Information Fusion 42 (2018) 102–110

Contents lists available at ScienceDirect

Information Fusion

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

Full Length Article

Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT

Xie Yutong[a,b], Zhang Jianpeng[a,b], Xia Yong[a,b,*], Michael Fulham[b,c,d,e], Zhang Yanning[a]

[a] Shaanxi Key Lab of Speech & Image Information Processing (SAIIP), School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
[b] Centre for Multidisciplinary Convergence Computing (CMCC), School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
[c] Department of PET and Nuclear Medicine, Royal Prince Alfred Hospital, NSW 2050, Australia
[d] Sydney Medical School, University of Sydney, NSW 2006, Australia
[e] Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, University of Sydney, NSW 2006, Australia

ARTICLE INFO

Keywords:
Lung nodule classification
Chest CT
Deep convolutional neural network (DCNN)
Back propagation neural network (BPNN)
AdaBoost, information fusion

ABSTRACT

The separation of malignant from benign lung nodules on chest computed tomography (CT) is important for the early detection of lung cancer, since early detection and management offer the best chance for cure. Although deep learning methods have recently produced a marked improvement in image classification there are still challenges as these methods contain myriad parameters and require large-scale training sets that are not usually available for most routine medical imaging studies. In this paper, we propose an algorithm for lung nodule classification that fuses the texture, shape and deep model-learned information (Fuse-TSD) at the decision level. This algorithm employs a gray level co-occurrence matrix (GLCM)-based texture descriptor, a Fourier shape descriptor to characterize the heterogeneity of nodules and a deep convolutional neural network (DCNN) to automatically learn the feature representation of nodules on a slice-by-slice basis. It trains an AdaBoosted back propagation neural network (BPNN) using each feature type and fuses the decisions made by three classifiers to differentiate nodules. We evaluated this algorithm against three approaches on the LIDC-IDRI dataset. When the nodules with a composite malignancy rate 3 were discarded, regarded as benign or regarded as malignant, our Fuse-TSD algorithm achieved an AUC of 96.65%, 94.45% and 81.24%, respectively, which was substantially higher than the AUC obtained by other approaches.

1. Introduction

The 2015 global cancer statistics showed that there are approximately 14.1 million new cancer cases each year. Lung cancer has an incidence of 13% and death rate of 19.5%, which are the highest rates across all cancers [1]. Early diagnosis and treatment are the most effective means to improve survival of lung cancer patients; the 5-year survival for those with an early diagnosis is approximately 54 % when compared to 4 % if the diagnosis is late when the patient has stage IV disease [2]. A "spot" on the lung, detected by chest computed tomography (CT), which measures less than 3 cm in diameter is defined as a lung nodule and may be benign or malignant. The National Lung Screening Trial showed that screening with CT will result in a 20% reduction in lung cancer deaths, by detecting early disease [3]. Radiologists globally typically visually analyze chest CT scans on a slice-by-slice basis, which is time-consuming, expensive and prone to reader bias and requires a high degree of skill and concentration.

Computer-aided diagnosis (CAD), however, avoids many of these issues and is increasingly being investigated as an alternative and complementary approach to conventional reading [4]. Many automated lung nodule classification approaches have been proposed in the literature and most of them consist of image preprocessing, nodule detection, nodule segmentation, feature extraction and classification. Among them, feature extraction is a critical step. The features used for lung nodule classification can be divided into hand-crafted features and features learned by deep neural networks (DNNs). Hand-crafted features include texture and shape descriptors, since there is a high correspondence between nodule malignancy and heterogeneity in voxel values and shape [5]. Once hand-crafted features are extracted, a variety of classification techniques can be used including the support vector machine (SVM) [6,7], decision tree [8,9], K-nearest neighbor (KNN) [10], back propagation neural network (BPNN) [11,12], random forest (RF) [13] and Adaboost [14,15].

The most commonly used usual texture descriptor is based on the

* Corresponding author at: Shaanxi Key Lab of Speech & Image Information Processing (SAIIP), School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China.
E-mail address: yxia@nwpu.edu.cn (Y. Xia).

http://dx.doi.org/10.1016/j.inffus.2017.10.005
Received 22 October 2016; Received in revised form 12 October 2017; Accepted 17 October 2017
Available online 25 October 2017
1566-2535/

gray level co-occurrence matrix (GLCM) [16]. Wang et al. [17] reported on 14 GLCM features that were extracted from CT images and a multilevel binomial logistic prediction model to classify 2171 benign or malignant lung nodules. Wu et al. [18] calculated 13 GLCM features and 12 radiological features and applied them to a BPNN and they found that incorporating radiological features into texture descriptors improves the accuracy of differentiating malignant from benign nodules. Zhao et al. [19] used GLCM texture features of CT images and metabolic features of PET images for describing solitary pulmonary nodules (SPNs) and used SVM to classify SPNs. Han et al. [20] further extracted 3D GLCM features to train a SVM for lung nodule classification.

Several shape descriptors, including the Feret shape measure, roundness, moment invariants, point distance histogram and a Fourier descriptor, can be used to characterize the heterogeneous shape of nodules. Frejlichowski [21] showed that the Fourier descriptor has an excellent performance in general shape analysis. Yoshimasu et al. [22] applied fast Fourier transform (FFT) analysis to quantify the complexity of nodule outlines. Ciompi et al. [23] reported a novel descriptor that sampled intensity profiles along circular patterns on spherical surfaces centered on the nodule in a multi-scale fashion to capture information on nodule morphology. These investigators used the Fourier transform to obtain a spectrum for each intensity profile. Nakamura et al. [24] selected Fourier transformation-based shape features to describe the nodule outline and adopted the radial gradient index to measure nodule speculation.

Texture and shape descriptors, although widely used, merely characterize one perspective of nodule heterogeneity. DNNs [25] are now increasingly being employed to learn image representation from a large training dataset. The deep convolutional neural network (DCNN) provides a uniform framework for learning-based joint feature extraction and classification and can learn highly representative, hierarchical image features with sufficient training data [26,27]. Krizhevsky et al. [27] developed an eight-layer AlexNet which reduced the top-5 error rate of image classification in the 2012 ImageNet Challenge. Anthimopoulos et al. [28] reported a nine-layer DCNN to classify patterns of interstitial lung disease. Wei et al. [29] proposed a multi-crop convolutional neural network (MC-CNN) to capture nodule heterogeneity. Hua et al. [30] applied a DCNN and deep belief network (DBN) and reported that deep learning achieved better discrimination between benign and malignant nodules.

Despite improved accuracy, these deep models have not achieved the same performance on routine lung nodule classification as they have on the ImageNet data. The suboptimal performance is attributed mainly to the over-fitting of deep models due to inadequate training data, as there is usually a relatively small dataset in medical image analysis, which relates to the work involved in image acquisition and annotation. To overcome this limitation, we suggest using the information extracted from CT images by a DCNN combined with traditional texture and shape descriptors in a fusion approach [31]. There are many approaches to data fusion [32] and the main differences relate to the level where the fusion takes place [31,33–36]. Feature fusion combines different types of features to form a new feature representation [33]. Xie et al. [37] fused texture, shape and deep features learned by a nine-layer DCNN and applied the combined features to a BPNN classifier. Decision fusion combines the decisions reached by using different data or features independently [38].

