Medical Image Analysis 57 (2019) 237–248

Contents lists available at ScienceDirect

Medical Image Analysis

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

# Semi-supervised adversarial model for benign–malignant lung nodule classification on chest CT

Yutong Xie [a], Jianpeng Zhang [a], Yong Xia [a,b,*]

[a] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[b] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China

ARTICLE INFO

*Article history:*
Received 25 November 2018
Revised 20 May 2019
Accepted 4 July 2019
Available online 10 July 2019

*Keywords:*
Lung nodule classification
Semi-supervised learning
Adversarial learning
Deep learning

ABSTRACT

Classification of benign–malignant lung nodules on chest CT is the most critical step in the early detection of lung cancer and prolongation of patient survival. Despite their success in image classification, deep convolutional neural networks (DCNNs) always require a large number of labeled training data, which are not available for most medical image analysis applications due to the work required in image acquisition and particularly image annotation. In this paper, we propose a semi-supervised adversarial classification (SSAC) model that can be trained by using both labeled and unlabeled data for benign–malignant lung nodule classification. This model consists of an adversarial autoencoder-based unsupervised reconstruction network *R*, a supervised classification network *C*, and learnable transition layers that enable the adaption of the image representation ability learned by *R* to *C*. The SSAC model has been extended to the multi-view knowledge-based collaborative learning, aiming to employ three SSACs to characterize each nodule's overall appearance, heterogeneity in shape and texture, respectively, and to perform such characterization on nine planar views. The MK-SSAC model has been evaluated on the benchmark LIDC-IDRI dataset and achieves an accuracy of 92.53% and an AUC of 95.81%, which are superior to the performance of other lung nodule classification and semi-supervised learning approaches.



## 1. Introduction

Lung cancer is the leading cause of cancer death (Torre et al., 2016). The 5-year survival rate for patients with advanced stage IV lung cancer is less than 5%, but it is at least 60% if the diagnosis is made early when the primary tumor is small and asymptomatic (Wu and Raz, 2016). Early lung cancer detection and effective treatment therefore offers the best chance for cure (Wang et al., 2017). The National Lung Screening Trial shows that screening with computed tomography (CT) results in a 20% reduction in lung cancer deaths through the identification of early disease (Wu and Raz, 2016; Bach et al., 2012). A spot on the lung on a chest CT is defined as a lung nodule, and it can be benign or malignant (Slatore et al., 2016). Most lung cancers arise from small malignant nodules. Radiologists typically read chest CT scans for malignant nodules on a slice-by-slice basis, and such an approach is time-consuming, expensive and can be prone to operator bias. Computer-aided diagnosis (CAD) systems avoid many of these issues and have been employed to assist radiologists in reading chest CT scans.

Most current CAD systems focus on extracting handcrafted (Han et al., 2015; Dhara et al., 2016; Alilou et al., 2017), learned (Shen et al., 2015), or combined nodule features (Xie et al., 2018b; 2017b; Buty et al., 2016), and then training a feature classifier such as the support vector machine (SVM) (Cortes and Vapnik, 1995), back propagation neural network (BPNN) (Rojas, 1996; Hecht-Nielsen, 1992; Zhang et al., 2018) and random forest (Buty et al., 2016; Breiman, 2001). Recently, deep convolutional neural networks (DCNNs) have achieved great success in many image classification tasks (Litjens et al., 2017), since they offer a unified end-to-end solution for feature extraction and classifier construction and free users from the troublesome handcrafted feature extraction (Xie et al., 2017a; Shen et al., 2017; Hussein et al., 2017a; Sakamoto et al., 2018; Dey et al., 2018). Although being more accurate than handcrafted features-based methods, DCNNs have not achieved the same performance on routine lung nodule classification as they have done in the ImageNet challenge. The suboptimal performance is attributed mainly to the fact that a DCNN may over-fit the lung nodule data, which is far from adequate to train a deep learning model.

* Corresponding author.
*E-mail address:* yxia@nwpu.edu.cn (Y. Xia).

https://doi.org/10.1016/j.media.2019.07.004
1361-8415/

In fact, an essential challenge in most deep learning-based medical image analysis tasks is the issue of small data, that relates to the work required in acquiring the image data and then in image annotation. Many research efforts have been devoted to address issue, including data augmentation, deep ensemble learning (Jia et al., 2018), combining traditional shallow models with deep ones (Zhang et al., 2018; Xie et al., 2018b), incorporating domain knowledge into the deep model (Xie et al., 2017a), and extracting patches on multiple planar views (Xie et al., 2018a; Setio et al., 2016). Although achieved improved performance, these methods still rely on the number of training data.

