Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology) *2022 23(11):957-967*
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Correspondence

https://doi.org/10.1631/jzus.B2101009

A method for distinguishing benign and malignant pulmonary nodules based on 3D dual path network aided by *K*-means clustering analysis

Dachuan GAO[1*], Xiaodan YE[2*], Xuewen HOU[1], Yang CHEN[1], Xue KONG[3], Yuanzhong XIE[3✉], Shengdong NIE[1✉]

[1]*School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*
[2]*Department of Radiology, Shanghai Chest Hospital, Shanghai 200030, China*
[3]*Medical Imaging Center, Tai'an Center Hospital, Tai'an 271000, China*

In the USA, there were about 1 806 590 new cancer cases in 2020, and 606 520 cancer deaths are expected to have occurred in 2021. Lung cancer has become the leading cause of death from cancer in both men and women (Siegel et al., 2020). Clinical studies show that the five-year survival rate of lung cancer patients after early diagnosis and treatment intervention can reach 80%, compared with that of patients having advanced lung cancer. Thus, the early diagnosis of lung cancer is a key factor to reduce mortality.

In general, the early stage of lung cancer manifests as pulmonary nodules (Gould et al., 2013; Cao et al., 2017). These blob-like structures with a diameter ranging from 3 to 30 mm can be categorized as juxtavascular, well-circumscribed, pleural tail, or juxtapleural (Zhang et al., 2013). However, not all pulmonary nodules are necessarily malignant. In clinical treatment, the radiologist makes a diagnosis according to the characteristics of the nodules in a computed tomography (CT) image, such as their shape, intensity, and texture. Additionally, as the number of patients increases, the radiologist has to analyze many CT images with the naked eye. This task is time-consuming, laborious, and intensive. Furthermore, the diagnostic result is significantly correlated with the clinical experience of the radiologists; the ability of different radiologists to diagnose nodules relates to their qualifications. Thus, a computer-aided diagnosis (CAD) system could be used as a tool for a "second opinion" to assist radiologists in clinical diagnosis.

Existing benign-malignant pulmonary nodule CAD classification methods can be sorted into two main categories. Those in the first category use traditional handcrafted features and usually include the following steps: nodule segmentation, feature extraction and optimization, and classification identification. Those in the second category use deep learning, which uses various deep neural networks to realize an end-to-end automatic classification of benign and malignant pulmonary nodules. Over the years, many studies have attempted to improve the performance of classification models. Dhara et al. (2016) proposed a malignant-benign classification model that combines shape and texture features based on CT images. Semi-automated technology was used to segment the pulmonary nodules, and multiple handcrafted features were calculated to represent them. The Lung Image Database Consortium (LIDC) dataset was divided into three different sample configurations, and the features were input into a support vector machine (SVM). The areas under the curve (AUCs) of the three configurations were 0.9505, 0.8822, and 0.8488, respectively. Gong et al. (2018) extracted 66 three-dimensional (3D) handcrafted features of 243 pulmonary nodules and trained multiple machine learning

✉ Yuanzhong XIE, xie01088@126.com
Shengdong NIE, nsd4647@163.com
* The two authors contributed equally to this work
Dachuan GAO, https://orcid.org/0000-0002-6399-1708
Yuanzhong XIE, https://orcid.org/0000-0003-2593-4806
Shengdong NIE, https://orcid.org/0000-0001-7825-4455

Received Dec. 13, 2021; Revision accepted May 3, 2022;
Crosschecked Nov. 11, 2022

classifiers under three sample configurations. The average AUCs of the SVM, naive Bayes, and linear discriminant analysis (LDA) classifiers were 0.94, 0.90, and 0.99, respectively. Shen W et al. (2017) proposed a multi-scale convolutional neural network (CNN) to classify benign and malignant pulmonary nodules. Three scale blocks were intercepted from nodules with a dimensionality of 32×32×32, 64×64×64, and 96×96×96, respectively. CNN models were then trained to produce multi-scale features, and random forest (RF) and SVM were used to complete the classification task. A promising accuracy (83.21%) and AUC score (0.89) were achieved on the LIDC-Image Database Resource Initiative (IDRI) database after applying the SVM. Zhu et al. (2018) designed two deep 3D dual path networks (DPNs) for the detection and classification of pulmonary nodules. CNN features were extracted based on the 3D DPN, and a gradient boosting machine (GBM) was used to classify benign and malignant pulmonary nodules. The average AUC reached 0.9044 when using the LIDC-IDRI database. Causey et al. (2018) proposed a classification and identification system for benign and malignant pulmonary nodules (NoduleX), which combined 200 deep CNN features and 103 handcrafted features, and identified them using an RF classifier. The AUC and accuracy of NoduleX were 0.9710 and 0.9320, respectively.

In recent years, many malignant-benign classification methods have been proposed, and classification accuracy has improved. Nevertheless, the following problems still remain: (1) the traditional handcrafted feature methods are cumbersome, and features with strong generalization ability rely on the joint extraction of experienced radiologists and engineering technicians; (2) most of the classification methods that are based on deep learning focus on the optimization and modification of the existing network model structure, while few studies have refined the classification scheme to improve accuracy. Previous classification methods (either traditional or deep learning) simply divided pulmonary nodules into benign or malignant. However, they did not consider that there may be several types of benign and malignant nodules; therefore, the classification accuracy was not satisfactory. Thus, the classification scheme plays a significant role in the overall performance of a model.