In this paper, we propose a novel algorithm that fuses texture, shape and deep model-learned information (Fuse-TSD) at the decision level for differentiating benign from malignant lung nodules. Three types of features are calculated on each lung nodule image patch. One is learned by an eight-layer DCNN, and the other two are the GLCM-based texture descriptors and Fourier shape descriptor, which characterize the heterogeneity in voxel values and shape for the nodule. For each feature type, we construct an ensemble classifier based on BPNN and AdaBoost. The decisions made by the three classifiers are fused via computing a weighted sum of likelihood, where the weight of each decision is estimated according to its accuracy on the validation set. Our proposed Fuse-TSD algorithm can effectively separate malignant from benign nodules when tested on the LIDC-IDRI dataset.

2. Description of datasets

LIDC-IDRI [39–41] is an open database in the cancer imaging archive (TCIA) for lung cancer diagnosis that contains 1018 clinical chest CT scans from seven institutions. Each scan has an associated XML file that details the locations and boundaries of nodules on each 512 × 512 slice that were read by up to four experienced thoracic radiologists. Each suspicious lesion was categorized as a non-nodule, a nodule $<$ 3 mm, or a nodule $\geqq$3 mm diameter in the long axis. Nodules are rated on a 5 point scale, from benign to malignant, by up to four experienced thoracic radiologists. Nodules $<$ 3 mm are not considered to be clinically relevant by current screening protocols [42,43]. Hence, we only considered nodules $\geqq$ 3 mm in diameter that were annotated by at least one radiologist and, therefore, we included 2669 lung nodules. For each nodule, we define the malignancy level given by one radiologist or the median malignancy level given by multiple radiologists as the composite malignancy rate [44]. The distribution of composite malignancy rate over 2669 selected nodules is shown in Table 1.

We used the schema suggested by Han et al. [20] and Wei et al. [29], where a lung nodule with a composite malignancy rate 1 or 2 is regarded as benign, a rate 4 or 5 is regarded as malignant and a rate of 3 has uncertain malignancy. Nodules rated at 3 can be discarded or regarded as benign or malignant. Accordingly, there are three possibilities to label these nodules for this study. If the nodules with a composite malignancy rate of 3 are discarded, there are 1324 benign and 648 malignant nodules (denoted by the Dataset D1); if they are regarded as benign, there are 2021 benign and 648 malignant nodules (denoted by the Dataset D2); and if they are regarded as malignant there are 1324 benign and 1345 malignant nodules (denoted by the Dataset D3) (see Table 2).

3. Fuse-TSD algorithm

Our Fuse-TSD algorithm consists of offline training and online testing. The offline training involves four major steps: (1) extracting the nodule region of interest (ROI) from each chest CT slice; (2) training a DCNN for deep feature extraction and extracting texture and shape features; (3) using each feature type to train an AdaBoosted BPNN; and (4) estimating the weight of each AdaBoosted BPNN to fuse the classification. The online testing has four steps: (1) extracting the nodule ROI from each slice of the to-be-tested nodule; (2) applying each ROI to the trained DCNN to generate its deep features and extracting its texture and shape features; (3) applying deep, texture and shape features to the trained ensemble classifier; and (4) classifying the entire nodule based on the label of each nodule slice. The algorithm is summarized in Fig. 1.

3.1. Lung nodule ROI extraction and segmentation

We extract the lung nodule(s) on a slice-by-slice basis using the location and boundary information provided by LIDC-IDRI. Since the largest nodule diameter is 64 mm, we first crop a 64 × 64 square region centered on the middle of the nodule centers provided by the radiologists, and then detect the nodule area, which is defined as the intersection of the areas marked by radiologists. Next, we draw a square

Table 1
Composite rate of malignancy in LIDC-IDRI for lung nodules ≥ 3 mm.

Composite Malignancy Rate	1	2	3	4	5
Number of Lung Nodules	389	935	697	464	184

Table 2
Three datasets used for this study.

Dataset		D1	D2	D3
Benign	Composite Malignancy Rate	1, 2	1, 2, 3	1, 2
	Number of Nodules	1324	2021	1324
Malignant	Composite Malignancy Rate	4, 5	4, 5	3, 4, 5
	Number of Nodules	648	648	1345

bounding box around the nodule as a ROI and set non-nodule voxels in this ROI to zero to eliminate the influence of no-nodule voxels [20,29]. This process is illustrated in Fig. 2.

3.2. Feature extraction

3.2.1. DCNN-based deep feature extraction

Based on the LeNet-5 model [45], an eight-layer DCNN was constructed to automatically learn the representation of each ROI. Since nodule ROIs have a variable size, we resized each ROI to 32 × 32 using the bicubic interpolation algorithm before applying them to the DCNN. Thus, the input of the network was 32 × 32 image patches. In this DCNN model, three convolutional layers consisted of 32, 32 and 64 kernels of size 5 × 5, respectively. Each kernel produced a 2D feature map, and hence the output of the first convolutional layer, for instance, was 32 32 × 32 feature maps. Kernels may contain different matrix values that are initialized randomly and are updated during training to optimize the classification accuracy. Each convolutional layer is followed by a 3 × 3 average-pooling layer with a stride of 2, which reduces the size of feature maps into 1/4. The last two layers are fully connected layers, which is denoted as F7 and F8 in Fig. 3. The rectified linear units (ReLU) are used in the convolutional layers and fully connected layers as the activation function to implement a non-linear transformation from input to output. Dropout is employed in the F7 layer to avoid over-fitting, which means to set the output of each hidden neuron to zero with a probability of 0.5 [27].

We defined the output of the F7 layer of the trained network as a 64-dimensional feature vector, namely the deep feature. We empirically set the learning rate to 0.001 and the maximum iteration number to 50, and chose the batch training style with the batch size of 100. We adopted the default settings suggested in the MatConvNet Toolbox [46] for the other parameters.

3.2.2. GLCM-based texture feature extraction

We adapted the GLCM-based texture descriptor to characterize the heterogeneity of the nodule voxel values. To quantify the spatial dependence of voxel values [47] we calculated the energy, contrast, entropy and inverse difference, which have been proven to be irrelevant but effective for image classification [48]. We used four GLCMs counted at 0°, 45°, 90° and 135°and obtained a 16-dimensional GLCM texture descriptor for each image patch.

