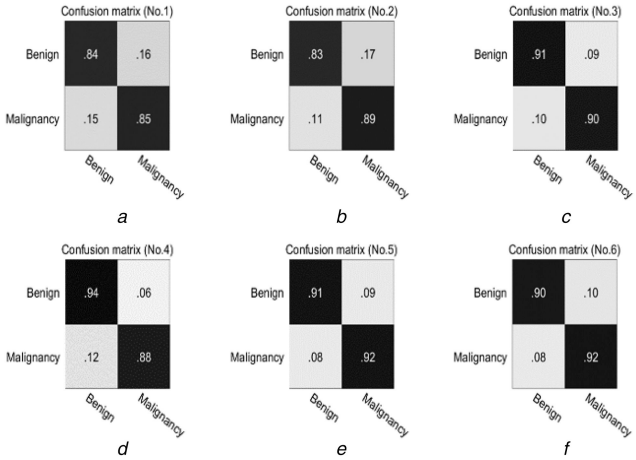


Fig. 11 *Results of confusion matrices on different numbers of testing samples*

(a) Confusion matrix on 100 testing samples, ***(b)*** Confusion matrix on 200 testing samples, ***(c)*** Confusion matrix on 300 testing samples, ***(d)*** Confusion matrix on 400 testing samples, ***(e)*** Confusion matrix on 500 testing samples, ***(f)*** Confusion matrix on 600 testing samples

Table 6 Sensitivity and specificity of the proposed RF on the GHGMC dataset

Training set	Number of benign	Number of malignant	Sensitivity	Specificity
no. 1	80	120	0.91	0.84
no. 2	100	100	0.89	0.86
no. 3	180	20	0.81	0.80
no. 4	140	60	0.80	0.79
mean	—	—	0.85	0.82

the number of benign nodules correctly predicted is 360 and the remaining benign nodules are wrongly predicted as malignant nodules by the proposed classification method. The number of malignant nodules correctly predicted is 184 and the remaining malignant nodules are wrongly predicted as benign nodules by the proposed classification method. As can be seen from Fig. 11*f*, the first row indicates a specificity of 0.90 and an FPR of 0.10, and the second row indicates an FNR of 0.08 and a sensitivity of 0.92. Therefore, the proposed classification method yields the stable classification performances in terms of sensitivity and specificity when the number of testing samples increases.

To evaluate the classification performance of the proposed method, an additional evaluation is performed on the GHGMC dataset. We randomly selected a set data, consisting of 19 benign nodules and 11 malignant nodules, as the testing dataset. Then, we also randomly selected four training sets from the remaining data, respectively. The detailed information is listed in Table 6. Table 6 shows the sensitivity and specificity of different training datasets. A mean sensitivity and a mean specificity are 0.85 and 0.82, respectively. Therefore, the proposed classification method yields the desirable classification performances in terms of sensitivity and specificity.

4.4 Statistical test for feature selection

4.4.1 Statistical intensity features: To assess the effectiveness of intensity features for distinguishing benign and malignant pulmonary nodules, the mean values of intensity features are calculated from the benign and malignant pulmonary nodules. About 200 samples with benign and malignant nodules are randomly selected from the LIDC dataset in this experiment and are divided into two cohorts: benign nodules (100) and malignant nodules (100). Table 7 lists the mean values of intensity features for benign and malignant pulmonary nodules. As shown in Table 7, the mean values of kurtosis for benign and malignant pulmonary nodules are 5.0709 and 2.0316. It is clear that the mean values of kurtosis for benign pulmonary nodules are the higher than that of

Table 7 Intensity feature values for benign and malignant pulmonary nodules

Intensity feature	Mean value of benign nodules	Mean value of malignant nodules
Max_Intensity	0.7549	0.8962
Min_Intensity	0.5130	0.4301
u_Intensity	0.6144	0.5929
σ_Intensity	0.0631	0.1013
skewness	1.0494	1.4452
kurtosis	5.0709	2.0316
$\Delta I(x, y)$	0.3142	0.2392
intensity variance	0.1835	0.1245
RDSm	0.7587	0.7226
RDSv	0.1207	0.1386

