TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 06/11 pp368–383
DOI: 10.26599/TST.2019.9010010
Volume 25, Number 3, June 2020

An Inception Module CNN Classifiers Fusion Method on Pulmonary Nodule Diagnosis by Signs

Guangyuan Zheng*, Guanghui Han, and Nouman Qadeer Soomro

Abstract: A “sign” on a lung CT image refers to a radiologic finding that suggests a pathological progression of some specific disease. Analysis of CT signs is helpful to understand the pathological origin of the lesion. In-depth study of lung nodules classification with different CT signs will help to distinguish benign and malignant nodules more clearly and accurately. To this end, we propose an Inception module-based ensemble classification method for pulmonary nodule diagnosis with different nodule signs. We first construct a Convolutional Neural Network (CNN) classifier adopting Inception modules and pre-train it on ImageNet. We then fine-tune this pre-trained classifier on 10 different lung nodule sign sample sets, and fuse these 10 classifiers with an artificial immune ensemble algorithm. The overall sensitivity, specificity, and accuracy of our proposed Artificial Immune Algorithm-based Inception Networks Fusion (AIA-INF) algorithm are 82.22%, 93.17%, and 88.67%, respectively, which are significantly higher than those of the alternative Bagging and Boosting methods. The experimental results show that our Inception-based ensemble classifier offers promising performance, and compared with other CADx systems, this scheme can offer a more detailed reference for diagnosis, and can be valuable for junior radiologist training.

Key words: sign; lung cancer; pulmonary nodule; Convolutional Neural Network (CNN); Artificial Immune Algorithm (AIA)

1 Introduction

Lung cancer is among the most fatal of diseases. Early diagnosis is considered to be crucial for better treatment and curability[1]. Significant progress has been made in screening breast, cervical, and prostate cancer[2], however, much progress still needs be made in lung cancer screening[3]. When symptoms of lung cancer appear, due to the peculiarities of the pulmonary anatomy, it has been growing for a considerable period and has potentially spreaded outside the lung. Therefore, at the initial presentation, most patients are already at an advanced stage of the disease. Out of every 100 newly diagnosed patients with lung cancer, 80 patients will be inoperable due to the progression of the cancer to an advanced stage, and only 20 patients can proceed for resection[4]. Most lung cancer manifests as nodules in the early stages[5]. The 10 years survival rate can be as high as 88% if lung cancer is diagnosed at Stage I and resected[6]. On the other hand, many benign diseases, such as hamartoma and carcinoid tumors, can also appear as nodules in CT slices[7], which results in varied interpretations among radiologists, and makes it difficult to decide whether a nodule is benign or

• Guangyuan Zheng is with the Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China, and also with the School of Information Technology, Shangqiu Normal University, Shangqiu 476000, China. E-mail: zhengguangyuan@bit.edu.cn.

• Guanghui Han is with the School of Biomedical Engineering, Sun Yat-sen University, Guangzhou 510006, China. E-mail: hanguanghui@outlook.com.

• Nouman Qadeer Soomro is with the Department of Software Engineering, Mehran University of Engineering and Technology, SZAB Campus, Khairpur Mir’s, 66020, Pakistan. E-mail: noumansoomro@gmail.com.

∗ To whom correspondence should be addressed.

Manuscript received: 2018-11-08; revised: 2019-03-11; accepted: 2019-03-12

malignant[8, 9]. Misinterpretation leads to unnecessary physical and mental pain, and extra financial burden for patients. The results reported in Refs. [10, 11] showed that interpretation errors could be as high as 42% when diagnosing pulmonary nodules on CT.

A "sign" on a lung CT image refers to a radiologic finding that suggests a pathological progression of some specific disease. Clinicians often infer the possible pathological processes from the sign of a lung focal[12]. Sign recognition from CT scans is a critical stage for the early diagnosis of lung cancer[13]. Some signs of nodules are known to be highly associated with lung cancer, such as certain internal structures[14, 15], spiculation, lobulation, and some texture[12, 14]. Radiologists usually assess the malignancy likelihood of a nodule by reference to these characteristics[10]. Wood et al.[15] found that whether a nodule on a CT scan was identified as suspected lung cancer was mainly due to its type and size (e.g., solid, part-solid, and ground-glass texture, etc). Today, Computer Aided Diagnosis (CAD) has become more important for both clinical diagnosis and for the training of junior doctors[16]. Therefore, CADx for sign recognition is deserving of further study.

Most existing studies focus on the detection of pulmonary nodules, and this has resulted in significant boosts in pulmonary nodule detection. The most commonly used detection approaches in literature are LDA[17], k-Nearest Neighbor (k-NN)[18–21], SVM[22–28], ANN[29–32], and Decision Tree (DT)[33]. Ensemble classification combines several weak classifiers into one strong classifier to improve the classification accuracy[34]. In the field of CADx, there is limited work on the fusion of multiple classifiers. In Refs. [35–37], the authors combined multiple traditional classifiers together to improve detection performance. Compared with the traditional classifier, Convolutional Neural Network (CNN) network can automatically extract the features of Regin Of Interest (ROI) images, achieve an end-to-end operation, and avoid information loss caused by intermediate processing steps. In this paper, we propose a deep CNN ensemble scheme for pulmonary nodule sign recognition. In our ensemble scheme, 10 classifiers are trained by transfer learning on 10 uncrossed sample sets selected randomly from a dataset. Individual networks are integrated using the artificial immune method to come up with a final classification. A visual depiction of the fusion scheme is given in Fig. 1. We name our approach as Artificial Immune Algorithm-based Inception Networks Fusion (AIA-INF) algorithm. We apply AIA-INF to a public pulmonary nodule dataset, that is Lung Image Database Consortium (LIDC), and compare this method to the ensemble approaches of Bagging and Boosting.

The main contributions of our work are as follows:

- We introduced Inception modules into AIA-INF which resulted in a higher recognition accuracy;
- Utilizing an Artificial Immune Algorithm (AIA) integration optimization method, we produced more superior and robust system performance; and
- We conducted a novel further exploration of malignant nodule CADx.

The rest of this paper is organized as follows. Section 2 reviews the related work on fusion classification in medical imaging field, Section 3 presents our ensemble classification method, and Section 4 describes our experimental scheme. We evaluate the efficiency of the method in Section 5 and discuss the results in Section 6. Section 7 concludes.

2 Related Work

With the development of computer technology and ralated research efforts, many excellent CADx schemes have emerged. Hussein et al.[38] proposed a 3D CNN-based nodule characterization strategy. They utilized the volumetric information from a CT scan to address the need for a large amount of training data for CNN and resorted to transfer learning to

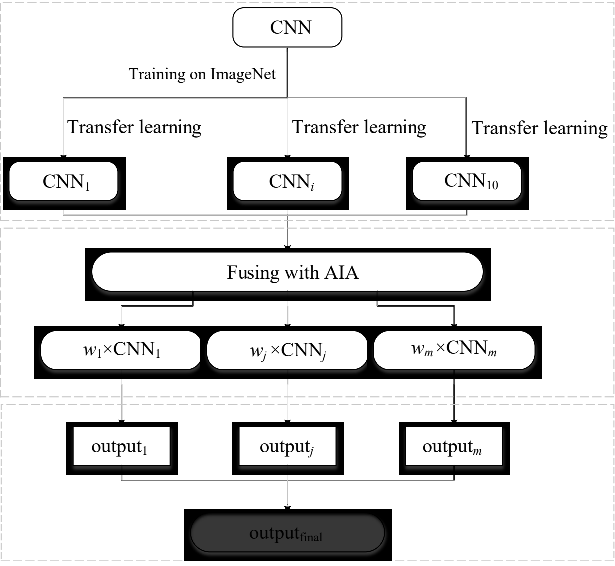


Fig. 1 Principle diagram and flow chart of AIA-INF.

obtain highly discriminative features. They also acquired the task dependent feature representation for six high-level nodule attributes and fused this complementary information via a Multi-Task Learning (MTL) framework. Finally, they proposed the incorporation of potential disagreement among radiologists while scoring different nodule attributes in graph regularized sparse multi-task learning. Shen et al.[39] proposed a method directly modeling raw nodule patches with an end-to-end machine-learning architecture for classifying lung nodule malignancy suspiciousness. They constructed a Multi-Crop CNN (MC-CNN) to automatically extract salient nodule information by employing a novel multi-crop pooling strategy which crops different regions from convolutional feature maps and then applied max-pooling at different times. Experimental results showed that the proposed method could not only fulfil nodule suspiciousness classification, but also could effectively characterize nodule semantic attributes (subtlety and margin) and nodule diameter. Wei et al.[40] presented a novel unsupervised spectral clustering algorithm to distinguish between benign and malignant nodules. In this algorithm, a new Laplacian matrix was constructed by using Local Kernel Regression Models (LKRM) and incorporating a regularization term. In this way, their algorithm can not only achieve a higher classification performance, but also tackle the out-of-sample problem. In Ref. [41], three types of deep neural networks (CNN, DNN, and SAE) were applied to lung cancer classification. Those networks were applied to the CT image classification task with some modifications for benign and malignant lung nodules and were evaluated on the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) database. The experimental results showed that the CNN network achieved the best performance with an accuracy of 84.15%, sensitivity of 83.96%, and specificity of 84.32%, which was the best result among the three networks used for comparison. Multi-View CNNs (MV-CNN) were proposed in Ref. [42] for lung nodule classification. Liu and Kang[42] carried out a binary classification (benign and malignant) and a ternary classification (benign, primary malignant, and metastatic malignant) on the LIDC-IDRI database using MV-CNN. The results revealed that the deep features learned by their proposed model had a higher separability than features from the image space. Most of these existing works focus on binary-classification of lung nodules. Further study of different nodule signs analysis is needed to understand the pathological origin of the lesion, and more accurately distinguish different types of nodules from the pathological perspective.