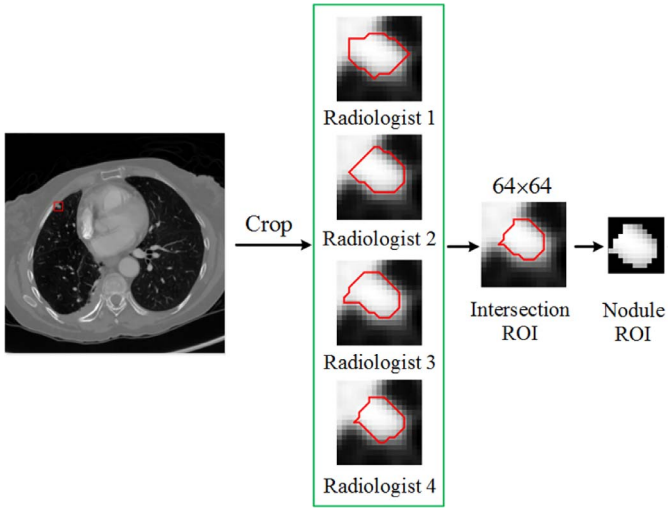

Fig. 2. Transaxial chest CT image with a small right middle lobe lung nodule adjacent to the pleura with ROI extraction and segmentation.

3.2.3. Fourier shape descriptor

On each image patch, we computed the Fourier descriptor of the nodule boundary to characterize the heterogeneity of the nodule shape. The computation had three steps: (1) identifying the center of gravity on the binary image patch, (2) moving a point along the nodule boundary and plotting the distance between the point and gravity center versus the geodesic distance that the point moved; and (3) applying the Fourier transform to this plot. We selected 52 low frequency coefficients as the Fourier descriptors of the nodule boundary.

3.3. AdaBoosted BPNN for patch classification

We use the AdaBoost algorithm to build an ensemble classifier, in which BPNN is the weak learner. Let the training data be denoted by$\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, where $x_i \in R^D$ is either the deep, texture or shape feature of the ith image patch, y_i representing the corresponding class label, and N is the number of training image patches. To construct a one-hidden-layer BPNN weak learner h_t, we sample 90% of

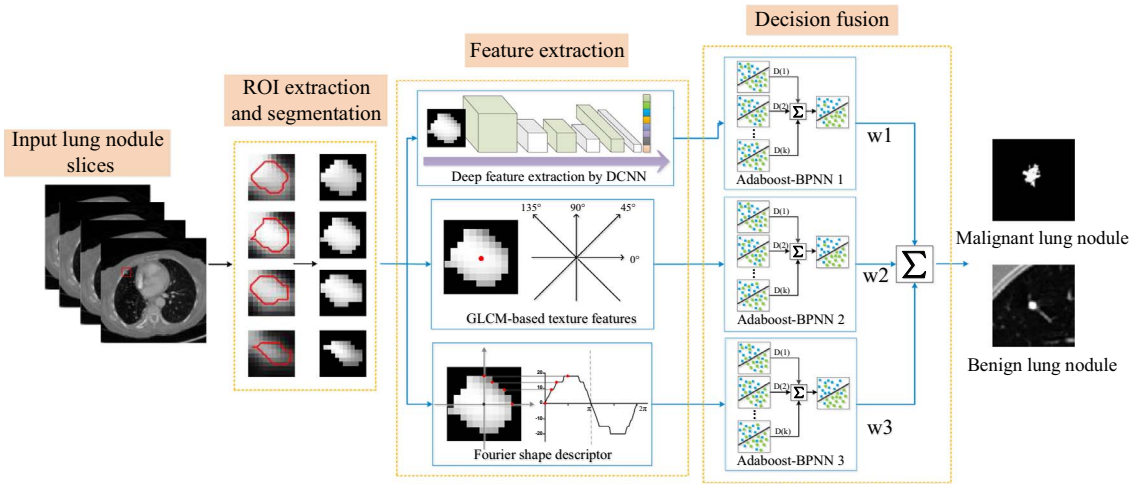

Fig. 1. Diagram of our proposed Fuse-TSD lung nodule classification algorithm.

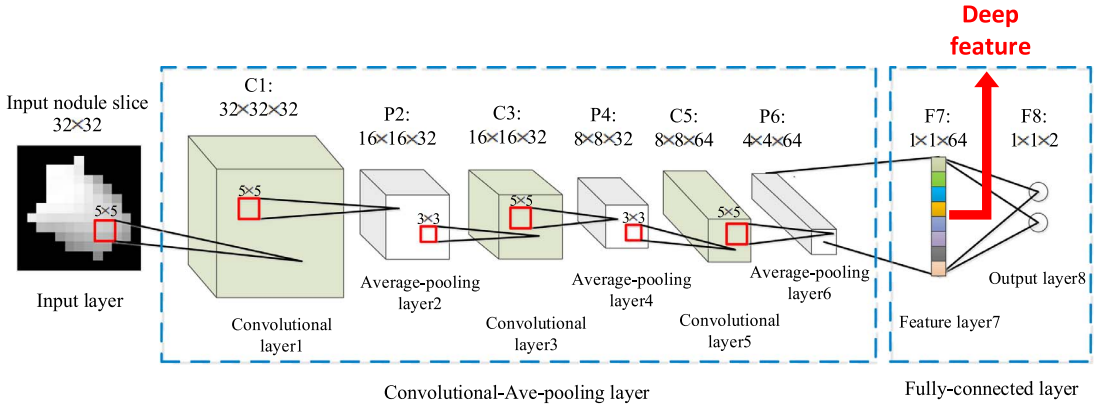

Fig. 3. Structure of the eight-layer DCNN.

training data according to the distribution of their weights, which are assumed to follow initially an uniform distribution, i.e.$\{w_0(i) = \frac{1}{N}, i = 1, 2, ..., N\}$, and left the other 10% of training data to form a validation set. In the BPNN, the number of neurons in the input layer is set to D, the number of output neurons is 2, and the number of hidden neurons is set to *log(D)* whose reasonability has been proved in [49]. We set the convergence error to 0.0004 which is the default in the BPNN toolbox in Matlab. During training the BPNN, if the error on the validation dataset was lower than 0.0004, the training stops. The learning rate is also a critical parameter. A small learning rate may lead to high training time; whereas a large learning rate may result in less accurate BPNN that is plunged into a local optimum. Hence, we set the learning rate to 0.001, representing a trade-off between time cost and performance. We adopted the default settings of other parameters suggested in the BPNN toolbox.

The trained BPNN was then tested on the validation set. Based on the obtained classification error $\boldsymbol{\varepsilon_t}$, the weight of this BPNN can be calculated as:

$$\alpha_t = \frac{1}{2} ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{1}$$

and the weight of each training data can be updated as:

$$w_{t+1}(i) = \begin{cases} \frac{w_t(i) exp(-\alpha_t y_i h_t(x_i))}{C_t}, & i \in m \\ \frac{w_t(i)}{C_t}, & others \end{cases} \tag{2}$$

where C_t is a normalization constant, and $\sum^{w_{t+1}}(i) = 1$. After training T BPNN in this way, the ensemble classifier can be defined as:

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t \tag{3}$$

We empirically set the number of weak BPNN classifiers T to 10. The construction of this ensemble classifier is summarized in the following Algorithm below. Since there are three groups of image features, we can train three AdaBoosted BPNNs.