Table 8 Mean values of texture features for benign and malignant pulmonary nodules from 100 nodule samples

Texture feature	Mean value of benign nodules	Mean value of malignant nodules
autocorrelation	1.0350	1.0829
contrast	0.0579	0.0101
correlation1	0.5437	0.6226
correlation2	0.5437	0.6226
cluster Prominence	20.8976	49.7619
dissimilarity	0.0008	0.0014
energy	0.9987	0.9970
entropy	0.0053	0.0117
homogeneity1	0.9978	0.9963
homogeneity2	0.9974	0.9959
maximum probability	0.9993	0.9985
sum of squares	1.0067	1.0566
sum average	2.0084	2.0195
sum variance	4.1247	4.2948
sum entropy	0.0052	0.0116
difference variance	0.0058	0.0101
difference entropy	0.0011	0.0019
information measure of correlation1	−0.7544	−0.8370
information measure of correlation2	−0.1269	−0.1785
INN	0.0045	0.0098
inverse difference moment normalised	0.9999	0.9998

malignant nodules. However, the mean value of skewness for malignant nodules is larger than that of benign nodules. Since the mean values of intensity u_Intensity, the variance σ_Intensity and the minimal intensity value Min_Intensity for benign and malignant pulmonary nodules have very minimal differences, they are excluded from intensity features. The statistical results of intensity feature values show that most of the intensity features can distinguish benign and malignant pulmonary nodules to some extent.

4.4.2 Statistical texture features: The GLCM texture features are selected based on MI, in which the relevant GLCM texture features are ranked by mRMR criterion. We employ the 0.632+ bootstrap estimator [60] to determine a low-variance measure of the prediction error for any texture feature subset. First, 22 GLCM texture features are computed by using six quantisation levels $Q = 8$, 16, 32, 64, 128, 256, four orientations $\theta = 0^{\circ}$, 45°, 90°, 135° and ten distances $d = 1, 2, \ldots, 10$. Texture values of the same distance are averaged over all orientations. We select $Q = 8$, $\theta = 90^{\circ}$ and $d = 1$ to obtain the mean values of GLCM texture features on 200 nodule samples, as shown in Table 8.

Table 9 Comparison of the subset of GLCM texture features (single and average), considering the number of features obtained from the minimum 0.632+ bootstrap estimated errors

Quantisation level	Number of average features	Minimum estimated errors
8 (average)	96	0.231
16 (average)	125	0.217
32 (average)	106	0.201
64 (average)	104	0.212
128 (average)	86	0.211
256 (average)	92	0.228
8 (single)	15	0.184
16 (single)	22	0.172
32 (single)	16	0.176
64 (single)	15	0.189
128 (single)	21	0.184
256 (single)	21	0.184

Table 10 Classification performance of GLCM texture features obtained from 200 bootstrap samples

Quantisation level	Number of average features	AUC	Accuracy	Sensitivity
8 (average)	96	0.82	0.7625	0.7264
16 (average)	125	0.82	0.7813	0.7289
32 (average)	106	0.83	0.7884	0.7465
64 (average)	104	0.81	0.7642	0.7212
128 (average)	86	0.81	0.7624	0.7286
256 (average)	92	0.81	0.7638	0.7294
8 (single)	15	0.90	0.8762	0.7847
16 (single)	22	0.84	0.8425	0.8054
32 (single)	16	0.86	0.8942	0.8423
64 (single)	15	0.90	0.8975	0.8075
128 (single)	21	0.90	0.8483	0.8045
256 (single)	21	0.90	0.8976	0.8468