Since medical image annotation requires a high degree of skill and concentration and is not always available, semi-supervised learning (SSL) has been adopted to enable us to use both labeled and unlabeled data for model training. The deep learning community has explored a large variety of SSL techniques (Cheng et al., 2016; Hinton and Salakhutdinov, 2006; Ranzato and Szummer, 2008; Springenberg, 2015; Radford et al., 2015; Makhzani et al., 2015; Rasmus et al., 2015; Zhang et al., 2017; Baur et al., 2017; Haeusser et al., 2017). As a typical example, a deep autoencoder (DAE) trained with unlabeled data can be altered into a classifier via replacing the decoder part with fully connected layers and fine-tuning with labeled data. However, DAE is a generative model, which learns the image representation that is suitable for reconstruction, but may not be suitable for discrimination. Sharing the parameters between the encoder part of DAE and the feature extraction part of a classification network may lead to limited discriminatory power (Rasmus et al., 2015). Therefore, we suggest jointly using a generative model trained with both labeled and unlabeled data and a discriminative model trained with only labeled data for semi-supervised medical image classification.

In this paper, we propose the semi-supervised adversarial classification (SSAC) model which can be trained by jointly using unlabeled and labeled data in a non-parameter-sharing manner. This model is composed of an adversarial autoencoder-based unsupervised reconstruction network *R*, a supervised classification network *C*, and learnable transition layers (T layers), which enable the adaption of the image representation ability learned by *R* to *C*. The SSAC model has been further extended to the multi-view knowledge-based collaborative (MV-KBC) learning (Xie et al., 2018a), denoted by MK-SSAC model, for benign–malignant lung nodule classification using chest CT. The proposed MK-SSAC model consists of 27 SSAC submodels, each characterizing a nodule's overall appearance (OA), heterogeneity in voxel values (HVV) and heterogeneity in shapes (HS) from each of sagittal, coronal, axial, and six diagonal planar views, respectively.

### 1.1. Related work

**DCNN-based nodule classification.** The success of DCNNs on several popular image classification benchmarks such as the ImageNet database has prompted many investigators to apply DCNNs to benign–malignant lung nodule classification. Hua et al. (2015) applied a DCNN and deep belief network (DBN) to separate benign lung nodules from malignant ones and reported that deep learning achieved better discrimination than traditional methods. Shen et al. (2017) proposed a multi-crop CNN to extract nodule salient information by cropping different regions from convolutional feature maps and then applying max-pooling multiple times. Hussein et al. (2017a) combined a 3D DCNN with graph regularized sparse multi-task learning to stratify the malignancy of lung nodules.

**Small data problem in medical image classification.** There are many attempts to address the issue of small data in medical image classification. First, although it is straightforward to design 3D DCNN (Hussein et al., 2017a; Dou et al., 2017b; 2017c; Li et al., 2018; Dou et al., 2017a; Yan et al., 2017), extending the use of 2D DCNN to the analysis of volumetric medical images on a slice-by-slice basis, together with data augmentation (Shen et al., 2017; Setio et al., 2016; Hussein et al., 2017b; Vigneault et al., 2018), enables us to have more training samples. Second, the prior domain knowledge, such as there is a high correspondence between a nodule's malignancy and its heterogeneity (see Fig. 1) (Xie et al., 2017a, 2018a; Metz et al., 2015), can be used to regularize deep models. In our previous work (Xie et al., 2018a), we proposed the MV-KBC learning model to separate malignant nodules from benign ones using limited chest CT data. We decomposed a 3D nodule into nine fixed views and constructed, for each view, a knowledge-based collaborative (KBC) submodel, which consists of three pre-trained ResNet-50 networks. We designed three types of image patches to fine-tune those three ResNet-50 in each KBC submodel, enabling them to characterize the nodule's OA, HVV, and HS, respectively. We jointly used nine KBC submodels to classify lung nodules with an adaptive weighting scheme learned during the error back propagation, which enables us to train the MV-KBC model in an end-to-end manner.

**Semi-supervised deep learning.** SSL methods, which leverage unsupervised learning with unlabeled data to support the learning of supervised model, enable us to train deep models using both labeled and unlabeled data, and hence reduce the work related to image annotation (Chapelle et al., 2009; Zhu, 2006). There are different choices for unsupervised learning models. Ranzato and Szummer (2008) used an unsupervised autoencoder to reconstruct the input, and shared parameters between the encoder and a classification network. Goodfellow et al. (2014) introduced an unsupervised generative adversarial network (GAN) which consists of two adversarial models: a generative model *G* that captures the data distribution and a discriminative model *D* that estimates the probability that a sample comes from the training data rather than *G*. Considering GAN as unsupervised learning model, researchers have developed a series of SSL methods. Makhzani et al. (2015) turned an autoencoder into the adversarial autoencoder (AAE) and exploited the encoder to predict the discrete class label. Springenberg (2015) modified the objective function of the discriminator, and thus proposed the categorical GAN (CatGAN), which takes into account the mutual