Algorithm: AdaBoosted BPNN as the baseline classifier

1. Input: A set of training set $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), ..., (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$

2. Initialization: Set the initial weights of training examples to $\{\boldsymbol{w_0(i)} = \frac{1}{N}, \boldsymbol{i} = 1, 2, ..., \boldsymbol{N}\}$, and randomly initialize BPNN parameters

3. For t = 1:T

(1) Sample 70% of training examples according to the weights $\boldsymbol{w_t(i)}$, $\boldsymbol{i} = 1, 2, ..., \boldsymbol{N}$ for training and others for validation;

(2) Train the BPNN as a weak classifier h_t using those sampled data;

(3) Calculate the classification error ε_t on the validation set;

(4) Set the weight of the BPNN to $\boldsymbol{\alpha_t} = \frac{1}{2}\boldsymbol{ln}\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$.

(5) Update the weight of training samples:

$$\boldsymbol{w_{t+1}(i)} = \begin{cases} \frac{\boldsymbol{w_t(i) exp(-\alpha_t y_i h_t(x_i))}}{\boldsymbol{C_t}}, & i \in m \\ \frac{\boldsymbol{w_t(i)}}{\boldsymbol{C_t}}, & others \end{cases}$$

where $\boldsymbol{C_t}$ is a normalization constant, and $\sum \boldsymbol{w_{t+1}(i)} = 1$.

4. Construct the ensemble strong classifier: $\boldsymbol{H} = \sum \boldsymbol{\alpha_t h_t}$. H is the probability output of lung nodule image patches.

3.4. Nodule classification

Let a lung nodule $\boldsymbol{\Psi}$ consist of S slices, denoted by $\Psi = \{\Psi_1, \Psi_2, ..., \Psi_S\}$. Input the ith slice Ψ_i into the jth AdaBoosted BPNN classifier, and we obtain a two-dimensional prediction vector $\boldsymbol{H'_j}(\Psi_i), j \in \{1, 2, 3\}$. To avoid the preference for any specific model, we normalized each prediction vector as:

$$H_{jk}(\Psi_i) = \frac{H'_{jk}(\Psi_i)}{\sum_k H'_{jk}(\Psi_i)}, \quad k \in \{'benign', 'malignant'\} \tag{4}$$

where each element $H_{jk}(\Psi_i)$ represents the likelihood of nodule $\boldsymbol{\Psi}$ being benign or malignant predicted by using the jth group of features extracted on the ith slice. Let the classification error on the validation set obtained during training the jth AdaBoosted BPNN classifier be denoted by ε_j. The class label of lung nodule $\boldsymbol{\Psi}$ is assigned based on the weighted sum of likelihood generated respectively by using three types of features calculated on each slice, shown as follows:

$$arg \max_k \sum_{i=1}^{S} \sum_{j=1}^{3} \omega_j H_{jk}(\Psi_i) \tag{5}$$

where the weight ω_j is calculated as follows:

$$\omega_j = \frac{1}{2} \ln\left(\frac{1-\varepsilon_j}{\varepsilon_j}\right) \tag{6}$$

4. Experiments and results

We evaluated our Fuse-TSD algorithm on the three LIDC-IDRI lung nodule datasets given in Table 2 using 10-fold cross validation. The performance of this algorithm was assessed by using the area under the receiver operator curve (AUC), accuracy, sensitivity and specificity. To further evaluate the robustness of our algorithm, the experiment was performed 10 times independently, and the mean and standard deviation (std) of obtained metrics are given in Table 3.

Nodule classification on dataset D1 is relatively easy, since the

Table 3
Performance of different decision fusions on three datasets.

Features	Deep Model		X				X	X	X
	Texture			X		X	X		X
	Shape				X	X		X	X
D1	AUC (%)	Mean	95.67	88.89	96.24	95.83	95.98	96.52	**96.65**
		std	0.05	0.04	0.05	0.01	0.01	0.02	0.01
	Accuracy (%)	Mean	88.02	85.44	87.33	88.22	88.95	88.43	**89.53**
		std	0.11	0.07	0.12	0.09	0.06	0.07	0.09
	Sensitivity (%)	Mean	80.43	70.57	**85.83**	82.63	82.34	83.73	84.19
		std	0.03	0.27	0.35	0.20	0.06	0.16	0.09
	Specificity (%)	Mean	90.41	**92.67**	87.56	90.88	92.08	90.62	92.02
		std	0.07	0.17	0.43	0.21	0.14	0.27	0.01
D2	AUC (%)	Mean	92.91	84.39	87.38	92.96	94.4	94.41	**94.45**
		std	0.03	0.02	0.06	0.03	0.01	0.01	0.01
	Accuracy (%)	Mean	86.92	83.39	85.43	85.77	87.62	**87.74**	**87.74**
		std	0.09	0.05	0.09	0.08	0.06	0.07	0.03
	Sensitivity (%)	Mean	79.37	73.79	71.36	80.95	80.27	80.23	**81.11**
		std	0.63	0.40	1.39	0.20	0.56	0.86	0.85
	Specificity (%)	Mean	89.34	87.14	**89.88**	87.31	89.79	**89.88**	89.67
		std	0.30	0.14	0.52	0.38	0.07	0.11	0.09
D3	AUC (%)	Mean	76.64	76.66	76.51	77.55	80.66	81.18	**81.24**
		std	0.03	0.01	0.04	0.05	0.09	0.01	0.01
	Accuracy (%)	Mean	70.9	67.86	70.82	70.11	71.43	71.67	**71.93**
		std	0.06	0.04	0.08	0.08	0.04	0.08	0.04
	Sensitivity (%)	Mean	54.21	43.62	54.7	55.26	55.92	58.25	**59.22**
		std	0.04	0.43	0.15	0.07	0.05	0.07	0.04
	Specificity (%)	Mean	88.42	**93.58**	87.66	90.83	87.55	85.49	84.85
		std	0.15	0.10	0.29	0.19	0.10	0.20	0.10

troublesome “composite malignancy rate 3” cases were excluded. It shows that, by fusing the decisions made by using DCNN feature, GLCM-based texture descriptor and Fourier shape descriptor, the Fuse-TSD algorithm had the highest AUC and accuracy on this dataset. Its sensitivity was slightly lower than that achieved by using just the shape descriptor and its specificity was slightly lower than that achieved by using the deep and texture descriptor and using texture descriptor alone. For datasets D2 and D3, “composite malignancy rate of 3” cases were labelled as benign or malignant and this increased the classification difficulty. All approaches performed less well. Although our algorithm had the 3rd best specificity on D2 and 2nd best specificity on D3, it had the highest AUC, accuracy and sensitivity on D2 and D3. This means that fusing the decisions made by using three types of image features is still the optimal choice for nodule classification.

Next, we evaluated Fuse-TSD against three state-of-the-art lung nodule classification algorithms proposed by Han et al. [20], Dhara et al. [50] and Wei et al. [29]. For the first two algorithms, we acquired the source code and tested them on our datasets using the 10-fold cross validation. For the algorithm from Wei et al., we did not have the source code, and hence adopted its performance reported in Ref. [29], which was calculated on a dataset of 2618 nodules without using the troublesome “composite malignancy rate 3” cases.