Ensemble classification is one of the most important method to improve the efficiency of nodule recognition, because it is able to integrate multiple classifiers into a unified classifier, and thereby provide more accurate, stable, and robust classification results. There are a variety of commonly used integration algorithms, among which Bagging and Boosting are two of the most popular. Tartar et al.[43] compared single SVM with Bagging SVMs. Results showed that in Total Classification Accuracy (TCA) and False Positive Ratio (FPR) metrics, Bagging SVM was significantly superior to single SVM. In another work, Tartar and Akan[44] proposed a classification approach for pulmonary nodule detection from CT imagery using morphological features of nodule patterns. The ensemble learning approaches, Bagging, Boosting, and Subspace were used for the classification process. The proposed system with random forest-based ensemble learning approaches resulted in the highest performance. Jaffar et al.[45] used ensemble Bagging trees to classify lung nodules, employing a multi-coordinate histogram of gradient and an intensity-based features descriptor. They verified the CADx system on the lung image consortium database, and attained superior results comparing with other existing CADx systems. Xie et al.[46] proposed a Transferable Multi-Model Ensemble (TMME) algorithm to separate malignant from benign lung nodules using limited chest CT data. This algorithm transferred the image representation abilities of three ResNet-50 models, pre-trained on the ImageNet database. Experimental results on the LIDC-IDRI dataset showed TMME's classification accuracy was markedly higher than seven state-of-the-art approaches. Farahani et al.[47] presented a computer-aided classification method using computed tomography lung images, based on an ensemble of three classifiers (MLP, KNN, and SVM). In their system, morphological features are used for the classification process in such a way that each classifier makes its own decision, and a majority-voting method is used to combine the decisions of the ensemble system. Results show a good level of improvement in the diagnosis of pulmonary nodules. Ma et al.[36] proposed a new weighted-sum method of multiple classifier fusion to recognize the common CT Imaging Signs of Lung (CISL) diseases. The classifiers are fused through a weighted method combining five types of classifiers for CISL recognition, namely SVM, Back-Propagation Neural Network (BPNN), Naïve Bayes (NB), *k*-NN, and DT. Through a comparison with each single classifier of its own, and two well-known methods of classifier fusion (Bagging and Boosting), the method was shown to be effective and promising. In the process of integrating multiple classifiers, the type and number of the classifiers involved are very important factors. For a specific classification problem, a particular ensemble scheme is used to combine a certain number of classifiers together. In order to attain the best overall classification performance, the number of the classifiers participating in the integration and their organizing mode need to be determined with an optimization strategy. Existing methods have used experiments or experience to make this determaination, but this method of combination is suitable only for specific training datasets. The performance of the ensemble body may not be optimal, and its generalized performance and stability may also fall short.

3 Method

3.1 AIA-INF algorithm

Using a classifier ensemble scheme can improve the overall classification performance[48]. Two critical factors determining the final accuracy of the integrated classifiers are: (a) the average generalization error of each individual classifier, and (b) the diversity of the individual classifiers in the ensemble.

Aiming to achieve superior performance from the fewest individual components, Zhou et al.[49] proposed a selective ensemble method named as Genetic Algorithm-based Selective ENsemble (GASEN). GASEN first trains a number of neural networks and assigns them random weights, then employs a genetic algorithm to evolve the weightings. In the process of ensemble generation, GASEN can find the optimal weight of the best matching degree between sub-networks by genetic crossover and mutation operations. Finally, according to the evolved weights, neural networks with high weights are selected to make up the final integrated network.

When selecting sub-networks, GASEN does not take the differences between individual networks into account. Moreover, genetic algorithm in Ref. [49] is liable to fall into local optima. The AIA-INF ensemble algorithm we proposed in this paper considers both the average generalization error and the diversity of component networks and, to the greatest extent possible, chooses classifiers with high output heterogeneity.

To apply AIA-INF, firstly we pre-train an Inception network on a natural image database and conduct transfer learning on 10 different sign sample sub-sets as the original individual classifiers. Then, we adopt the AIA method to integrate the sub-networks together while enhancing the difference between individual networks, thereby improving the classification accuracy of the ensemble network. This method actively chooses individual networks with a high degree of difference and high individual classification accuracies into an ensemble through immune cloning and their weights mutation.

Corresponding to the idea of AIA, in this paper, we look at each sub-network as an antibody, and the label value of the real category as an antigen. The principle of our fusion Inception network algorithm is illustrated in Fig. 2, which shows the step-by-step fusion procedure of the AIA-INF.

In Fig. 2, each rectangle stacked structure represents an individual sub-network. $\boldsymbol{O}$ represents the output of an individual, and is looked as an antibody; af is the affinity between values of the predicted category and the real label of an input sample. The lines between each two sub-network entities represent the similarity of the two predicted outputs, denoted as sm. For a sub-network net_i in the ensemble, when it is removed from the ensemble, the mean similarity between the remained sub-networks is called the residual average similarity $\overline{\text{sm}}$ of net_i, if the $\overline{\text{sm}}$ of net_i is lower than that of the other sub-networks, and the predictive affinity af of net_i is the lowest, then $\overline{\text{sm}} \times \text{af}$ of net_i must be the smallest in the ensemble. If this net_i is removed from the ensemble in the process of classifier fusion, it will be also beneficial to improve the overall classification accuracy and the diversity of individual classifiers in the ensemble. According to this idea, in the fusion process, if the value of $\overline{\text{sm}} \times \text{af}$ of a subnetwork is less than a certain threshold θ, the subnetwork will be excluded from the ensemble classifier.

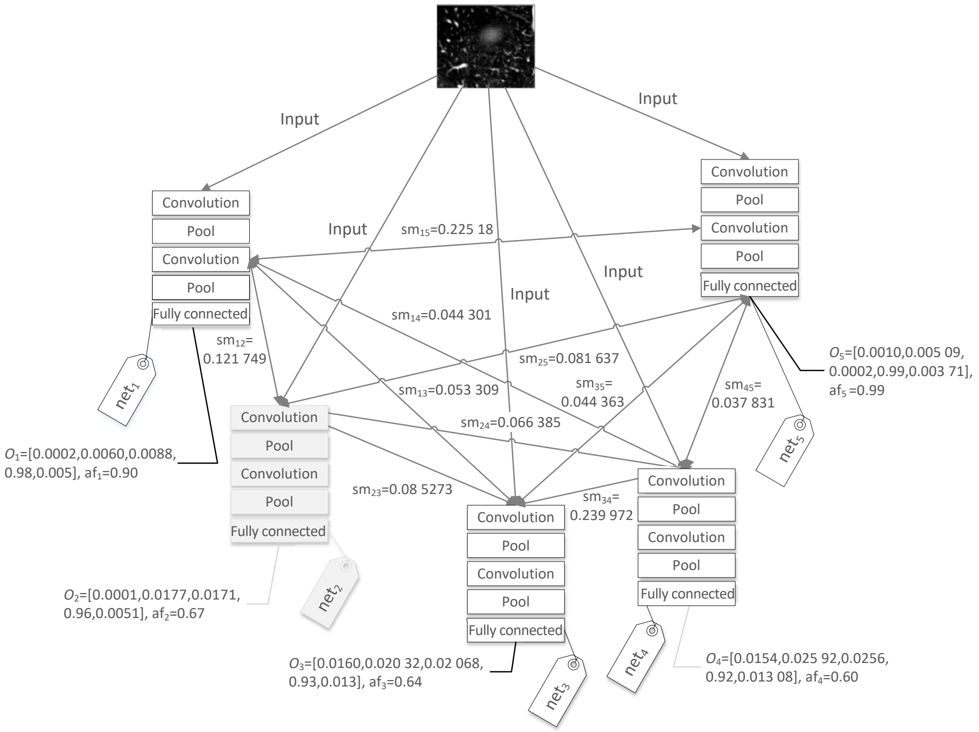


Fig. 2 Fusion scheme of Inception networks.

For an input ROI image patch, each individual classifier in an ensemble predicts a normalized vector $\boldsymbol{O}$, in which each element corresponds to a class. The category corresponding to the highest element in the vector will be regarded as the predicted category of the classifier. For example, the highest value of the elements in $O_1([0.0002, 0.0060, 0.0088, 0.98, 0.005])$, is the 4th element (0.98), so classifier net_1 places the input ROI patch into Category "3". In Fig. 2, the residual average similarities of net_1 to net_5 are 0.188 24, 0.175 972, 0.195 633 3, 0.223 353, and 0.216 218, respectively, their predicted affinity is 0.90, 0.67, 0.64, 0.60, and 0.99. The product of their residual average similarities and predicted affinity $\text{sm} \times \text{AF}$ is 0.169 942, 0.117 901, 0.125 205, 0.134 012, and 0.214 055, respectively. Obviously, the product of $\text{sm} \times \text{AF}$ of net_2 is the smallest, and net_2 will be excluded from the ensemble.