Benign nodules are usually located at the edge of the chest cavity and have a smooth surface. Malignant nodules are typically accompanied by lobules, burrs, vascular fusion, and cystic air spaces (Snoeckx et al., 2018). Moreover, irregular boundaries and a high growth rate are the main characteristics of malignant nodules (Albert and Russell, 2009). Fig. S1a presents two-dimensional (2D)-slice patches of different types of pulmonary nodules with the same malignancy score from the LIDC dataset. The top row in Fig. S1a shows four nodules (labelled (1) to (4)) that are malignant but have different morphological manifestations. We speculate that there are many potential types of malignant nodules. Furthermore, benign and uncertain nodules may also have several potential types. Owing to the complex internal structure of pulmonary nodules, the difference between samples can be very large. Their complex distribution poses a great challenge in the malignant-benign classification of pulmonary nodules, which severely affects classification performance. To resolve this issue, a novel classification model is proposed herein, which identifies benign and malignant pulmonary nodules based on 3D DPN with *K*-means clustering. The findings of this study indicate that this approach represents a great improvement in classification performance.

First, the pulmonary nodule block was extracted from CT images according to the annotation drawn by four radiologists, and the pre-3D DPN model was trained to extract CNN features. Then, feature selection was implemented using the RF model. Next, the *K*-means clustering algorithm was used to generate new subclass labels for benign and malignant classes, which were named cluster labels. Finally, the 3D DPN multi-classification model was retrained using data with cluster labels, and the final prediction results were obtained according to the mapping relationship between subclasses and real classes. For test data, the preliminary prediction class was determined based on the 3D DPN; afterwards, the class weight value was calculated according to the distance relationship between the test sample and the multiple cluster centers. The preliminary class was multiplied by the weight value, and the final prediction class was obtained through the mapping table.

We collected lung CT images of the study subjects from two sources. First, we downloaded the LIDC-IDRI dataset (Armato et al., 2011) to evaluate the proposed method. This dataset has 1010 cases, each including images from a clinical thoracic CT scan and an associated Extensible Markup Language

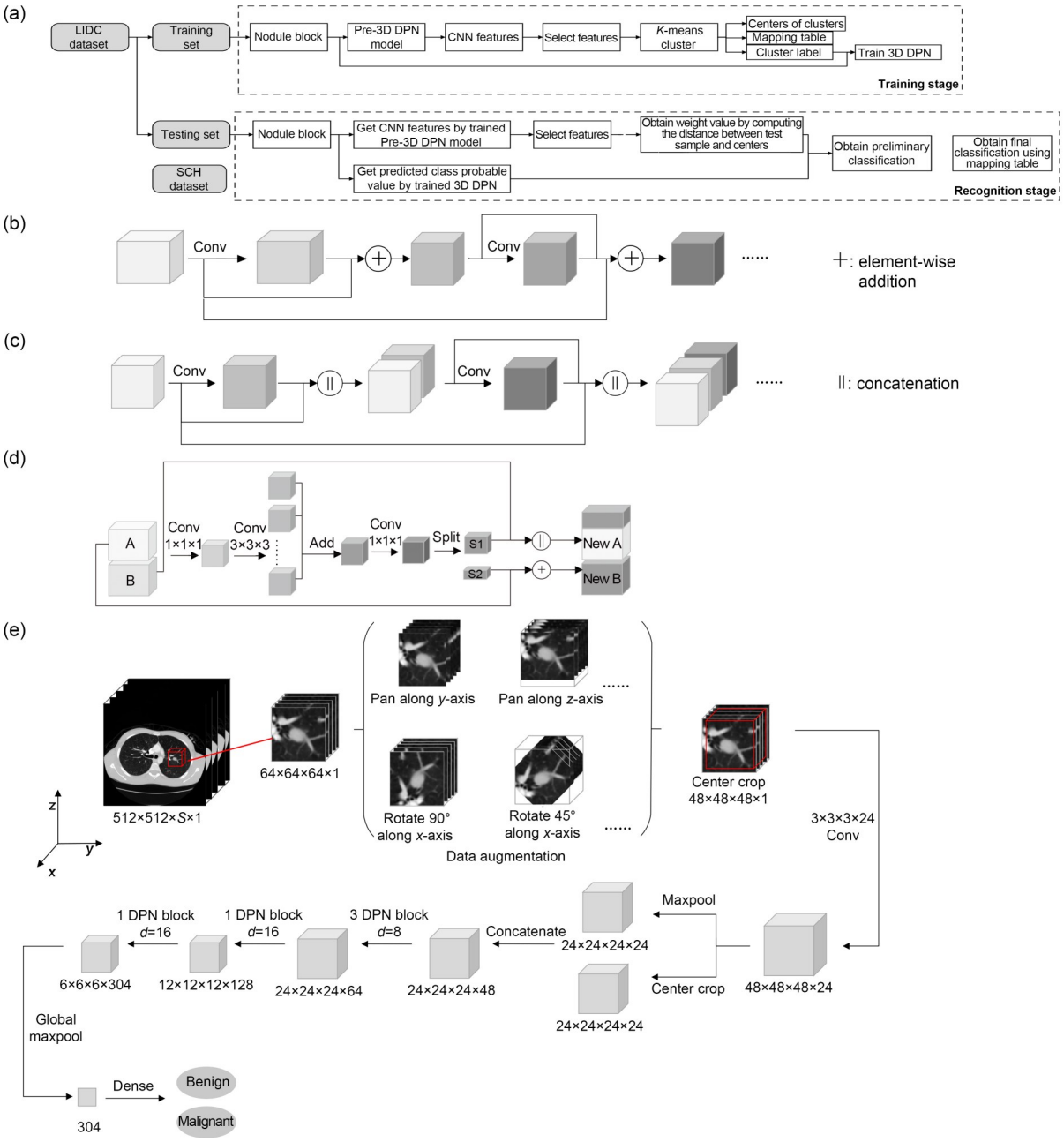
(XML) file that records the annotations from four radiologists. We discarded CT scans with a slice thickness greater than 3 mm, inconsistent slice spacing or a missing slice, and we considered the following types of nodules: (1) those equal to or larger than 3 mm in size; (2) those marked by most of the radiologists (at least three out of four radiologists). The annotations from more than one radiologist for the same nodule were merged if the spatial distance was less than a certain threshold. We averaged the positions, malignancy, and other attributes for each candidate. After merging the annotations, each patch was centered on the nodule center according to the annotation. To capture most of the nodule morphology and surroundings, we set the nodule patch size to 64×64×64. In total, 966 nodules were considered, namely, 427 benign nodules (average score≤2.5), 253 malignant nodules (average score≥3.5), and 286 ambiguous nodules (2.5<average score<3.5), to develop a new classification scheme. The malignancy label of nodules used in our work was the result of the joint labelling of four radiologists, not biopsy results. We also performed a consistency analysis using the annotation results. Fig. S1b shows the results of the consistency test of four experts in pairs. We excluded labels that were marked by experts as ambiguous (malignancy score=3), so there were some missing cases in the analysis results. Usually, Cohen's kappa score is used to evaluate the consistency of diagnosis of multiple experts (McHugh, 2012). When the kappa score is in the range of 0.4 to 0.6, it indicates that the judgments of two experts show moderately strong agreement, and when in the range of 0.6 to 0.8, the judgments have strong agreement. Fig. S1b shows that, except for radiologist 4, other radiologists showed strong consistency in their identification of benign and malignant nodules. We integrated the annotation results of the four experts, minimized the judgment bias caused by the individual differences of the radiologists, and replaced the biopsy results of pulmonary nodules with the integrated annotation as much as possible.