The performance of the four algorithms is shown in Table 4. The algorithm from Han et al. [20] had the lowest accuracy, since they used only texture features to describe nodule appearance. Dhara et al. [50] carried out extensive mining of shape, margin sharpness and GLCM-based texture features and hence had higher accuracy. Although Wei et al. [29] used the MC-CNN model, their algorithm underperformed when compared to the multi-feature approach reported by Dhara et al [50]. Our Fuse-TSD algorithm achieved the highest AUC and accuracy on all three datasets by virtue of the joint use of deep, texture and shape features that better charactise nodule heterogeneity. It should be noted that the methods of Han et al. and Dhara et al. are prone to achieving bias performance, and hence sometimes have the highest sensitivity or specificity. We suggest that our algorithm, which can fuse the decisions made using three image features, is still the preferred choice for nodule classification with its better performance across the ‘clean’ D1 and the ‘problematic’ D2 and D3 datasets.

Table 4
Performance of the four algorithms across D1–D3 datasets.

Dataset	Measures	Han et al. [20]	Dhara et al. [50]	Wei et al. [29]	Proposed (Mean ± std)
D1	AUC (%)	89.25	95.76	93.00	**96.65 ± 0.01**
	Accuracy (%)	85.59	88.38	87.14	**89.53 ± 0.09**
	Sensitivity (%)	70.62	**84.58**	77.00	84.19 ± 0.09
	Specificity (%)	**93.02**	90.03	93.00	92.02 ± 0.01
D2	AUC (%)	93.79	94.44	/	**94.45 ± 0.01**
	Accuracy (%)	87.36	87.69	/	**87.74 ± 0.03**
	Sensitivity (%)	73.75	80.00	/	**81.11 ± 0.85**
	Specificity (%)	**93.37**	89.30	/	89.67 ± 0.09
D3	AUC (%)	76.26	79.74	/	**81.24 ± 0.01**
	Accuracy (%)	70.97	71.17	/	**71.93 ± 0.04**
	Sensitivity (%)	53.61	53.47	/	**59.22 ± 0.04**
	Specificity (%)	89.41	**89.74**	/	84.85 ± 0.10

5. Discussion

5.1. Feature fusion versus decision fusion

Information fusion can take place at the feature or decision level. In our algorithm, we extracted three features from a different perspective, used each of the features to train an Adaboosted BPNN, and fused the decision made by each Adaboosted BPNN to classify each nodule. We also attempted to perform fusion at the feature level, i.e. concatenating the DCNN feature, GLCM-based texture descriptor and Fourier shape descriptor into a big feature vector and applying it to an AdaBoosted BPNN, as shown in Fig. 4.

We compared the performance of these two fusion strategies and the results (see Table 5) show that our Fuse-TSD algorithm, where extracted information was fused at the decision level, achieved better performance. The inferiority of feature fusion can be largely ascribed to the troublesome mutual influence of different groups of features and the substantially increased dimensionality of the fused features.

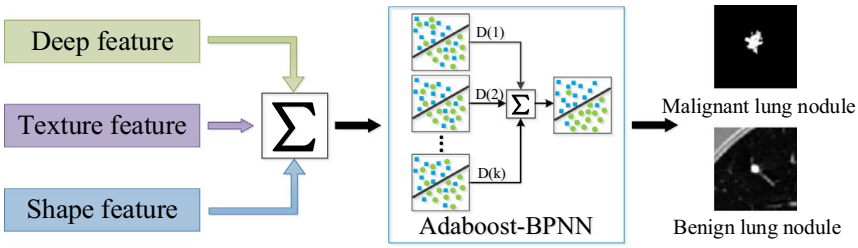

Fig. 4. Diagram of the feature fusion approach.

Table 5
Performance of feature fusion and decision fusion.

	Method	AUC (%) (Mean ± SD)	Accuracy (%) (Mean ± SD)	Sensitivity (%) (Mean ± SD)	Specificity (%) (Mean ± SD)
D1	Feature fusion	96.45 ± 0.02	89.05 ± 0.03	**84.33 ± 0.02**	91.12 ± 0.19
	Fuse-TSD	**96.65** ± 0.01	**89.53** ± 0.09	84.19 ± 0.09	**92.02** ± 0.01
D2	Feature fusion	92.55 ± 0.09	86.92 ± 0.08	77.78 ± 0.15	**89.85 ± 0.25**
	Fuse-TSD	**94.45** ± 0.01	**87.74** ± 0.03	**81.11** ± 0.85	89.67 ± 0.09
D3	Feature fusion	78.77 ± 0.02	71.59 ± 0.13	58.64 ± 0.03	**85.10 ± 0.11**
	Fuse-TSD	**81.24 ± 0.01**	**71.93 ± 0.04**	**59.22 ± 0.04**	84.85 ± 0.10

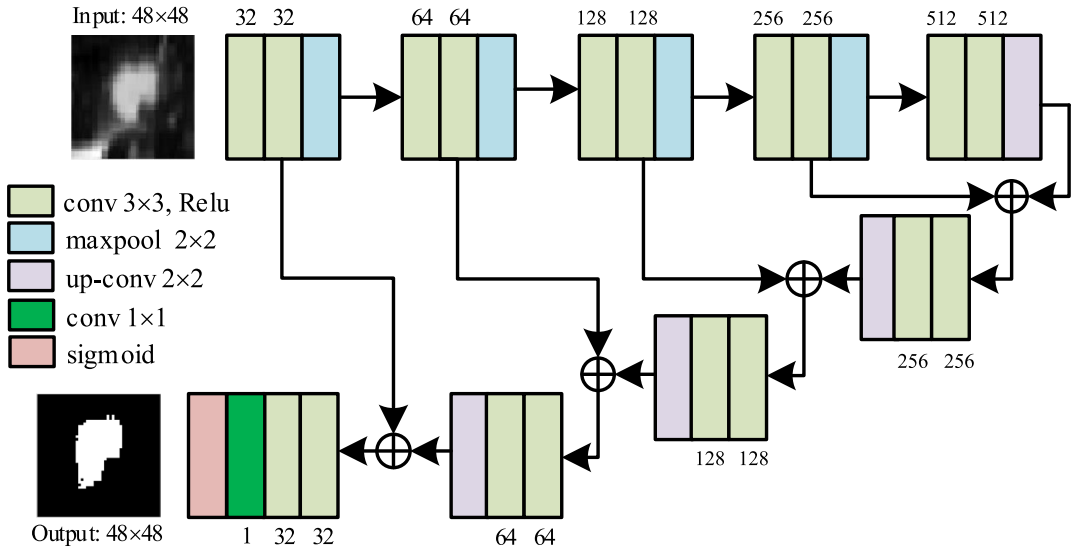

Fig. 5. Architecture of the U-Net for lung nodule segmentation.

Table 6
Performance of our algorithm when using manual or automated nodule segmentation.