Then, we select a GLCM texture feature that used the largest dependency on the nodule classification. Then, others are iteratively added to this feature until all features are considered. For each iteration, the selected subset of GLCM texture features is used to classify benign and malignant nodules. Table 9 shows the minimum estimated errors by considering the six quantisation levels and feature sets (single and average). As shown in Table 9, when features of the same distance are average, the estimated error increases compared with the single feature set. The subsets of GLCM texture features are selected: the AUROC, accuracy, sensitivity, specificity, positive predictive value and negative predictive value are listed in Table 10. As observed from Table 10, the quantisation level cannot improve or degenerate the discrimination power of GLCM texture features. Therefore, we employ quantisation level $Q = 8$ in the GLCM texture feature extraction process. Thus, the dimensionality of texture features and the time consumption are reduced. When the number of GLCM features ranges from 15 to 22, we achieve an AUC of 0.90. Herein, 15 GLCM texture features are employed for classifying benign and malignant pulmonary nodules.

There are three types of texture features in total. To verify the effect of different types of texture features of the proposed classification method, we perform a compared experiment with respect to different types of texture features. Three types of texture features' classification performance in terms of sensitivity, specificity and accuracy on a group of pulmonary nodule samples are listed in Table 11. As shown in Table 11, we obtain a sensitivity of 0.89 and a specificity of 0.86 by incorporating these texture features. It is obvious that the proposed classification method can

Table 11 Classification sensitivity, specificity and accuracy on three types of texture features

Texture feature	Sensitivity	Specificity	Accuracy
GLCM	0.78	0.79	0.78
LBP	0.82	0.81	0.82
Gabor	0.84	0.80	0.83
GLCM + LBP + Gabor	0.89	0.86	0.90

Table 12 Classification sensitivity, specificity and accuracy on different feature sizes

Number of features	Sensitivity	Specificity	Accuracy
20	0.74	0.76	0.74
40	0.76	0.77	0.76
60	0.80	0.79	0.80
80	0.81	0.86	0.82
100	0.84	0.80	0.83
120	0.87	0.82	0.87
140	0.90	0.83	0.89
160	0.92	0.84	0.90

Table 13 Classification accuracy with different numbers of training samples from LIDC dataset

Samples	Accuracy
200	0.7031
400	0.7654
600	0.8232
800	0.8562
1000	0.8837
1200	0.8849
1400	0.8912
1600	0.8953
1800	0.8972
2000	0.9114

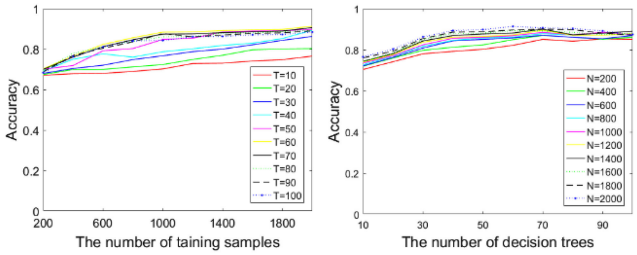
Table 14 Classification accuracy with ten different numbers of decision trees

Trees	Accuracy
10	0.7652
20	0.8025
30	0.8624
40	0.8932
50	0.8976
60	0.9131
70	0.9061
80	0.9032
90	0.8913
100	0.8742

significantly improve the sensitivity and accuracy. We also verify the effectiveness of the number of features for the proposed classification method. Classification sensitivity, specificity and accuracy on different feature sizes on a group of pulmonary nodule samples are shown in Table 12. As shown in Table 12, when feature sizes are 160, our proposed classification method obtains a sensitivity of 0.92 and a specificity of 0.84.