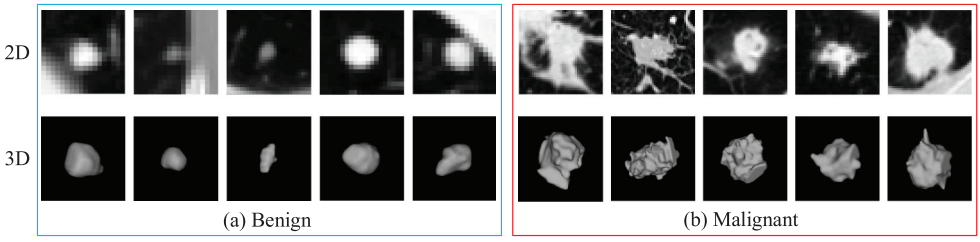


**Fig. 1.** Transaxial 2D patches and the corresponding 3D visualization of benign and malignant lung nodules in chest CT scans. It shows that there is a high correspondence between a nodules malignancy and its heterogeneity in voxel values and heterogeneity in shapes.

**Table 1**
Distribution of the 2557 lung nodules over other eight attributes and annotations. "C" and "N" are the category and the number of nodules, respectively.

| Attributes | | Scoring | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Subtlety | C | Extremely Subtle | Moderately Subtle | Fairly Subtle | Moderately Obvious | Obvious | – |
| | N | 140 | 297 | 333 | 849 | 938 | 0 |
| Sphericity | C | Linear | – | Ovoid | – | Round | – |
| | N | 5 | 249 | 395 | 1128 | 780 | 0 |
| Calcification | C | Popcorn | Laminated | Solid | Non-central | Central | Absent |
| | N | 0 | 0 | 189 | 82 | 101 | 2185 |
| Texture | C | Non-solid | – | Part Solid | – | Solid | – |
| | N | 190 | 85 | 107 | 261 | 1914 | 0 |
| Internal structure | C | Soft Tissue | Fluid | Fat | Air | – | – |
| | N | 2531 | 14 | 9 | 2 | 1 | 0 |
| Margin | C | Poorly Defined | – | – | – | Sharp | – |
| | N | 133 | 280 | 200 | 726 | 1218 | 0 |
| Lobulation | C | Marked | – | – | – | None | – |
| | N | 1792 | 490 | 119 | 120 | 36 | 0 |
| Spiculation | C | Marked | – | – | – | None | – |
| | N | 1965 | 345 | 85 | 113 | 49 | 0 |

information between observed examples and their predicted class distribution. Rasmus et al. (2015) proposed the ladder network to alleviate the adaptation problem between unsupervised generative and supervised discriminative models. This network has an auxiliary supervised output in the encoder part and skip connections from the encoder to decoder, aiming to reduce the burden on the encoding layers during unsupervised learning. Despite improved accuracy, this model still uses shared parameters in encoder part for both generative and discriminative tasks.

### 1.2. Contributions

The main contributions of this work include: (a) the proposed SSAC model uses learnable T layers to transfer the representation ability learned by the reconstruction network *R* to the classification network *C*, abandoning the parameter sharing and feature adaption strategy, (b) the adversarial training is used in *R* to minimize the discrepancy between the distributions of input lung nodules and reconstructed one, and (c) the extended MK-SSAC model has been evaluated on the LIDC-IDRI dataset and achieved the state-of-the-art performance in benign–malignant lung nodule classification.

## 2. Experimental dataset

The LIDC-IDRI dataset (Armato et al., 2011; Armato III et al., 2015b; Clark et al., 2013) in the Cancer Imaging Archive (TCIA) contains 1018 clinical chest CT scans with lung nodules obtained from seven institutions. Each CT scan was evaluated for annotation by up to four radiologists and has an associated XML file that details the nodules locations and nine sematic attributes of subtlety, sphericity, calcification, texture, internal structure, margin, lobulation, spiculation and malignancy (Qin et al., 2019). We sampled the nodules, whose diameters range from 3 mm to 30 mm, for this study, since nodules $< 3$ mm are not considered to be clinically relevant by current screening protocols (Han et al., 2015; Dhara et al., 2016; Xie et al., 2018b; 2017a). The malignancy of each nodule was evaluated over five levels, from benign to malignant, by up to four experienced thoracic radiologists. Following the settings in Han et al. (2015); Dhara et al. (2016) and Xie et al. (2018b, 2017a), we treated the nodules with a median malignancy level $< 3$ as benign, the nodules with a median malignancy level $> 3$ as malignant and the nodules with a median malignancy level $= 3$ as unlabeled data. Thus, we have 1301 benign, 644 malignant, and 612 unlabeled lung nodules in this dataset. We also summarized the detail distribution of nodules over other attributes and annotations in Table 1.

