IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Pulmonary nodule risk classification in adenocarcinoma from CT images using deep CNN with scale transfer module

ISSN 1751-9659
Received on 1st March 2019
Revised 12th December 2019
Accepted on 15th January 2020
E-First on 14th May 2020
doi: 10.1049/iet-ipr.2019.0248
www.ietdl.org

Jie Zheng[1], Dawei Yang[2,3], Yu Zhu[1] ✉, Wanghuan Gu[1], Bingbing Zheng[1], Chunxue Bai[2,3], Lin Zhao[4], Hongcheng Shi[5], Jie Hu[2,3], Shaohua Lu[6], Weibin Shi[7], Ningfang Wang[2]

[1]*School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, People's Republic of China*
[2]*Department of Pulmonary Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, People's Republic of China*
[3]*Shanghai Respiratory Research Institute, Shanghai 200032, People's Republic of China*
[4]*Department of Respiratory and Critical Care Medicine, Rizhao People's Hospital, Jining Medical University, Rizhao, Shandong 276800, People's Republic of China*
[5]*Department of Nuclear Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, People's Republic of China*
[6]*Department of Pathology, Zhongshan Hospital, Fudan University, Shanghai 200032, People's Republic of China*
[7]*Medical Examination Center, Zhongshan Hospital, Fudan University, Shanghai 200032, People's Republic of China*
✉ *E-mail: zhuyu@ecust.edu.cn*

Abstract: Pulmonary nodules risk classification in adenocarcinoma is essential for early detection of lung cancer and clinical treatment decision. Improving the level of early diagnosis and the identification of small lung adenocarcinoma has been always an important topic for imaging studies. In this study, the authors propose a deep convolutional neural network (CNN) with scale-transfer module (STM) and incorporate multi-feature fusion operation, named STM-Net. This network can amplify small targets and adapt to different resolution images. The evaluation data were obtained from the computed tomography (CT) database provided by Zhongshan Hospital Fudan University (ZSDB). All data have a pathological label and their lung adenocarcinomas risk are classified into four categories: atypical adenomatous hyperplasia, adenocarcinoma in situ, minimally invasive adenocarcinoma, and invasive adenocarcinoma. The authors' deep learning network STM-Net was trained and tested for the risk stage prediction. The accuracy and the average area under the receiver operating characteristic curve achieved by their method are 95.455% and 0.987 for the ZSDB dataset. The experimental results show that STM-Net largely boosts classification accuracy on the pulmonary nodules classification compared with state-of-the-art approaches. The proposed method will be an effective auxiliary to help physicians diagnosis pulmonary nodules risk classification in adenocarcinoma in early-stage.

1 Introduction

Lung cancer is the most malignant tumour with great prevalence in many countries all over the world [1]. As early lung cancer usually presents as asymptomatic pulmonary nodules, and the current diagnosis and treatment level is difficult to make a timely and accurate diagnosis of the pulmonary nodule, many patients are already in the advanced stage of lung cancer at the time of diagnosis which greatly reduces their survival. Therefore, it is important to improve lung cancer detection and diagnosis in the early stage. In the clinical medical field, ground glass opacity and ground glass nodules (GGNs) appear as tiny turbid areas in computed tomography (CT) images [2]. For such small nodules, screening for lung cancer with low-dose CT promotes the detection and diagnosis of the early stage of lung cancer, especially adenocarcinoma to a certain extent [3].

For doctors, screening pulmonary nodule from substantial CT images is a tedious and subjective task, which can easily lead to misdiagnosis and missed diagnosis, so using computer-aided diagnosis (CAD) [4] technology is especially important. The meaning of CAD includes two aspects, CADe (Computer-Aided Detection) and CADx (Computer-Aided Diagnosis). The main function of the former is to assist the radiologist to recognise and detect lung cancer, and the latter is to assist the radiologist to analyse benign or malignant of the detected lung lesions.

The main workflow of the CAD system for lung cancer includes pulmonary nodule segmentation, feature extraction and classification. Our main task is pulmonary nodules classification and identification. It is significant to learn the CAD system pulmonary nodule classification algorithm, which can effectively analyse the characteristics of the tumour such as benign and malignant judgment and better to assist the doctor or radiologist in the diagnosis.

In the past several years, many studies on the pulmonary nodule classification are based on the benign and malignant two-classification algorithm [5–11]. In fact, lung cancer generally has a long growth process of substitution, migration, evolution and transformation, so pulmonary nodule can be classified more detailed. According to the update from IASCL and summary of Travis *et al*. [12] and Zheng *et al*. [13], since 2011, lung cancer has a new international risk classification standard, the GGNs ≤ 30 mm are divided into four categories: (i) atypical adenomatous hyperplasia (AAH), (ii) adenocarcinoma in situ (AIS), (iii) minimally invasive adenocarcinoma (MIA), (iv) invasive adenocarcinoma (IAC). Among them, AAH is a benign nodule, while the other three are malignant nodules, and the malignancy risk degree increases in the order of classification.

At present, the algorithms for pulmonary nodule classification and recognition are mainly divided into traditional radiomics methods and deep learning (DL) methods [14, 15]. The traditional radiomics use image composite features combined with classifiers for classification and recognition [8]. Cornell University's Reeves *et al.* [9] utilised a 46-dimensional 3D feature which including pulmonary nodule morphology, density, surface curvature and edge gradient, and used SVM, K-nearest Neighbours and logistic regression to classify benign and malignant pulmonary nodules on International Early Lung Cancer Action Program and National Lung Screening Trail datasets. The results show that nearly 70% of the classification accuracy was achieved with optimal parameters. However, the non-uniform distribution of pulmonary nodule size in

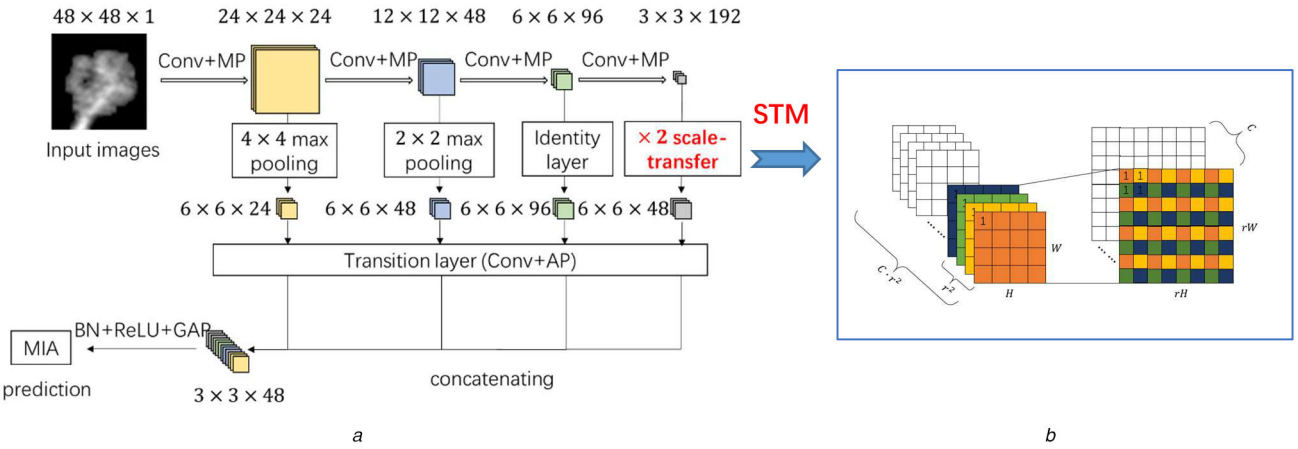

Fig. 1 *Architecture of STM-Net*

(a) STM-Net: Conv denotes Convolution, MP denotes Max Pooling, AP denotes Average Pooling, BN denotes Batch Normalisation, and GAP denotes Global Average Pooling. The red bold mark is the STM. In this figure, we use an MIA sample as the example, ***(b)*** Detail design of STM