4.5 Sensitivity to the number of training samples

Since the RF method depends on the number of training samples, we want to validate how sensitive the proposed classification algorithm is to the number of training samples. To achieve a fair comparison, the number of trees T and the tree depth D are same in this experiment. The RF code is run by using different numbers of training samples. The performance of the proposed method is evaluated in terms of accuracy. Their respective accuracies are summarised in Table 13. As shown in Table 13, the classification accuracy is improved from 0.7031 to 0.9114 when we add the number of training samples. The curve of classification accuracies by using ten different numbers of training samples is plotted in Fig. 6.



a *b*

Fig. 12 *Accuracy of the proposed classification method with varying the number of decision trees and training samples*

(a) Curves of accuracy with ten different numbers of decision trees by using ten different numbers of training samples, ***(b)*** Curves of accuracy with ten different numbers of training samples by using ten different numbers of decision trees

4.6 Sensitivity to the number of decision trees

It is well known that increasing the number of decision trees can improve the accuracy of classification because comprehensive predictions are detailed including more discriminative information. We validate how sensitive the proposed classification algorithm is to the number of decision trees. The number of training samples and the tree depth D are same in this experiment. The classification results are shown in Table 14.

Fig. 7 illustrates the accuracy of training RF with ten different numbers of decision trees. As shown in Fig. 7, the increase in the number of decision trees is beneficial until 60 trees and beyond the additional increase of accuracy is slightly decreasing. In addition, the proposed RF method has the relative stability with an increased number of decision trees. To justify the effect of the number of decision trees and training samples for classification performance, we perform experiments on different numbers of decision trees and training samples. Fig. 12 shows the curves of the accuracy of the proposed classification method with varying the number of decision trees and training samples. Fig. 12*a* shows the curves of accuracy with ten different numbers of decision trees by using ten different numbers of training samples. Fig. 12*b* plots the curves of accuracy with ten different numbers of training samples by using ten different numbers of decision trees. As shown in Fig. 12*a*, the accuracy of the proposed classification method is rather low due to overfitting when N is small and increases with an increased number of training samples and the proposed classification method also obtains a higher accuracy when the number of decision trees increases. As shown in Fig. 12*b*, the accuracy of the proposed classification method increases with an increased number of decision trees and the proposed classification method also obtains a higher accuracy when the number of training samples increases.

5 Discussion

To justify the effectiveness and robustness of the proposed RF method, the proposed method is compared with other state-of-the-art classification methods. It is very difficult to perform fair comparisons using the same dataset because the source codes of these state-of-the-art classification methods were not been published. Although some methods were validated by using the same LIDC dataset, they were not exactly the same samples; therefore, results can have the large variability. To a rough assessment of the performance of the proposed classification method, several existing methods are summarised in Table 15. Han's method [7] employed texture features and geometry-related features of pulmonary nodules for differentiating malignant and benign pulmonary nodules. 2D texture features were calculated in this paper such as Haralick, Gabor and LBP features by using the

Table 15 Comparisons of the terminologies of the confusion matrix between the proposed method and other the state-of-the-art methods

Method	Sensitivity	Specificity	Accuracy	AUC
Han *et al.* [7]	0.89	0.86	—	0.94
Dhara *et al.* [56]	0.90	0.86	—	0.95
Chen *et al.* [5]	—	—	0.90	0.95
Sen [60]	0.77	0.93	0.87	0.93
our method	0.92	0.83	0.90	0.95

LIDC dataset. The SVM classifier was employed to classify malignant and benign pulmonary nodules. As can be seen from Table 15, Han's method [7] obtained a sensitivity of 0.89 and a specificity of 0.86. Dhara's method represented the shape of a nodule in terms of few diagnostic characteristics such as lobulation, sphericity and spiculation and sphericity [56]. A differential geometry-based technique was introduced to compute the spiculation of pulmonary nodules. Dhara's method obtained a sensitivity of 0.90 and a specificity of 0.86 [56]. Dhara's method [56] yielded no significant performance improvement compared with Han's method [7]. Chen *et al.* [5] also performed a comparison between ANNs and multivariable LR for differentiating malignant pulmonary nodules from benign pulmonary nodules. ANNs had a higher classification performance than LR, which obtained an AUROC of 0.955 and an accuracy rate of 0.90. Sen *et al.* proposed a multi-crop CNN (MC-CNN) for malignant and benign nodules classification. The learned deep features were captured by the MC pooling strategy [58]. Sen's method [58] obtained a sensitivity of 0.77 and a specificity of 0.93. However, Chen's method [5] achieved a substantial improvement over Sen's method [58] in terms of accuracy. Sen's method [58] yielded a significantly higher specificity than do the listed other methods, but as indicated by sensitivity, Sen's method [58] yielded a relative lower specificity than that of Han's method [7] and Dhara's method [56]. In nodule diagnosis, the sensitivity is more important than the specificity for classifying the malignant and benign nodules. The proposed classification method achieves the higher sensitivity and the lower specificity compared with other methods. As is known to all, the high sensitivity and the high specificity are desirable generally. The ROC reflects the trade-off of classifiers between hit rates and false alarm rates. For a better explanation of ROC, an AUROC is calculated, which is listed in the last column of Table 15. In particular, the proposed RF classification method achieves an AUC value of 0.95.