We use cosine similarity to measure the affinity between the classifier prediction and the sample label. Assuming that at a given time, the number of the sub-networks in an ensemble is *N*, and we want to learn an Inception ensemble network to approximate function *f*: $R^m \rightarrow \mathbf{C}$, where *m* refers to the number of the samples to be classified, and $\mathbf{C}$ means the set of categories used in the classification task. Let "$\boldsymbol{f}_i$" represent the prediction of the *i*-th classifier, such that $\mathbf{F}_i = [f_{i1}, \boldsymbol{f}_{i2}, \ldots, \boldsymbol{f}_{ij}, \ldots, \boldsymbol{f}_{im}]$, where $\boldsymbol{f}_{ij}$ means the prediction of the *i*-th classifier for the *j*-th example. Let $\boldsymbol{c}_j$ present the real label of the j-th example. For an individual network *i*, the similarity between $\boldsymbol{f}_{ij}$ and $\boldsymbol{c}_j$ can be attained by

$$\text{cos\_dist}_{ij}(\boldsymbol{f}_{ij}, \boldsymbol{c}_j) = \frac{\sum\limits_{k=0}^{9}(\boldsymbol{e}_{ik}\boldsymbol{e}_{jk})}{\sqrt{\sum\limits_{k=0}^{9}(\boldsymbol{e}_{ik})^2} + \sqrt{\sum\limits_{k=0}^{9}(c_{jk})^2}} \tag{1}$$

where $\boldsymbol{e}_{ik}$ is the *k*-th element of classifier f_i's prediction vector, and c_{jk} is $\boldsymbol{c}_j$'s *k*-th element, with *k* in the range of 0 to 9.

The number of involved sub-networks will change after each evolution. So, we use a normalized reciprocal Euclidean to measure the similarity between two subnet outputs,

$$\text{sm}_{ij} = \frac{1/\,\text{Euclidean\_dist}(\text{net}_i, \text{net}_j)}{\sum\limits_{k}^{m}\sum\limits_{t}^{m}(1/\text{Euclidean\_dist}(\text{net}_k, \text{net}_t))} \tag{2}$$

where $i, j, k, t = 1, \ldots, m;\ i \neq j;\ k \neq t$.

Then the residual average similarities of net_i is:

$$\overline{\text{sm}_i} = \frac{\sum\limits_{1}^{N}\text{sm}_n}{N-1}\quad (n \neq i) \tag{3}$$

According to the immune algorithm theory of the survival of the fittest, and suppressing the highdensity antibodies, the main goals of our AIA-INF algorithm are to: (a) clone and mutate individual networks with higher output affinity, and (b) moderately exclude individual networks with lower affinity and high similarity between. For the same input, if no two individual networks are able to produce the same correct prediction, we at least want to ensure they are not providing the same wrong output at least. In this way, the diversity of the networks is enhanced.

The affinity between the prediction value (antibody) and the sample label (antigen) — "AF", which indicates the matching degree between the output and the class label. Assuming there are n samples to be tested, the expected output is $\text{EX} = [\text{ex}_{11}, \ldots, \text{ex}_{ij}, \ldots, \text{ex}_{\text{nm}}]^{\text{T}}$. Here, ex_{ij} means the expected i-th output of the j-th sub-network. Then

$$\text{AF} = [\text{af}_{11}, \ldots, \text{af}_{ij}, \ldots, \text{af}_{nm}] = \text{cos\_dist}(\text{EX}, f_i) \tag{4}$$

where af_{ij} represents the affinity metric of the j-th individual network on the i-th sample. The individual networks whose affinity is lower than the threshold are discarded, and individual networks having a value higher than the threshold will remain for cloning and mutation. The mutation rules is expressed as

$$M' = M + \alpha N(0,1)\text{e}^{-\overline{\text{sm}}_i \alpha f_{ij}} \tag{5}$$

where M' means the weight vector after the mutation of M, and α is a parameter used to control the mutation rate; by changing α, we can tune the mutation latitude manually. $N(0,\ 1)$ is a Gaussian random function, the goal of which is to produce many near-0 variations. αf_{ij} is a [0, 1] number, measuring the distance between the output of the ensemble network and the label of the input; the higher the affinity, the smaller the variation, which makes convergence faster. The network integration algorithm of networks is summarized in Algorithm 1.

Algorithm 1 Fusing algorithm of AIA-INF

For an Inception network framework:

(1) Train an original Inception classifier on the ImageNet database.

(2) Randomly select 10 non-intersecting sub datasets from the nodule sign sample database. Then fine-tune this pre-trained network on 10 sign subsets separately to gain 10 different sub-networks.

(3) Use these 10 sub-networks as initial antibody population. Then, select randomly another non-used training dataset from the original database. For each sample image patch, the 10 sub-networks output 10 different prediction vectors. Assign a random weight w_k to each output; here $0 \leqslant w_k \leqslant 1$, $\sum_{k=0}^{N} w_k = 1$, $N = 10$.

(4) Compute the affinity af_{ij} for the prediction vector of each sub-network and their corresponding actual label, and the residual average similarity matrix $\overline{\text{sm}}_i$ of sub-networks in the ensemble.

(5) If the affinity between the output of the sub-network and the label is higher than the threshold then stop, otherwise:

(a) Clone the individual sub-network according to $\overline{\text{sm}}_i \times \text{af}_{ij}$.

(b) Generate new weight values by the variant formula given in Eq. (5).

(c) Compute the affinity of the new weighted individual network.

(d) Select all individual networks whose $\overline{\text{sm}}_i \times \text{af}_{ij}$ value is higher than the threshold, and add them to the ensemble.

(e) Compute the similarity of every two individual networks in new ensemble and compute new residual average similarity $\overline{\text{sm}}_i$.

(f) Eliminate the sub-networks whose $\overline{\text{sm}}_i \times \text{af}_{ij}$ is lower and similarity is higher than the threshold.

(g) Compute AF of the new networks.

(6) Repeat the above process until AF meets the requirement or the iteration count reaches the preset maximum value.

3.2 Individual classifier design

Figure 3 shows the length distribution of nodule diameters in LIDC, which mostly sits between 2 mm and 33 mm. The average value of CT image pixel intervals is 0.6761 mm; converting to pixels, the diameters are between approximately 3 pixels and 49 pixels. To arrive at a robust ensemble classifier that can handle ROIs of all sizes, we construct a deep learning framework with an input size of 32×32 pixels.

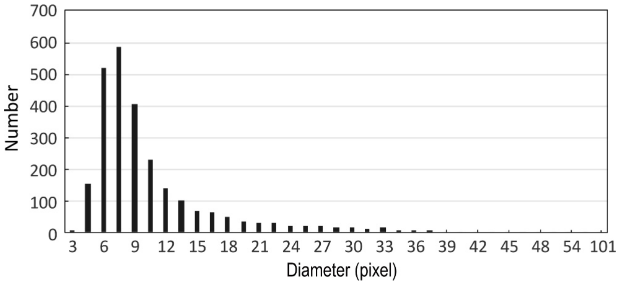


Fig. 3 Distribution of nodule diameters in LIDC.

CNN is inspired by biological mechanism and designed to emulate the behavior of visual systems. CNN can automatically learn representations of input data by multiple levels of feature extraction[44]. In the ImageNet recognition challenge, the use of CNN dramatically improved the results[50–52]. ImageNet[53] is a database launched by L. Fei-Fei for object recognition in the field of computer vision. ImageNet is a very comprehensive database organized according to the WordNet hierarchy (currently only the nouns), including more than 1.2 million categorized natural images in 1000+ classes. In 2012, AlexNet's success in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) started the revival of CNN. Since then, GoogLeNet[54], VGG[55], ResNet[56], Inception[54], and Inception-ResNet-v2[57] moved into the spotlight each in turn. Inception-ResNet-v2 obtained the highest recognition accuracy on ImageNet among all competitors in 2016, which led to the adoption of Inception modules in net framework design. From LeNet to Inception-ResNet-v2, CNN increased in depth and its convolutional ability became more powerful, with Inception-ResNet-v2 obtaining a 96.92% recognition accuracy. In the process of network design and construction, GoogLeNet-ResNet-v2 employed many Inception modules. An Inception module contains convolution kernels of different scale in one layer, which greatly improves the feature extraction ability of the net, and avoids the loss of essential feature information because of repeated convolution and down sampling. 1×1, $1\times n$, and $n\times 1$ convolution are applied in the Inception modules, which accelerates the network training, and computing efficiency is improved. GoogLeNet-ResNet-v2 adds a residual connection scheme, which further enhances the convergence speed and improves the accuracy of the network. For these reasons, we adopted this idea in designing our network framework, and employed three Inception modular techniques, which are shown in Fig. 4.