datasets also affects the classification accuracy. Li *et al.* [10] applied an improved semi-supervised FCM clustering algorithm to cluster 11 features including grey variance, boundary roughness and moment invariant, and accuracy reached 77.6% on the Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) dataset. Alahmari *et al.* [11] combined delta radiomics with conventional (non-delta) features, using machine learning to classify the benign and malignant pulmonary nodules, found the best improved average area under the receiver operating characteristic curve (AUC) was 0.822 when delta features were similar with conventional features versus an AUC 0.773 for conventional features only. In the previous study, we proposed a method for calculating the grey density distribution characteristics, and classified the nodules into four categories (AAH, AIS, MIA, IAC) in Zhongshan Hospital Fudan University (ZSDB) dataset, the overall classification accuracy reached 89.2% [16]. After that, we also proposed a new exponential function to weight the angular histogram for each component of the distribution map [17]. The accuracy was increased to 90.8%.

However, traditional methods have several limitations. First, the diversity of the pulmonary nodule causes poor robustness of traditional methods. Secondly, the hand-defined features often fail to fully reflect the target information. These factors affect the final classification and recognition performance. In order to address these problems, previous work has concentrated on applying DL model to pulmonary nodules classification and recognition. DL model simulates the form of human brain neurons to transmit information for self-learning and training, and thus it has become very effective in the field of image and video in recent years [14, 15]. DL also plays an important role in biomedicine domain, such as U-Net [18], 3D-Unet [19], HIC-Net [20] and Agile-CNN [21].

Setio *et al.* [22] designed a pulmonary nodule depth-assisted diagnosis network, which used multi-view to reduce false positive rate. This method was validated on 888-frame LIDC-IDRI dataset, achieving 85.4% sensitivity with 1FPs/scan false positive rate and 90.1% sensitivity with 4FPs/scan false positive rate; Shen *et al.* [23] used multi-crop convolutional neural network to classify suspected malignant pulmonary nodules, dividing the nodules into benign nodules and malignant nodules. The experiment also tested on LIDC-IDRI dataset, and the accuracy of the classification reached 87.14%, and AUC of the ROC curve reached 0.93; Dai *et al.* [24] proposed a new 3D network ALNC-3D, which combined pulmonary nodule benign–malignant classification and pulmonary nodule image attributes classification to improve the accuracy of pulmonary nodule classification, and obtained 91.47% classification accuracy on LIDC-IDRI dataset; Zhao *et al.* [25] proposed a new network DenseSharp, which classified their own dataset HHDB (from Huadong Hospital Affiliated to Fudan University) into three categories (AAH-AIS, MIA, IAC) in the form of 3D maps and the accuracy reached 64.1% on the dataset collected by themselves.

For lung adenocarcinomas risk classification, the most difficulties are the representation of small nodules features. The idea of multi-scale module fusion is effective in CNNs for image super-resolution [26], semantic segmentation [27] and object detection [28]. In this paper, we propose a DL algorithm based on scale-transferrable. The algorithm applies the scale-transfer module (STM) to the classification networks, and incorporates with multi-feature fusion algorithm. It can solve the scale diversification of the pulmonary nodule CT images and the classification of the small nodule target problems, and has achieved remarkable performance in the classification strategy according to IASCL and Travis *et al.* [12] and Zheng *et al.* [13].

The contributions of this paper are as follows: (a) we propose a new deep CNN network (STM-Net) which incorporates STM and multi-feature fusion algorithm to design a variable-scale DL method; (b) STM-Net can amplify the features of the small targets clearly without adding any parameters or inserting 0 when up-sampling. (c) STM-Net has an advanced pulmonary nodule risk classification performance on ZSDB dataset, which reaches more than 95% accuracy for four-classification and 0.987 AUC value of (AAH, AIS, MIA and IAC). (d) We also visualise and analyse the feature distribution of pulmonary nodules from the output of the network middle layer.

The rest of this paper is organised as follows. In Section 2, we introduce ZSDB dataset and our network architecture. Section 3 introduces the implementation details of experiments, evaluates experimental results and compares to the other two networks. The discussion and conclusion of this paper are given in Sections 4 and 5, respectively.

2 Methods and materials

2.1 Network

Due to the nodule size in the segmented pulmonary nodule picture being small, we incorporate the STM and multi-feature algorithm to design the STM-Net. The overall network architecture of STM-Net can be seen in Fig. 1*a*.

As seen in Fig. 1*a*, we first input $48 \times 48 \times 1$ segmented pulmonary nodule images into four convolution layers and four pooling layers, the kernel size of these four convolution layers is 3×3 and stride is 1, and four pooling layers use the 2×2 max-pooling with stride 2. After that, we obtain four sets of feature maps with different sizes. Then we use max-pooling to obtain the low-level feature map with large receptive field and use STM to obtain the deep feature map with high-level semantic information, which also can make all feature maps to the same size. Thirdly, these same size feature maps through a transition layer which can uniform channel dimension of each feature group. The transition layer consists of 1×1 convolution layer and 2×2 average pooling layer. After the transition layer, we have the four groups feature map with the same channel size, then we use channel-fusion to

Table 1 STM-Net architecture. Conv_1–Conv_4 are the four convolution layers on the upper part of the network. Conv_1_S–Conv_4_S are the sampling layers of Conv_1–Conv_4. Conv_1 and Conv_2 use down-sampling, and Conv_3 uses the identity map. The Conv_4 uses the STM. TL (1–4) represents the transition layer of Conv_1_S–Conv_4_S. CL is the concatenating layer, and LL is the linear layer. The output size of LL represents the final number of classifications

Layers	Output Size(Input $48 \times 48 \times 1$)	STM-Net
Conv_1	$24 \times 24 \times 24$	3×3 conv, stride 1; 2×2 max pooling, stride 2;
Conv_2	$12 \times 12 \times 48$	3×3 conv, stride 1; 2×2 max pooling, stride 2
Conv_3	$6 \times 6 \times 96$	3×3 conv, stride 1; 2×2 max pooling, stride 2
Conv_4	$3 \times 3 \times 192$	3×3 conv, stride 1; 2×2 max pooling, stride 2
Conv_1_S	$6 \times 6 \times 24$	4×4 max pooling, stride 4; BN; ReLU
Conv_2_S	$6 \times 6 \times 48$	2×2 max pooling, stride 2; BN; ReLU
Conv_3_S	$6 \times 6 \times 96$	Identity layer; BN; ReLU
Conv_4_S	$6 \times 6 \times 192$	$\times 2$ scale-transfer module; BN; ReLU
TL(1–4)	$3 \times 3 \times 12$	1×1 conv, stride 1; 2×2 avg pooling, stride 2;
CL	$3 \times 3 \times 48$	concatenating; BN; ReLU
LL	4/3	1×1, stride 1

concatenate these feature maps. Then outputs are put in the linear layer which includes batch normalisation operation, ReLU and global average pooling. Finally, through softmax classifier to obtain recognition results.