6 Conclusions and future works

In this paper, we have developed an improved RW and RF methods for classifying the benign and malignant pulmonary nodules. The improved RW achieved the better segmentation performance than that of the traditional RW model. This was because texture feature and shape index were incorporated. The advantage of the improved RW was that it could automatically segment various types of pulmonary nodules by using the proposed seed acquisition method. In addition, intensity, texture and geometry features were extracted from the segmented nodules. The combination of GLCM-based, rotation invariant uniform LBP-based and Gabor-based methods for texture extraction could improve the classification performance. An improved RF classifier was proposed to predict the class label of benign or malignant pulmonary nodules. We took into account the probability density function at each feature vector of each tree for improving the classification performance. The experiments on the LIDC dataset and the GHGMC dataset demonstrated the effectiveness of the proposed method for the segmentation and classification of pulmonary nodules.

In the future, we will further improve the classification performance of pulmonary nodules and optimise the proposed model. In addition, our further work will grade the images based on the degree of the malignancy of pulmonary nodules, which is of valuable significance for the diagnosis and treatment of lung cancer in clinic applications.

7 Acknowledgments

The authors thank Dr. W.B. Zhu, Dr. P. Chen, Dr. R. Bai, Dr. L. Zhang, Dr. F. Long, Radiologist G.Q. Qiao and Engineer L. Tang for their helpful comments and advice which contributed much to this paper. This work was supported by the National Natural Science Foundation of China (61305038, 61273249), the Natural Science Foundation of Guangdong Province, China (8451064101000631), the Public Science and Technology Research Funds Projects of Ocean (201505002) and the Fundamental Research Funds for the Central Universities, Key Laboratory of Autonomous Systems and Network Control of Ministry of Education, Doctoral Fund of Ministry of Education of China (20130172110028).