Second, we collected CT images and histopathological test data of 36 patients who underwent lung cancer diagnosis at Shanghai Chest Hospital (SCH) in Shanghai, China. Among the 67 nodules, 32 were benign and 35 were malignant. They were used only as test data to evaluate the performance of the proposed model. An institutional review board approved this study, and patient agreement was obtained under the condition that all data were anonymized. Fig. S1c provides a detailed description of the experimental data.

Here, we propose a novel classification method that identifies benign and malignant pulmonary nodules based on 3D DPN combined clustering analysis. First, a volume of interest (VOI) with a dimensionality of 64×64×64 that was centered on a pulmonary nodule was extracted from CT images according to the annotation drawn by radiologists. The pre-3D DPN model was then trained using nodule samples to extract CNN features. Next, feature selection was implemented based on the RF model, and a *K*-means clustering algorithm was used to generate malignant-benign cluster labels. Finally, the 3D DPN was retrained to identify malignant-benign pulmonary nodules using the data with new labels. Fig. 1a presents a flowchart of the proposed classification method, which consists of two main stages: training and recognition. Each step is described in the following sections.

With the continuously increasing amount of training data and rapid development of computer hardware performance, deep learning is considered one of the best solutions for medical image processing and medical diagnostics (Shen DG et al., 2017). It can effectively realize the end-to-end automatic classification and identification of pulmonary nodules. Generally, the depth of the network has a great influence on classification performance, i.e., a network with deep layers has more abstract and advanced learned features. However, a deep network causes gradient vanishing and gradient exploding. Thus, the classification model was inadequately trained. The shortcut connection that is used in residual network (ResNet) is conducive to gradient backpropagation, which can realize the training and learning of the deeper network. Furthermore, the output of each residual module is based on the offset of the previous residual module, which can effectively retain the previous features and reduce feature redundancy. The dense connection structure that is used in a dense convolutional network (DenseNet) enhances feature propagation through cascading the feature maps of all previous layers as the input of the next layer. The structures of ResNet and DenseNet are shown in Figs. 1b and 1c.

DPN shows superior performance in ImagNet-1k, Places365 classification tasks, and Pattern Analysis, Statical Modeling and Computational Learning-Visual



(f)

DPN module in the DPN network

Layer	Sub-layer	Input	Kernel	Quantity of kernel	Output
DPN block 1	CNN	24×24×24×48	$\begin{Bmatrix}1\times1\times1\\3\times3\times3,\ d=8\\1\times1\times1\end{Bmatrix}\times3$	64	24×24×24×64
DPN block 2	CNN	24×24×24×64	$\begin{Bmatrix}1\times1\times1\\3\times3\times3,\ d=16\\1\times1\times1\end{Bmatrix}\times1$	128	24×24×24×128
	Maxpool	24×24×24×128	2×2×2		12×12×12×128
DPN block 3	CNN	12×12×12×128	$\begin{Bmatrix}1\times1\times1\\3\times3\times3,\ d=16\\1\times1\times1\end{Bmatrix}\times1$	304	12×12×12×304
	Maxpool	12×12×12×304	2×2×2		6×6×6×304

Fig. 1 Schematic diagram and network architecture of the proposed method. (a) Schematic diagram of the proposed method; (b) Structure of a residual block; (c) Structure of a dense block; (d) Structure of the DPN block (proposed method); (e) 3D DPN architecture of our proposed method; (f) A detailed description of the DPN module in the DPN network. LIDC: Lung Image Database Consortium; DPN: dual path network; CNN: convolutional neural network; SCH: Shanghai Chest Hospital; Conv: convolution; *S*: number of CT images; *d*: number of group.