Methods	AUC (%) (Mean ± std)	Accuracy (%) (Mean ± std)	Sensitivity (%) (Mean ± std)	Specificity (%) (Mean ± std)
Fuse-TSD (U-Net)	96.46 ± 0.02	89.16 ± 0.01	80.66 ± 0.19	**93.25** ± 0.03
Fuse-TSD	**96.65** ± 0.01	**89.53** ± 0.09	**84.19** ± 0.09	92.02 ± 0.01

5.2. Manual versus automated nodule segmentation

The aim of this research was to investigate if incorporating traditional visual features which were extracted under the guidance of domain knowledge into the image representation and deep features learned by a CNN could improve lung nodule classification. Although lung nodule segmentation is an indispensable step in our algorithm, the segmentation on each image patch can be achieved by using many available methods. As a case study, we adopted the U-Net, a fully convolutional network (FCN) model [51], to segment the lung nodule in each 64 × 64 image patch on a slice-by-slice basis. As shown in Fig. 5, the U-Net consists of a contracting path and an expansive path. The contracting path follows the typical architecture of a CNN, in which there is the repeated application of two 3 × 3 padded convolutional layers, each followed by a ReLU function and a 2 × 2 max pooling operation with stride 2 for downsampling. Every step in the expansive path consists of an upsampling of the feature map followed by a 2 × 2 convolutional layer, a concatenation with the correspondingly feature map from the contracting path and two 3 × 3 convolutions, each followed by a ReLU. At the final layer, a 1 × 1 convolution is used to map each 32 component feature vector to the desired number of classes.

We show the performance of our Fuse-TSD algorithm with and without the U-Net based nodule segmentation on the D1 dataset D1 in Table 6. We found that replacing themanual segmentation of nodules with an automated approach leads only to a slight drop in classification performance, since the U-Net produces relatively accurate nodule segmentation. These findings suggest that our algorithm still works when manual segmentation of nodules is replaced by an automated segmentation approach.

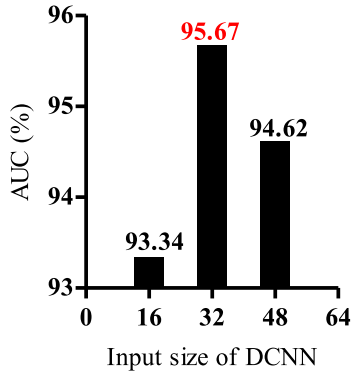

Fig. 6. Performance of our algorithm on the dataset D1 when using different input sizes of DCNN.

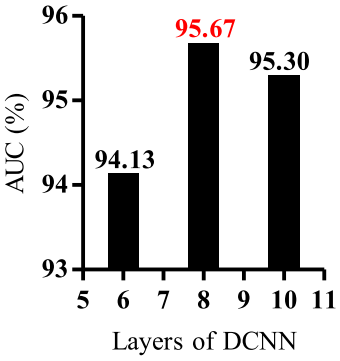

Fig. 7. Performance of our algorithm on the dataset D1 when using the DCNN with different depth.

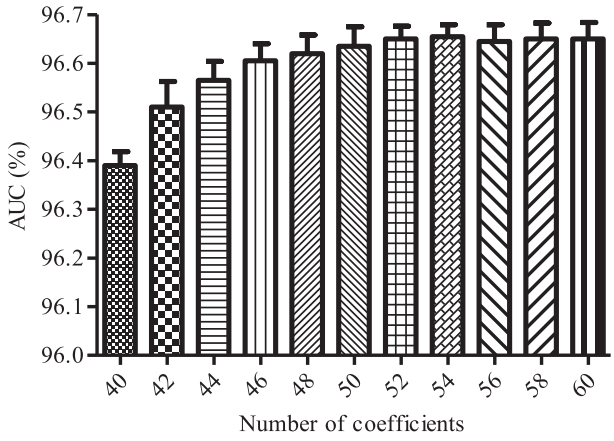

Fig. 8. Variation of AUC on the dataset D1 over the dimensionality of Fourier shape descriptor.

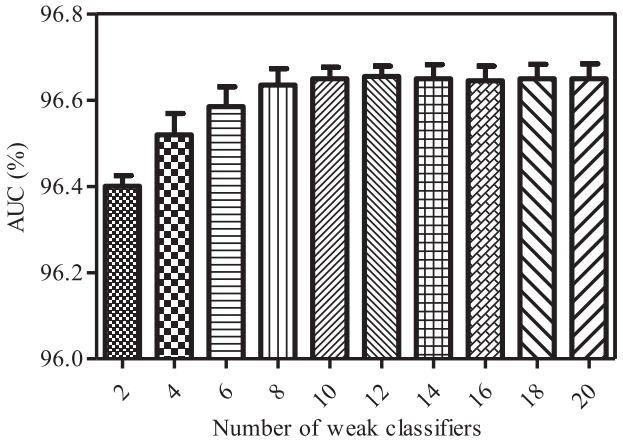

Fig. 9. Variation of our algorithm's AUC on the dataset D1 over the number of weak classifiers.

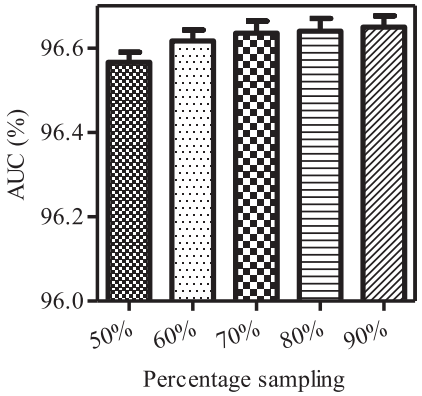

Fig. 10. Variation of our algorithm's AUC on the dataset D1 over the percent of training data used for training.

Table 8
Time cost of each major step of the proposed algorithm.

Steps	Feature Extraction			
	Deep Features	Texture Features	Shape Features	Ensemble Classification
Train (h)	2.04	1.12	0.33	0.30
Test (s)	0.03	0.27	0.05	0.0002

5.3. Field of view and depth of DCNN

Field of view (FoV) is an important concept in DCNN, as it determines how much local information can be captured by convolution masks at a time. To investigate the appropriate size of FoV, we chose dataset D1 as a case study and repeated the nodule classification

Table 7
Performance of our algorithm when using GLCM features or UPLBP as the texture descriptor.

Dataset	Method	Mean ± SD (%)			
		AUC	Accuracy	Sensitivity	Specificity
D1	DCNN + Fourier + UPLBP	95.70 ± 0.30	88.28 ± 0.05	84.15 ± 0.09	90.12 ± 0.11
	DCNN + Fourier + GLCM	**96.65** ± 0.01	**89.53** ± 0.09	**84.19** ± 0.09	**92.02** ± 0.01
D2	DCNN + Fourier + UPLBP	93.05 ± 0.10	86.92 ± 0.09	80.95 ± 0.05	88.83 ± 0.06
	DCNN + Fourier + GLCM	**94.45** ± 0.01	**87.74** ± 0.03	**81.11** ± 0.85	**89.67** ± 0.09
D3	DCNN + Fourier + UPLBP	78.81 ± 0.03	71.86 ± 0.09	58.32 ± 0.04	**86.07** ± 0.08
	DCNN + Fourier + GLCM	**81.24** ± **0.01**	**71.93** ± **0.04**	**59.22** ± **0.04**	84.85 ± 0.10

experiment with the same DCNN while setting the size of field of view to 16 × 16, 32 × 32 and 48 × 48. The variation of the AUC over the size of FoV is seen in Fig. 6 and shows that our algorithm had the highest AUC when the size of FoV is 32 × 32. Therefore, we empirically set the size of FoV to 32 × 32 and resized each ROI into this size for this study.