8 References

[1] Siegel, R.L., Miller, K.D., Jemal, A.: 'Cancer statistics, 2015', *CA Cancer J. Clin.*, 2015, **65**, pp. 5–29
[2] Girvin, F., Ko, J.P.: 'Pulmonary nodules: detection, assessment, and CAD', *Am. J. Roentgenol.*, 2008, **191**, pp. 1057–1069
[3] Iwano, S., Nakamura, T., Kamioka, Y., *et al.*: 'Computer-aided differentiation of malignant from benign solitary pulmonary nodules imaged by high-resolution CT', *Comput. Med. Imaging Graph.*, 2008, **32**, pp. 416–422
[4] Chen, H., Xu, Y., Ma, Y.J., *et al.*: 'Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on CT images clinical evaluation', *Acad. Radiol.*, 2010, **17**, pp. 595–602
[5] Chen, H., Zhang, J., Xu, Y., *et al.*: 'Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans', *Expert Syst. Appl.*, 2012, **39**, pp. 11503–11509
[6] Lin, P.-L., Huang, P.-W., Lee, C.-H., *et al.*: 'Automatic classification for solitary pulmonary nodule in CT image by fractal analysis based on fractional Brownian motion model', *Pattern Recognit.*, 2013, **46**, pp. 3279–3287
[7] Han, F., Wang, H., Zhang, G., *et al.*: 'Texture feature analysis for computer-aided diagnosis on pulmonary nodules', *J. Digit. Imaging*, 2015, **28**, pp. 99–115
[8] Cheng, J.-Z., Ni, D., Chou, Y.-H., *et al.*: 'Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans', *Sci. Rep.*, 2016, **6**, pp. 24454–24466
[9] Dhara, A.K., Mukhopadhyay, S., Dutta, A., *et al.*: 'A combination of shape and texture features for classification of pulmonary nodules in lung CT images', *J. Digit. Imaging*, 2016, **29**, pp. 466–475
[10] Liu, Y., Balagurunathan, Y., Atwater, T., *et al.*: 'Radiological image traits predictive of cancer status in pulmonary nodules', *Clin. Cancer Res.*, 2017, **23**, pp. 1442–1449
[11] Tajbakhsh, N., Suzuki, K.: 'Comparing two classes of end-to-end machine-learning models in lung nodule detection and classification: MTANNs vs. CNNs', *Pattern Recognit.*, 2017, **63**, pp. 476–486
[12] Grady, L.: 'Random walks for image segmentation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, pp. 1768–1783
[13] Eslami, A., Karamalis, A., Katouzian, A., *et al.*: 'Segmentation by retrieval with guided random walks: application to left ventricle segmentation in MRI', *Med. Image Anal.*, 2013, **17**, pp. 236–253
[14] Xu, Z., Bagci, U., Foster, B., *et al.*: 'A hybrid method for airway segmentation and automated measurement of bronchial wall thickness on CT', *Med. Image Anal.*, 2015, **24**, pp. 1–17
[15] Tan, J.H., Acharya, U.R., Lim, C.M., *et al.*: 'An interactive lung field segmentation scheme with automated capability', *Digit. Signal Process.*, 2013, **23**, pp. 1022–1031
[16] Wang, Q., Lu, L., Wu, D., *et al.*: 'Automatic segmentation of spinal canals in CT images via iterative topology refinement', *IEEE Trans. Med. Imaging*, 2015, **34**, pp. 1694–1704
[17] Mi, H., Petitjean, C., Vera, P., *et al.*: 'Joint tumor growth prediction and tumor segmentation on therapeutic follow-up PET images', *Med. Image Anal.*, 2015, **23**, pp. 84–91
[18] Ju, W., Xiang, D., Zhang, B., *et al.*: 'Random walk and graph cut for co-segmentation of lung tumor on PET-CT images', *IEEE Trans. Image Process.*, 2015, **24**, pp. 5854–5867
[19] Patz, T., Preusser, T.: 'Segmentation of stochastic images with a stochastic random walker method', *IEEE Trans. Image Process.*, 2012, **21**, pp. 2424–2433
[20] Park, S.H., Lee, S., Yun, I.D., *et al.*: 'Structured patch model for a unified automatic and interactive segmentation framework', *Med. Image Anal.*, 2015, **24**, pp. 297–312
[21] Dong, X.P., Shen, J.B., Shao, L., *et al.*: 'Sub-Markov random walk for image segmentation', *IEEE Trans. Image Process.*, 2016, **25**, pp. 516–527
[22] Breiman, L.: 'Random forests', *Mach. Learn.*, 2001, **45**, pp. 5–32
[23] Mursalin, M., Zhang, Y., Chen, Y.H., *et al.*: 'Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier', *Neurocomputing*, 2017, **241**, pp. 204–214
[24] Kostoglou, K., Michmizos, K.P., Stathis, P., *et al.*: 'Classification and prediction of clinical improvement in deep brain stimulation from intraoperative microelectrode recordings', *IEEE Trans. Biomed. Eng.*, 2017, **64**, pp. 1123–1130
[25] Kostis, W.J., Reeves, A.P., Yankelevitz, D.F., *et al.*: 'Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images', *IEEE Trans. Med. Imaging*, 2003, **22**, pp. 1259–1274