Object Classes (PASCAL-VOC) detection tasks (Chen et al., 2017). It integrates the advantages of an aggregated residual network (ResNeXt) for feature multiplexing and DenseNet for feature cascading, which can effectively resolve the aforementioned problems. The DPN structure in our work is shown in Fig. 1d. The module includes a cascading branch, a superposition branch, and a micro-convolution structure. First, the input data are divided into A and B using the split operation. Then, the integral data are inputted into a micro-convolution structure consisting of 1×1×1, 3×3×3, and 1×1×1 consecutive convolutional layers. Next, the output is divided into two parts, S1 and S2. S2 is automatically added to the residual path to form New B, and S1 is cascaded to the densely connected path to form New A. Finally, New A and New B are combined to yield the input of the next DPN block.

The 3D DPN classification model for identifying benign and malignant pulmonary nodules (Fig. 1e) includes five DPN modules. The network inputs are 64×64×64 image patches that are centered on nodules. After data augmentation and clipping, samples are fed into the 3D DPN model. Finally, the classification result is obtained from the output layer. Fig. 1f shows a diagram of the DPN block structure used in our work. Note that the 3×3×3 convolution layer adopts the idea of group convolution in ResNeXt, the number of groups is d, and the convolution results of each group are added together.

Based on the annotation of nodules marked by radiologists in both the LIDC-IDRI and SCH databases, 64×64×64 blocks were intercepted from the 512×512×S×1 CT images, where S is the number of CT images (usually S>64). Then, the training nodule samples were augmented in three dimensions (x-axis, y-axis, and z-axis), mainly including rotation and flipping. Disproportionate data augmentation was adopted to balance the number of samples in each cluster. Next, the nodule blocks were cropped to 48×48×48 to eliminate the difference caused by the change in the boundary pixel value. Finally, our pre-3D DPN architecture was trained using pre-processed nodule blocks with a malignancy label.

Generally, the output of a binary CNN classification model is a 2D vector (y_0, y_1). The function Softmax(y_i) represents the probability distribution of two classes, defined as follows:

$$\mathrm{Softmax}\left(y_i\right)=\frac{\exp\left(y_i\right)}{\exp\left(y_0\right)+\exp\left(y_1\right)},\ i=0,1. \quad (1)$$

The cross-entropy loss function is used to measure the degree of difference between the actual one-hot classification label and the predicted class probability of the model. To train the network through minimizing the loss function, we used the binary cross-entropy loss function Loss, defined as follows:

$$\mathrm{Loss}=-\frac{1}{N}\sum\left(q\lg y_1+\left(1-q\right)\lg y_0\right), \quad (2)$$

where q indicates the benign or malignant label of each sample (q=0, 1), and N is the batch size, which was set to 16 in our work.

We used the Adam optimization algorithm (Kingma and Ba, 2015), which is based on a first-order gradient, to optimize the objective Eq. (2). The initial learning rate was set to 0.003. When the verification loss did not decay within 5 epochs, it was multiplied by 1/3, and a total of 120 epochs were implemented. In addition, we used early stopping and regularization strategies. We extracted 304 features from the dense layer of the 3D DPN model. Fig. S2 illustrates the feature set.

Accurate and effective features can improve the classification accuracy of pulmonary nodules. However, when the feature redundancy reaches a certain level, the computational complexity of the entire model and the accuracy of malignant-benign classification decrease. Therefore, we used the RF model to exclude CNN features that were not significantly related to malignancy (Díaz-Uriarte and de Andrés, 2006). We selected features with a weight value higher than 0.01 to form a feature subset. Finally, those features that were moderately or strongly correlated with the malignant-benign classification were retained for the subsequent clustering analysis.

As an unsupervised machine learning technique, a clustering algorithm is used in data analysis to effectively improve classification performance (Lu and Weng, 2007). A clustering algorithm divides all data samples into several clusters according to the similarity of features and attributes. It satisfies the condition that the similarity of samples within the clusters should be higher, while the similarity of samples between clusters should be lower. The K-means clustering algorithm uses mainly feature distance to measure the similarity

of different samples. The samples are assigned to one class according to the distance. This has the advantage of high efficiency and simplicity.

After feature selection, we performed *K*-means clustering analysis of the training data to form new cluster labels. Eq. (3) presents the mapping relationship of the sample label, where $\boldsymbol{Y}_{\mathrm{tr}}$ represents the corresponding label of the training set, $\boldsymbol{Y}_{\mathrm{tr}}^{*}$ represents the cluster label of the training set, 0 is the benign nodule, 1 is the malignant nodule, and c and d indicate the number of clusters of benign and malignant pulmonary nodules, respectively. Specifically, y_{0i} and y_{1i} represent the samples with the original label of 0 and 1, respectively, and i is the current label ($0<i\leqslant c$) obtained via clustering. Since the potential subtypes are unknown, we will discuss the number of optimal clusters (c and d) in the next section.

$$\boldsymbol{Y}_{\mathrm{tr}}=\begin{bmatrix}0\\0\\\cdots\\0\\1\\1\\\cdots\\1\end{bmatrix}\Rightarrow\boldsymbol{Y}_{\mathrm{tr}}^{*}=\begin{bmatrix}y_{01}\\y_{02}\\\cdots\\y_{0c}\\y_{11}\\y_{12}\\\cdots\\y_{1d}\end{bmatrix}. \tag{3}$$

The 3D DPN multi-classification model follows the setting of the pre-3D DPN, which was retrained using data with cluster labels. For multi-classification, the output of the CNN is a multi-dimensional vector (y_0, y_1, ⋯, y_c, ⋯, y_{c+d}) for each sample with label q, where q=0, 1, ⋯, c, ⋯, $c+d$, and c and d are the numbers of clusters within a malignant or benign class, respectively. The function Softmax(y_i) is defined as:

$$\mathrm{Softmax}\left(y_i\right)=\frac{\exp\left(y_i\right)}{\sum_{j=1}^{c+d}\exp\left(y_j\right)},\ i=1,2,...,c+d. \tag{4}$$

The categorical cross-entropy Loss function is selected for model optimization, and is defined by the cross-entropy of batch size N:

$$\mathrm{Loss}=-\frac{1}{N}\sum_{i=1}^{c+d}q_i\lg y_i,\ i=1,2,...,c+d. \tag{5}$$

In the recognition phase, given a test sample, the trained 3D DPN classification model outputs the initial predicted class distribution vector. To further improve classification accuracy, we weighted the initial vector to obtain the final predicted result.