In this network, the input feature map size of the transition layer is 6×6, and the STM with up-sampling factor 2 is applied to the fourth convolution layer. After the STM, the feature map size is changed from $3 \times 3 \times 192$ to $6 \times 6 \times 48$. Then through the transition layer, the feature map becomes $3 \times 3 \times 12$, and finally fuses into $3 \times 3 \times 48$. The specific implementation details of our convolutional network are shown in Table 1.

Fig. 1*b* shows the STM, which is the key to achieve multi-feature fusion. Assume that the input of the upper layer is $H \times W \times C \cdot r^2$, where r is the up-sampling factor. When performing up-sampling, the number of channels is equally divided into C parts, each part has r^2 channels. Feature maps in one part are mapped into a new feature map, the elements in original feature maps were rearranged periodically along channels. The output tensor of the STM is $rH \times rW \times C$. In contrast with traditional up-sampling and deconvolution, the STM is a more effective up-sampling operation, because the scale-transfer operation achieves up-sampling by compressing the number of channels to expand width and height, there are no additional noise points and parameters. And comparing with upscale methods [29, 30] which need convolution with a stride 1 in large-scale feature maps, STM can implement up-sampling without train.

Therefore, the STM is r^2 time faster than the methods up-sampling before convolution. The principle of STM can be summary as (1)

$$I^{\mathrm{LS}}_{xr,yr,c} = I^{\mathrm{SS}}_{x,y,c \cdot r^2} \tag{1}$$

where I^{LS} is the large-scale feature maps, and I^{SS} is the small-scale feature maps. The channel number of I^{SS} must satisfy an integer multiple of r^2. After STM, every four pixels share the same receptive fields. So the features get enhanced, especially for the small nodules since the STM can amplify small target features.

2.2 Dataset

We evaluate our approach on a lung CT image dataset – ZSDB which is provided by the cooperative grade-A tertiary hospital – Zhongshan hospital, Fudan University.

ZSDB dataset: Considering the diversity of pulmonary nodules, researchers try to create a new targeted dataset to illustrate the universality and robustness of classification methods. The Medical Image Processing Laboratory of East China University of Science and Technology used the rich materials provided by the cooperative grade-A tertiary hospital to establish a large-scale lung CT image dataset – ZSDB. The dataset has 1971 samples, including pulmonary nodules with the diameter ≤ 32 mm, and all nodules were labelled based on pathological diagnosis, which are more accurate. According to IASLC criteria, pulmonary nodules are classified into four categories (AAH, AIS, MIA and IAC). The imaging parameters of the ZSDB dataset are as follows: electric parameters are 500 mA, 120 kV, the image size is 512×512, the image type is normal CT, the chest image pixel distribution density is 0.703125 mm, and the single sheet thickness is 0.625 mm. The LIDC-IDRI dataset is a combination of incidentally found or screening cohort, and the scan condition is not limited by the 1 mm thickness. In our study, we mainly focus on establishing a DL model designed for assisting current screening project. So we collect the cases from a local CT screening program, and collect the CT scan with the thickness lower than 1 mm.

2.3 Dataset pretreatment

The original image on ZSDB dataset is untrainable for the classification task, pulmonary nodule part must be segmented out as input data for training. ZSDB dataset has physician-labelled auxiliary documentation for us to segment the nodule section conveniently. However, simply extracting the nodule area is not enough for classification because there are numerous impurities and noise that affect the classification effect. Therefore, the lung parenchyma segmentation is an important pre-treatment work before the pulmonary nodule classification.

There are many effective methods for lung parenchyma segmentation, such as grey segmentation including threshold method [31], region growing method [32], clustering method [33] and random field method [34] boundary segmentation including Sobel operator model and Prewitt operator model, active contour model segmentation including Snake model [35]. In this paper, we use the KMEANS unsupervised clustering algorithm to distinguish the pixel-level grey in different regions of lung CT images [16], retain pixel values of the nodule and clear redundant pixel values of other parts, the lung parenchyma segmentation results are shown in Fig. 2.

The number of samples after lung parenchyma segmentation is as follows: 34 samples of AAH, 312 samples of AIS, 242 samples of MIA, and 1383 samples of IAC, total 1971 samples. In experiments, the ratio of training samples and testing samples is 4:1, so the samples number of the training set and test set is 1577 and 394, respectively. And we use the five-fold cross-validation to train networks, in each fold training samples account for 80% and validation samples account for 20%. The sample images of ZSDB dataset after lung parenchymal segmentation and pulmonary nodule extraction are shown in Fig. 3.

In addition, because the number of samples on ZSDB dataset is insufficient and the number imbalance between different classes is too large. More specifically, the maximum number of IACs is more than 50 times that of the minimum number of AAHs. Therefore, we consider use data augmentation to assist in training. The data augmentation methods used in this paper include: (a) random rotate image of 0–15° clockwise or counter-clockwise; (b) random translate of 2 pixels in each axis.

2.4 Training

This paper uses Python 3.6 and Tensorflow 1.2 with GPU (1080Ti) to collate data and train network. The main task is training the multi-classification of pulmonary nodules in adenocarcinoma on our own dataset ZSDB. The batch size and epoch in training is set

to 64 and 50, respectively. The initial learning rate is set to 0.001 which is divided by 10 at 25th and 40th epoch, and the dropout rate is set to 0.5 for reducing overfitting. In addition, we use stochastic gradient descent (SGD) and cross entropy loss [36] to train STM-Net, the loss function is shown in (2):

$$l_{\text{cls}} = -\frac{1}{n}\sum_{n}\sum_{c} y_{\text{c}} \log p_{\text{c}} \tag{2}$$

where n is the number of samples (the batch size in the specific training), c is the number of class, which is 4 or 3 on ZSDB dataset. y_{c} represents the distribution of real sample, when samples belong to the same category, $y_{\text{c}} = 1$; otherwise, $y_{\text{c}} = 0$. p_{c} is the final output of the network, it represents the predicted classification probability.

3 Results

3.1 Classification performances

In order to verify the effectiveness of STM-Net, we compare the classification results with four other networks: 2D-CNN [37], DenseNet [38], MSP-Net [39] and STM-SVM in Table 2. 2D-CNN is a basic 2D convolutional network with four convolution layers, DenseNet has the characteristics of multi-feature fusion, MSP-Net uses the multi-scale pooling, and STM-SVM uses SVM as classifier to classify the feature extracted from STM-Net. Due to the lack of sample data, especially the AAH which only has 34 samples, we carried out experiments on ZSDB dataset with four different training strategies: (a) the original data as training samples, divided into four categories (AAH, AIS, MIA and IAC); (b) data augmentation for AAH samples, AAH is expanded to 4 times, and still divided into four categories; (c) data augmentation for the other three types of samples except IAC, AAH is expanded to 20 times, AIS and MIA are expanded to 5 times, divided into four categories; (d) combining AAH and AIS into one category, divided into three categories (AAH-AIS, MIA and IAC). The four training strategies in this paper are abbreviated as 'Ori.' 'DA1', 'DA2', and '3c.' in the later part.

