Computer Methods and Programs in Biomedicine 160 (2018) 141–151

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Quantitative CT analysis of pulmonary nodules for lung adenocarcinoma risk classification based on an exponential weighted grey scale angular density distribution feature

Vanbang Le[a,1], Dawei Yang[b,c,1], Yu Zhu[a,*], Bingbing Zheng[a], Chunxue Bai[b,c], Hongcheng Shi[d], Jie Hu[b,c], Changwen Zhai[e], Shaohua Lu[e]

[a] School of Information Science and Engineering, East China University of Science and Technology, Shanghai, Postcode 200237, China
[b] Department of Pulmonary Medicine, ZhongShan Hospital, Fudan University, Shanghai, Postcode 200032, China
[c] Shanghai Respiratory Research Institute, Shanghai, Postcode 200032, China
[d] Department of Nuclear Medicine, ZhongShan Hospital, Fudan University, Shanghai, Postcode 200032, China
[e] Department of Pathology, ZhongShan Hospital, Fudan University, Shanghai, Postcode 200032, China

ARTICLE INFO

Article history:
Received 5 June 2017
Revised 21 February 2018
Accepted 2 April 2018

Keywords:
Lung nodule classification
K-means
Exponential weighted
Reference map
Angular histogram

ABSTRACT

Background and objectives: To improve lung nodule classification efficiency, we propose a lung nodule CT image characterization method. We propose a multi-directional feature extraction method to effectively represent nodules of different risk levels. The proposed feature combined with pattern recognition model to classify lung adenocarcinomas risk to four categories: Atypical Adenomatous Hyperplasia (AAH), Adenocarcinoma In Situ (AIS), Minimally Invasive Adenocarcinoma (MIA), and Invasive Adenocarcinoma (IA).

Methods: First, we constructed the reference map using an integral image and labelled this map using a K-means approach. The density distribution map of the lung nodule image was generated after scanning all pixels in the nodule image. An exponential function was designed to weight the angular histogram for each component of the distribution map, and the features of the image were described. Then, quantitative measurement was performed using a Random Forest classifier. The evaluation data were obtained from the LIDC-IDRI database and the CT database which provided by Shanghai Zhongshan hospital (ZSDB). In the LIDC-IDRI, the nodules are categorized into three configurations with five ranks of malignancy ("1" to "5"). In the ZSDB, the nodule categories are AAH, AIS, MIA, and IA.

Results: The average of Student's *t*-test *p*-values were less than 0.02. The AUCs for the LIDC-IDRI database were 0.9568, 0.9320, and 0.8288 for Configurations 1, 2, and 3, respectively. The AUCs for the ZSDB were 0.9771, 0.9917, 0.9590, and 0.9971 for AAH, AIS, MIA and IA, respectively.

Conclusion: The experimental results demonstrate that the proposed method outperforms the state-of-the-art and is robust for different lung CT image datasets.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Lung cancer is a disease with significant prevalence in several countries around the world [1]. Improving the level of early diagnosis and the identification of small lung adenocarcinoma has been always an important topic for imaging studies. In the field of clinical medicine, ground-glass opacity (GGO) and ground-glass nodules (GGN) appear as a tiny cloudy region in a CT image [2]. By measuring the tiny region of a GGN at a high resolution, the area can be classified as one of three types: a pure GGN (pGGN), a mixed GGN (mGGN) (or a part-solid nodule (PSN)) and a solid nodule (SN). When the pGGN shows a clear edge density, then translucent, non-solid nodules at the opposite edge are less clear than the part-solid nodule with regard to the low-energy region [3].

In 2011, the International Association for the Study of Lung Cancer (IASLC), the American Thoracic Society (ATS) and the European Respiratory Society (ERS) proposed a new international multi-disciplinary classification system for lung adenocarcinoma [4]. The possible classes are atypical adenomatous hyperplasia (AAH); adenocarcinoma in situ (AIS); minimally invasive adenocarcinoma (MIA); and invasive adenocarcinoma (IA). The degree of

* Corresponding author.
E-mail addresses: zhuyu@ecust.edu.cn (Y. Zhu), zbbddmail@gmail.com (B. Zheng), bai.chunxue@zs-hospital.sh.cn (C. Bai), shi.hongcheng@zs-hospital.sh.cn (H. Shi), zhai.changwen@zs-hospital.sh.cn (C. Zhai).
[1] These authors contributed equally to this article, and both should be considered first author.

https://doi.org/10.1016/j.cmpb.2018.04.001
0169-2607/© 2018 Elsevier B.V. All rights reserved.

risk in each of these classes is increasing. The classification of lung adenocarcinomas from AAH to IA will have great significance for clinical auxiliary diagnoses and have great significance in improving the 5-year survival rate.

Quantitative methods mainly use a combination of image features and classifiers for the classification and recognition of lung nodules. There are several features, namely, 2D shape, 3D shape [18], texture [5], wavelet transform [9], or density distribution features [11]. The common classifiers are ANN [16], SVM [13], CNN [24], Random Forest [27] and deep learning [17].

Reeves et al. [6] at Cornell University calculated the morphology, density features, surface curvature and edge gradient of nodule images to construct 46 dimensions of 3D features, and used classification methods to analyse the classification of the risk of pulmonary nodules (benign/malignant) from the Public Lung Image Database (ELCAP) [7] and the National Lung Screening Trial (NLST) [8] database. The results obtained showed almost 70% classification accuracy at optimal parameters. Lee HY et al. [15] performed a quantitative analysis of preoperative CT imaging metrics to distinguish invasive adenocarcinoma from AIS, MIA and showed that the classification performance is good. For the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [14], Han et al. [5] used image texture features to classify the risk of pulmonary nodules into two classes, benign and malignant. Dhara et al. [13] also classified the pulmonary nodules into benign and malignant categories for the LIDC-IDRI database. They used the same configurations as Han et al. [5]. The features include shape and texture of the image.

Maldonado et al. [11] proposed the CT value density distribution calculation method. First, they collected the image block set (block sizes as 9×9) from nodule CT images. Then, they used an affinity propagation clustering approach to classify the correlation matrix of the image block set. After that, they scanned the testing nodule image, calculated the density level for pixels and exported the feature vector. Finally, the different categories of pulmonary nodules were recognized by analysing the components of the feature vector. Mayo Clinic Medical Center [10–12,19] introduced an effective computer-aided nodule assessment and risk yield (CANARY) system, which proposed the determination of lung nodule classification and risk prediction based on the CT densities of the nodule images. The CANARY system which measured the classification of the images as good (G), intermediate (I), or poor (P) cases of benign or malignant lung nodules [11].

To optimize the classification performance for lung adenocarcinoma and improve their clinical significance, this article proposes an effective set of lung CT image density distribution features for pulmonary nodule risk classification. We proposed a nodule image feature of weighted grey scale angular density distribution. The remainder of this paper is organized as follows: In Section 2, we introduce the material preparation. In Section 3, we formulate the proposed framework of weighted grey scale angular density distribution feature extraction, which includes unsupervised feature representation and pattern recognition models. In Section 4, we present the experimental results of the process outlined in Section 3 and show an analysis of the results. In Section 5, we conclude the paper.

2. Materials

In this paper, the open data set LIDC-IDRI and the data set provided by Shanghai Zhongshan Hospital were used to validate the classification performance for the proposed algorithm. The robustness and the universality of the proposed method have been proved via different data sets for case area, imaging characteristic and classification modality.

2.1. Pre-processing for the LIDC-IDRI database

The CT images of 1018 patients were downloaded from The Cancer Imaging Archive (TCIA) Public Access Portal [20]. We used additional XML files to locate the nodule region and export its malignancy level. The statistics for the LIDC-IDRI database are shown in Table A1 (see Table A1 in the Appendix). The pixel spacing ranged from 0.5 to 0.8 mm, and the slice thickness ranged from 0.6 to 5.0 mm. The major diameter of the nodules ranged from 2.79 to 15.77 mm.