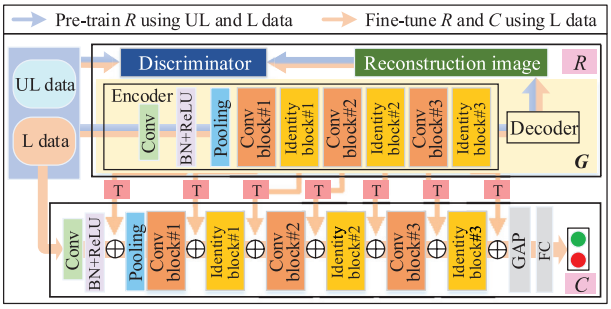


**Fig. 2.** SSAC model. "UL": unlabeled data; "L": labeled data; "*R*": adversarial autoencoder-based reconstruction network; "*C*": classification network; and "*G*": generator which contains an encoder and decoder. "T": learnable T layer.

The Tianchi Lung Nodule Detection[1] dataset was obtained from 16 of the leading cancer hospitals in China which supplied low-dose lung CT scans from some 3000 patients with lung nodules, including supporting clinical data as well. However, only 1000 patient data in the preliminary round are allowed to download. The location centroids and diameters of nodules are marked by three radiologists on the $512 \times 512$ slice. The diameters range from 5 mm to 30 mm which are the similar with those in the LIDC-IDRI dataset. For our study, we extracted 1227 unlabeled nodules from 1000 patients.

Combining both datasets, we have 1301 benign, 644 malignant, and 1839 unlabeled lung nodules.

## 3. Methodology

The proposed SSAC model consists of three major modules: an adversarial autoencoder-based unsupervised reconstruction network *R*, a supervised classification network *C*, and learnable T layers (see Fig. 2). This model has been extended to the MK-SSAC model (see Fig. 4(b)) for benign–malignant lung nodule classification, which contains 27 SSAC submodels, each characterizing a

[1] https://tianchi.aliyun.com/competition/introduction.htm?spm=5176.100066.333.6.RWZZWO&raceId=231601.

nodule's OA, HVV, and HS from each of three orthographic and six diagonal views, respectively.

### 3.1. SSAC model

The SSAC model includes a reconstruction network $R$, a classification network $C$, and learnable T layers between them. The reconstruction network $R$ consists of a generator $G$, which reconstructs each input image $I$, and a discriminator $D$, which distinguishes the input image $I$ from the reconstructed image. The generator $G$ contains an encoder and a decoder (see Fig. 3(a)). The encoder that comprises a conv-bn-relu block, three conv blocks, and three identity blocks is a ResNet-like network (He et al., 2016) and takes $64 \times 64$ patches as input. The decoder that includes four upsample-conv-bn-relu blocks and a conv-tanh block generates a reconstructed patch of the same size. The discriminator $D$ is a typical DCNN architecture (see Fig. 3(b)), containing four convolutional layers with $3 \times 3$ kernels, a fully connected (FC) layer with 512 neurons, and a FC layer with two neurons followed by the softmax activation function.

The classification network $C$ consists of two parts, which are separated by a global average pooling (GAP) layer (see Fig. 2). Before the GAP layer, it has the same architecture as the encoder of $R$; and after the GAP layer, it has a fully connected layer with two neurons followed by the softmax function.

Instead of sharing parameters between $R$ and $C$, we designed $t$ ($t = 7$) learnable T layers to bridge $R$ and $C$ by transferring the feature maps obtained by the conv-bn-relu block, three conv blocks, and three identity blocks of $R$ to the corresponding blocks of $C$. Each T layer is a $1 \times 1$ convolutional layer with a stride of 1 followed by batch normalization (BN) and the rectified linear unit (ReLU) activation.

The SSAC model can be trained in three steps. Step 1, we used the parameters converged on the ImageNet dataset (Deng et al., 2009) to initialize $C$, aiming to transfer the image representation ability learned on large-scale image datasets to the characterization of our nodule data. Step 2, we pre-trained $R$ by using both labeled and unlabeled data. Given the $n$th labeled or unlabeled input data $\mathbf{X}_n, n = 1, 2, \ldots, N$, we defined the loss function of $R$ as follow

$$L_R(\mathbf{X}_n) = l_{mse}(G(\mathbf{X}_n), \mathbf{X}_n) - \lambda[l_{bc}(D(G(\mathbf{X}_n), 0) + l_{bc}(D(\mathbf{X}_n), 1)] \tag{1}$$

where $l_{bc}$ is the cross-entropy adversarial loss for $D$, which measures the discrepancy between the distributions of an input image and the reconstructed one, and $l_{mse}$ is the mean square reconstruction loss for $G$. We adopted the Adam algorithm (Kingma and Ba, 2014) to optimize this loss function and set batch size to 32. Since the adversarial training part may be less useful before $G$ can produce reasonably good image reconstruction, we set $\lambda = 0.1$ initially, and set $\lambda = 1$ after 200 epochs.