The weight value (w_i) calculation is shown in Eq. (6), where w_i is determined by calculating the reciprocal of the distance between the test sample x_{te} and each cluster center (Center$_i$). x_{te} represents the feature vector of an input test sample, and dist represents the distance between the test sample and multiple cluster centers. The smaller the distance from a certain cluster center, the higher the probability that the test sample belongs to one sub-class. Additionally, $c+d$ represents the number of all sub-classes that are decomposed in the clustering process.

$$w_i=\frac{1/\mathrm{dist}\left(x_{\mathrm{te}},\mathrm{Center}_i\right)}{\sum_{j=1}^{c+d}1/\mathrm{dist}\left(x_{\mathrm{te}},\mathrm{Center}_i\right)},\ i=1,2,...,c+d. \tag{6}$$

The weighted predicted class is calculated according to Eq. (7) through multiplying each cluster center weight value w_i by the initial multi-class probability distribution Softmax(y_i) and selecting the class weights based on the maximum weighted subclasses distribution value. We named this post-processing process "weight adjustment." The final predicted class is determined based on the established cluster mapping table. Generally, the accuracy of the traditional binary classification model is obtained by summing the diagonal elements in the confusion matrix and dividing them by the total number of samples. The proposed method sums the diagonal elements as well as the elements in the same class. Thus, our proposed classification method can greatly improve the fault tolerance rate.

$$\mathrm{Class}_{\mathrm{weighted}}=\underset{i\in\{1,2,...,c+d\}}{\arg\max}\left(w_i\times\mathrm{Softmax}\left(y_i\right)\right). \tag{7}$$

Our model was implemented in Python using a custom version of the TensorFlow framework, which was enabled to perform volumetric convolution. All experiments were performed on a standard workstation equipped with an Intel Xeon E5-2673 central processing unit (CPU) working at 2.4 GHz and a NVIDIA Ge Force GT extreme (GTX) 1080 graphics card.

To validate the effectiveness of the proposed scheme, we used 966 pulmonary nodules from the LIDC database and 67 from the SCH database in the experiment. The training dataset was divided randomly, accounting for 85% of the LIDC database; the remaining 15% and the SCH database were considered as the test dataset. As shown in Fig. 2a, we carried out the following two sets of experiments under three different sample configurations: (1) exploring the influence of cluster number on the accuracy of benign-malignant classification; and (2) exploring the influence of different architectures on the accuracy of benign-malignant classification. Three sample configuration schemes regard the nodules marked by radiologists with a malignancy score of 3 in the LIDC database as discarded, benign, or malignant. In configurations 2 and 3, unknown data are placed in the benign and malignant groups. These uncertain nodules are used to validate the performance of the classification scheme in dealing with fault data. Generally, the evaluation criteria of a classification model include accuracy, sensitivity, specificity, and AUC. A detailed explanation of each evaluation indicator is given below and in Eqs. (8)–(10).

Accuracy (ACC) is the proportion of all samples correctly classified by the model. Sensitivity (SEN) is the proportion of malignant nodules correctly classified by the model. The higher the sensitivity, the lower the missed diagnosis rate. Specificity (SPE) is the proportion of benign nodules correctly classified by the model. The higher the specificity, the lower the misdiagnosis rate. True positive (TP) and true negative (TN) represent the numbers of samples of malignant nodules and benign nodules correctly classified by the model, respectively. On the contrary, false positive (FP) and false negative (FN) represent the numbers of samples of malignant and benign nodules, respectively, that are predicted incorrectly by the model.

$$\mathrm{ACC}=(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}), \quad (8)$$

$$\mathrm{SEN}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}), \quad (9)$$

$$\mathrm{SPE}=\mathrm{TN}/(\mathrm{TN}+\mathrm{FP}). \quad (10)$$

Furthermore, the area under the receiver operating characteristic (ROC) curve is a graphical method that shows the trade-off between the TP rate (sensitivity) and FP rate of the classification model (van Erkel and Pattynama, 1998).

If the number of clusters in a certain class is extremely large, there would be very few training samples of each subclass in the class. The classification model would be inadequately trained, despite implementing an unbalanced data augmentation method. Thus, we set the maximum cluster number to 4. As shown in Figs. 2b–2d, the bestValue represents the best cluster numbers of benign and malignant classes in the nodule dataset, which was obtained for the three configurations through multiple iterations using a grid search strategy (Bergstra and Bengio, 2012).

For configuration 1, compared with the binary classification, the best classification accuracy was obtained when the number of benign clusters was 1 and the number of malignant clusters was 3 (bestValue=[1, 3]). We also verified the effectiveness of our method applied to 67 clinical datasets from SCH. The accuracy was 86.57%. The reason for the poor experimental performance is that, unlike the data in the LIDC database, the data were obtained from clinics. Furthermore, each radiologist may make a different judgment for the same nodule sample. For configurations 2 and 3, there were more potential types of benign and malignant pulmonary nodules. When the bestValue was [3, 2] in configuration 2 and [3, 3] in configuration 3, the classification accuracy was 86.23% and 84.06%, respectively. The experimental results revealed multiple potential subtypes of both benign and malignant pulmonary nodules. The binary CNN classification model makes it difficult to extract effective features to distinguish between benign and malignant nodules. The clustering algorithm effectively achieves class decomposition. More benign and malignant subcategories are decomposed, which is beneficial for the training of the network. In addition, weight adjustment increases the connection between the test sample and multiple cluster centers in the training dataset. This plays an important role in the classification of benign and malignant pulmonary nodules.