For the nodule selection of LIDC-IDRI dataset, every nodule was evaluated for the degree of malignancy by up to four radiologists. There were many nodules that were assessed with differing threat levels by the radiologists. The malignancy levels were from rank “1” to rank “5” (five levels), rank “1” or “2” of nodule is regarded as benign, rank “4” or “5” is regarded as malignant and rank of “3” has uncertain malignancy [5]. Some nodules were ranked 1, 2, 3 or 3, 4, 5 at the same time. To ensure balance of the number of nodules containing different ranks, we preferred the rank which the sample number is less than the others. For example, a nodule that was assessed at the malignancy was given the smaller sample number of the ranks, and at the same time, the nodule was removed from the others group. Specially, we removed all the nodules which appeared to be calcified along with nodules with a major length diameter that was >16 mm. The remaining data set contained 1318 nodules, where the samples of rank “1” to “5” are 139, 392, 393, 257 and 137, respectively. Han et al. [5] defined the risk levels of nodules by three configurations. Configuration 1 classifies the rank of “1” or “2” as benign and “4” or “5” as malignant. Configuration 2 classifies the rank of “1”, “2”, or “3” as benign and “4” or “5” as malignant. Configuration 3 classifies the rank of “1” or “2” as benign and “3”, “4” or “5” as malignant [5,13]. The same configurations are used for the evaluation of the proposed classification scheme. If the nodules with malignancy rank “3” are discarded, there are 531 benign and 394 malignant nodules; if they are regarded as benign, there are 924 benign and 394 malignant nodules; and if they are regarded as malignant there are 531 benign and 787 malignant nodules (see Table A2 in the Appendix).

2.2. Lung CT dataset provided by Shanghai Zhongshan Hospital (ZSDB)

The dataset was collected from radiology data of 350 patients, imaging time since 2014 to 2015. All ground-truth samples were pathologically defined by four clinical experts at Shanghai Zhongshan Hospital. These nodules (4 mm $<$ major length diameter $<$32 mm) were divided into four categories: AAH, IAS, MIA, and IA. The sample numbers per class are 92, 157, 158, and 188, respectively. The imaging parameters of the ZSDB are the following: the electric settings are 500 mA and 120 Kv; the size of the image is 512×512; the type is PET/CTQSZLXX; the protocol is 10.1 CT CHEST; the pixel spacing is 0.703125 mm; and the slice thickness is 0.625 mm. However, the CAD framework proposed in this paper will be more effective in the enlarged dataset. More details of the ZSDB database are shown in Table A1 in the Appendix.

3. Methods

The framework of the proposed method is shown in Fig. 1.

3.1. Pulmonary nodule segmentation

Several algorithms for pulmonary nodule segmentation have been proposed, such as Hessian matrix-based methods [21], the 3D fuzzy connectivity-based approach [22] and the active contour-based method [23]. In this paper, K-means [25] ($K = 3$) clustering

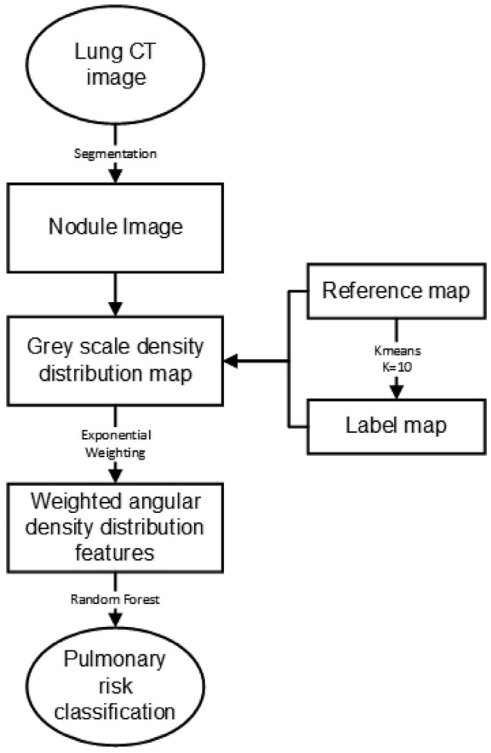

Fig. 1. The framework of the proposed method.

is performed in the designated area of the region of interest (ROI). The proposed segmentation method is applied to extract the nodule image of the ZSDB dataset.

For an image pixel set (x_1, x_2,..., x_n), K-Means clustering aims to partition the n observations into K ($\leq$ n) sets $S = \{S_1, S_2, ..., S_K\}$ so as to minimize the within-cluster sum of squares (sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find:

$$\arg\min_{S} \sum_{i=1}^{K} \sum_{x \in S_i} ||x - \mu_i||^2 \quad (1)$$

The lung parenchyma segmentation method based on K-Means and morphological CLOSING-operation was applied to extracted lung region in CT image. Firstly, K-Means unsupervised approach was performed to clustering the lung CT image, and then the thoracic and course lung area were extracted. Considering the case of the juxta-pleural and pleural-tail nodules, this paper uses the closed operation to repair and optimize the edge of the lung parenchyma binary image and export the exquisitely segmented image.

After that, K-Means approach was performed once again to segment the lung nodule from region of interest (ROI). The background and redundant area are eliminated by analysing the clustering results, and the lung nodule area in the ROI is extracted. The image sequence of the nodules is constituted by segmenting the continuous CT image frames. In each image, the pixels of the nodule region are preserved. The image is then described by the geometric aspects and the density distributions of the nodules.

The locations of the lung nodules in pulmonary parenchyma are random. The experimental results showed the efficiency of the present method with different complex levels of lung nodules. There are some instances that belong to a cloudy and ground-glass-like appearance of the part-solid nodules, and in addition, there are nodules that include juxta-vascular or infiltration vessels.

3.2. Pulmonary nodule image feature extraction

3.2.1. Reference map construction and labelling

We propose an image grey scale density distribution calculation method based on a rescaled integration, which we named as *reference map* ($H(x,y)$). The reference map was defined by an integration image which generated by a ones matrix (Ones(x,y)), with sizes of (w,h). We calculated the integration of this ones matrix, as following:

$$G(x, y) = sum(Ones[: x, : y]) \quad (2)$$

Where $G(x, y)$ is the integration of ones matrix.

Then, we rescaled the range of $G(x, y)$ to [−824,176] and the proposed reference map was extracted, as following:

$$H(x, y) = \frac{G(x, y)}{\max(G(x, y))} \times 1000 - 824 \quad (3)$$

In this paper, the unit for each point in reference map would be set as HU, the range of reference map were from-824 to 176 (HU). The value range was set up manually by consulting the opinion of radiologists. The minimal and maximal values could be the range for lung nodule images, where the value of the upper left point stands for the lowest value (−824 HU) and the maximum value (176 HU) was located at the lower right point. After that, a K-means (K = 10) method for the clustering process of the reference map was performed, and then the label map was calculated. Fig. 2 shows the visualization of the reference map, the histogram image, and the label map.

Considering the stability of the distribution of reference map clustering center, the size of the reference map should not be too small. This ensures the optimization for the distribution map of the nodule image. So, the size of reference map [w, h] = [200, 200] is chosen in the experiment. By clustering for the reference map, the class center tends to stable while the class number is too large. Hence, in order to ensure the optimality of the feature vector and program running time, K = 10 is chosen.

3.2.2. The grey scale density distribution map of lung nodule image

We used the reference map and the label map to calculate the density distribution level for pixels of nodule images. For each testing pixel in the nodule region, we selected a neighbourhood as a size was [5, 5] of testing image block and then found the matching area in the reference map. The distance $d(Block, Matched)$ from the testing window to the matching area is minimized. The distance follows:

$$d(Block, Matched) = min \|Block - Matched\|_2 \quad (4)$$

where *Block* and *Matched* are matrices with the same size of the testing image block and the matched region in the reference map, respectively. The corresponding location on the label map of the *Matched* area (the pixel locations and the points of matched area have a one-to-one correspondence) is within the region of one category or in the cross-region of two categories, which are labelled by a continuous value. Therefore, we calculated the round value of the average value of the corresponding region as the density distribution level of the testing pixels.

The density distribution map of the nodule image (D) is as follows:

$$D(i, j) = \{g_{i,j} | i \in [0, w], j \in [0, h]\} \quad (5)$$

Where $g_{i,j}$ is the grey scale density distribution level of the pixels in the nodule image. These values describe the difference in solid degree of the pixels in the pulmonary nodule image.