Step 3, we further used the labeled samples to fine-tune the SSAC model. Given the $m$th labeled input data $\mathbf{X}_m$ and corresponding label $\mathbf{Y}_m, m = 1, 2, \ldots, M$, we defined the loss function of SSAC model as

$$\begin{aligned} &L_{SSAC}(\mathbf{X_m}) \\ &= \lambda_1\{l_{mse}(G(\mathbf{X_m}), \mathbf{X_m}) - [l_{bc}(D(G(\mathbf{X_m})), 0) + l_{bc}(D(\mathbf{X_m}), 1)]\} \\ &\quad + l_{bc}(C(\mathbf{X_m}), \mathbf{Y_m}) \end{aligned} \tag{2}$$

where the first term is the reconstruction loss for $G$, the second term is the adversarial loss for $D$, the third term is the supervised classification loss, and the weighting parameter $\lambda_1$ that balances the contributions of the generative model and discriminative model was empirically set to 0.01. Similarly, we adopted the Adam algorithm (Kingma and Ba, 2014) to optimize this loss function and set batch size to 32.

Moreover, when pre-training $R$ and fine-tuning the SSAC model, we randomly chose 10% of the training patches to form a validation set and terminated the training process if the error on the other 90% of training patches continues to decline but the error on the validation set starts to rise, even before reaching the maximum epoch number 500.

### 3.2. MK-SSAC model

The SSAC model can be extended to MV-KBC learning (Xie et al., 2018a) for benign–malignant lung nodule classification in three steps: (1) extracting multi-view OA, HVV and HS patches from each lung nodule, (2) constructing nine knowledge-based collaborative SSAC (KBC-SSAC) models and training each of them using the patches extracted on each of nine planar views, and (3) constructing and training the MK-SSAC model for nodule classification.

#### 3.2.1. Extracting multi-view OA, HVV and HS patches

We resampled all chest CT scans to a unified voxel size of $1.0 \times 1.0 \times 1.0$ mm$^3$ using the spline interpolation. Let the location of a nodule be the middle of the nodule's centers given by radiologists. For each lung nodule, we first cropped a $64 \times 64 \times 64$ cube that is centered on its location such that the nodule is always contained completely in the cube, and then extracted 2D slices on the transverse, sagittal, coronal and six diagonal planes, respectively (Setio et al., 2016). Thus, we obtained nine views of slices for each nodule.

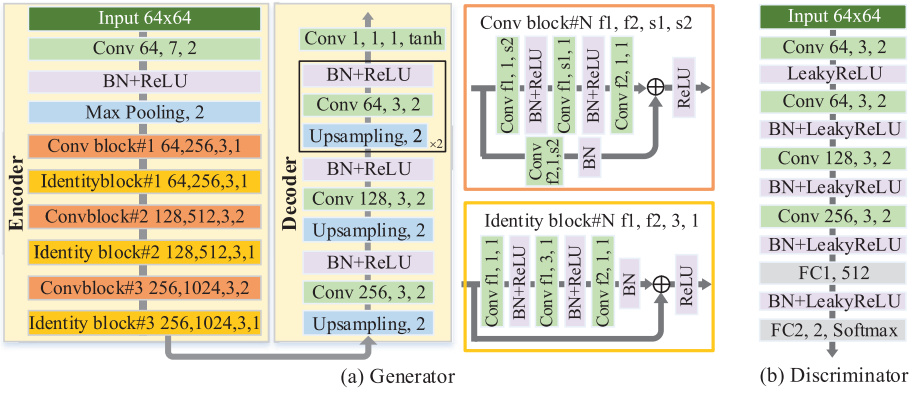


**Fig. 3.** Detailed architecture of reconstruction network R: (a) Generator, and (b) Discriminator. Green "Conv": convolutional layers with parameters indicated as the number of filters, kernel size and stride; Blue: $2 \times 2$ maxpooling or upsampling layers; Violet "BN+ReLU": batch normalization (BN) and ReLU activation; Gray "FC1" or "FC2": fully connected layers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

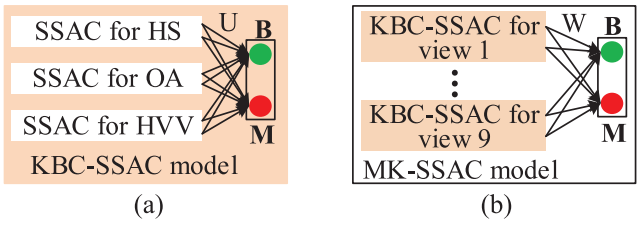


**Fig. 4.** (a) KBC-SSAC model. (b) MK-SSAC model. "U" is the assemble of weights between the output of SSAC and the output of the KBC-SSAC model. "W" is the assemble of weights between the output of KBC-SSAC and the output of the MK-SSAC model. "B" and "M" mean benign and malignant nodules, respectively.