The architecture of a sub-CNN network is shown in Fig. 5.

4 Experiment

4.1 Datasets

We selected 4 categories of signs from LIDC-IDRI, a public reference database of lung nodules on CT scans established by the National Cancer Institute

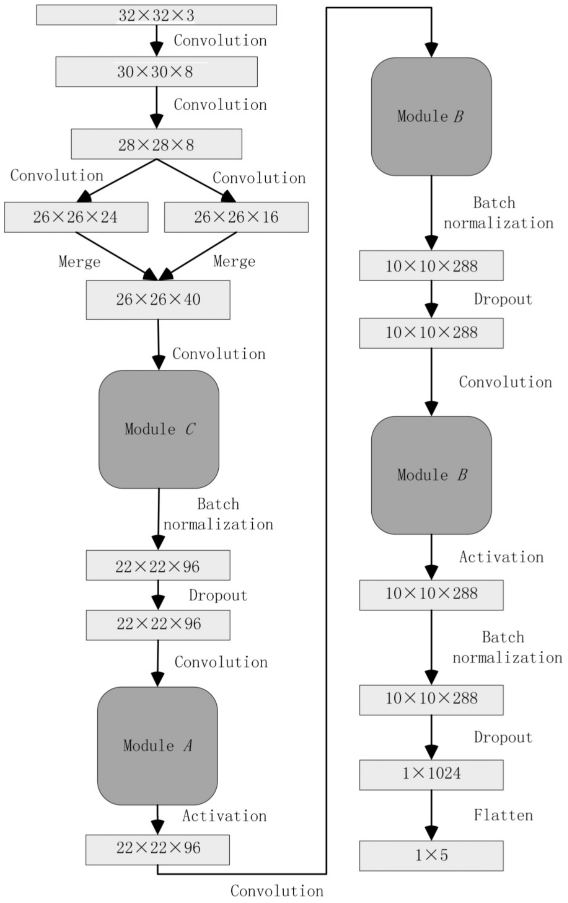
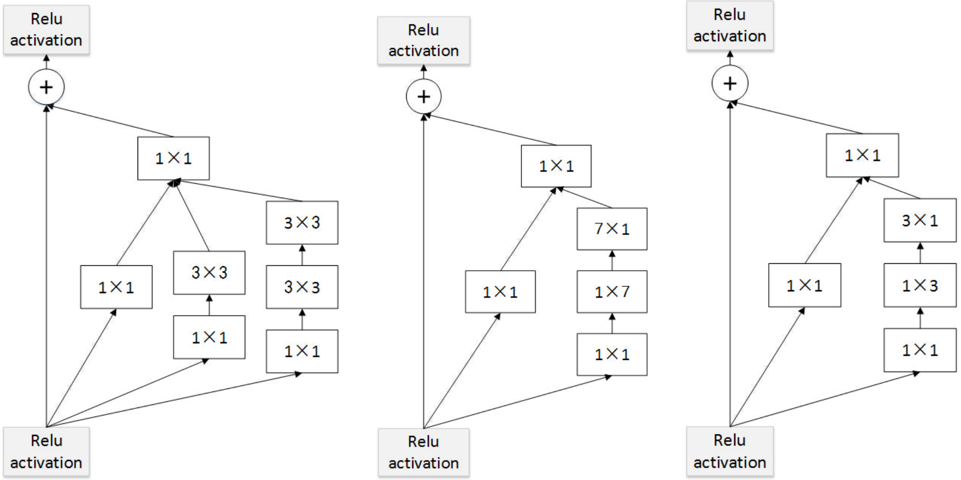


Fig. 5 Architecture of the sub-CNN network.



(a) Module *A* (b) Module *B* (c) Module *C*

Fig. 4 Inception modules used in CNN framework.

(NCI) in collaboration with the University of Chicago, Cornell University, Iowa State University, University of Michigan, and University of California. This database is an essential reference for CADx research and experiments.

The LIDC-IDRI[16] database contains 1018 cases, and 200 000 CT slices. The thickness of these slices is between 1.25 mm and 3 mm, and the length and width of each image is 512 pixels. Each case includes images from a clinical thoracic CT scan, and an associated XML file that records the results of a two-phase image annotation process performed by four experienced thoracic radiologists. In the initial blinded-read phase, each radiologist independently reviewed each CT scan and marked lesions belonging to one of three categories (“nodule$\geqslant$3 mm”, “nodule$\leqslant$3 mm”, and “non-nodule$\geqslant$3 mm”). In the subsequent unblinded-read phase, each radiologist independently reviewed their own marks along with the anonymized marks of the other three radiologists to produce a final opinion. The database contains 7371 3D lesions marked as nodules by at least one radiologist; 2669 lesions have been marked as nodule $\geqslant$3 mm by at least one radiologist, of which 928 (34.7%) have been marked as such by all four radiologists. These 2669 lesions include nodule outlines and subjective nodule characteristic ratings[58]. The characteristics of the nodules are presented in Table 1.

4.2 Selection of signs

From experience[59], the greater the degree of lobulation, spiculation, or non-solidity of texture signs, the greater the possibility the nodules are malignant; on the other hand, signs of calcification mostly indicate a benign nodule, except for cases of non-central appearance. There is no certain evidence showing that signs of subtlety[60], internal structure[61], sphericity, and margin[14] have strong

Table 1 LIDC nodule characteristics with corresponding notes and possible scales.

Feature	Radiologist’s description	Scale
Subtlety	Subtlety of nodule	1 Extremely subtle 5 Obvious
Internal structure	Internal structure score of nodule	1 Extremely subtle 2 Fluid 3 Fat 4 Air
Calcification	Internal calcification of nodule	1 Popcorn appearance 2 Laminated appearance 3 Solid appearance 4 Non-central appearance 5 Central calcification 6 Absent
Sphericity	Shape in terms of nodule’s roundness/sphericity, three terms defined	1 Linear appearance 3 Ovoid appearance 5 Round appearance
Margin	Margin of nodule, with only the extreme values explicitly defined	1 Poorly defined 5 Sharp margin
Lobulation	Nodule lobulation, with only the extreme values explicitly defined	1 No lobulation 5 Marked lobulation
Spiculation	Nodule spiculation, with only the extreme values explicitly defined	1 No spiculation 5 Marked spiculation
Texture	Nodule internal texture, with only three terms defined	1 Non-solid/GGO 3 Part solid/mixed 5 Solid texture
Malignancy	Malignancy level of this nodule (assuming 60-year-old male smoker)	1 Highly unlikely for cancer 2 Moderately unlikely for cancer 3 Indeterminate likelihood 4 Moderately suspicious for cancer 5 Highly suspicious for cancer

relationship with the malignancy of a nodule. In order to gain a significant comparative result, we chose nodules with Calcification$=4$, lobulation$\geqslant 4$, spiculation$\geqslant 4$, texture$\leqslant 2$, and malignancy$\geqslant 3$, to serve as experimental samples, according to the scheme shown in Table 2.

We performed a statistical analysis of 21 057 ROI regions with characteristic descriptions in XML files. A histogram of the selected signs distribution is shown as Fig. 6.

We selected the 4 categories of signs which are highly prevalent in malignant nodules as experimental objects—non-central calcification, lobulation, spiculation, and non-solid/Ground-Glass Opacity (GGO) texture. Considering the normal patches as a negative class, this results in a total of 5 classes. Samples of the 4 categories positive signs are shown in Table 3.

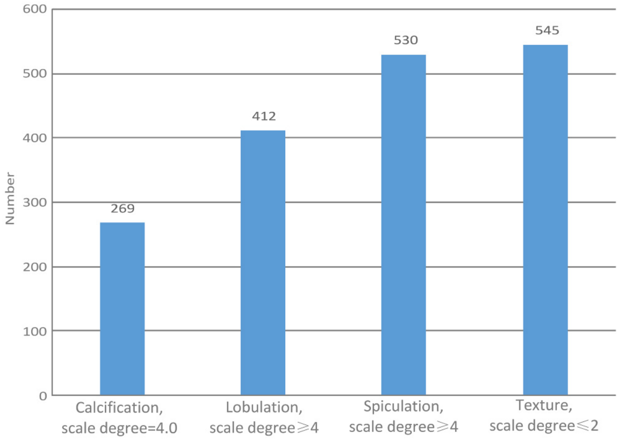


Fig. 6 Number of every kind of selected signs.

4.3 ROI segmentation

Based on the central point of the merged lesion regions outlined by 4 experts, we cut out image patches of 32×32 pixels as experimental images, and selected only positive patches as experimental objects, as shown in Table 3. From these samples, we randomly selected 10 non-intercross subsets for the pre-trained network to perform transfer learning. Moreover, in order to make the trained individual networks as specific as possible, we ensured that one category of ROI patches from a single of one same patient will not be included in any two training sets at the same time.

4.4 Taining of networks

CNN is a data-driven learning algorithm, and training a CNN requires a huge number of samples. The number of samples in LIDC-IDRI and LISS[62] are limited and cannot meet the training requirement of the classifier. Therefore, we employed a transfer learning method to train individual networks.