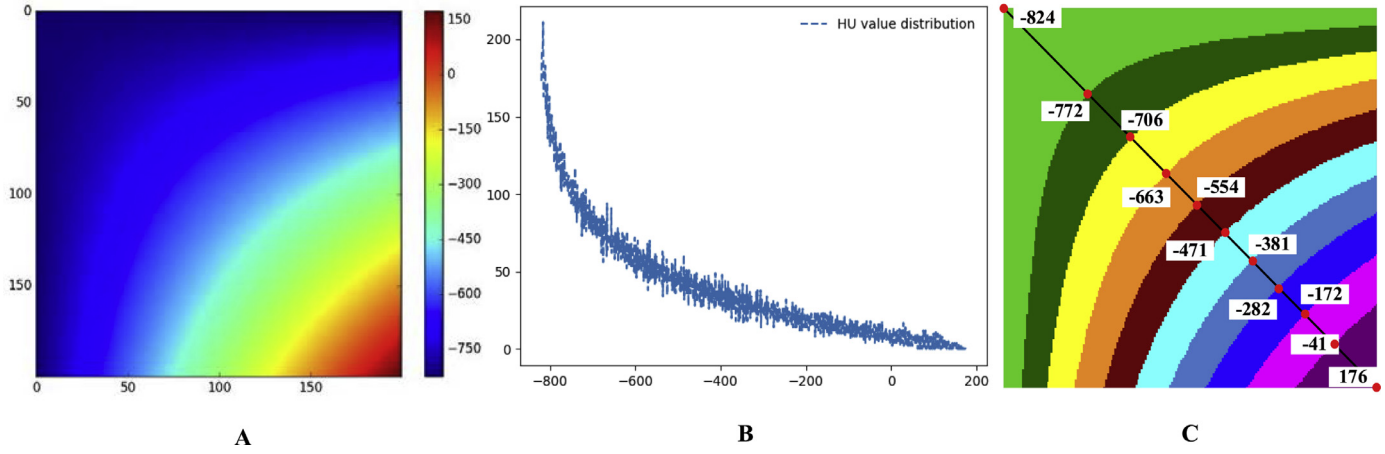

Fig. 2. Reference map and label map. (A) Reference map (sizes as 200×200); (B) The histogram of the reference map. The horizontal and vertical are the grey scale levels (HU) and its frequency (the number of occurrences of grey scale level) in reference map. This curve appears to follow the exponential distribution. (C) The categories resulted from the clustering process and its HU value range (10 categories). The colours of Lime Green, Dark Green, Cyan, Dodger Blue, Navy, Yellow, Peru, Red, Magenta and Purple describe the different classes from 1 to 10, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

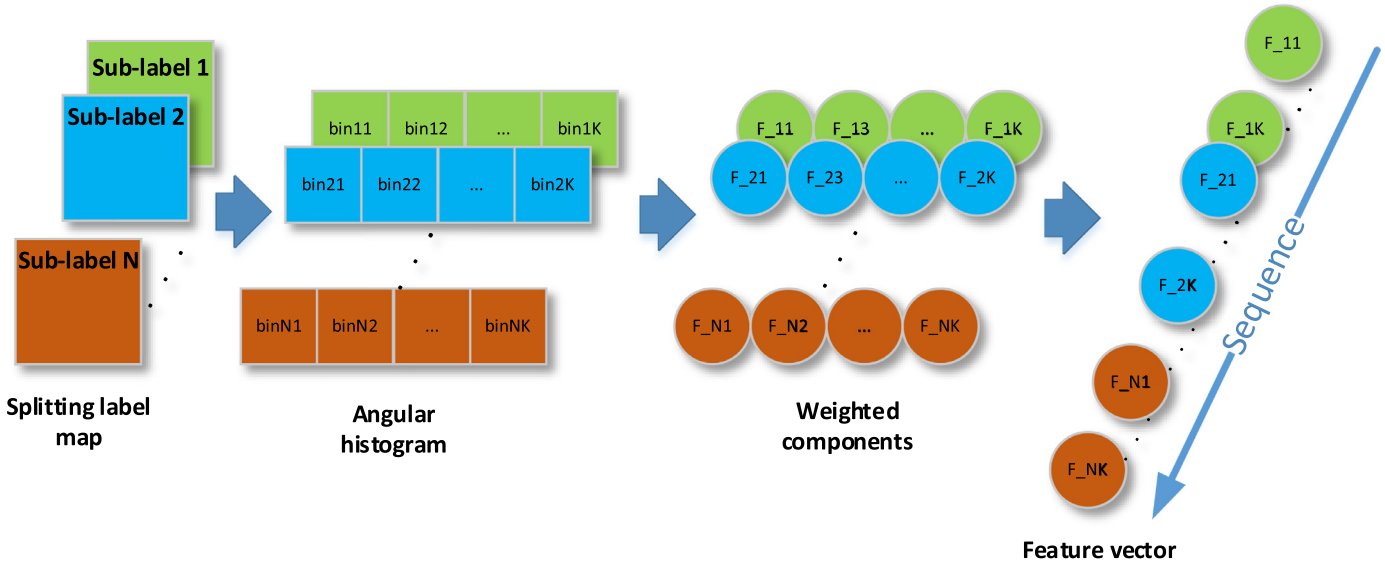

Fig. 3. The schematic for the weighted angular density distribution feature extraction from the label map of nodule image. Where N and K are the sub-label maps and bins number, respectively.

3.2.3. Extracting the grey scale angular density distribution feature based on exponential weighting

In Section 3.2.1 and 3.2.2, we proposed a 10-level grey scale density distribution map to describe nodule images. This type of feature map is insufficient to reflect the directional distribution information for solid or partly solid elements in a lung nodule. However, the elemental directional distribution of a nodule is also important to get more effective features. To solve this problem, we propose an angular density distribution feature based on exponential weighting function. The calculation steps are shown in Fig. 3:

In the framework, firstly, we calculated the centroid of the nodule image, and split the grey scale density distribution map D into a 10 sub-label map. Then, the locations of the pixels in a sub-label map were transformed from planar coordinates to angular coordinates and the angular histograms for each sub-label map were generated. Finally, we calculated the exponential weighted histogram to obtain the proposed weighted angular grey scale density distribution of nodule image.

The local centre $(\bar{x}, \bar{y})$ of the nodule image is:

$$(\bar{x}, \bar{y}) = \left(\frac{\sum_{x=1}^{h} x(D_{x,y} > 0)}{\sum(D_{x,y} > 0)}, \frac{\sum_{y=1}^{w} y(D_{x,y} > 0)}{\sum(D_{x,y} > 0)} \right) \tag{6}$$

where D is label map (grey scale density distribution map) of nodule image was extracted by the proposed method in Section 3.2.2.

We resolved the label map of nodule image into 10 *sub-label maps (S)*, and then calculated the angular histogram for each sub-label map. In this way, the image point (x,y) was transformed to (r, α), where α was calculated by the following:

$$\alpha = \arg\cos\left(\frac{y}{\sqrt{(x-\bar{x})^2 + (y-\bar{y})^2}} \right) \tag{7}$$

When the pixel (x,y) is planning to direction bin(b) if angle α is followed:

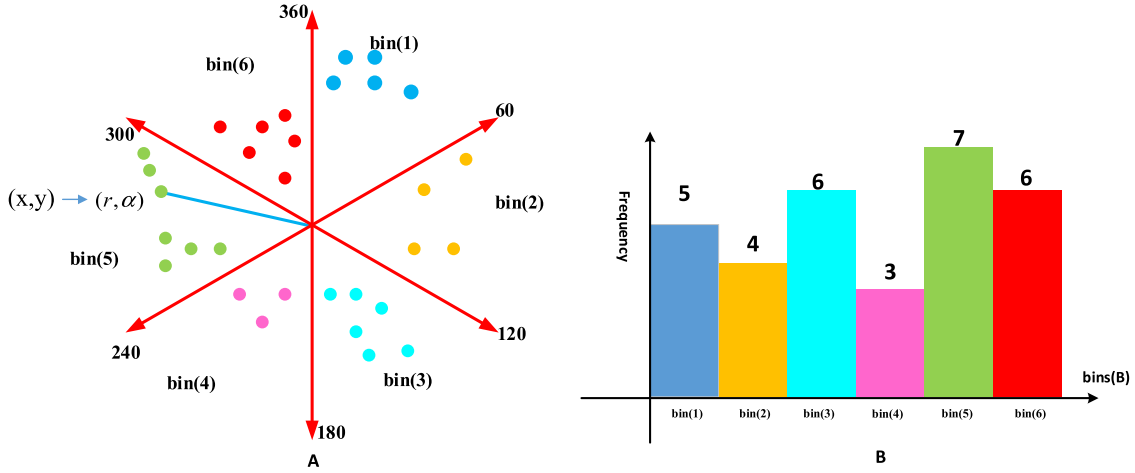

Fig. 4. Example of the transformation to the angular coordinate system with angle step $\theta = 30$. (A) Points of nodule image after transform to angle coordinate system; (B) Angular histogram of density distribution map. In Fig. 4 (A), (r, α) is the location in angle coordinate of the image point (x,y).

$$\frac{180 \times \cos(\alpha)}{\theta \times \pi} \in (b-1, b] \tag{8}$$

The angular histogram illustrates the image points in the bins, and the angle step (θ) and bins (bins number $B = \frac{360}{\theta}$) are also directly related to the nodule size. When the bin spacing is too small, the number of pixels in the bins is rather small and will reduce the distinction between the bins, so we chose minimum angle step to evaluate the proposed method as 30.