On each slice, a square region of interest (ROI) that encapsulates a nodule was identified as an OA patch to represent the nodule's overall appearance. To characterize the nodule's HVV, non-nodule voxels inside the OA patch were set to 0 and, if the OA patch is larger than $16 \times 16$, a $16 \times 16$ patch that contains the maximum nodule voxels was extracted as an HVV patch. To generate the nodule's HS patch, nodule voxels inside the OA patch were set to 0. We also generated four augmented data for each training patch using rotation, horizontal or vertical flip like (Xie et al., 2018a). The translation step was in the range of $[-6, 6]$ voxels, and the rotation angle was randomly selected from $\{90^\circ, 180^\circ, 270^\circ\}$. Then, all OA, HVV and HS patches were resized to $64 \times 64$.

In the training stage, the region of a nodule was defined as the intersection of the areas marked by radiologists. In the testing stage, nodule segmentation was performed on a slice-by-slice basis by using the U-Net (Xie et al., 2018a) trained on the training data with the mini-batch stochastic gradient descent (mini-batch SGD) algorithm, a batch size of 32, maximum epoch number of 100, and learning rate of 0.001. We also used the trained U-Net to segment the Tianchi data, which have no segmentation annotation.

#### 3.2.2. KBC-SSAC model

The OA, HVV and HS patches extracted and augmented on each planar view were used to train a KBC-SSAC model, which contains three SSAC submodels that characterize the nodule's OA, HVV, and HS, respectively (see Fig. 4(a)). Each SSAC submodel has been pre-trained following the steps described in Section 3.1. For the $k$th KBC-SSAC model, given the $m$th labeled input patch triplet $\{\mathbf{X}_{mk}^{(OA)}, \mathbf{X}_{mk}^{(HVV)}, \mathbf{X}_{mk}^{(HS)}\}$ and label $\mathbf{Y}_m$, we denoted the output of $C$ in each SSAC as $\mathbf{O}_{mk}^{(\#)} \in R^2, \# \in \{OA, HVV, HS\}, m = 1, 2, \ldots, M, k = 1, 2, \ldots, 9$. Hence, the output of the $k$th KBC-SSAC model can be calculated as

$$M_{mkj} = f\left(\sum_{\#}\sum_{i=1}^{2} U_{kij}^{(\#)} O_{mki}^{(\#)}\right) \tag{3}$$

where $\{U_{kij}^{(\#)} \in R^2, k = 1, 2, \ldots, 9, i \in \{1, 2\}, j \in \{1, 2\}\}$ is the assemble of weights between the output of each SSAC $O_{mki}^{(\#)}$ and the output of the $k$th KBC-SSAC $M_{mkj}$, the subscript $i \in \{1, 2\}$ indicates the $i$th neuron in the output layer of each SSAC, the subscript $j \in \{1, 2\}$ indicates the $j$th neuron in the output layer of each KBC-SSAC, and $f(\cdot)$ is a softmax activation.

#### 3.2.3. MK-SSAC model

The MK-SSAC model includes nine KBC-SSAC submodels, whose predictions are combined by two output neurons. The two-neuron output layer of each KBC-SSAC model is connected to the same two-neuron classification layer followed by the softmax function (see Fig. 4(b)). The output of this classification layer $\mathbf{P}_m \in R^2$ is the prediction made by the MK-SSAC model, which can be formulated as

$$\mathbf{P}_m = f\left(\sum_{k=1}^{9}\sum_{j=1}^{2} W_{kj} M_{mkj}\right) \tag{4}$$

where $\{W_{kj} \in R^2, k = 1, 2, \ldots, 9, j \in \{1, 2\}\}$ are the assemble of weights between the output of the $k$th KBC-SSAC $M_{mkj}$ and the output of MK-SSAC $\mathbf{P}_m$, the subscript $j \in \{1, 2\}$ represents for the $j$th neuron in the output layer of each KBC-SSAC, and $f(\cdot)$ is the softmax function. In summary, the total loss of our MK-SSAC model is

$$L_{MK-SSAC}(\mathbf{X}_{mk}^{(\#)}) = \sum_{m=1}^{M}\sum_{k=1}^{9}\sum_{\#} \lambda_1 \{l_{mse}(G(\mathbf{X}_{mk}^{(\#)}), \mathbf{X}_{mk}^{(\#)}) - [l_{bc}(D(G(\mathbf{X}_{mk}^{(\#)})), 0) + l_{bc}(D(\mathbf{X}_{mk}^{(\#)}), 1)]\} + \sum_{m=1}^{M} l_{bc}(\mathbf{P}_m, \mathbf{Y}_m) \tag{5}$$

The setting of $\lambda_1$ and training strategies are the same to those for Eq. (2).