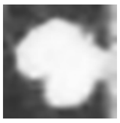
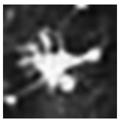
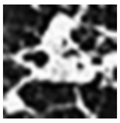
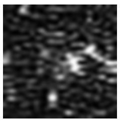
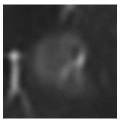
Firstly, we trained an Inception network on the ImageNet database[53]. For the pre-training network model, different from the framework shown in Fig. 5, we set the output dimension to 27, corresponding to the number of high synset categories in ImageNet, and named it Inception-PreTraining-net. The accuracy of Inception-PreTraining-net converged to 95.8%.

We then randomly selected 10 non-crossed subsets from LIDC-IDRI for transfer learning. Through rotation, zooming, and random clipping, the number of ROI patches was multiplied by 200. Based on the

Table 2 Scheme of signs selection.

Selected	Subtlety	Internal structure	Calcification	Sphericity	Margin	Lobulation	Spiculation	Texture	Malignancy
Yes	–	–	$=4$	–	–	$\geqslant 4$	$\geqslant 4$	$\leqslant 2$	$\geqslant 3$
No	–	–	$\neq 4$	–	–	<4	<4	>2	<3

Table 3 Sample of non-central calcification, lobulation, spiculation, and non-Solid/GGO texture signs.

	Non-central calcification		Lobulation		Spiculation		Non-solid/GGO texture	
								
Subtlety degree	4.67	5.0	5.0	4.0	4.0	5.0	3.0	3.67
Internal structure degree	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Calcification degree	**4.0**	**4.0**	6.0	6.0	6.0	6.0	6.0	6.0
Sphericity degree	3.33	2.67	4.25	4.0	5.0	5.0	4.0	4.33
Margin degree	3.67	4.67	4.5	4.0	2.0	4.0	1.0	2.67
Lobulation degree	1.67	2.0	**4.25**	**4.0**	4.0	1.0	1.0	2.67
Spiculation degree	1.33	1.33	2.25	4.0	**4.0**	**5.0**	2.0	1.67
Non-solid of texture degree	5.0	4.67	4.75	4.0	5.0	4.0	**1.0**	**1.33**

trained Inception-PreTraining-net from ImageNet, we conducted transfer learning based on the 10 augmented datasets, separately. We used network as shown in Fig. 5 to conduct transfer learning, naming this Inception-FineTuning-net. The only difference between this and Inception-PreTraining-net is that the output dimension of the full connection layer of the former is 5. We copied the pre-trained parameters of all the other layers prior to the fully connected layer in the trained Inception-PreTraining-net network to the Inception-FineTuning-net network model. We adopted an iterative process to fine-tune the last n-layer of the Inception-FineTuning-net. If the verification performance improved compared with the previous $n-1$ layer, then we continued to fine-tune the $n+1$ layer, and so on. Finally, we found that the network achieved the best classification performance when we fine-tuned the full connection layer of Inception-FineTuning-net and its previous B model units are fine-tuned. By fine-tuning in this way, 10 different sub-classifiers are generated.

In the next stage, we gave a random weight w to each classifier and integrated 10 classifiers together. At the beginning of this stage, the performance of the generated ensemble classifier is not optimal and the number of the sub-networks involved is also not the most appropriate. The optimal weight and number of component networks in the ensemble are difficult to determine. To address this problem, we introduced AIA into our scheme. Compared to two common optimization algorithms, the Simulated Annealing[63] and the Genetic Algorithm[64], AIA is faster and does not as easily to fall into local optima. We randomly selected another 1000 unused samples for sub-classifiers selection and integration. We then applied AIA, including cloning, mutation, superior network selection, and weighted fusion on the trained sub-networks. After the number of sub-networks was optimized and the optimal weights were determined, the ensemble was generated. The final classification result was then obtained by a weighted sum method, which is illustrated in Fig. 1.

We set α in the mutation Eq. (5) to 0.9, the max iteration number to 50, and the initial ensemble prediction stop affinity threshold value to 0.85, which will be gradually increased as the iteration number increases.

4.5 Signs recognition and classification

In the recognition stage, for a given input image patch, each neural network component of an ensemble gives a confidence vector "$\boldsymbol{vc}$" as the output of the sub-classifier. In the next classification stage, every $\boldsymbol{vc}$ multiplies a corresponding weight "w_i" generated in Section 4.4. We summed up all weighted vectors by category as the output of the ensemble. The highest element in the sum vector represents the corresponding class produced by the ensemble.

5 Result

In order to measure the performance of AIA-INF, three metrics were adopted: sensitivity & specificity, classification accuracy, and confusion matrix. Sensitivity is also named as True Positive Rate (TPR) or recall rate, and it represents the percentage of samples recognized as positive that are actually positive. Specificity is also called True Negative Rate (TNR), and it reflects the percentage of recognized negative samples that are actually negative. Let *"TP"* (True Positive) denote the positive samples that are correctly identified, *"TN"* (True Negative) denote the negative samples that are correctly identified, *"FP"* (False Positive) represent wrongly identified positive samples, and *"FN"* (False Negative) represent wrongly identified negative samples. The sensitivity and specificity can then be formulated as $TP/(TP+FN)$ and $TN/(TN+FP)$. Classification accuracy means the percentage of correctly recognized samples in terms of both sensitivity and specificity. A confusion matrix is a commonly used method to evaluate the performance of classification. Compared to Receive Operating Characteristic (ROC) curve and Area Under the Curve (AUC), a confusion matrix can also describe the misclassification rate between multiple classes. The row-wise elements in a confusion matrix correspond to the actual category of the object, and the column-wise elements correspond to the recognized category. Elements on the diagonal indicate the percentage rate of patterns that have been recognized correctly, whereas non-diagonal elements indicate the percentage of the corresponding row patterns that have been wrongly identified in a column. In the model recognition procedure, if two patterns are very similar, their samples can easily be misidentified with each other. Ideally, if the categories of all samples are correctly predicted, the confusion matrix will form a diagonal matrix.

To ensure the independence of the validation set, we sought to avoid introducing training data in a circular fashion. As such, we randomly selected 450 unused image patches of signs to conduct a 5-fold cross-validation experiment.

We made an overall performance comparison between the ensemble and its sub-classifiers. This ensemble was eventually made up of a fusion of 6 individual classifiers. The outcome of the performance comparison is shown in Table 4, in which "SE", "SP", and "ACC" represent "sensitivity", "specificity", and "accuracy", respectively.

As shown in Table 4, our proposed ensemble Inception network obtained significant improvement in performance compared to the 6 individual classifiers. In addition, the sensitivity and specificity also recorded a corresponding improvement. The increase in the rate of sensitivity is 18.35% to 22.61%, the minimum increase in the rate of specificity is 9.91% as compared with that of Sub-net$_2$, and 14.35% compared with Sub-net$_1$. The accuracy is improved by a maximum of 23.52% and a minimum of 14.44%.

Table 4 Performance comparison between sub-nets & ensemble.

(%)

Classifier	SE	SP	ACC
Sub-net$_1$	63.87	83.07	73.67
Sub-net$_2$	59.61	78.82	72.92
Sub-net$_3$	61.52	82.45	65.15
Sub-net$_4$	62.23	83.76	71.32
Sub-net$_5$	62.15	82.31	70.72
Sub-net$_6$	63.73	81.36	74.23
Net-ensemble	**82.22**	**93.17**	**88.67**

6 Discussion

Bagging and Boosting are two popular prevailing ensemble approaches. Bagging[34], also called bootstrap aggregating, is an integrated technique that trains sub-classifiers on new data sets selected from an original dataset. Samples for the new data sets are sampled with a replacement method, with reuse permitted. The trained classifier ensemble is used to classify new samples, and then the majority votes or average methods are used to count the results of all classifiers. The final prediction class is the one with the most votes. Boosting[65] is an iteration-based algorithm, in which a new weak classifier is added in each round until the ensemble's classification error rate is reduced to a preset small value. Each training sample is assigned a weight denoting the probability of the sample being selected into a classifier training set. If one sample is classified correctly, its probability of being selected will be decreased when constructing dataset for the next training round. By using this method, Boosting "focuses on" the samples that are difficult to correctly classify.