An example for the transformation to the angular coordinate system of the local (x,y) in the sub-label map S(k) and angular histogram extraction are show in Fig. 4 with $\theta = 60^o$ and bins = 6.

In the sub-label map S(k), the feature component in $f_{(k,b)}$ for the *bin(b)* direction is as follows:

$$f_{(k,b)} = \underbrace{\sum\sum \left|(r,\alpha)_{k,b}\right|}_{\text{Angular histogram of density distribution}} \times \underbrace{w_{\text{exponential}}}_{\text{weighted}} \tag{9}$$

The angular histogram shows the spatial distribution for pixels of the nodule image in different directions. We analysed statistically the number of pixels of the different directions on the same sub-label map so that the feature vector shows the spatial distribution relative to the nodule centre of these pixels. The image pixel $(r, \alpha)_{k,b}$ belongs to the direction *bin(b)* in sub-label map S(k). In this way, the weighted $w_{\exp onential}$ is the exponential function to the amplitude of mean location in the direction of *bin(b)*. The average coordinate of the pixels in direction *bin(b)* is $(x_{average}, y_{average}) = (\frac{\sum_x x}{n}, \frac{\sum_y y}{n})$, where (x,y) is pixel in the direction bin(b) of sub label map S(k), and n is pixel number in this region $(n = \sum_x \sum_y |(x,y)|)$. Therefore, the weighted $w_{\exp onential}$ is as follow:

$$w_{\text{exponential}} = \exp\left(-\sqrt{\frac{3 \times \left(x_{average}^2 + y_{average}^2\right)}{\max\left(x_{average}, y_{average}\right)^2}} + \xi\right) \tag{10}$$

where ξ is the repair parameter. Parameter ξ adjusts the value of the weight for the small nodules (nodules where the pixel coordinates are close to zero). The commonly selected value of ξ in these experiments is from 0.001 to 0.1. This weight enhances the effect of the pixels closer to the centre and reduces the effect of the pixels on the borders at the same time. The main purpose is to improve the fault tolerance rate for the lung nodule segmentation, and remove the interference factors in the border zone, i.e., vessel, bronchus, chest wall. Meanwhile, improve the significant influence of the solid region which located at the centre of the nodule. In addition to the selection of angle step, θ depends on the size of the nodule image. In this way, a bigger step is a smaller image, and vice versa.

Finally, the normalized exponential weighting angular density distribution feature F_G was extracted as follows:

$$F_G = \{100 \times \frac{f_{(k,b)}}{\sum_k \sum_b f_{(k,b)}} | k \in [1,K], b \in [1,B]\} \tag{11}$$

The factors were computed in different directions for each sub-label map. All the components were sequenced to construct the weighted grey scale angular density distribution feature of the nodule image. The feature was used to characterize the risk level of the pulmonary nodule. We have taken the concept of the density distribution and its calculation from the method of Maldonado et al. [10–12] and optimized that concept. The method presented by Maldonado et al. [10–12] is mainly based on collected image blocks to extract the density distribution feature of the nodule image. The image block set is randomly collected from the pulmonary nodule image. The histogram distribution of all image blocks is not stable. The grey scale density distribution of a lung nodule image may not achieve the optimal results. Moreover, there is lack of robustness to the process for different databases.

In the proposed method, the reference map generated based on the integration image was used instead of the distance matrix of image cells. The HU range covers overall values of the grey scale, which range from the value of the air to the calcification region in the lung CT image. The histogram of the reference map is more flat and smooth than the distance matrix, and it basically obeys an exponential distribution. In this paper, the calculated density distribution of the nodule image will achieve the optimal result. The density distribution calculated from the labelled reference map is unified for different datasets. We also calculated the spatial relationship between the factors of feature vectors.

3.3. Pulmonary nodule image classification

We used the *p* values to measure the significance of the feature set. A scalar was calculated from the multi-dimension feature vector by the sum of the gradient. And then, the cross-validation *p*-values between classes is generated via a two-tailed Student's test. For the nodule classification, the quantitative measurement was performed using the well-known Random Forest classifier. The area under the receiver operating curve characteristic (AUC) value and cross-validation score were used to evaluate the training pa-

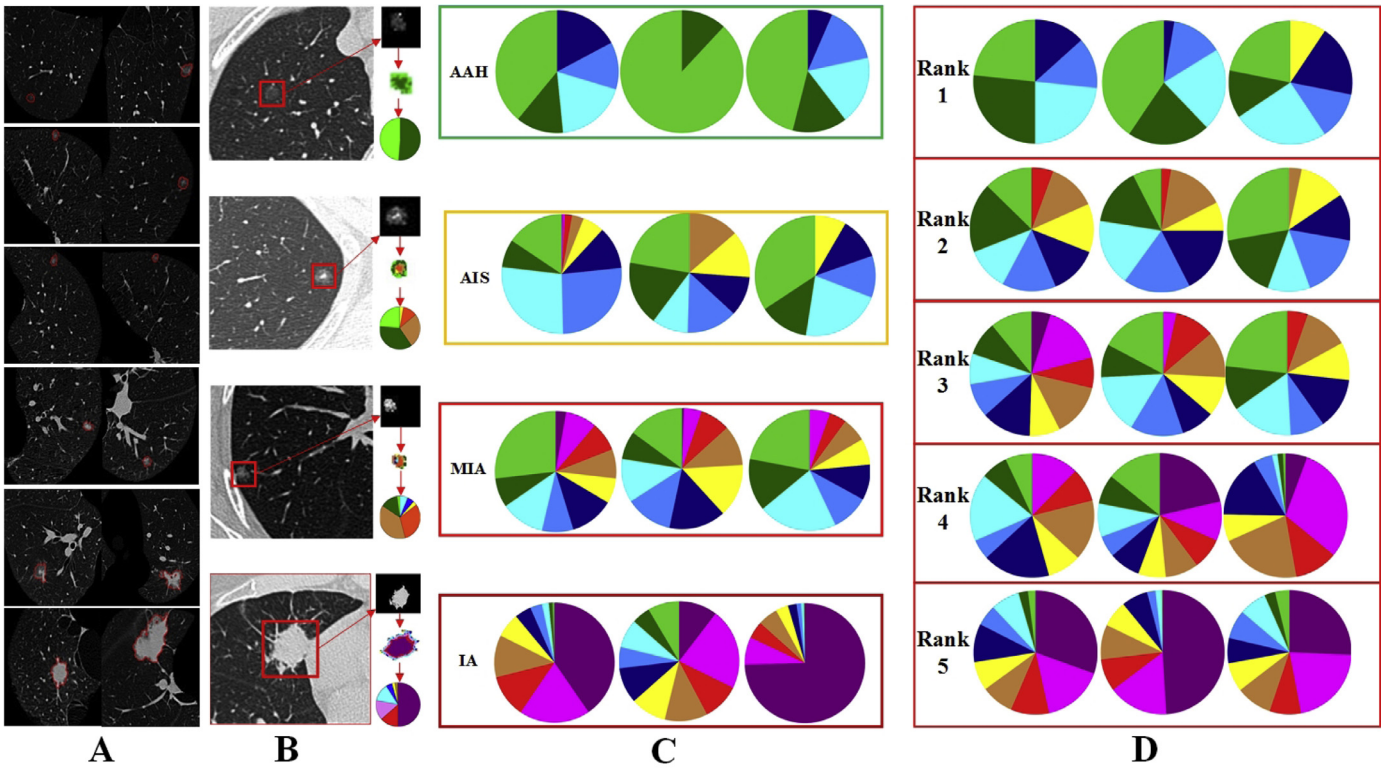

Fig. 5. Pulmonary nodule image segmentation and characterization process for the ZSDB and the LIDC-IDRI databases. (A) Nodule image segmentation; (B) Image characterization. (C) and (D) show the examples of the density distribution feature (Percentage) for ZSDB and LIDC-IDRI databases, respectively. (C) Rows 1 to 4 are the features of AAH, AIS, MIA, and IA, respectively; (D) Rows 1 to 5 are the features of the nodule from rank 1 to 5, respectively.

rameters [26]. The AUC of the ROC is computed using the Sensitivity and 1-Specificity, where

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$Specify = \frac{TN}{TN + FP} \tag{13}$$

Cross-validation is scored to multi-class classification where the accuracy is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{14}$$

where:
TP (True Positive rate) = correctly classified positive cases;
TN (True Negative rate) = correctly classified negative cases;
FN (False Negative) = incorrectly classified positive cases; and
FP (False Positive) = incorrectly classified negative cases.