## 4. Experiments and evaluation

### 4.1. Experimental design

We applied the proposed MK-SSAC model to the LIDC-IDRI dataset five times independently, with the 10-fold cross validation. The Tianchi dataset that includes 1839 unlabeled nodules was used to train each MK-SSAC model. It took about 24 h to train the MK-SSAC model and less than 0.5 s to use it to classify each nodule (Intel Xeon E5- 2640 V4 CPU, 4 NVIDIA Titan X GPU, 512GB RAM, Keras and Tensorflow). The training process is time consuming, but can be performed offline. The very fast testing suggests that our model could be used in a routine clinical workflow.

**Comparison to other SSL methods.** Our MK-SSAC model has been compared to three SSL models - AAE (Makhzani et al., 2015), CatGAN (Springenberg, 2015), and the ladder network (Rasmus et al., 2015), in which parameters are shared between the unsupervised and supervised components. Specifically, both AAE and ladder network share parameters between their encoder and classification network, and CatGAN shares parameters between the discriminator and classification network. For a fair comparison, we used the same MV-KBC model but replaced our SSAC with AAE, CatGAN, and ladder network, respectively, in this experiment. Besides, we used the same encoder, decoder and inputs, the same discriminator in AAE and CatGAN, and the same training settings, including the optimizer, learning rate, weighting coefficient, and batch size.

**Comparison to hand-crafted feature-based methods.** Our MK-SSAC model has been compared to two state-of-the-art hand-crafted feature-based methods, which were abbreviated as Method I and II. Method I uses 3D gray-level co-occurrence matrix (GLCM)-based texture features to describe nodule appearance (Han et al., 2015), and Method II performs massive mining of multiply visual features (MVF), including the shape, margin sharpness, and GLCM-based texture features, for better representation of nodules (Dhara et al., 2016). Both methods were tested on our dataset five times independently using the 10-fold cross validation.

**Comparison to DCNN-based methods.** Our MK-SSAC model was also compared to two state-of-the-art deep learning methods, which were abbreviated as Method A (Xie et al., 2018b) and B (Xie et al., 2018a). Method A fuses the texture, shape and deep model-learned information (Fuse-TSD) at the decision level, and Method B provides an ensemble of 27 pre-trained and fine-tuned ResNet-50 models, each of which characterizes the OA, HVV and HS of nodules from one of nine planar views, respectively. It should

**Table 2**
Performance of our MK-SSAC model and three SSL models on the LIDC-IDRI dataset. "L" and "UL" are the number of labeled and unlabeled nodules.

| Methods | Nodules | Results (%) (mean+standard deviation) | | | |
|---|---|---|---|---|---|
| | L/UL | Accuracy | Sensitivity | Specificity | AUC |
| MK-CatGAN | 1945/1839 | 90.89 ± 0.15 | 81.61 ± 0.12 | 95.48 ± 0.10 | 93.76 ± 0.25 |
| MK-AAE | 1945/1839 | 91.13 ± 0.15 | 82.92 ± 0.20 | 95.19 ± 0.12 | 94.00 ± 0.21 |
| MK-Ladder Network | 1945/1839 | 92.11 ± 0.27 | 83.07+0.19 | **96.59** ± 0.10 | 95.36 ± 0.20 |
| MK-SSAC | 1945/1839 | **92.53** ± 0.05 | **84.94** ± 0.17 | 96.28 ± 0.08 | **95.81** ± 0.19 |

**Table 3**
Performance of our MK-SSAC model and two state-of-the-art hand-crafted feature-based methods on the same number of labeled lung nodules.

| | Methods | Nodules | Results (%) (mean standard deviation) | | | |
|---|---|---|---|---|---|---|
| | | L/UL | Accuracy | Sensitivity | Specificity | AUC |
| I | 3D GLCM+SVM (Han et al., 2015) | 1945/0 | 85.38 ± 0.10 | 70.20 ± 0.15 | 92.80 ± 0.20 | 88.19 ± 0.16 |
| II | MVF+SVM (Dhara et al., 2016) | 1945/0 | 87.90 ± 0.17 | 84.50 ± 0.19 | 89.09 ± 0.25 | 93.77 ± 0.15 |
| | MK-SSAC | 1945/1839 | **92.53** ± 0.05 | **84.94** ± 0.17 | **96.28** ± 0.08 | **95.81** ± 0.19 |

be noted that the performance reported in Method A was obtained on the same LIDC-IDRI dataset, but the model was trained with a different group of nodules. For a fair comparison, we ran the source code of Method A on our dataset five times independently using the 10-fold cross validation. As for Method B, since it was evaluated on the same LIDC-IDRI dataset with the same data selection scheme, we adopted its performance reported in (Xie et al., 2018a).