The reason of 3-categories classification is that the AAH samples with pathology label are too few for fairly training the deep neural networks. In clinical, the AAH lesions are usually considered as benign, and they rarely undergo surgical treatment unless obvious malignant signs are presented in the CT images [25]. However, fortunately, it is still reasonable in the clinical context, the two subtypes of AAH and AIS lesions ($\lesssim$3 cm) are reported to have a 100% disease-specific survival if they are completely resected [12]. A 3D DL network DenseSharp was

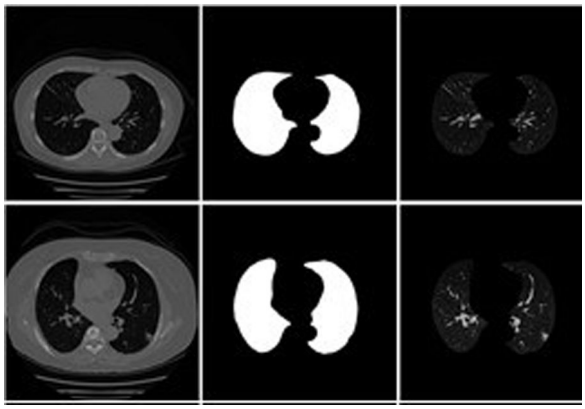

Fig. 2 *Examples of lung parenchymal segmentation*

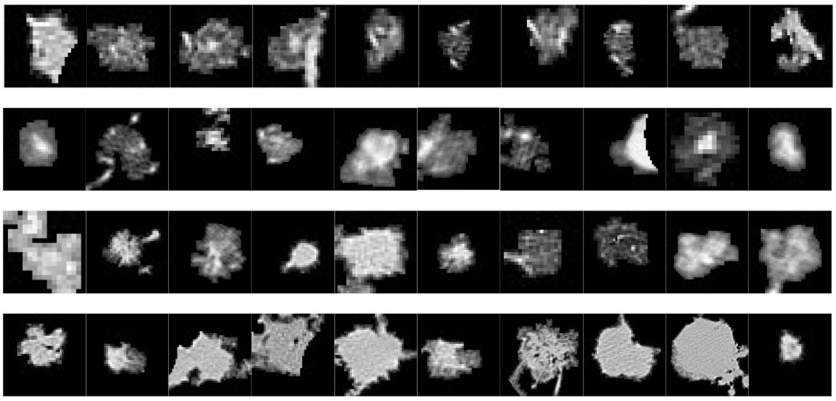

Fig. 3 *Image samples of ZSDB dataset after lung parenchymal segmentation and pulmonary nodule extraction. The rows show the AAH, AIS, MIA and IAC image samples respectively. For the convenience, the sample size here is processed in the same size*

Table 2 Classification accuracy on ZSDB dataset for four training strategies. Ori. is the abbreviation of Original, which represents training data without data augmentation. DA1 and DA2 are 4-classification with data augmentation. On the DA1 training strategy, AAH is expanded to 4 times. On the DA2, AAH is expanded to 20 times, and AIS and MIA are expanded to 5 times. 3c. represents 3-classifications on experiment, which classify AAH and AIS into one category

Method	Ori., %	DA1, %	DA2, %	3c., %
2D-CNN [37]	88.070	90.171	89.868	91.081
DenseNet [38]	90.657	90.910	90.152	92.545
MSP-Net [39]	92.172	93.687	91.162	95.630
STM-SVM (ours)	92.920	93.950	92.117	95.981
STM-Net (ours)	**94.697**	**95.455**	**94.444**	**97.429**

Table 3 Sensitivity and specificity of STM-Net testing on ZSDB dataset for DA1. Avg. is the abbreviation of average. The average sensitivity and specificity are calculated by the number of each class and their scores

Class	AAH	AIS	MIA	IAC	Avg.
sensitivity (with 95% CI)	0.857 (0.887–0.827)	0.952 (0.970–0.934)	0.857 (0.887–0.827)	0.971 (0.985–0.957)	0.952 (0.970–0.934)
specificity (with 95% CI)	0.995 (1.000–0.989)	0.982 (0.993–0.971)	0.991 (0.999–0.983)	0.982 (0.993–0.971)	0.983 (0.994–0.972)