In addition, the depth of DCNN also impacts classification accuracy. Usually, a deeper DCNN performs better than a shallow one. However, the training dataset for this study was extremely small. Hence the DCNN may over-fit quickly as its architecture goes deeper. We used different DCNNs with the number of layers ranging from 6 to 10, and the AUCs of our algorithm are shown in Fig. 7. The DCNN with eight layers achieved the highest AUC.

5.4. Dimensionality of the Fourier shape descriptor

Dimensionality of the Fourier shape descriptor is determined by how many low frequency coefficients of the Fourier transform are selected. We set this number from 40 to 60 with an interval of 2 and plotted the AUC of the proposed algorithm on the D1 dataset versus the number of low frequency coefficients (see Fig. 8). The mean of AUC improves with the increasing number of coefficients and then plateaus at 52. Therefore, we suggest using 52 low frequency coefficients as the Fourier descriptor of the nodule boundary.

5.5. Characterizing the heterogeneity in voxel values

The high correspondence between malignancy and heterogeneity in voxel values, underlines the importance of texture descriptors. Besides GLCM-based statistics, other texture descriptors such as LBP descriptor that can characterize lung nodule voxel value heterogeneity. Uniform pattern [52], a compact extension to LBP that describes each pixel by thresholding its neighbors and encodes the result as a binary array, reduces the dimensionality of the texture descriptor from 256 to 59 and is invariant to rotations. We repeated our experiments using the GLCM features and uniform pattern LBP (UPLBP) as the texture descriptor. We show the performance in Table 7. The GLCM features resulted in higher AUC, classification accuracy and sensitivity on all three datasets and higher classification specificity on D1 and D2 when compared to the UPLBP descriptor. Therefore, we suggest using GLCM-based texture descriptors to represent the heterogeneity of lung nodules.

5.6. Number of weak learners in the AdaBoosted BPNN

In the AdaBoosted BPNN classifier, each BPNN plays the role of an independent and weak decision maker. The number of weak learners is one of the most critical parameters in an ensemble learning model. We chose D1 as a case study and exhaustively searched all possible values of this parameter, ranging from 2 to 20 with an interval 2. We repeated the experiment 10 times and recorded the mean and standard deviation of the AUC in Fig. 9. As the number of BPNNs increases, the mean of AUC first improves and then plateaus when the number exceeds 10. Thus, we suggest using 10 BPNNs in the AdaBoosted BPNN classifier.

5.7. Percentage of training data for BPNN

We evaluated the performance of our algorithm, when it was trained using 50%, 60%, 70%, 80% and 90% of the nodules randomly sampled from D1. The experiment was conducted 10 times and the mean and standard deviation of the obtained AUCs were plotted in Fig. 10. Our results show that the proposed algorithm is relatively robust to the variable size of the training dataset, and its average AUC only drops slightly from 96.65% to 96.57% when the ratio of training images reduces from 90% to 50%. We empirically used 90% of training data for training and the other 10% for validation.

5.8. Computational complexity

Our algorithm performs DCNN-based deep feature / texture / shape feature extraction, ensemble learning, training the LeNet-5 model and calculates the GLCM-based texture features and Fourier shape descriptor on about 15,000 patches, all of which are time consuming. In addition, training the ensemble classifier means training three AdaBoosted BPNNs, each containing 10 BPNNs. Therefore, our algorithm has high computational complexity, particularly during the off-line training. In our experiments, it takes about 8 h to train the proposed model (Intel Xeon E5-2678 V3 2.50 GHz X2, NVIDIA Tesla K40c GPU, 128GB Memory, 120GB SSD and Matlab 2014). The application of the trained model to nodule classification, however, is relatively fast, with less than 0.4 s to classify each nodule. Hence we suggest that our algorithm could be used in a routine clinical workflow. The duration of each main step in our proposed algorithm is shown in Table 8.

6. Conclusions

We propose a novel Fuse-TSD lung nodule classification algorithm that uses texture, shape and deep model-learned information at the decision level for distinguishing malignant from benign lung nodules. Our results suggest that combining the image representation learned by deep models with traditional visual features at the decision level improves the performance of nodule classification and produce more accurate results than three current state-of-the-art approaches. Our future work will focus on discovering more effective deep learning methods for characterizing lung nodules.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 61471297, 61771397 and 61572405, in part by the Key Projects of National Natural Science Foundation of China under Grants 61231016, in part by the China 863 program under Grants 2015AA016402, in part by the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University under Grants Z2017041, and in part by the Australian Research Council (ARC) Grants. We acknowledged the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this work.

References

[1] A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman, Global cancer statistics, 2012, Ca Cancer J. Clin. 61 (2) (2011) 69–90.
[2] P.B. Bach, J.N. Mirkin, T.K. Oliver, C.G. Azzoli, D.A. Berry, O.W. Brawley, et al., Benefits and harms of CT screening for lung cancer: a systematic review, Jama J. Am. Med. Assoc. 307 (22) (2012) 2418–2429.
[3] J. Abraham, Reduced lung-cancer mortality with low-dose computed tomographic screening, N. Engl. J. Med. 365 (5) (2011) 395–409.
[4] W. Huang, S. Zeng, M. Wan, G. Chen, Medical media analytics via ranking and big learning: a multi-modality image-based disease severity prediction study, Neurocomputing 204 (2016) 125–134.
[5] S. Metz, C. Ganter, S. Lorenzen, S.V. Marwick, K. Holzapfel, K. Herrmann, et al., Multiparametric MR and PET imaging of intratumoral biological heterogeneity in patients with metastatic lung cancer using voxel-by-voxel analysis, PLoS ONE 10 (7) (2014).
[6] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[7] V.N. Vapnik, A. Lerner, Pattern recognition using generalized portrait method, Autom. Remote Control 24 (6) (1963) 774–780.
[8] L. Rokach, Decision forest: twenty years of research, Inf. Fusion 27 (2016) 111–125.
[9] A. Sankaran, A. Jain, T. Vashisth, M. Vatsa, R. Singh, Adaptive latent fingerprint segmentation using feature selection and random decision forest classification, Inf. Fusion 34 (2016) 1–15.
[10] J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy K-nearest neighbor algorithm, IEEE Trans. Syst. Man Cybern. 15 (4) (1985) 580–585.
[11] R. Hecht-Nielsen, Theory of the backpropagation neural network, Neural Netw. 1 (1) (1988) 65–93.
[12] H. Chen, J. Zhang, Y. Xu, B. Chen, K. Zhang, Performance comparison of artificial