To further demonstrate the effectiveness of our classification scheme, we applied our method to 3D-Alex, 3D-DenseNet, and 3D-ResNet using the same dataset. The first 3D CNN architecture was modified from AlexNet. Compared with other non-deep learning methods, AlexNet achieved a significant improvement on ImageNet large-scale visual recognition tasks in 2012 (Krizhevsky et al., 2017). The second 3D CNN architecture was modified from DenseNet (Huang

(a)

Pulmonary nodules in three configurations

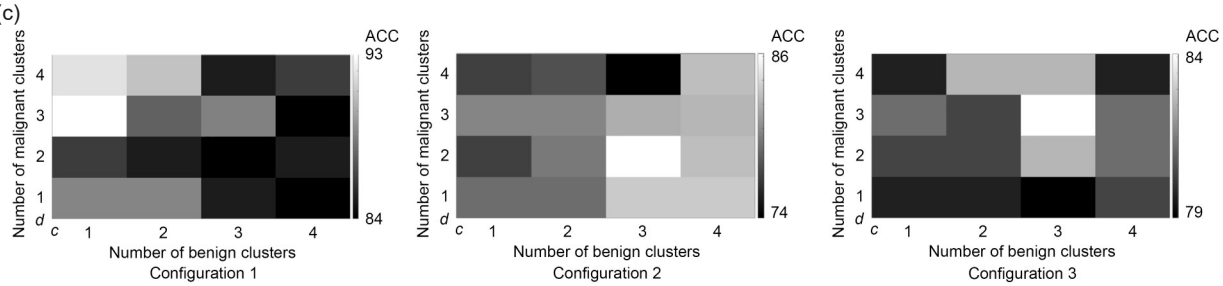
Configuration	Description	Benign	Malignant
1	LIDC: 1, 2 as benign; 4, 5 as malignant	427	253
	SCH: 1 as benign; 5 as malignant	32	35
2	LIDC: 1, 2 as benign; 3, 4, 5 as malignant	427	539
3	LIDC: 1, 2, 3 as benign; 4, 5 as malignant	713	253

(b)

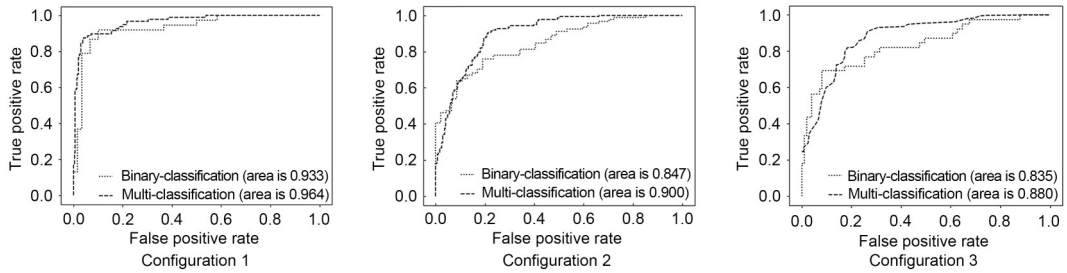
Classification performance in three configurations

Configuration	bestValue	3D-DPN aided by *K*-means clustering analysis				Binary classification ACC (%)
		ACC (%)	SEN (%)	SPE (%)	AUC (%)	
1	[1, 3]	92.86	94.44	91.94	96.43	89.80
1 (SCH)		86.57	91.43	81.25		79.10
2	[3, 2]	86.23	92.94	75.47	90.03	78.26
3	[3, 3]	84.06	61.29	90.65	87.96	78.99

(c)



(d)



(e)

Different predictive models

Architecture	Number of parameters	ACC (%)	SEN (%)	SPE (%)	AUC (%)
3D-AlexNet	2.83×10^4	86.73	83.78	88.52	89.55
3D-AlexNet with clustering		91.83	86.49	95.08	94.88
3D-DenseNet	5.00×10^5	87.76	86.84	88.33	91.74
3D-DenseNet with clustering		89.79	78.95	96.67	92.87
3D-ResNet	1.26×10^6	84.69	75.00	89.39	87.51
3D-ResNet with clustering		91.84	87.50	93.94	95.30
3D-DPN	1.85×10^5	89.80	88.10	91.07	93.32
3D-DPN with clustering		92.86	94.44	91.94	96.43

(f)

Comparison of different methods

Work	Method	Number of nodules	Configuration	ACC (%)	SEN (%)	SPE (%)	AUC (%)
Wu et al. (2019)	Handcrafted features+RF with clustering	952	1	92.37	94.28	90.27	97.02
			2	85.62	87.23	82.54	91.90
			3	82.05	93.56	56.59	86.62
Zhu et al. (2018)	Deep 3D DPN+GBM	1004	1	90.44			
Lyu et al. (2020)	ML-xResNet	865	1	92.19	92.10	91.50	97.05
Dey et al. (2018)	MoDenseNet	686	1	90.40	90.47	90.33	95.48
Mastouri et al. (2021)	[VGG16, VGG19]+SVM	3186 (2D)	1	91.99	91.85	92.27	95.90
Our method	3D DPN with clustering	966	1	92.86	94.44	91.94	96.43
			2	86.23	92.94	75.47	90.03
			3	84.06	61.29	90.65	87.96