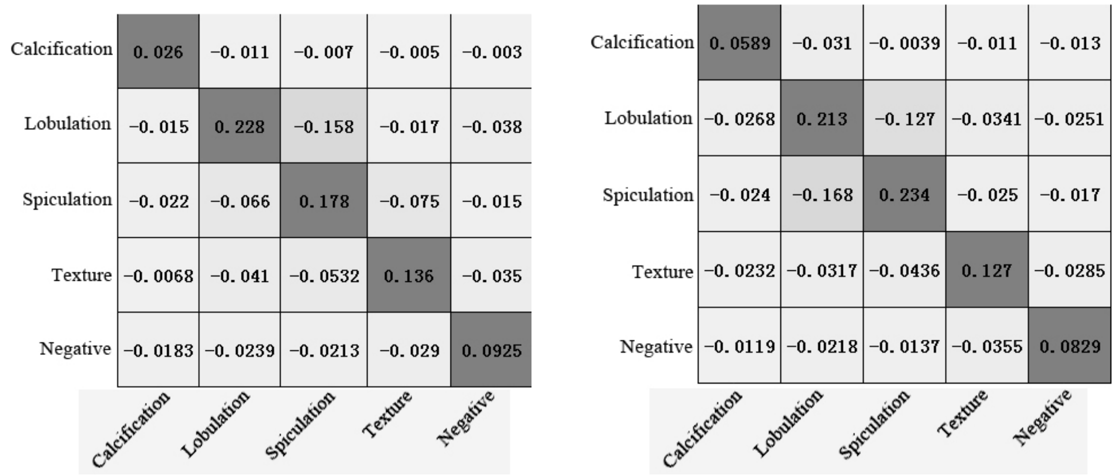
We compared the performance of our proposed ensemble classifier with Bagging and Boosting. We implemented our methods in a TensorFlow environment. We adopted 6 sub-networks as weak classifiers and conducted Bagging and Boosting classification integration experiments. The average recognition result of each sign is recorded, and is shown in Table 5. From the experimental results, it is clear that our algorithm offers superior performance to Bagging or Boosting.

For a more comprehensive analysis, we compared our ensemble Inception network, Bagging, and Boosting with a confusion matrix. Firstly, we computed the classification confusion matrix of the Inception ensemble network, Bagging, and Boosting, then performed subtraction operations on a confusion matrix of Inception ensemble network and Bagging, and a confusion matrix of Inception ensemble network and Boosting. Thus, we obtained two different confusion matrices, which are illustrated in Fig. 7. As can be seen from Fig. 7, all the values of diagonal elements are positive in the two confusion matrices, while all the non-diagonal values are negative. The average of diagonal elements are 0.1321 and 0.1431, respectively, while the sum of non-diagonal values are inversely −0.1321 and −0.1431. These results indicate that the

Table 5 Performance comparison between our method, Bagging, and Boosting.

(%)

Sign	Our method			Bagging			Boosting		
	SE	SP	ACC	SE	SP	ACC	SE	SP	ACC
Calcification	88.89	94.72	–	77.78	90.56	–	82.22	89.72	–
Lobulation	77.78	92.5	–	56.67	86.39	–	56.67	85.83	–
Spiculation	75.56	90.83	–	55.56	82.78	–	55.56	84.17	–
Texture	81.11	93.61	–	66.67	88.89	–	57.78	86.67	–
Negative	87.78	94.17	–	71.11	89.72	–	68.89	88.33	–
Average	**82.22**	**93.17**	**88.67**	65.56	87.67	75.46	64.22	86.94	74.36



(a)

	Calcification	Lobulation	Spiculation	Texture	Negative
Calcification	0.026	−0.011	−0.007	−0.005	−0.003
Lobulation	−0.015	0.228	−0.158	−0.017	−0.038
Spiculation	−0.022	−0.066	0.178	−0.075	−0.015
Texture	−0.0068	−0.041	−0.0532	0.136	−0.035
Negative	−0.0183	−0.0239	−0.0213	−0.029	0.0925

(b)

	Calcification	Lobulation	Spiculation	Texture	Negative
Calcification	0.0589	−0.031	−0.0039	−0.011	−0.013
Lobulation	−0.0268	0.213	−0.127	−0.0341	−0.0251
Spiculation	−0.024	−0.168	0.234	−0.025	−0.017
Texture	−0.0232	−0.0317	−0.0436	0.127	−0.0285
Negative	−0.0119	−0.0218	−0.0137	−0.0355	0.0829

Fig. 7 (a) Accuracy comparison between our method and Bagging, and (b) our method and Boosting.

classification accuracy rate of AIA-INF is superior to Bagging and Boosting, and its classification error rate is lower.

We also computed efficiency between AIA-INF, Bagging, and Boosting, with the corresponding recognition being shown in Table 6. We can see that our Inception ensemble network has the highest computational efficiency. The main reason behind this efficient result is that Bagging needs resampling in training and Boosting needs retraining of the weight of wrongly classified samples in each iteration, meaning that the Inception network consumes the least time in performing recognition. For learning time, AIA-INF consumes more time because the initial number of our sub-networks is 10, and the artificial immune system algorithm needs to perform many matrix operations to obtain the optimal weights of each sub-network.

Table 6 Time consumption of three methods.

	Learning (h)	Recognition (s)
Bagging	1.32	0.009 41
Boosting	1.65	0.010 26
AIA-INF	2.27	0.007 65

In order to verify the effectiveness of the Inception module network, we designed a common CNN with three convolution layers, as shown in Table 7. Instead of doing transfer learning, we simply conducted a performance comparison experiment between the two classifiers on a pulmonary nodule sign database. The performance is illustrated in Fig. 8.

Using the same transfer training pipeline, we grouped the four types of sign samples into one class as positive nodules and randomly selected 2-time non-pathological lung parenchyma patches as negative samples to conduct a 2-class distinguishing experiment. The

Table 7 Structure of the common CNN.

Layer	Image size	Kernel size	Stride
1	32×32	7	1
2	27×27	7	2
3	11×11	5	1
4	8×8	–	–

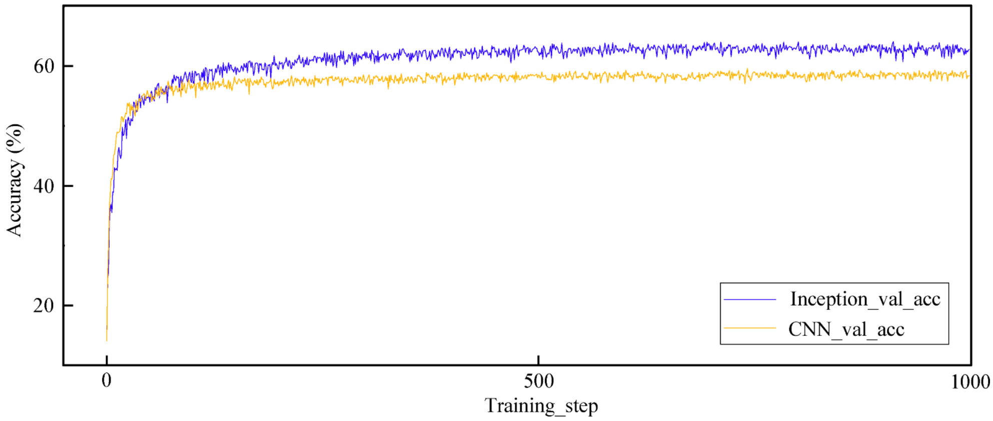


Fig. 8 Performance comparison between our Inception network and a common CNN.

sensitivity, specificity, and accuracy of our algorithm are 89.75, 94.21, and 93.78, respectively. Compared with the existing state-of-the-art algorithms, our algorithm achieved similar classification performance, as shown in Table 8.

From Fig. 8, we can see that a common CNN converges faster than the Inception network, but the final classification performance is clearly lower than Inception network. This shows that the Inception network can obtain more discriminanting features than a common CNN network.

The α in the mutation Eq. (5) controls the latitude of the maximum mutations. Its value should not be too small, which would lead to a drastic increase in the number of iterations. In our experiments, the best integration performance is usually achieved in around 30 to 40 steps, as shown in Fig. 9. Therefore, we set the maximum number of iterations to 50. An appropriate initial affinity threshold value can speed up the convergence of the integration. So we set it to 0.85 for this experiment.

7 Conclusion

In this paper, we presented a new AIA-INF and applied it to lung CT signs recognition. This repersents an in-depth study of pulmonary nodule computer-aided diagnosis, and it is the first attempt to analyze four types of pulmonary nodule signs closely related to lung cancer using a deep neural network. Given a ROI image patch, our Inception network ensemble outputs a weighted sum of class confidence vectors. Experiments showed that our Inception network ensemble holds great advantages for lung CT signs recognition, after optimization by an artificial immune algorithm.

Table 8 Performance comparison of six algorithms. (%)

Algorithm	SE	SP	ACC
Zhou et al.[66]	–	–	88.40
Chen et al.[67]	–	–	78.70
Xie et al.[46]	**91.43**	94.09	93.40
Shen et al.[39]	77.00	93.00	87.14
Paul et al.[68]	–	–	89.45
Our ensemble	89.75	**94.21**	**93.78**

Compared with individual networks, our ensemble gives a 12.49% improvement in sensitivity and 7.67% in specificity; compared with Bagging and Boosting, the improvement rates are 16.66% and 5.50%, 18.00% and 6.23%, respectively. The recognition time is 0.007 65 s, which shows the Inception ensemble to be a real time classification method. Experiments demonstrated that the ensemble classifier integrated with an artificial immune algorithm gives good results for lung sign recognition. This CADx scheme can offer a more detailed reference for diagnosis, and can also be of value for junior radiologist training.

In the process of experimentation, we found that the lack of labeled samples is the key problem affecting the performance of lung sign recognition systems. Hospital Radiology Information System (RIS) hold many diagnostic reports on CT images. In future work, we plan to combine the qualitative or quantitative analysis results in the diagnosis report with the corresponding CT images to obtain a large number of weakly labeled samples, and study dedicated algorithms to further improve the recognition performance of CADx systems.