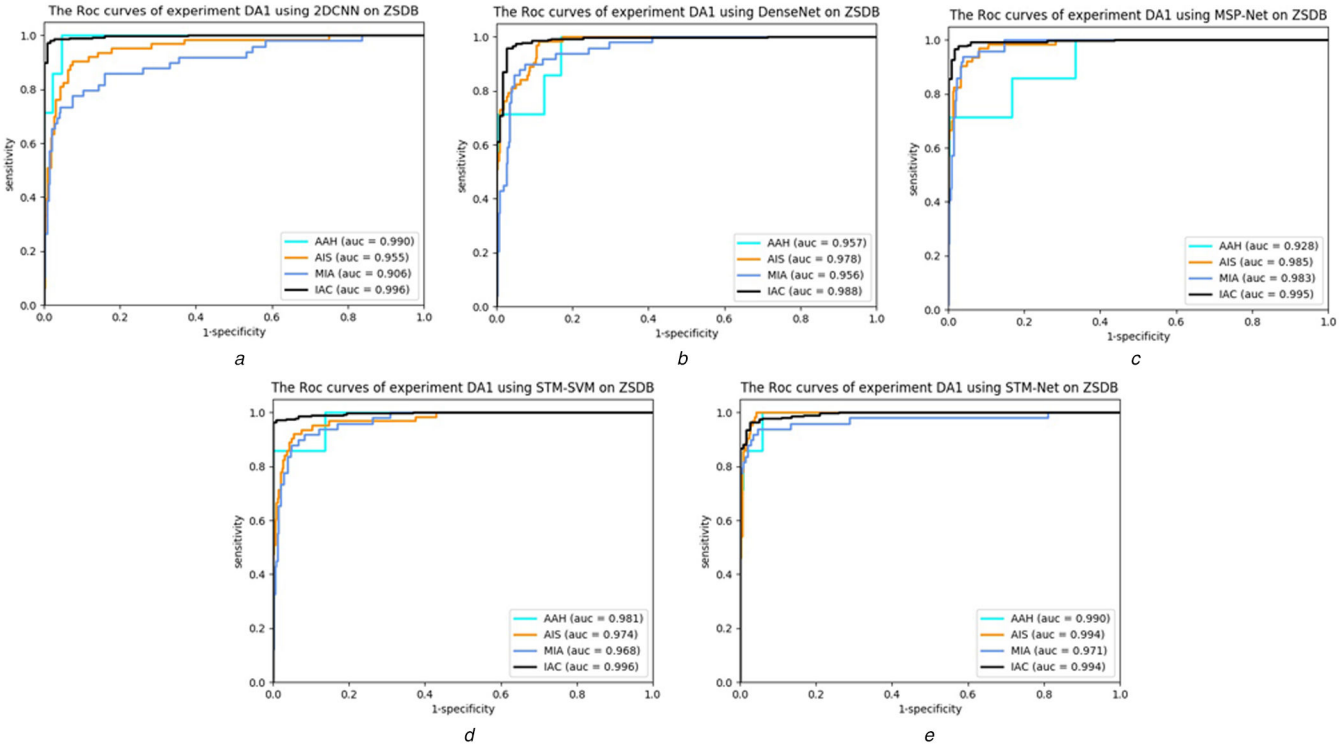

Fig. 4 *ROC curve of DA1 using 2D-CNN, DenseNet, MSP-Net, STM-SVM and STM-Net on ZSDB dataset*
(a) 2D-CNN, ***(b)*** DenseNet, ***(c)*** MSP-Net, ***(d)*** STM-SVM, ***(e)*** STM-Net

proposed to classify three categories (AAH-AIS, MIA, IAC) on dataset HHDB (from Huadong Hospital Affiliated to Fudan University) and the accuracy reached 64.1% [25]. Although the STM-Net DL method proposed in this paper also get the best classification accuracy in the 3-categories, it has significant improvement in 4-categories classification task to satisfy the demand of auxiliary clinical diagnosis, shown in Table 2.

In Table 2, we can observe the following three points. First, the classification accuracy of STM-Net is the best in all four training strategies than 2D-CNN, DenseNet, MSP-Net and STM-SVM, which achieve 94.697% [95% CI 96.607–92.787%], 95.455% [95% CI 97.225–93.685%], 94.444% [95% CI 96.394–92.494%] and 97.429% [95% CI 98.779–96.079%] on Ori., DA1, DA2 and 3c, respectively. Secondly, the classification accuracy of STM-SVM is better than the DenseNet and MSP-Net, which means the feature extracted from the STM-Net contains better classification information of pulmonary nodule. Thirdly, the accuracy results from the last column show that the (AAH-AIS, MIA, IAC) three classification (3c.) has the best training performance compared with other three strategies. However, the overall accuracy of DA2 is slightly lower than Ori., it can be caused by the data augmentation for the other three categories except for IAC. After this data augmentation, the classification accuracy of these three categories increase, while the accuracy of IAC decreases slightly, and the overall accuracy also shows a downward trend. The sensitivity and specificity of STM-Net on DA1 strategy are shown in Table 3. The average sensitivity and specificity of STM-Net testing on ZSDB dataset for DA1 are 0.952 and 0.983, respectively. The average sensitivity and specificity are obtained through multiplying the number of each category by their respective scores, and then dividing by the total number of samples.

Fig. 4 shows the multi-classification ROC curves and AUC of the five networks with DA1 training strategies. It can be found that with the same data augmentation (DA1), the AUC of STM-Net is higher than other networks. The average AUC of STM-Net on DA1 is 0.987, while 2D-CNN [37], DenseNet [38], MSP-Net [39] and STM-SVM are 0.962, 0.970, 0.973 and 0.980, respectively. The ROC curve of STM-Net is also the best among these five networks. In particular, the improvement of AAH in STM-Net is significant, and the AUC of AAH has achieved 0.990 for DA1.

Fig. 5 is the Confusion matrixes of four training strategies using STM-Net on ZSDB dataset. The confusion matrix, also known as the error matrix, is a standard format for accuracy evaluation, where the abscissa is the prediction category and the ordinate is the real category. The confusion matrix is a good representation of classification performance. For example, the DA1 in Fig. 5 indicates that 86% of AAH data are classified correctly, and the remaining 14% are misclassified into MIA. For Ori., we can see that the classification accuracy of AAH with this training strategy is relatively low. This is because without data augmentation, the number of AAH samples is very rare. In comparison, the classification performance of AAH with data augmentation (DA1 and DA2) is significantly better. Although the overall accuracy of DA2 is not so satisfactory (see Table 2), the confusion matrix in Fig. 5 indicates that the classification result of DA2 is more balanced than the other three training strategies. It is because the number imbalance between the four samples types of DA2 is the smallest. In addition, the confusion matrix shows that the confusion between AIS and MIA is serious. It may prove that the characteristics of AIS and MIA are relatively close, so the classifier can misjudge easily when classifying these two types. Overall, the performance of the three-classifications (3c.) is the best, which can also be confirmed from Table 2.

Fig. 6 presents some predicted results, which including four correct predicted results and one incorrect predicted result. The pulmonary nodule area is in the rectangle, the yellow is the correct predicted result, and the red is the incorrect predicted result. The ground truth of these pulmonary nodule images is AAH, AIS,

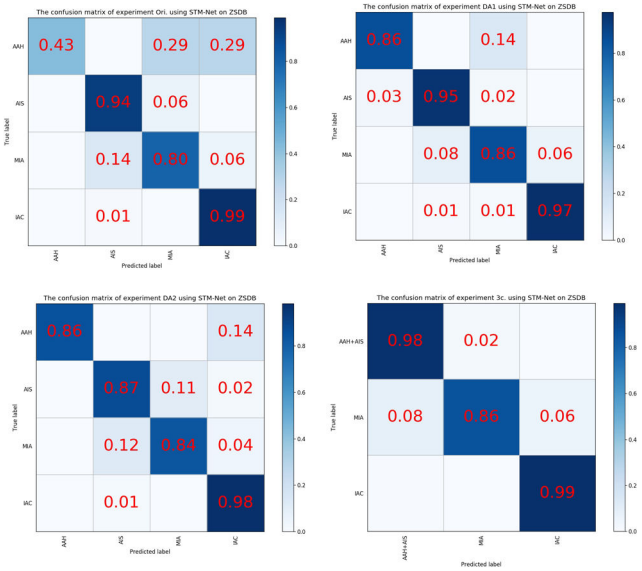

Fig. 5 *Confusion matrixes of four training strategies using STM-Net on ZSDB dataset. Confusion matrixes of four training strategies using STM-Net on ZSDB dataset. Ori. represents training data without data augmentation. DA1 and DA2 are 4-classification with data augmentation strategy. For DA1 training strategy, AAH is expanded to 4 times. For DA2, AAH is expanded to 20 times, and AIS and MIA are expanded to 5 times. 3c. represents 3-classifications on experiment, which classify AAH and AIS into one category*

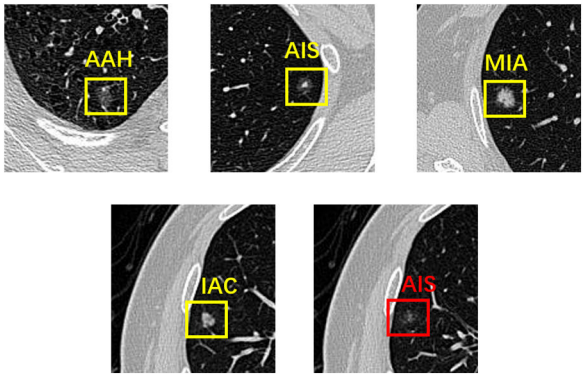

Fig. 6 *Examples of the predicted classification results on ZSDB dataset*

MIA, IAC and IAC respectively. The two images in the second row belong to the same lung, but the second image is the edge area of this pulmonary nodule, its solid part is very small, so it is misclassified into AIS.