Fig. 2 Experimental configurations and experimental results. (a) Description of pulmonary nodules in three configurations; (b) Comparison of classification performance in three experimental configurations; (c) Heat maps between bestValue and classification accuracy (%) under configurations 1, 2, and 3; (d) ROC curves for configurations 1, 2, and 3; (e) Performance of different predictive models based on 3D CNNs aided by clustering analysis; (f) Comparison of our method with the results of related studies. LIDC: Lung Image Database Consortium; SCH: Shanghai Chest Hospital; ACC: accuracy; SEN: sensitivity; SPE: specificity; AUC: area under the curve; DPN: dual path network; RF: random forest; GBM: gradient boosting machine; VGG: visual geometry group network; SVM: support vector machine.

et al., 2017), which achieved a significant improvement on four highly competitive object recognition benchmark tasks (Canadian Institute for Advanced Research (CIFAR)-10, CIFAR-100, Street View House Number (SVHN), and ImageNet). The third 3D CNN architecture was based on ResNet (He et al., 2016), which won first place in ImageNet detection and positioning, Common Object in Context (COCO) detection and segmentation, and ImageNet Large Scale Visual Recognition Challenge (ILSVRC) classification tasks. Some adjustments based on the original architectures were implemented to better adapt to the LIDC dataset. The results of each experiment with the settings described above are presented in Fig. 2e.

Note that our work did not compare and verify the hyperparameters of the network model used. Our aim was to explain that, under a better training condition, the clustering algorithm can improve classification accuracy effectively compared to the binary classification model. Fig. 2e shows that the accuracy of the different predictive models combined with a clustering algorithm was improved. This indicates that the class decomposition and weight adjustment were effective.

We compared our proposed classification model with two published methods (Fig. 2f). Overall, it is difficult to make comparisons with other studies on pulmonary nodule classification, since most studies did not employ the whole LIDC dataset. Wu et al. (2019) proposed a scheme for nodule classification based on RF with clustering analysis, using the same experimental configurations as in our study. Zhu et al. (2018) designed two deep 3D DPNs for the detection and classification of pulmonary nodules. CNN features were extracted for the task of malignant-benign classification, and the GBM was used to identify benign and malignant pulmonary nodules. The average ACC was 0.9044 on the LIDC-IDRI database.

Evidently, the overall classification performance of our proposed scheme in each configuration was better than that in the study of Wu et al. (2019). Wu et al. (2019) extracted handcrafted features and used clustering algorithms and the RF classifier for classification, whereas we used deep CNNs to extract CNN features. Compared with handcrafted features, CNN features can better reflect nodule information when the amount of data is sufficient. Comparison of our results with those of Zhu et al. (2018) revealed that the 3D DPN greatly improved classification performance after using the clustering algorithm.

The experimental results revealed that the overall classification performance of configuration 1 was better than those of configurations 2 and 3. This indicated that the classification improved after the uncertain nodules were removed. The overall classification performance of configuration 2 was better than that of configuration 3, indicating that nodules with a rank of malignancy 3 were much more likely to be categorized as benign.

We propose a novel classification method to differentiate between malignant and benign pulmonary nodules, based on 3D DPN combined with *K*-means clustering. In this study, 966 pulmonary nodules from the LIDC database and 67 from the SCH database were used for experimental verification. We also compared the method with classic deep learning architecture and state-of-the-art classification models.

In recent years, many traditional deep learning methods for the malignant-benign classification of pulmonary nodules have been proposed to optimize model structure. Considering that the internal structure of pulmonary nodules is complex, the difference between samples is large. Additionally, there are several types of benign and malignant nodules, and the overlap of multiple sub-classes may make it difficult to differentiate between benign and malignant nodules. Compared with existing deep learning research, the advantage of our proposed method is the incorporation of clustering analysis into the classification process. Moreover, the fault tolerance and accuracy of the network model are improved through class decomposition and weight adjustment. Experiments were performed with three different pulmonary nodule configurations. The three configurations regard the data in the LIDC as discarded, benign, and malignant, respectively. The experimental results revealed that, compared with the binary classification CNN model, the clustering classification method yields a certain improvement in accuracy, and can handle faulty data well.

Furthermore, we verified the advantages of three classic CNN architectures (3D-Alex, 3D-ResNet, and 3D-DenseNet) combined with clustering algorithms for the classification of malignant-benign pulmonary nodules. This comparison demonstrated the effectiveness of our proposed scheme. Compared with published methods, this scheme is comparable in terms of

the classification of malignant-benign pulmonary nodules and can extract more advanced CNN features to represent the inherent potential spatial distribution of the sample. The accuracy, sensitivity, specificity, and AUC of our proposed method were 92.86%, 94.44%, 91.94%, and 96.43%, respectively.

Lastly, our proposed method has certain limitations. First, we used only CNN features for clustering from the LIDC database and did not consider other clinical information. Second, we verified the four CNN architectures under the premise of ensuring that the model was fully learned, but did not consider the impact of hyperparameter changes on the classification results. Finally, we used only a total of 1033 nodule samples from the LIDC and SCH databases. In deep learning, a better classification performance is often obtained when the amount of data used for training increases. Therefore, the use of a larger training dataset may be a possible research direction in the future to improve the accuracy of the classification of pulmonary nodules.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 81830052), the Science and Technology Innovation Action Plan of Shanghai (No. 18441900500), and the Shanghai Natural Science Foundation of China (No. 20ZR1438300).