neural network and logistic regression model for differentiating lung nodules on CT scans, Expert Syst. Appl. 39 (13) (2012) 11503–11509.
[13] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[14] M.P. Sesmero, J.M. Alonso-Weber, G. Gutierrez, A. Ledezma, A. Sanchis, An ensemble approach of dual base learners for multi-class classification problems, Inf. Fusion 24 (2015) 122–136.
[15] Y. Freund, R.E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (7) (1999) 119–139.
[16] L.K. Soh, C. Tsatsoulis, Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices, IEEE Trans. Geosci. Remote Sens. 37 (2) (1999) 780–795.
[17] H. Wang, X.H. Guo, Z.W. Jia, H.K. Li, Z.G. Liang, K.C. Li, et al., Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image, Eur. J. Radiol. 74 (1) (2009) 124–129.
[18] H. Wu, T. Sun, J. Wang, X. Li, W. Wang, D. Huo, et al., Combination of radiological and gray level co-occurrence matrix textural features used to distinguish solitary pulmonary nodules by computed tomography, J. Digit. Imaging 26 (4) (2013) 797–802.
[19] J. Zhao, G. Ji, Y. Qiang, X. Han, B. Pei, Z. Shi, A new method of detecting pulmonary nodules with PET/CT based on an improved watershed algorithm, PLoS ONE 10 (4) (2015) e0123694.
[20] F. Han, H. Wang, G. Zhang, H. Han, B. Song, L. Li, et al., Texture feature analysis for computer-aided diagnosis on pulmonary nodules, J. Digit. Imaging 28 (1) (2015) 99–115.
[21] D. Frejlichowski, An experimental comparison of seven shape descriptors in the general shape analysis problem, International Conference on Image Analysis & Recognition, 2010, pp. 294–305.
[22] T. Yoshimasu, M. Kawago, Y. Hirai, T. Ohashi, Y. Tanaka, S. Oura, et al., Fast Fourier transform analysis of pulmonary nodules on computed tomography images from patients with lung cancer, Ann. Thorac. Cardiovasc. Surg. Off. J. Assoc. Thorac. Cardiovasc. Surg. Asia 21 (1) (2015) 1–7.
[23] F. Ciompi, C. Jacobs, E.T. Scholten, M.M. Wille, P.A. de Jong, M. Prokop, et al., Bag-of-frequencies: a descriptor of pulmonary nodules in computed tomography images, IEEE Trans. Med. Imaging 34 (4) (2015) 962–973.
[24] Z. Huo, M.L. Giger, C.J. Vyborny, U. Bick, P. Lu, D.E. Wolverton, et al., Analysis of spiculation in the computerized classification of mammographic masses, Med. Phys. 22 (10) (1995) 1569–1579.
[25] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[26] H.C. Shin, H.R. Roth, M. Gao, L. Lu, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298.
[27] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2) (2012).
[28] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, S. Mougiakakou, Lung pattern classification for interstitial lung diseases using a deep convolutional neural network, IEEE Trans. Med. Imaging 35 (2016) 1207–1216.
[29] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang, J. Tian, Multi-scale convolutional neural networks for lung nodule classification, Inf. Process. Med. Imaging (2015) 588–599.
[30] K.L. Hua, C.H. Hsu, S.C. Hidayati, W.H. Cheng, Y.J. Chen, Computer-aided classification of lung nodules on computed tomography images via deep learning technique, Oncotargets Ther. 8 (2015) 2015–2022.
[31] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art, Inf. Fusion 14 (2013) 28–44.
[32] M.M. Kokar, J.A. Tomasik, J. Weyman, Formalizing classes of information fusion systems, Inf. Fusion 5 (3) (2004) 189–202.
[33] L.A. Alexandre, Gender recognition: a multiscale decision fusion approach, Pattern Recognit. Lett. 31 (11) (2010) 1422–1425.
[34] M. Fauvel, J. Chanussot, J.A. Benediktsson, Decision fusion for the classification of urban remote sensing images, IEEE Trans. Geosci. Remote Sens. 44 (10) (2006) 2828–2838.
[35] A. Ross, A. Jain, Information fusion in biometrics, International Conference on Audio- and Video-Based Biometric Person Authentication, 2001, pp. 2115–2125.
[36] Y.J. Chin, T.S. Ong, A.B.J. Teoh, K.O.M. Goh, Integrated biometrics template protection technique based on fingerprint and palmprint feature-level fusion, Inf. Fusion 18 (1) (2014) 161–174.
[37] Y. Xie, J. Zhang, S. Liu, W. Cai, Y. Xia, Lung nodule classification by jointly using visual descriptors and deep features, Lect. Notes Comput. Sci. 10081 (2017) 116–125.
[38] A.M. Aziz, A new multiple decisions fusion rule for targets detection in multiple sensors distributed detection systems with data fusion, Inf. Fusion 18 (18) (2014) 175–186.
[39] S.G. Armato III, G. Mclennan, L. Bidaut, M.F. Mcnittgray, C.R. Meyer, A.P. Reeves, 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. 38 (2) (2011) 915–931.
[40] S.G. Armato III, G. McLennan, L. Bidaut, M.F. McNitt-Gray, C.R. Meyer, A.P. Reeves, L.P. Clarke, Data from LIDC-IDRI. Cancer Imaging Arch.
[41] B.V.K. Clark, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, et al., The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository, J. Digit. Imaging 26 (2013) 1045–1057.
[42] A.A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, R.S. Van, et al., Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks, IEEE Trans. Med. Imaging 35 (5) (2016) 1160–1169.
[43] D.M. Xu, H. Gietema, H.D. Koning, R. Vernhout, K. Nackaerts, M. Prokop, et al., Nodule management protocol of the NELSON randomised lung cancer screening trial, Lung Cancer 54 (2) (2006) 177–184.
[44] G.M. Dasovich, R. Kim, D.S. Raicu, J.D. Furst, A Model for the Relationship Between Semantic and Content Based Similarity Using LIDC, SPIE Medical Imaging, 2010, pp. 185–192.
[45] Y. Lécun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[46] A. Vedaldi, K. Lenc, MatConvNet - convolutional neural networks for MATLAB, Eprint Arxiv, (2016) 689–692.
[47] R.M. Haralick, K. Shanmugam, I.H. Dinstein, R.M. Haralick, K. Shanmuga, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 3 (6) (1973) 610–621.
[48] K. Manivannan, P. Aggarwal, V. Devabhaktuni, A. Kumar, D. Nims, P. Bhattacharya, Particulate matter characterization by gray level co-occurrence matrix based support vector machines, J. Hazard. Mater. 223-224 (2) (2012) 94–103.
[49] G. Mirchandani, W. Cao, On hidden nodes for neural nets, IEEE Trans. Circuits Syst. 36 (5) (1989) 661–664 5.
[50] 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 (2016) 1–10.
[51] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, 9351 (2015) 234–241.
[52] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.