3.2 Results analysis

The output dimension of the linear layer before softmax is 48×1, and these 48 features are self-learned by the network. For a DL algorithm, although it is difficult to know what these 48 features specifically represent, we can analyse the differences and connections between them by comparing the 48 features among different categories. As shown in Fig. 7, from top to bottom correspond to AAH, AIS, MIA and IAC, respectively. The first image of each row is a partial screenshot from the original CT image, and the yellow arrow in this image points to the pulmonary nodule. The second image in each row is a nodule sample taken from the CT image which after the lung parenchyma segmentation, and also is the input image of the network. The last is the output line chart of the linear layer before the softmax layer, which is the 48-dimensional feature mentioned before. From the distribution of the line chart, we can see that the other three categories except IAC fluctuate from low to high, while the tendency of IAC is from high to low. In addition, the peaks of AAH is narrow and high in the rear of the line chart, while the peaks of AIS move a little bit forward and relatively flat. The peaks of MIA have the same shift trend as AIS but distribute in a wider range. Overall, the distribution of AIS and MIA is relatively close, which confirms the previous conclusion: AIS and MIA samples are easily confused and misjudged.

Table 4 shows the different pulmonary nodule classification methods and their performances, where HHDB is collected by Zhao *et al.* [25] cooperate with the Huadong Hospital Affiliated to Fudan University. Compared with the literature which used the same ZSDB dataset [17], the performance of this paper is obviously more impressive, and our classification accuracy is about 4.5% higher than [17].

4 Discussion

Early detection of lung cancer increases the chances of patients' survival, which increases the motivation in developing accurate and fast diagnostic tools to detect lung cancer earlier. Automated classification and recognition of pulmonary nodule can effectively assist physicians in the diagnosis and analysis of disease. We design an end-to-end DL CNN network STM-Net with STM to classify GGNs to (AAH, AIS, MIA and IAC) in ZSDB dataset according to the update from IASCL and Travis *et al.* [12] and Zheng *et al.* [13].

Radiomics and DL methods are usually established to calculate different kinds of features to predict clinical cancer status. There are lots of work to screen pulmonary nodule into binary categories as benign and malignancy [9–11, 22–24]. For instance, Jacobs *et al.* [40] proposed a 128-dimensional radiomics feature for semi-substantial nodules, and verified it on NLST dataset, which achieved 80% classification accuracy. More recently, Li *et al.* [41] effectively fusion the intensity, geometric and texture features, rotation invariant uniform local binary pattern and Gabor filter methods to generate valid eigenvectors, then used the random forest method to modify the eigenvectors to classify benign and

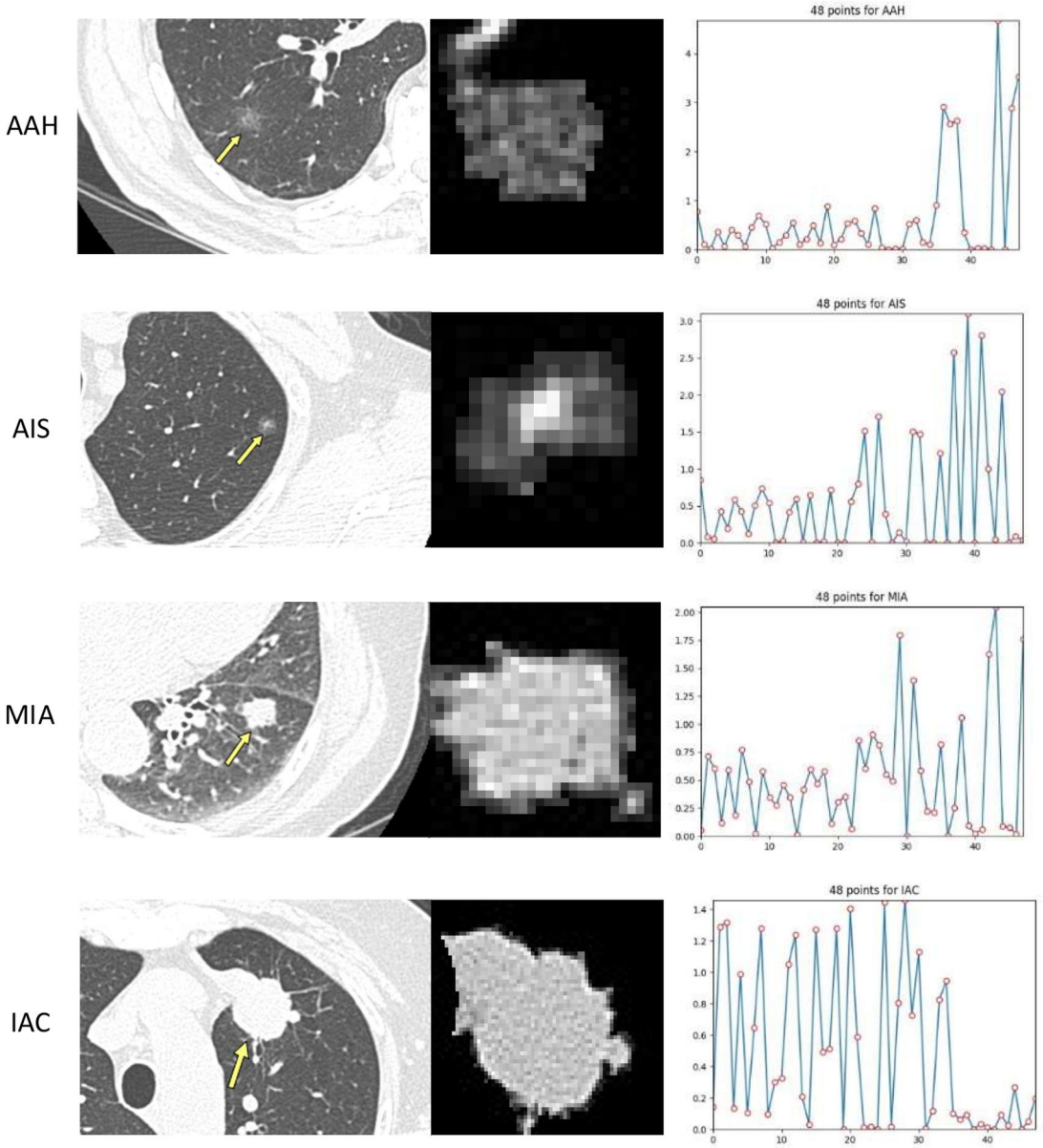

Fig. 7 *Examples of the linear layer features of STM-Net. The first column is the original CT screenshot, and the yellow arrow points to the pulmonary nodule. The second column is the nodule sample after lung parenchyma segmentation, which is also used for network training. The third column is the output line chart of the linear layer before softmax. The output dimension of the linear layer is* 48×1*, corresponding to 48 points in the line chart*

Table 4 Different classification methods and their classification performances

Methods	Datasets	Performance
DenseSharp [25]	HHDB	accuracy is 64.1% for 3-classification task (AAH + AIS,MIA,IAC).
angular density feature [17]	ZSDB	accuracy is increased to 90.8% compared to the above.
STM-Net(ours)	ZSDB	accuracy reaches 95.455% for 4-classification and 97.429% for 3-classification.

malignant nodules and achieved 0.92 sensitivity results on LIDC-IDRI dataset. For DL methods, such as multi-scale CNN model [23], advanced DenseNet [42] and 3D DCNN models [24], are also developed to classify benign and malignant nodules.