Author contributions

Shengdong NIE contributed to the conception of the study. Dachuan GAO performed the experiment and wrote the manuscript. Xiaodan YE provided clinical data from Shanghai Chest Hospital (SCH). Xuewen HOU, Yang CHEN, and Xue KONG contributed significantly to analysis and manuscript preparation. Yuanzhong XIE helped perform the analysis with constructive discussions. All authors have read and approved the final manuscript, and therefore, have full access to all the data in the study and take responsibility for the integrity and security of the data.

Compliance with ethics guidelines

Dachuan GAO, Xiaodan YE, Xuewen HOU, Yang CHEN, Xue KONG, Yuanzhong XIE, and Shengdong NIE declare that they have no conflict of interest.

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008 (5). This study was approved by the Ethics Committee of Shanghai Chest Hospital (approval No. KS1832). Informed consent was obtained from all patients for being included in the study.

References

Albert RH, Russell JJ, 2009. Evaluation of the solitary pulmonary nodule. *Am Fam Physician*, 80(8):827-831.

Armato SG III, McLennan G, Bidaut L, et al., 2011. 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):915-931. https://doi.org/10.1118/1.3528204

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

Cao P, Liu XL, Yang JZ, et al., 2017. A multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules. *Pattern Recogn*, 64:327-346. https://doi.org/10.1016/j.patcog.2016.11.007

Causey JL, Zhang JY, Ma SQ, et al., 2018. Highly accurate model for prediction of lung nodule malignancy with CT scans. *Sci Rep*, 8:9286. https://doi.org/10.1038/s41598-018-27569-w

Chen YP, Li JN, Xiao HX, et al., 2017. Dual path networks. Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, p.4470-4478.

Dey R, Lu ZJ, Hong Y, 2018. Diagnostic classification of lung nodules using 3D neural networks. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA. IEEE, p.774-778. https://doi.org/10.1109/ISBI.2018.8363687

Dhara AK, Mukhopadhyay S, Dutta A, et al., 2016. A combination of shape and texture features for classification of pulmonary nodules in lung CT images. *J Digit Imaging*, 29(4):466-475. https://doi.org/10.1007/s10278-015-9857-6

Díaz-Uriarte R, de Andrés SA, 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3. https://doi.org/10.1186/1471-2105-7-3

Gong J, Liu JY, Sun XW, et al., 2018. Computer-aided diagnosis of lung cancer: the effect of training data sets on classification accuracy of lung nodules. *Phys Med Biol*, 63(3):035036. https://doi.org/10.1088/1361-6560/aaa610

Gould MK, Donington J, Lynch WR, et al., 2013. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *CHEST*, 143(5 Suppl):e93S-e120S. https://doi.org/10.1378/chest.12-2351

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. IEEE, p.770-778. https://doi.org/10.1109/CVPR.2016.90

Huang G, Liu Z, van der Maaten L, et al., 2017. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA. IEEE, p.2261-2269. https://doi.org/10.1109/CVPR.2017.243

Kingma DP, Ba J, 2015. Adam: a method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations, San Diego. https://doi.org/10.48550/arXiv.1412.6980

Krizhevsky A, Sutskever I, Hinton GE, 2017. ImageNet classification with deep convolutional neural networks. *Commun ACM*, 60(6):84-90. https://doi.org/10.1145/3065386.

Lu D, Weng Q, 2007. A survey of image classification methods and techniques for improving classification performance. *Int J Remote Sens*, 28(5):823-870. https://doi.org/10.1080/01431160600746456

Lyu J, Bi XJ, Ling SH, 2020. Multi-level cross residual network for lung nodule classification. *Sensors*, 20(10): 2837. https://doi.org/10.3390/s20102837

Mastouri R, Khlifa N, Neji H, et al., 2021. A bilinear convolutional neural network for lung nodules classification on CT images. *Int J Comput Assist Radiol Surg*, 16(1):91-101. https://doi.org/10.1007/s11548-020-02283-z

McHugh ML, 2012. Interrater reliability: the kappa statistic. *Biochem Med*, 22(3):276-282.

Shen DG, Wu GR, Suk HI, 2017. Deep learning in medical image analysis. *Annu Rev Biomed Eng*, 19:221-248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Shen W, Zhou M, Yang F, et al., 2017. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recogn*, 61:663-673. https://doi.org/10.1016/j.patcog.2016.05.029

Siegel RL, Miller KD, Goding Sauer A, et al., 2020. Colorectal cancer statistics, 2020. *CA Cancer J Clin*, 70(3):145-164. https://doi.org/10.3322/caac.21601

Snoeckx A, Reyntiens P, Desbuquoit D, et al., 2018. Evaluation of the solitary pulmonary nodule: size matters, but do not ignore the power of morphology. *Insights Imaging*, 9(1):73-86. https://doi.org/10.1007/s13244-017-0581-2

van Erkel AR, Pattynama PM, 1998. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *Eur J Radiol*, 27(2):88-94. https://doi.org/10.1016/S0720-048X(97)00157-5

Wu WH, Hu HH, Gong J, et al., 2019. Malignant-benign classification of pulmonary nodules based on random forest aided by clustering analysis. *Phys Med Biol*, 64(3):035017. https://doi.org/10.1088/1361-6560/aafab0

Zhang F, Song Y, Cai WD, et al., 2013. Context curves for classification of lung nodule images. 2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Hobart, TAS, Australia. IEEE, p.1-7. https://doi.org/10.1109/DICTA.2013.6691494

Zhu WT, Liu CC, Fan W, et al., 2018. DeepLung: deep 3D dual path nets for automated pulmonary nodule detection and classification. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA. IEEE, p.673-681. https://doi.org/10.1109/WACV.2018.00079

Supplementary information

Figs. S1 and S2