[26] Kuhnigk, J.M., Dicken, V., Bornemann, L., *et al.*: 'Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans', *IEEE Trans. Med. Imaging*, 2006, **25**, pp. 417–434

[27] Diciotti, S., Picozzi, G., Falchini, M., *et al.*: '3-D segmentation algorithm of small lung nodules in spiral CT images', *IEEE Trans. Inf. Technol. Biomed.*, 2008, **12**, pp. 7–19

[28] Dehmeshki, J., Amin, H., Valdivieso, M., *et al.*: 'Segmentation of pulmonary nodules in thoracic CT scans: a region growing approach', *IEEE Trans. Med. Imaging*, 2008, **27**, pp. 467–480

[29] Kubota, T., Jerebko, A.K., Dewan, M., *et al.*: 'Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models', *Med. Image Anal.*, 2011, **15**, pp. 133–154

[30] Farag, A.A., Abd El Munim, H.E., Graham, J.H., *et al.*: 'A novel approach for lung nodules segmentation in chest CT using level sets', *IEEE Trans. Image Process.*, 2013, **22**, pp. 5202–5213

[31] Netto, S.M.B., Silva, A.C., Nunes, R.A., *et al.*: 'Automatic segmentation of lung nodules with growing neural gas and support vector machine', *Comput. Biol. Med.*, 2012, **42**, pp. 1110–1121

[32] Chen, K., Li, B., Tian, L.F., *et al.*: 'Vessel attachment nodule segmentation using integrated active contour model based on fuzzy speed function and shape-intensity joint Bhattacharya distance', *Signal Process.*, 2014, **103**, pp. 273–284

[33] Sun, S.S., Guo, Y., Guan, Y.B., *et al.*: 'Juxta-vascular nodule segmentation based on flow entropy and geodesic distance', *IEEE J. Biomed. Health Inf.*, 2014, **18**, pp. 1355–1362

[34] Messay, T., Hardie, R.C., Tuinstra, T.R.: 'Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset', *Med. Image Anal.*, 2015, **22**, pp. 48–62

[35] Li, B., Chen, Q.L., Peng, G.M., *et al.*: 'Segmentation of pulmonary nodules using adaptive local region energy with probability density function-based similarity distance and multi-features clustering', *Biomed. Eng. Online*, 2016, **15**, pp. 49–76

[36] Diciotti, S., Lombardo, S., Falchini, M., *et al.*: 'Automated segmentation refinement of small lung nodules in CT scans by local shape analysis', *IEEE Trans. Biomed. Eng.*, 2011, **58**, pp. 3418–3428

[37] Zhang, F., Song, Y., Cai, W., *et al.*: 'Lung nodule classification with multilevel patch-based context analysis', *IEEE Trans. Biomed. Eng.*, 2014, **61**, pp. 1155–1166

[38] Ye, X.J., Lin, X.Y., Dehmeshki, J., *et al.*: 'Shape-based computer-aided detection of lung nodules in thoracic CT images', *IEEE Trans. Biomed. Eng.*, 2009, **56**, pp. 1810–1820

[39] Muramatsu, C., Hara, T., Endo, T., *et al.*: 'Breast mass classification on mammograms using radial local ternary patterns', *Comput. Biol. Med.*, 2016, **72**, pp. 43–53

[40] Beura, S., Majhi, B., Dash, R.: 'Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer', *Neurocomputing*, 2015, **154**, pp. 1–14

[41] Sethi, G., Saini, B.S.: 'Computer aided diagnosis system for abdomen diseases in computed tomography images', *Biocybern. Biomed. Eng.*, 2015, **36**, pp. 42–55

[42] Torheim, T., Malinen, E., Kvaal, K., *et al.*: 'Classification of dynamic contrast enhanced MR images of cervical cancers using texture analysis and support vector machines', *IEEE Trans. Med. Imaging*, 2014, **33**, pp. 1648–1656