However, the four categories standard (AAH, AIS, MIA and IAC) of lung cancer are rarely discussed with radiomics and DL methods. In our previous work, the proposed characteristics of grey density distribution with exponential weighted angular histogram classified the nodules into four categories (AAH, AIS, MIA and IAC) in ZSDB dataset and get an accuracy of 90.8% [17]. This, however, still depends on hand-craft feature engineering. In this paper, we designed a new CNN for this task. Since the size of pulmonary nodules is usually small, the features are not easy to detect, which makes the classification task become difficult. We propose the pulmonary nodule classification network STM-Net which incorporate STM and scale fusion to achieve the multi-classification task. The mechanism of STM increases the inception field of convolution layer output and effectively amplifies small target features. The proposed DL model STM-Net achieves 95.455% accuracy with proper data augmentation. Compared with other popular used DL models, such as 2D-CNN [37], DenseNet [38], multi-scale pooling [39], the STM-Net significantly increases the classification accuracy of small pulmonary nodules.

Although it is still a black-box of the internal mechanism of DL, we try to further analyse the features learned from our DL network. The linear layer output with 48-dimensional features was discussed in Fig. 7. Obviously, the distribution of learned features can characterise different categories effectively. This is a useful hint for future research about how to combine DL and radiomics to achieve better results.

Additionally, the majority of collected patients' pulmonary nodule samples are malignant, which lead to the number imbalance in different categories on ZSDB dataset. More specifically, AAH

patients are healthier than other patients and they do not need surgical treatment, which causes a serious lack of AAH data with pathology. By contrast, the number of IAC data on ZSDB dataset is the most. Therefore, we use data augmentation to address this problem of insufficient data, and the experimental results show that data augmentation is indispensable. However, only relying on data augmentation to expand the number of data can be temporary, because data augmentation only expands the original data and keep the same distribution of it. However, the characteristics of pulmonary nodules are very complex, the rare samples of AAH cannot fully present the distribution of AAH features. Therefore, in order to achieve more reliable and favourable performance, we still need to collect more train samples in the following works. In clinical, there is some work to combine radiomics analysis with genomic variation [43, 44], relapse of the disease [45], and post-treatment progression-free survival [7].

Although resection of pulmonary nodules is the ideal and reliable way for diagnosis, there is a crucial need for developing non-invasive diagnostic CADs to eliminate the risks associated with the surgical procedure. It is increasingly important to achieve pulmonary nodules detection and segmentation automatically as well as classification. The current aim of combination of the DL method and oncogene variation detection is to increase diagnostic accuracy. We will also focus on the optimisation of DL network with radiomics and genomics, and realise an end-to-end CAD system to help lung cancer screening and auxiliary diagnosis. It is significant to support doctors making clinical decisions for lung cancer patients and increasing their survival rate when detecting early-stage lung cancer.

5 Conclusion

In this paper, we proposed a new pulmonary nodule risk classification network, STM-Net. This network incorporates STM and multi-feature fusion to achieve pulmonary nodule classification. The input pulmonary nodule images first pass through four convolution and pooling layers to extract the four different size features. Then using max-pooling and STM to unify the size of feature maps, and through the transition, layer to unify the channel size of these four groups of feature map. After that, we use channel fusion to combine different semantic level features and achieve final classification. Experimental results show that our proposed network can achieve more effective classification performance on ZSDB dataset, the accuracy and the AUC are 95.455% and 0.987, respectively.

However, the pulmonary nodule classification is only a part of the CAD system. Therefore, in the subsequent research, we will try to integrate pulmonary nodule segmentation and feature extraction to achieve end-to-end CAD system to help lung cancer screening and auxiliary diagnosis.

6 Acknowledgments

The authors greatly appreciate the financial supported by the Zhongshan Hospital Clinical Research Foundation Nos. 2016ZSLC05 and 2016ZSCX02, the National Key Scientific and Technology Support Program No. 2013BAI09B09, the Natural Science Foundation of Shanghai No. 15ZR1408700, the Shandong Medical and Health Science and Technology Development Plan Project (2017WS717) and Support Project for Young Teachers of Jining Medical University (JY2016KJ053Y). The first two authors contributed equally to this article, and both should be considered first authors.

7 References

[1] Dhara, A.K., Mukhopadhyay, S., Khandelwal, N.: 'Computer-aided detection and analysis of pulmonary nodule from CT images: a survey', *IETE Tech. Rev.*, 2012, **29**, (4), pp. 265–275

[2] Jin, M.G., Chang, M.P., Hyun, J.L.: 'Ground-glass nodules on chest CT as imaging biomarkers in the management of lung adenocarcinoma', *Am. J. Roentgenol.*, 2011, **196**, (3), pp. 533–543

[3] Kishi, K., Homma, S., Kurosaki, A., *et al.*: 'Small lung tumors with the size of 1 cm or less in diameter: clinical, radiological, and histopathological characteristics', *Lung Cancer*, 2004, **44**, (1), pp. 43–51

[4] Kunio, D.: 'Computer-aided diagnosis in medical imaging: historical review, current status and future potential', *Comput. Med. Imaging Graph.*, 2007, **31**, (4–5), pp. 198–211

[5] Junji, S., Hiroyuki, A., Roger, E., *et al.*: 'Effect of the computer output on radiologists' decision-making for classification of solitary pulmonary nodules in chest radiographs'. CARS 2002 Computer Assisted Radiology and Surgery, Paris, France, 2002, pp. 588–595

[6] Tan, Y., Schwartz, L.H., Zhao, B.: 'Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field', *Med. Phys.*, 2013, **40**, (4), pp. 043502–043502

[7] Maldonado, F., Duan, F., Raghunath, S.M., *et al.*: 'Noninvasive computed tomography–based risk stratification of lung adenocarcinomas in the national lung screening trial', *Am. J. Respir. Crit. Care Med.*, 2015, **192**, (6), pp. 737–744

[8] Lambin, P., Rios-Velazquez, E., Leijenaar, R., *et al.*: 'Radiomics: extracting more information from medical images using advanced feature analysis', *Eur. J. Cancer*, 2012, **48**, (4), pp. 441–446

[9] Reeves, A.P., Xie, Y., Jirapatnakul, A.: 'Automated pulmonary nodule CT image characterization in lung cancer screening', *Int. J. Comput. Assist. Radiol. Surg.*, 2016, **11**, (1), pp. 73–88