4. Experimental

The proposed classification scheme and the competing technique are evaluated on the data set of 1318 and 595 nodules of the LIDC-IDRI and ZSDB, respectively. The classification uses the Random Forest algorithm with two classes (Benign/Malignant) for LIDC-IDRI and four classes (AAH, AIS, MIA, IA) for ZSDB databases.

4.1. Pulmonary nodule segmentation

The K-means clustering is used to segment the lung nodule in the region of interest. There are some instances that belong to a cloudy and ground-glass-like appearance of the part-solid nodules, and there are nodules that include vascular infiltration. The segmentation method based on K-means clustering has a good segmentation effect with clear edges on (part) solid or ground-glass nodules, which obtain well-circumscribed, juxta-vascular, juxta-pleural or pleural tail nodules. Segmentation images of lung nodules are shown in Fig. 5(A).

4.2. Weighted grey scale angular density distribution feature extraction

For LIDC-IDRI, we found the location of nodules using the coordinates marked in the attached XML file. For ZSDB, the method described in Section 3.1 was used to extract the correct coordinates of the nodule edges. The feature extraction method proposed in Sections 3.2 and 3.3 was used to calculate the feature vector of nodule images. Our work used a large number of small size nodules, so the window size is small. Choosing the number of bins is also directly related to the nodule size. When the bin spacing is too small, the number of pixels in the bins is rather small and will reduce the distinction between the bins.

We set three types of the angle step are (30◦, 60◦, 90◦). The features extracted using the proposed method are named as the Exponential Weighting based feature (Ew). The different angle steps constructed 3 sets of weighted angular density distribution feature, such as Ew-30, Ew-60 and Ew-90. The features have difference responses to the identification process of nodule image set.

The characterization process using the proposed method of the segmented nodule images are shown in Fig. 5(B). The meaning of these colours is the same as the colour code in Fig. 2. The features of AAH, AIS, MIA were mainly composed of lower density levels. These factors corresponded to the GGO regions of nodule images. Otherwise, the nodules of the IA category have a large number of higher density level components. The result accorded with the radiologic characteristics, since IA nodule images are composed of a large area of the solid region (Fig. 5(C)). Fig. 5(D) shows the visualization of normalization density distribution map for LIDC-IDRI.

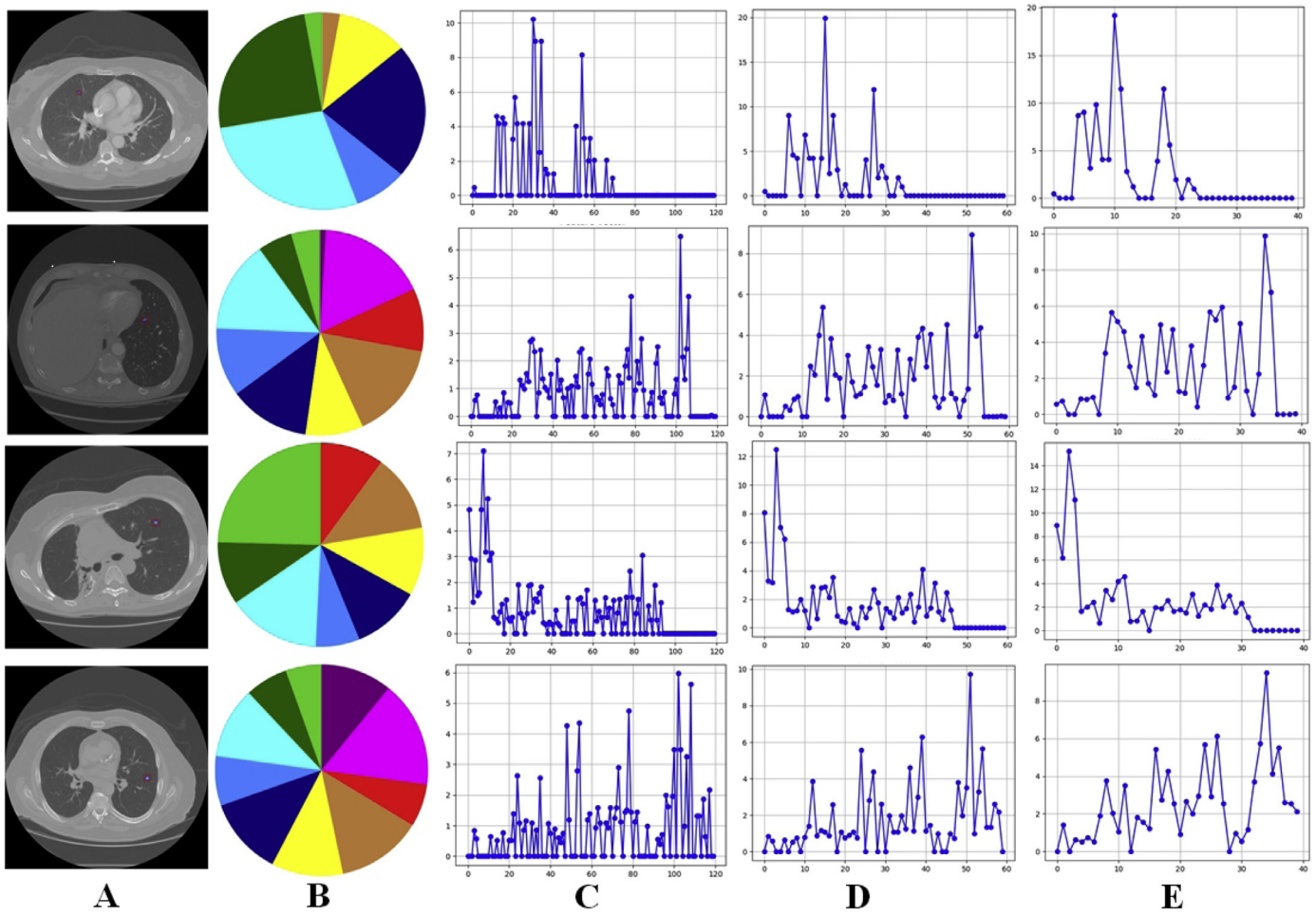

Fig. 6. Lung nodule angular histogram exponential weighted features extraction. (A)Source image of the pulmonary nodule, the red line is the nodule region; (B) Normalization of density distribution feature vector; (C) to (E) are Ew-30, Ew-60, Ew-90, respectively.

Table 1
p-values of 3 features for different LIDC-IDRI data configurations and for the ZSDB database.

	LIDC-IDRI			ZSDB					
	Config. 1	Config. 2	Config. 3	AAH-AIS	AAH-MIA	AAH-IA	AIS-MIA	AIS-IA	MIA-IA
Ew-30	0.0043	0.0043	0.0120	0.0005	0.0311	0.0009	1.96e−06	4.47e−06	3.21e−05
Ew-60	5.88e−07	1.00e−06	0.0019	0.0003	0.0260	0.0004	6.14e−08	6.69e−07	4.79e−06
Ew-90	1.68e−08	2.05e−08	0.0012	0.0003	0.0103	3.88e−08	2.14e−14	2.09e−14	6.99e−12
Average	**0.0014**	**0.0014**	**0.0050**	**0.0003**	**0.0224**	**0.0004**	**6.74e−07**	**1.71e−06**	**1.23e−05**

The examples of exponential weighted grey scale angular density distribution feature extraction are shown in Fig. 6. There are three types of features computed from the segmented nodule are shown in Fig. 6(C) to (E). The proposed feature vectors can describe the spatial correlation between pixels in the nodule image. The different distribution and dimensions of feature vectors were extracted from the same density distribution map. Varying bins or weighting functions can make a difference to the amplitude of factor vectors. In this process, the smaller the bins get, the longer the vector increases and the smaller the amplitude for each factor. This is a more in-depth study compared to Maldonado et al. [11].

4.3. Feature set analysis

The *results of Sensitivity and Specificity f* values for classification from the LIDC-IRDI and ZSDB databases are shown in Table 1 for all features. The averages of *p* values are less than 0.02 for the evaluation data. In the evaluation for the feature set of LIDC-IDRI, the *p* values of Configuration 3 are higher than those of Configuration 1 and Configuration 2. Where, the highest *p* value is 0.0120 with the Ew-90 feature. The result shows the rank 3 nodules are more inclined to be characterized as benign. For the feature set of ZSDB, the larger *p* values are for AAH and MIA. The best saliency value is that of AIS with IA. That is, because of the interference of the juxta-vascular nodules in ZSDB.