References

[1] A. Jones, D. Stockton, A. Simpson, and J. Murchison, Idiopathic venous thromboembolic disease is associated with a poorer prognosis from subsequent malignancy, *British Journal of Cancer*, vol. 101, no. 5, p. 840, 2009.

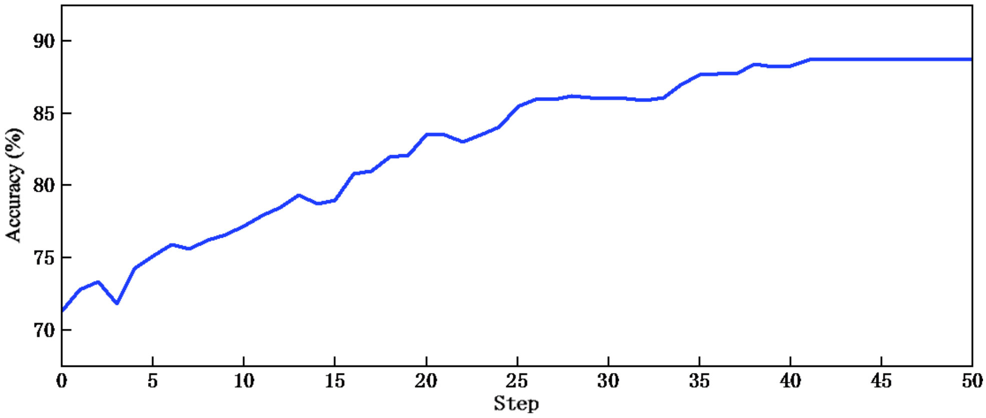


Fig. 9 Accuracy with iterations steps.

[2] G. K. Singh and A. Jemal, Socioeconomic and racial/ethnic disparities in cancer mortality, incidence, and survival in the United States, 1950–2014: Over six decades of changing patterns and widening inequalities, *Journal of Environmental and Public Health*, vol. 2017, pp. 1–19, 2017.

[3] T. B. Richards, S. J. Henley, M. C. Puckett, H. K. Weir, B. Huang, and T. C. Tucker, Abstract B29: Trends in racial disparities in five-year net survival for lung cancer, 2001–2009, in *AACR*, Washington, DC, USA, 2017, p. B29.

[4] G. V. Scagliotti, Symptoms, signs and staging of lung cancer, *European Respiratory Monograph*, vol. 6, no. 17, pp. 86–119, 2001.

[5] K. Alzahouri, M. Velten, P. Arveux, M.-C. Woronoff-Lemsi, D. Jolly, and F. Guillemin, Management of SPN in France. Pathways for definitive diagnosis of solitary pulmonary nodule: A multicentre study in 18 French districts, *BMC Cancer*, vol. 8, no. 1, p. 93, 2008.

[6] International Early Lung Cancer Action Progrom Investigators, Survival of patients with stage I lung cancer detected on CT screening, *New England Journal of Medicine*, vol. 355, no. 17, pp. 1763–1771, 2006.

[7] J. Okami, Y. Ito, M. Higashiyama, T. Nakayama, T. Tokunaga, J. Maeda, and K. Kodama, Sublobar resection provides an equivalent survival after lobectomy in elderly patients with early lung cancer, *The Annals of Thoracic Surgery*, vol. 90, no. 5, pp. 1651–1656, 2010.

[8] M.-L. Chabi, I. Borget, R. Ardiles, G. Aboud, S. Boussouar, V. Vilar, C. Dromain, and C. Balleyguier, Evaluation of the accuracy of a computer-aided diagnosis (CAD) system in breast ultrasound according to the radiologist's experience, *Academic Radiology*, vol. 19, no. 3, pp. 311–319, 2012.

[9] D. S. Gierada, T. K. Pilgram, M. Ford, R. M. Fagerstrom, T. R. Church, H. Nath, K. Garg, and D. C. Strollo, Lung cancer: Interobserver agreement on interpretation of pulmonary findings at low-dose CT screening, *Radiology*, vol. 246, no. 1, pp. 265–272, 2008.

[10] S. G. Armato, F. Li, M. L. Giger, H. MacMahon, S. Sone, and K. Doi, Lung cancer: Performance of automated lung nodule detection applied to cancers missed in a CT screening program, *Radiology*, vol. 225, no. 3, pp. 685–692, 2002.

[11] F. Li, S. Sone, H. Abe, H. MacMahon, S. G. Armato, and K. Doi, Lung cancers missed at low-dose helical CT screening in a general population: Comparison of clinical, histopathologic, and imaging findings, *Radiology*, vol. 225, no. 3, pp. 673–683, 2002.

[12] J. Collins, CT signs and patterns of lung disease., *Radiol Clinics*, vol. 39, no. 6, pp. 1115–1135, 2001.

[13] S. G. Spiro, M. K. Gould, and G. L. Colice, Initial evaluation of the patient with lung cancer: Symptoms, signs, laboratory tests, and paraneoplastic syndromes: ACCP evidenced-based clinical practice guidelines (2nd edition), *Chest*, vol. 132, no. 3, p. 149S–160S, 2007.

[14] C. V. Zwirewich, S. Vedal, R. R. Miller, and N. L. Müller, Solitary pulmonary nodule: High-resolution CT and radiologic-pathologic correlation, *Radiology*, vol. 179, no. 2, pp. 469–476, 1991.

[15] D. E. Wood, G. A. Eapen, D. S. Ettinger, L. Hou, D. Jackman, E. Kazerooni, D. Klippenstein, R. P. Lackner, L. Leard, A. N. Leung, et al., Lung cancer screening, *Journal of the National Comprehensive Cancer Network*, vol. 10, no. 2, pp. 240–265, 2012.

[16] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans, *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[17] T. W. Way, B. Sahiner, H.-P. Chan, L. Hadjiiski, P. N. Cascade, A. Chughtai, N. Bogot, and E. Kazerooni, Computer-aided diagnosis of pulmonary nodules on CT scans: Improvement of classification performance with nodule surface features, *Medical Physics*, vol. 36, no. 7, pp. 3086–3098, 2009.

[18] I. C. Sluimer, P. F. van Waes, M. A. Viergever, and B. van Ginneken, Computer-aided diagnosis in high resolution CT of the lungs, *Medical Physics*, vol. 30, no. 12, pp. 3081–3090, 2003.

[19] P. D. Korfiatis, A. N. Karahaliou, A. D. Kazantzi, C. Kalogeropoulou, and L. I. Costaridou, Texture-based identification and characterization of interstitial pneumonia patterns in lung multidetector CT, *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 675–680, 2010.

[20] L. S∅rensen, P. Lo, H. Ashraf, J. Sporring, M. Nielsen, and M. De Bruijne, Learning COPD sensitive filters in pulmonary CT, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, London, UK, 2009, pp. 699–706.

[21] C. Jacobs, E. M. van Rikxoort, E. T. Scholten, P. A. de Jong, M. Prokop, C. Schaefer-Prokop, and B. van Ginneken, Solid, part-solid, or non-solid?: Classification of pulmonary nodules in low-dose chest computed tomography by a computer-aided diagnosis system, *Investigative Radiology*, vol. 50, no. 3, pp. 168–173, 2015.

[22] X. Ye, X. Lin, J. Dehmeshki, G. Slabaugh, and G. Beddoe, Shape-based computer-aided detection of lung nodules in thoracic CT images, *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 7, pp. 1810–1820, 2009.

[23] M. Firmino, G. Angelo, H. Morais, M. R. Dantas, and R. Valentim, Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy, *Biomedical Engineering Online*, vol. 15, no. 1, p. 2, 2016.

[24] T. Sun, R. Zhang, J. Wang, X. Li, and X. Guo, Computer-aided diagnosis for early-stage lung cancer based on longitudinal and balanced data, *PloS One*, vol. 8, no. 5, p. e63559, 2013.

[25] T. Sun, J. Wang, X. Li, P. Lv, F. Liu, Y. Luo, Q. Gao, H. Zhu, and X. Guo, Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set, *Computer Methods and Programs in Biomedicine*, vol. 111, no. 2, pp. 519–524, 2013.

[26] W. Sun, B. Zheng, and W. Qian, Computer aided lung cancer diagnosis with deep learning algorithms, in *Medical*

Imaging 2016: Computer-Aided Diagnosis, San Diege, CA, USA, 2016, vol. 9785, p. 97850Z, 2016.

[27] F. Han, H. Wang, B. Song, G. Zhang, H. Lu, W. Moore, H. Zhao, and Z. Liang, A new 3D texture feature-based computer-aided diagnosis approach to differentiate pulmonary nodules, in *Medical Imaging 2013: Computer-Aided Diagnosis*, Dresden, Germany, vol. 8670, p. 86702Z, 2013.

[28] F. Han, H. Wang, G. Zhang, H. Han, B. Song, L. Li, W. Moore, H. Lu, H. Zhao, and Z. Liang, Texture feature analysis for computer-aided diagnosis on pulmonary nodules, *Journal of Digital Imaging*, vol. 28, no. 1, pp. 99–115, 2015.