[43] Gomez, W., Pereira, W.C.A., Infantosi, A.F.C.: 'Analysis of co-occurrence texture statistics as a function of gray-level quantization for classifying breast ultrasound', *IEEE Trans. Med. Imaging*, 2012, **31**, pp. 1889–1899

[44] Liu, L., Lao, S., Fieguth, P.W., *et al.*: 'Median robust extended local binary pattern for texture classification', *IEEE Trans. Image Process.*, 2016, **25**, pp. 1368–1381

[45] Rastghalam, R., Pourghassem, H.: 'Breast cancer detection using MRF-based probable texture feature and decision-level fusion-based classification using HMM on thermography images', *Pattern Recognit.*, 2016, **51**, pp. 176–186

[46] Guo, Z., Wang, X., Zhou, J., *et al.*: 'Robust texture image representation by scale selective local binary patterns', *IEEE Trans. Image Process.*, 2016, **25**, pp. 687–699

[47] Bianconi, F., Fernandez, A.: 'Evaluation of the effects of Gabor filter parameters on texture classification', *Pattern Recognit.*, 2007, **40**, pp. 3325–3335

[48] Ivanovici, M., Richard, N.: 'Fractal dimension of color fractal images', *IEEE Trans. Image Process.*, 2011, **20**, pp. 227–235

[49] Gangeh, M.J., Tadayyon, H., Sannachi, L., *et al.*: 'Computer aided theragnosis using quantitative ultrasound spectroscopy and maximum mean discrepancy in locally advanced breast cancer', *IEEE Trans. Med. Imaging*, 2016, **35**, pp. 778–790

[50] Jacobs, C., van Rikxoort, E.M., Twellmann, T., *et al.*: 'Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images', *Med. Image Anal.*, 2014, **18**, pp. 374–384

[51] Liu, J., Wang, S., Linguraru, M.G., *et al.*: 'Computer-aided detection of exophytic renal lesions on non-contrast CT images', *Med. Image Anal.*, 2015, **19**, pp. 15–29

[52] Guo, Z.H., Zhang, L., Zhang, D.: 'Rotation invariant texture classification using LBP variance (LBPV) with global matching', *Pattern Recognit.*, 2010, **43**, pp. 706–719

[53] Jaeger, S., Karargyris, A., Candemir, S., *et al.*: 'Automatic tuberculosis screening using chest radiographs', *IEEE Trans. Med. Imaging*, 2014, **33**, pp. 233–245

[54] Xu, D., Li, H.: 'Geometric moment invariants', *Pattern Recognit.*, 2008, **41**, pp. 240–249

[55] Wee, C.-Y., Paramesran, R., Mukundan, R.: 'Fast computation of geometric moments using a symmetric kernel', *Pattern Recognit.*, 2008, **41**, pp. 2369–2380

[56] Dhara, A.K., Mukhopadhyay, S., Saha, P., *et al.*: 'Differential geometry-based techniques for characterization of boundary roughness of pulmonary nodules in CT images', *Int. J. Comput. Assist. Radiol. Surg.*, 2016, **11**, pp. 337–349

[57] Armato, S.G., McLennan, G., Bidaut, L., *et al.*: 'The lung image database consortium, (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans', *Med. Phys.*, 2011, **38**, pp. 915–931

[58] Aminikhanghahi, S., Shin, S., Wang, W., *et al.*: 'A new fuzzy Gaussian mixture model (FGMM) based algorithm for mammography tumor image classification', *Multimedia Tools Appl.*, 2017, **76**, pp. 10191–10205

[59] Dora, L., Agrawal, S., Panda, R., *et al.*: 'Optimal breast cancer classification using Gauss–newton representation based algorithm', *Expert Syst. Appl.*, 2017, **85**, pp. 134–145

[60] Shen, W., Zhou, M., Yang, F., *et al.*: 'Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification', *Pattern Recognit.*, 2017, **61**, pp. 663–673