4.4. Pulmonary nodule classification

4.4.1. Classification for LIDC-IDRI database

In the experiments, all the features are used together to evaluate the proposed method. The best classification effect of the proposed system was a combination of the performance for all the features. The training and testing process used ten-fold modes for LIDC-IRDC and five-fold modes for ZSDB. The parameters of Sensitivity, Specificity, AUC and Accuracy are the top of 500 times for each evaluated data set.

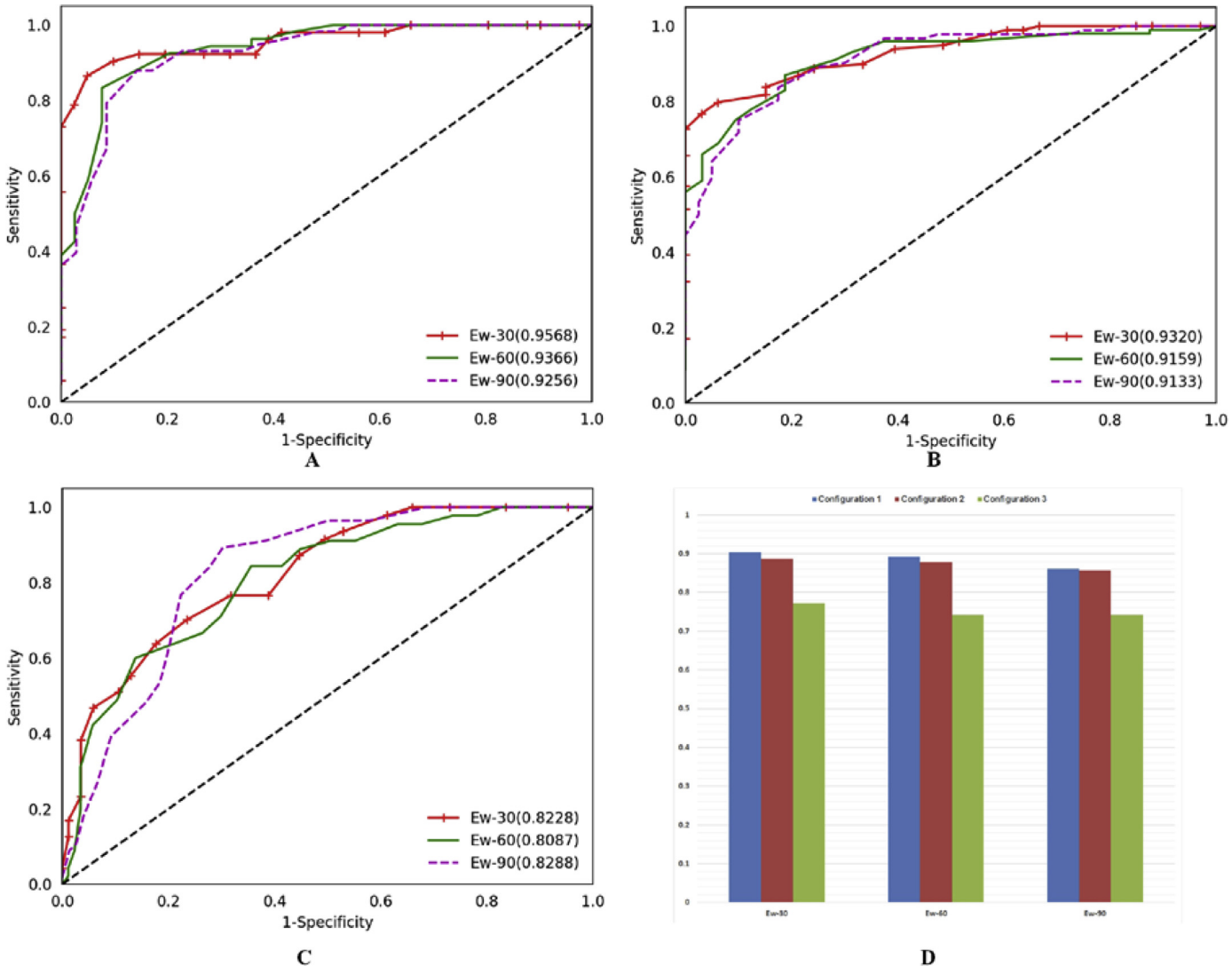

Fig. 7. The ROC curves and classification accuracy for different configurations from the LIDC-IDRI. (A) Configuration 1; (B) Configuration 2; (C) Configuration 3; (D) Classification accuracy for different configurations from the LIDC-IDRI with the proposed three different features.

Table 2
Sensitivity and Specificity of 3 features for Configuration 1, Configuration 2 and Configuration 3.

	Configuration1		Configuration2		Configuration3		ZSDB	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Ew-30	0.9433	0.9428	0.9065	0.9200	0.7804	0.8289	0.8409	0.9535
Ew-60	0.9230	0.9375	0.8990	0.9166	0.7608	0.8266	0.8550	0.9599
Ew-90	0.9318	0.9655	0.9000	0.9411	0.7380	0.8518	0.8280	0.9523
Average	0.9327	0.9486	0.9018	0.9259	0.7597	0.8358	0.8413	0.9552

The proposed method in this paper used the nodule image features of exponential weighting grey scale angular density distribution, and a Random Forest algorithm is performed to classify the pulmonary nodules. This method automatically increases the significant components of feature vectors and reduces the interference factors. The results of Sensitivity and Specificity for the LIDC-IDRI and ZSDB datasets are shown in Table 2. The comparisons with the methods of Dhara [13] and Han [5] are provided in Table 3 for different configurations. The statistics show that the sensitivity is more statistically significant than the result of Dhara et al. The AUC and classification accuracy for different configurations are shown in Fig. 7. The best values of AUC and Accuracy of the 3 features are (0.9568, 0.9320, 0.8288), (0.9032, 0.8863, 0.7727) for Configurations 1, 2, and 3, respectively. The classification accuracy of Configuration 1 is higher than that of Configuration 2 and 3. The worst was the result of Configuration 3. The effectiveness of the proposed method is just a little less than Dhara et al. [13] for the AUC value in Configuration 3. However, the other evaluation statistics are higher for the three configurations. The classification accuracy of the proposed method outperforms the most recent classification work.

4.4.2. Classification for ZSDB database

The boxplots for sensitivity, specificity, AUC and accuracy of multiple classes with 3 features in 500 evaluation times for the ZSDB dataset are provided in Fig. 8. The average of Sensitiv-