[29] M. G. Penedo, M. J. Carreira, A. Mosquera, and D. Cabello, Computer-aided diagnosis: A neural-network-based approach to lung nodule detection, *IEEE Transactions on Medical Imaging*, vol. 17, no. 6, pp. 872–880, 1998.

[30] M. Tan, R. Deklerck, B. Jansen, M. Bister, and J. Cornelis, A novel computer-aided lung nodule detection system for CT images, *Medical Physics*, vol. 38, no. 10, pp. 5630–5645, 2011.

[31] S. E. Darmanayagam, K. N. Harichandran, S. R. R. Cyril, and K. Arputharaj, A novel supervised approach for segmentation of lung parenchyma from chest CT for computer-aided diagnosis, *Journal of Digital Imaging*, vol. 26, no. 3, pp. 496–509, 2013.

[32] A. S. Abdalla, I. A. Yusuf, S. H. A. A. Mohammed, M. A. Mahmoud, and Z. A. Mustafa, A computer-aided diagnosis system for classification of lung tumors, *Journal of Clinical Engineering*, vol. 40, no. 3, pp. 130–134, 2015.

[33] R. Niehaus, D. S. Raicu, J. Furst, and S. Armato, Toward understanding the size dependence of shape features for predicting spiculation in lung nodules for computer-aided diagnosis, *Journal of Digital Imaging*, vol. 28, no. 6, pp. 704–717, 2015.

[34] L. Breiman, Bagging predictors, *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[35] B. Verma and J. Zakos, A computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques, *IEEE Transactions on Information Technology in Biomedicine*, vol. 5, no. 1, pp. 46–54, 2001.

[36] L. Ma, X. Liu, L. Song, C. Zhou, X. Zhao, and Y. Zhao, A new classifier fusion method based on historical and on-line classification reliability for recognizing common CT imaging signs of lung diseases, *Computerized Medical Imaging and Graphics*, vol. 40, pp. 39–48, 2015.

[37] J. Y. Choi, D. H. Kim, K. N. Plataniotis, and Y. M. Ro, Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography, *Expert Systems with Applications*, vol. 46, pp. 106–121, 2016.

[38] S. Hussein, K. Cao, Q. Song, and U. Bagci, Risk stratification of lung nodules using 3D CNN-based multi-task learning, in *International Conference on Information Processing in Medical Imaging*, Boone, NC, USA, 2017, pp. 249–260.

[39] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang, and J. Tian, Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification, *Pattern Recognition*, vol. 61, pp. 663–673, 2017.

[40] G. Wei, H. Ma, W. Qian, F. Han, H. Jiang, S. Qi, and M. Qiu, Lung nodule classification using local kernel regression models with out-of-sample extension, *Biomedical Signal Processing and Control*, vol. 40, pp. 1–9, 2018.

[41] Q. Song, L. Zhao, X. Luo, and X. Dou, Using deep learning for classification of lung nodules on computed tomography images, *Journal of Healthcare Engineering*, vol. 2017, pp. 1–7, 2017.

[42] K. Liu and G. Kang, Multiview convolutional neural networks for lung nodule classification, *International Journal of Imaging Systems and Technology*, vol. 27, no. 1, pp. 12–22, 2017.

[43] A. Tartar, N. Kilic, and A. Akan, Bagging support vector machine approaches for pulmonary nodule detection, in *2013 International Conference of Control, Decision and Information Technologies (CoDIT)*, Hammamet, Tunisia, 2013, pp. 047–050.

[44] A. Tartar and A. Akan, Ensemble learning approaches to classification of pulmonary nodules, in *2016 International Conference of Control, Decision and Information Technologies (CoDIT)*, Saint Julian's, Malta, 2016, pp. 472–477.

[45] M. A. Jaffar, A. B. Siddiqui, and M. Mushtaq, Ensemble classification of pulmonary nodules using gradient intensity feature descriptor and differential evolution, *Cluster Computing*, vol. 231, no. 1, pp. 393–407, 2018.

[46] Y. Xie, Y. Xia, J. Zhang, D. D. Feng, M. Fulham, and W. Cai, Transferable multi-model ensemble for benign-malignant lung nodule classification on chest CT, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Quebec City, Canada, 2017, pp. 656–664.

[47] F. V. Farahani, A. Ahmadi, and M. F. Zarandi, Lung nodule diagnosis from CT images based on ensemble learning, in *2015 IEEE Conference of Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Niagara Falls, Canada, 2015, pp. 1–7.

[48] L. K. Hansen and P. Salamon, Neural network ensembles, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 2002.

[49] Z. H. Zhou, J. Wu, and W. Tang, Ensembling neural networks: Many could be better than all, *Artificial Intelligence*, vol. 137, nos. 1&2, pp. 239–263, 2002.

[50] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki, Automated detection of pulmonary nodules in PET/CT images: Ensemble false-positive reduction using a convolutional neural network technique, *Medical Physics*, vol. 43, no. 6, p. 2821, 2016.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Proceedings of Advances in Neural Information Processing Systems*, Lake Tahoe, NE, USA, 2012, pp. 1097–1105.

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, ImageNet large scale visual recognition challenge, *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248–255.

[54] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1–9.

[55] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.

[56] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2016, pp. 770–778.

[57] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, Inception-v4, Inception-resnet and the impact of residual connections on learning, in *31th AAAI Conference on Artificaial Intelligence*, San Francisco, CA, USA, 2017, vol. 4, p. 12.

[58] B. Van Ginneken, S. G. Armato III, B. de Hoop, S. van Amelsvoort-van de Vorst, T. Duindam, M. Niemeijer, K. Murphy, A. Schilham, A. Retico, M. E. Fantacci, et al., Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study, *Medical Image Analysis*, vol. 14, no. 6, pp. 707–722, 2010.

[59] D. E. Ost and M. K. Gould, Decision making in patients with pulmonary nodules, *American Journal of Respiratory and Critical Care Medicine*, vol. 185, no. 4, pp. 363–372, 2012.

[60] P. Opulencia, D. S. Channin, D. S. Raicu, and J. D. Furst, Mapping LIDC, RadLexTM, and lung nodule image features, *Journal of Digital Imaging*, vol. 24, no. 2, pp. 256–270, 2011.

[61] M. F. Mcnitt-Gray, E. M. Hart, N. Wyckoff, J. W. Sayre, J. G. Goldin, and D. R. Aberle, A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: Preliminary results, *Medical Physics*, vol. 26, no. 6, p. 880, 1999.

[62] G. Han, X. Liu, F. Han, I. Santika, Y. Zhao, X. Zhao, and C. Zhou, The LISS—A public database of common imaging signs of lung diseases for computer-aided detection and diagnosis research and medical education, *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 648–656, 2015.

[63] D. Bertsimas and J. Tsitsiklis, Simulated annealing, *Statistical Science*, vol. 8, no. 1, pp. 10–15, 1993.

[64] L. B. Booker, D. E. Goldberg, and J. H. Holland, Classifier systems and genetic algorithms, *Artificial Intelligence*, vol. 40, no. 1–3, pp. 235–282, 1989.

[65] R. E. Schapire, The strength of weak learnability, *Machine Learing*, vol. 5, no. 2, pp. 1979–227, 1990.

[66] Z.-H. Zhou, Y. Jiang, Y.-B. Yang, and S.-F. Chen, Lung cancer cell identification based on artificial neural network ensembles, *Artificial Intelligence in Medicine*, vol. 24, no. 1, pp. 25–36, 2002.

[67] H. Chen, W. Wu, H. Xia, J. Du, M. Yang, and B. Ma, Classification of pulmonary nodules using neural network ensemble, in *International Symposium on Neural Networks*, Guilin, China, 2011, pp. 460–466.

[68] R. Paul, L. Hall, D. Goldgof, M. Schabath, and R. Gillies, Predicting nodule malignancy using a CNN ensemble approach, in 2018 *International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, France, 2018, pp. 1–8.

Guangyuan Zheng received the BS degree in 2010 from China University of Geosciences, and now is a PhD candidate in Beijing Institute of Technology. His major research interests include machine learning, medical image analysis, and computer safety.

Nouman Qadeer Soomro received the BEng and MEng degrees from the Mehran University of Engineering and Technology, Jamshoro, Pakistan, in 2008 and 2011, respectively, and the PhD degree from the Machine Learning and Multimedia Retrieval (MLMR) Laboratory, Department of Computer Science, Beijing Institute of Technology (BIT), China, in 2015. Currently, he is an assistant professor and the head of Department of Software Engineering, Mehran University of Engineering and Technology, SZAB Campus, Khairpur, Pakistan. His major research interests include machine learning, computer vision, medical imaging, remote sensing image processing, pattern recognition, multimedia processing, and data science. He has widely published at highly ranked international conferences and international journals such as *IEEE GRSL*, *Signal Processing: Image Communication*. He has obtained several honors and awards such as secured funding from Japan Science and Technology (JST) Agency, Japanese Government, and Government of Pakistan.

Guanghui Han received the PhD degree from Beijing Institute of Technology in 2018. He is now working as a postdoctoral researcher with the School of Biomedical Engineering in Sun Yat-sen University. His research interests cover medical image analysis and machine learning.