[10] Li, Q., Liu, H., Su, Z.: 'Modified fuzzy clustering with partial supervision algorithm in classification and recognition of pulmonary nodules', *J. Graphics*, 2015, **36**, (2), pp. 244–250

[11] Alahmari, S.S., Cherezov, D., Goldgof, D.B.: 'Delta radiomics improves pulmonary nodule malignancy prediction in lung cancer screening', *IEEE. Access.*, 2018, **6**, pp. 77796–77806

[12] Travis, W.D., Asamura, H., Bankier, A.A., *et al.*: 'The IASLC lung cancer staging project: proposals for coding T categories for subsolid nodules and assessment of tumor size in part-solid tumors in the forthcoming eighth edition of the TNM classification of lung cancer travis', *J. Thorac. Oncol.*, 2016, **11**, (8), pp. 1204–1223

[13] Zheng, X.P., Li, M., Zhang, G.Z.: '*Early-stage lung cancer: screening and management*' (Springer, Singapore, 2015)

[14] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'Imagenet classification with deep convolutional neural networks', *Commun. ACM*, 2012, **60**, (2), p. 2012

[15] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016, pp. 770–778

[16] Le, V., Yang, D., Zhu, Y., *et al.*: 'Automated classification of pulmonary nodules for lung adenocarcinomas risk evaluation: an effective CT analysis by clustering density distribution algorithm', *J. Med. Imaging. Health. Inform.*, 2017, **7**, (8), pp. 1753–1758

[17] Le, V., Yang, D., Zhu, Y., *et al.*: 'Quantitative CT analysis of pulmonary nodules for lung adenocarcinoma risk classification based on an exponential weighted grey scale angular density distribution feature', *Comput. Methods Programs Biomed.*, 2018, **160**, pp. 141–151

[18] Ronneberger, O., Fischer, P., Brox, T.: 'U-Net: convolutional networks for biomedical image segmentation'. Int. Conf. on Medical Image Computing and Computer-assisted Intervention, Munich, Germany, 2015, pp. 234–241

[19] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., *et al.*: '3D U-Net: learning dense volumetric segmentation from sparse annotation'. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 2016, pp. 424–432

[20] Öztürk, Ş., Akdemir, B.: 'HIC-net: A deep convolutional neural network model for classification of histopathological breast images', *Comput. Electr. Eng.*, 2019, **76**, pp. 299–310

[21] Zhao, X., Liu, L., Qi, S., *et al.*: 'Agile convolutional neural network for pulmonary nodule classification using CT images', *Int. J. Comput. Assist. Radiol. Surg.*, 2018, **13**, (4), pp. 585–595

[22] Setio, A.A.A., Ciompi, F., Litjens, G., *et al.*: 'Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks', *IEEE Trans. Med. Imaging*, 2016, **35**, (5), pp. 1160–1169

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

[24] Dai, Y., Yan, S., Zheng, B., *et al.*: 'Incorporating automatically learned pulmonary nodule attributes into a convolutional neural network to improve accuracy of benign-malignant nodule classification', *Phys. Med. Biol.*, 2018, **63**, Article number: 245004

[25] Zhao, W., Yang, J., Sun, Y., *et al.*: '3D deep learning from CT scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas', *Cancer Res.*, 2018, **78**, (24), pp. 6881–6889

[26] Zhou, P., Ni, B., Geng, C., *et al.*: 'Scale-transferrable object detection'. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Salt Lake City, Utah, 2018, pp. 528–537

[27] Shi, W., Caballero, J., Huszar, F., *et al.*: 'Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network'. IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016

[28] Kass, M., Witkin, A., Terzopoulos, D.: 'Snakes: active contour models', *Int. J. Comput. Vis.*, 1988, **1**, (4), pp. 321–331

[29] Zeiler, M.D., Fergus, R.: 'Visualizing and understanding convolutional networks'. European Conf. on Computer Vision, Zurich, Switzerland, 2014, pp. 818–833

[30] Osendorfer, C., Soyer, H., Smagt, P.V.D.: 'Image super-resolution with fast approximate convolutional sparse coding'. Int. Conf. on Neural Information Processing, Montreal, Canada, 2014, pp. 250–257

[31] Otsu, N.: 'A threshold selection method from gray-level histograms', *IEEE Trans. Syst. Man Cybernet.*, 2007, **9**, (1), pp. 62–66

[32] Hojjatoleslami, S.A., Kittler, J.: 'Region growing: a new approach', *IEEE Trans. Image Process.*, 1998, **7**, (7), pp. 1079–1084

[33] Jain, A.K, Dubes, R.C.: 'Algorithms for clustering data', *Technometrics.*, 1988, **32**, (2), pp. 227–229
[34] Lafferty, J., Mccallum, A., Pereira, F.: 'Conditional random fields: probabilistic models for segmenting and labeling sequence data', *Proc. ICML*, 2001, **3**, pp. 282–289
[35] Wang, P., Chen, P., Yuan, Y.*, et al.*: 'Understanding convolution for semantic segmentation'. IEEE Winter Conf. on Applications of Computer Vision, Lake Tahoe, NV, USA, 2018
[36] Shore, J., Johnson, R.: 'Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy', *IEEE Trans. Inf. Theory*, 1980, **26**, (1), pp. 26–37
[37] Yan, X., Pang, J., Qi, H.*, et al.*: 'Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: a comparison between 2D and 3D strategies'. Asian Conf. on Computer Vision, Amsterdam, The Netherlands, 2016
[38] Huang, G., Liu, Z., Maaten, L.*, et al.*: 'Densely connected convolutional networks'. IEEE Conf. on Computer Vision and Pattern Recognition, Hawaii, HI, USA, 2017
[39] Gu, W., Zhu, Y., Chen, X.*, et al.*: 'Hierarchical CNN based real-time fatigue detection system by visual-based technologies using multi-scale pooling model', *IET Image Process.*, 2018, **12**, (12), pp. 2319–2329
[40] Jacobs, C., Van Rikxoort, E.M., Twellmann, T.*, et al.*: 'Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images', *Med. Image Anal.*, 2014, **18**, (2), pp. 374–384
[41] Li, X.X., Li, B., Tian, L.F.*, et al.*: 'Automatic benign and malignant classification of pulmonary nodules in thoracic computed tomography based on RF algorithm', *IET Image Process.*, 2018, **12**, (7), pp. 1253–1264
[42] Liu, Y., Hao, P., Zhang, P.*, et al.*: 'Dense convolutional binary-tree networks for lung nodule classification', *IEEE. Access.*, 2018, **6**, pp. 49080–49088
[43] Rios Velazquez, E., Parmar, C., Liu, Y.*, et al.*: 'Somatic mutations drive distinct imaging phenotypes in lung cancer', *Cancer Res.*, 2017, **77**, (14), pp. 3922–3930
[44] Clay, R., Kipp, B.R., Jenkins, S.*, et al.*: 'Computer-aided nodule assessment and risk yield (CANARY) may facilitate non-invasive prediction of EGFR mutation status in lung adenocarcinomas', *Sci. Rep.*, 2017, **7**, (1), p. 17620
[45] Oikonomou, A., Khalvati, F., Tyrrell, P.N.*, et al.*: 'Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy', *Sci. Rep.*, 2018, **8**, (1), p. 4003