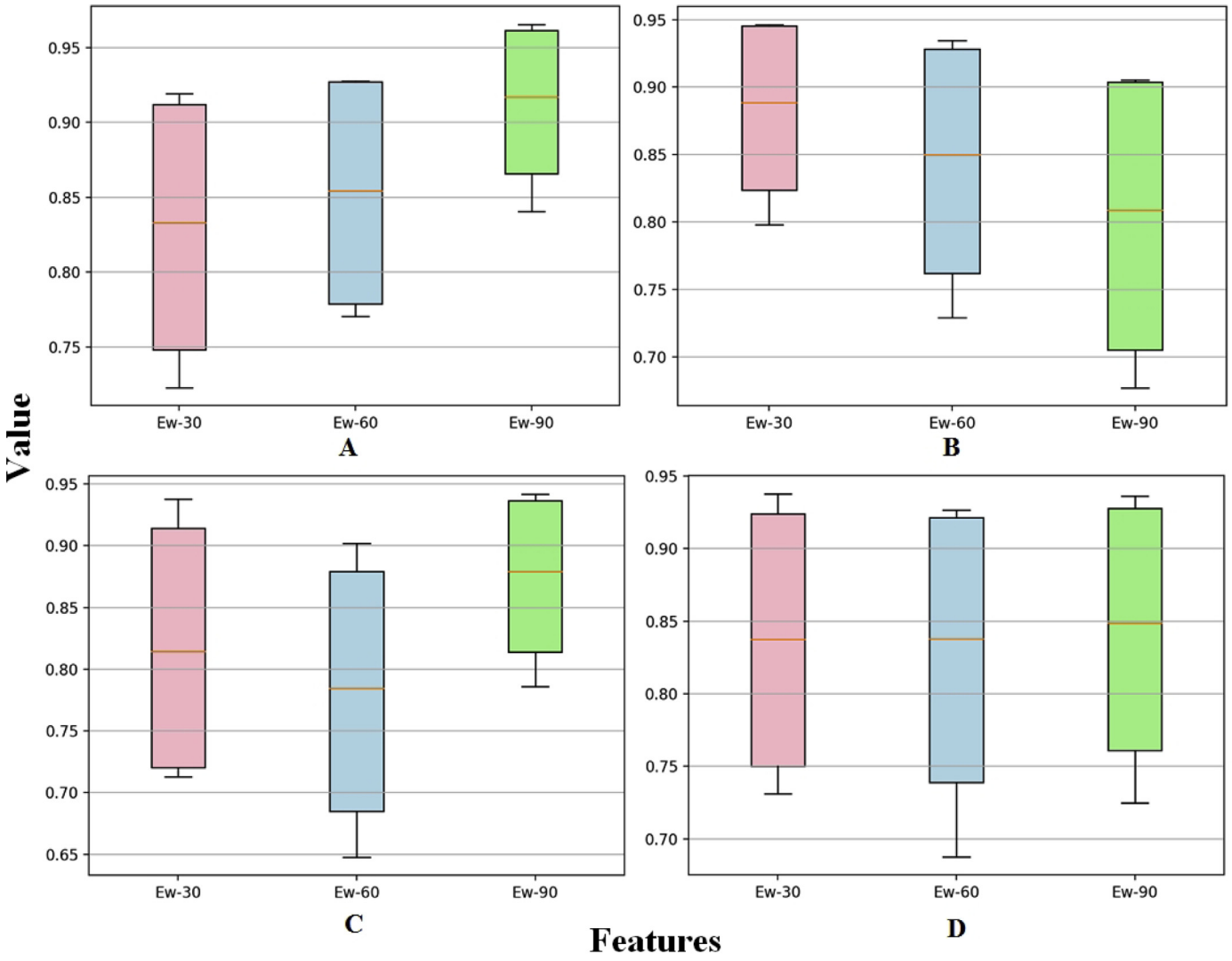

Fig. 8. Boxplots for sensitivity, specificity, AUC and accuracy of multiple classes with 3 features in 500 evaluation times for the ZSDB dataset. (A) Sensitivity; (B) Specificity; (C) ROC; (D) Accuracy.

Table 3
AUC of ROC values for three configurations from the LIDC-IDRI compared with Dhara et al. and Han et al.

	Samples	Configuration1			Configuration2			Configuration3		
		Sens.	Spec.	AUC	Sens.	Spec.	AUC	Sens.	Spec.	AUC
Proposed	1318	0.9327	0.9486	0.9568	0.9018	0.9259	0.9320	0.7597	0.8358	0.8288
Dhara et al.	0891	0.8973	0.8636	0.9505	0.8289	0.8073	0.8822	0.7614	0.7491	0.8488
Han et al.	1356	0.8935	0.8602	0.9450	0.8023	0.7914	0.8703	0.7467	0.7240	0.8315

*Sens.: Sensitivity
*Spec.: Specificity

ity, Specificity, AUC and Accuracy are 0.8704, 0.9697, 0.9715 and 0.9075, respectively. In the evaluation of multiple classes, the best of ROC and Accuracy are (0.9771, 0.9917, 0.9590, 0.9971), and (0.7478, 0.9167, 0.7450, 0.9567) for AAH, AIS, MIA, and IA, respectively. The true positive rate (Sensitivity) of the AIS and IA classes is higher than that of the other classes. At the same time, for the AAH category, the false positive rate is lower than that of the classes of AIS, MIA and IA. However, the nodule samples of MIA are easy to be confused with the AIS or IA classes. The main reason is the influence of the nodule surrounding the factors of vascular vessels. The experimental results show the high performance for an early cancer detection process.

The difference in the mean of the recognition rates is not significant for the angle steps (30, 60 and 90), and they are close up 0.85. However, the change trend of sensitivity and specificity are the opposite for the features, while the ROC of Ew-90 is better than the others, as shown in Fig. 8. The sensitivity is increasingly with the higher angle step, and the best of the mean of sensitivity is bigger than 0.95 (mean of standard deviation less than 0.05) with Ew-90 feature. At the same time in these cases, the highest specificity is close to 0.95 (standard deviation less than 0.05) with Ew-30 feature. In summary, the experiments show the high-performance for nodule risk classification of the proposed feature extraction methods for Chinese lung cancer cases.

5. Discussion

For the pulmonary nodule risk classification problem, we proposed an image characterization method based on a grey scale density distribution. We designed a clustered integral image (size of 200×200, with HU values ranging from −824 to 176) based on a K-means algorithm ($K = 10$) to extract the image density of nodules. For each nodule, we calculated the angular histogram features (angle steps are 30○, 60○, 90○), and weighted the features using an exponential distribution function. For the different angle steps, the classification performance of the feature set extracted from the smaller angle step is better than the others. Although the *p*-value of Ew-30 is larger than the other two features, but the ROC and accuracy are better. The best AUCs were 0.9568, 0.9320, and 0.8288 for Configurations 1, 2, and 3, respectively. The proposed method is compared with the recent classification work of Dhara et al. [13] and Han et al. [5] for the evaluated data of LIDC-IDRI. The configuration of the nodules is suggested by Han et al., they proposed a nodule classification based on image texture feature with 1356 samples. The results show that the differentiation rate and AUC are 90% and 92.7%, respectively. Dhara et al. also used this schema to validate their approach. They used 2D shape-based, 3D shape based, margin-based, and texture-based features of nodules image. The classification effect is good for complex texture features. The evaluated data was 891 nodules from the LIDC-IDRI, with most of data points being solid nodules. As a result, maximum A_z values were above 0.9505 for Configuration 1. In addition, the experiment data suggests that nodules with composite rank of malignancy "3" share more common features with the benign category, which fits with the discussion of Han [5] and Dhara [13]. The proposed classification scheme outperforms the state-of-the-art algorithms for all configurations for the LIDC-IDRI dataset.

At the same time, the evaluation of the ZSDB dataset shows the efficient performance of the proposed method. This method can distinguish the nodule categories into AAH, AIS, MIA and IA with high precision. The best average AUC of 0.9812 is achieved for the multi-classes validation. In this nodule recognition schema, most of the previous work did not consider the prediction of AAH with other categories for lung adenocarcinomas [12,15,19]. Foley et al. [19] also used CT density distribution features to characterize nodule images for 264 and 294 cases of the Mayo Clinic cohort and the National Lung Screening Trial study, respectively. The categories of the dataset are AIS, MIA, and IA. The results show a sensitivity of 95.4% (95% CI: 75.1%–99.7%) and specificity of 96.8% (95% CI: 82%–99.8%) in the training set and a sensitivity of 98.7% (95% CI: 91.8%–99.9%) and specificity of 63.6% (95% CI: 31.6%–87.6%) in the independent validation set. In the comparison with the state-of-art, we extend the category to four categories, as AAH, AIS, MIA and IA and achieve good results for the classification. The experimental results illustrate the robustness of the proposed method for different lung CT image datasets and different forms of classification.

6. Conclusion

The paper presents an angular gray scale density distribution feature weighted by exponential function for lung nodule image risk classification. In the proposed method, the pixels in nodule image are divided into different Fdistribution levels to generate the density distribution features. The experiments show a good classification effectiveness for different validation datasets such as LIDC-IDRI and ZSDB. A random forest classifier is performed for training and testing. The categories are Benign/Malignant and AAH/AIS/MIA/IA for LIDC-IDRI and ZSDB, respectively. A number of experiments were conducted with the proposed method as well as some other existing methods for the validation in LIDC-IDRI.

The experimental results provide a great reference in clinical diagnosis as well as the development of electronic diagnostic systems that detect early-stage lung cancer for patients in China or Asia. It has significance to support doctors in performing clinical diagnoses for lung cancer patients and increasing their survival rate when detecting first-stage lung cancer areas in time. Further work will be concerned about more effective feature extraction algorithms for proper representation of nodules in the feature space and will improve the performance of classification.

Conflict of interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgements

The authors greatly appreciate the financial supported by Zhongshan Hospital Clinical Research Foundation No. 2016ZSLC05, No.2016ZSCX02, National Key Scientific and Technology Support Program No.2013BAI09B09, National Natural Science Foundation of China No.81500078 and The Natural Science Foundation of Shanghai No. 15ZR1408700.

Appendix

Tables A1 and A2.

Table A1
Two datasets used to this study.

LIDC-IDRI database

Ranks	Sample Number	Grey Scale (min to max, HU)	Major length Diameter (min to max, mm)
"1"	139	−805 to −73	2.79 to 11.29
"2"	392	−821 to −25	3.53 to 13.65
"3"	393	−847 to 11	3.46 to 14.32
"4"	257	−743 to −4	4.28 to 15.77
"5"	137	−719 to 0	5.08 to 15.51
Total	1318		

ZSDB database

Classes	Sample number	Ages (Years, Y)	Gender (Male, M; Female, F)	Grey Scale (min to max, HU)	Diameter (min to max, mm)
AAH	092	63.5 ± 8.5	F:M = 25%:75%	−810.95 to −6.24	4.12 to 13.93
AIS	157	57 ± 21	F:M = 67%:33%	−869.56 to −453.33	4.82 to 20.88
MIA	158	56 ± 22	F:M = 77%:23%	−845.01 to 33.61	4.05 to 30.75
IA	188	58 ± 27	F:M = 51%:49%	−450.63 to 18.87	5.57 to 31.67
Total			595		

Table A2
The sample number for each configuration of the LIDC-IDRI database.

Configurations	Benign		Malignant	
	Ranks	Samples	Ranks	Samples
Configuration 1	'1' and '2'	531	'4' and '5'	394
Configuration 2	'1', '2', and '3'	924	'4' and '5'	394
Configuration 3	'1' and '2'	531	'3', '4' and '5'	787

References

[1] A.K. Dhara, S. Mukhopadhyay, N. Khandelwal, Computer-aided detection and analysis of pulmonary nodule from CT images: a survey, IETE Tech. Rev. 29 (2012) 265–275.

[2] J.M. Goo, C.M. Park, H.J. Lee, Ground-glass nodules on chest CT as imaging biomarkers in the management of lung adenocarcinoma, Am. J. Roentgenol. 196 (2011) 533–543.

[3] S. Noriyuki, Y. Hidetake, K. Masatoshi, K. Tsukasa, N. Kazuya, K. Satoshi, K. Takeshi, Y. Masato, N. Michinobu, H. Hiroshi, Volumetric measurement of artificial pure ground-glass nodules at low-dose CT: comparisons between hybrid iterative reconstruction and filtered back projection, Eur. J. Radiol. 84 (2015) 2654–2662.

[4] Y.-C. Yeh, K. Kadota, J.-i. Nitadori, C.S. Sima, N.P. Rizk, D.R. Jones, W.D. Travis, P.S. Adusumilli, International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society classification predicts occult lymph node metastasis in clinically mediastinal node-negative lung adenocarcinoma, Eur. J. Cardiothorac. Surg. 49 (2016) 9–15.

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

[6] A.P. Reeves, J.A. Xie, Automated pulmonary nodule CT image characterization in lung cancer screening, Int. J. Comput. Assist. Radiol. Surgery 11 (2016) 73–88.

[7] Cornell University, Vision and Image Analysis Group, ELCAP Public Lung Image Database. Available from: http://www.via.cornell.edu/lungdb.html.

[8] National cancer institute- Cancer Data Access System. Available from: https://biometry.nci.nih.gov/cdas/datasets/nlst/.

[9] H.M. Orozco, O.O.V. Villegas, V.G.C. Sánchez, H.D.J.O. Domínguez, M.D.J.N. Alfaro, Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine, Biomed. Eng. Online 14 (2015) 9.

[10] R. Sushravya, M. Fabien, R. Srinivasan, A.K. Ronald, S.D.P. Zackary, J.B. Brian, P. Tobias, A.R. Richard, Noninvasive risk stratification of lung adenocarcinoma using quantitative computed tomography, J. Thoracic Oncol. 9 (2014) 1698–1703.

[11] F. Maldonado, F. Duan, S.M. Raghunath, S. Rajagopalan, R.A. Karwoski, K. Garg, E. Greco, H. Nath, R.A. Robb, B.J. Bartholmai, T. Peikert, Noninvasive computed tomography–based risk stratification of lung adenocarcinomas in the National Lung Screening Trial, Am. J. Respir. Crit. Care Med. 192 (2015) 737–744.

[12] F. Maldonado, M.B. Jennifer, R. Sushravya, C.A. Marie, J.B. Brian, D.A. Mariza, E.H. Thomas, A.K. Ronald, R. Srinivasan, A.M. Sykes, P. Yang, S.Y. Eunhee, A.R. Richard, P. Tobias, Noninvasive characterization of the histopathologic features of pulmonary nodules of the lung adenocarcinoma spectrum using computer-aided nodule assessment and risk yield (CANARY)-a pilot study, J. Thoracic Oncol. 8 (2013) 452–460.

[13] A.K. Dhara, S. Mukhopadhyay, A. Dutta, M. Garg, N. Khandelwal, A combination of shape and texture features for classification of pulmonary nodules in lung CT images, J. Digit. Imaging 6 (2016) 1–10.

[14] S.G. Armato III, G. McLennan, L. Bidaut, M.F. McNitt-Gray, C.R. Meyer, A.P. Reeves, B. Zhao, D.R. Aberle, C.I. Henschke, E.A. Hoffman, E.A. Kazerooni, H. MacMahon, E.J. Van Beeke, D. Yankelevitz, A.M. Biancardi, P.H. Bland, M.S. Brown, R.M. Engelmann, G.E. Laderach, D. Max, R.C. Pais, D.P. Qing, R.Y. Roberts, A.R. Smith, A. Starkey, P. Batrah, P. Caligiuri, A. Farooqi, G.W. Gladish, C.M. Jude, R.F. Munden, I. Petkovska, L.E. Quint, L.H. Schwartz, B. Sundaram, L.E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A.V. Casteele, S. Gupte, M. Sallamm, M.D. Heath, M.H. Kuhn, E. Dharaiya, R. Burns, D.S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B.Y. Croft, The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans, Med. Phys. 38 (2011) 915–931.

[15] J.Y. Son, H.Y. Lee, K.S. Lee, J.H. Kim, J. Han, J.Y. Jeong, O.J. Kwon, Y.M. Shim, Quantitative CT analysis of pulmonary ground-glass opacity nodules for the distinction of invasive adenocarcinoma from pre-invasive or minimally invasive adenocarcinoma, PLoS One 9 (2014) e104066.

[16] Y. Zhu, Y. Tan, Y. Hua, M. Wang, G. Zhang, J. Zhang, Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography, J. Digit. Imaging 23 (2010) 51–65.

[17] A.A.A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S.J. van Riel, M.M.W. Wille, M. Naqibullah, C.I. Sánchez, B. van Ginneken, Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks, IEEE Trans. Med. Imaging 35 (2016) 1160–1169.

[18] A. El-Baz, M. Nitzken, F. Khalifa, A. Elnakib, G. Gimelfarb, R. Falk, M.A. El-Ghar, 3D shape analysis for early diagnosis of malignant lung nodules, in: Biennial International Conference on Information Processing in Medical Imaging, Springer Berlin Heidelberg, 2011, pp. 772–783.

[19] F. Foley, S. Rajagopalan, S.M. Raghunath, J.M. Boland, R.A. Karwoski, F. Maldonado, B.J. Bartholmai, T. Peikert, Computer-aided nodule assessment and risk yield risk management of adenocarcinoma: the future of imaging? in: Seminars in Thoracic and Cardiovascular Surgery, 28, 2016, pp. 120–126.

[20] The Cancer Imaging Archive. https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI.

[21] L. Gonçalves, J. Novo, A. Campilho, Hessian based approaches for 3D lung nodule segmentation, Expert Syst. Appl. 61 (2016) 1–15.

[22] P. Badura, E. Pietka, Soft computing approach to 3D lung nodule segmentation in CT, Comput. Biol. Med. 53 (2014) 230–243.

[23] EzhilE. Nithila, S.S. Kumar, Segmentation of lung nodule in CT data using active contour model and fuzzy C-mean clustering, Alexandria Eng. J. 55 (2016) 2583–2588.

[24] W. Li, P. Cao, D. Zhao, J. Wang, Pulmonary nodule classification with deep convolutional neural networks on computed tomography images, Comput. Math. Methods Med. (2016).

[25] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1, 1967, pp. 281–297.

[26] S. Bose, A. Pal, R. SahaRay, J. Nayak, Generalized quadratic discriminant analysis, Pattern Recognit. 48 (2015) 2676–2684.

[27] P. Boinee, A.D. Angelis, G. Foresti, Meta random forests, Int. J. Comput. Intell. 2 (2005) 138–147.