### 4.2. Evaluation criteria

The performance of benign–malignant lung nodule classification was assessed by the mean and standard deviation of obtained accuracy, sensitivity, specificity, and area under the receiver operator curve (AUC). Accuracy and AUC measure the overall classification performance, sensitivity and specificity give the proportion of malignant and benign nodules that are correctly identified, respectively, and AUC is sensitive to the imbalance between two classes. These metrics can be calculated as follows

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$AUC = \int_0^1 t_{pr}(f_{pr})df_{pr} = P(X_p > X_n) \tag{9}$$

where *TP, TN, FN* and *FP* represent the number of true positive, true negative, false negative and false positive, respectively. The true positive rate $t_{pr}$ is a function of the false positive rate $f_{pr}$ along the receiver operator curve, and $X_p$ and $X_n$ are the confidence scores for a positive and negative instance, respectively.

## 5. Results

### 5.1. Comparison to other SSL methods

The performance of the proposed MK-SSAC model and three SSL models was compared in Table 2, which shows that the MK-SSAC model achieved the highest accuracy, sensitivity, and AUC and the second highest specificity. Especially, the significant improvement in sensitivity (minimum 1.87%) is substantial. Since a higher sensitivity indicates a lower false negative rate, our model is more suitable for lung nodule screening and potentially more useful in clinical practice than other three SSL models. The results of this experiment suggest that using learnable T layers to transfer the image representation ability learned by the reconstruction network to the classification network outperforms the parameter sharing strategy, in which the image representation ability learned by a generative model is directly used in a discriminative model.

### 5.2. Comparison to hand-crafted feature-based methods

Table 3 presents the performance of our MK-SSAC model and two hand-crafted feature-based methods. Method I only uses a single feature, and hence achieved the lowest accuracy of 85.38%. Method II fuses three types of features and achieved a higher accuracy of 87.90%. Our model obtained the highest accuracy of 92.53% and substantially improved sensitivity, specificity, and AUC over other two methods. The results of this experiment show that the proposed deep learning method has distinct advantages over single and combined hand-crafted feature-based methods in lung nodule classification.

### 5.3. Comparison to DCNN-based methods

The performance of the proposed MK-SSAC model and two DCNN-based methods was compared in Table 4, which shows that

**Table 4**
Performance of our MK-SSAC model and two state-of-the-art DCNN-based methods on the same number of labeled lung nodules.

| | Methods | Nodules | Results (%) (mean standard deviation) | | | |
|---|---|---|---|---|---|---|
| | | L/UL | Accuracy | Sensitivity | Specificity | AUC |
| A | Fuse-TSD (Xie et al., 2018b) | 1945/0 | 88.73 ± 0.15 | 84.40 ± 0.20 | 90.88 ± 0.13 | 94.02 ± 0.20 |
| B | MV-KBC (Xie et al., 2018a) | 1945/0 | 91.60 ± 0.15 | **86.52** ± 0.25 | 94.00 ± 0.30 | 95.70 ± 0.24 |
| | MK-SSAC | 1945/1839 | **92.53** ± 0.05 | 84.94 ± 0.17 | **96.28** ± 0.08 | **95.81** ± 0.19 |

**Table 5**
Results of paired $t$-test of seven different methods' accuracy and AUC. "R/A" represents rejecting or accepting the hypothesis $H_0$.

| Methods | | $\Delta_{Accuracy}$ | | | | $\Delta_{AUC}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\mu_\Delta$ | $\sigma_\Delta$ | t | R/A | $\mu_\Delta$ | $\sigma_\Delta$ | t | R/A |
| 1 | MK-SSAC *vs* 3D GLCM+SVM (Han et al., 2015) | 7.15 | 0.10 | 159.88 | **R** | 7.63 | 0.30 | 56.87 | **R** |
| 2 | MK-SSAC *vs* MVF+SVM (Dhara et al., 2016) | 4.64 | 0.20 | 51.88 | **R** | 2.05 | 0.24 | 19.1 | **R** |
| 3 | MK-SSAC *vs* Fuse-TSD (Xie et al., 2018b) | 3.81 | 0.18 | 47.33 | **R** | 1.79 | 0.28 | 14.29 | **R** |
| 4 | MK-SSAC *vs* MV-KBC (Xie et al., 2018a) | 0.93 | 0.19 | 10.94 | **R** | 0.12 | 0.06 | 4.47 | **R** |
| 5 | MK-SSAC *vs* MK-CatGAN (Springenberg, 2015) | 1.64 | 0.17 | 21.57 | **R** | 2.06 | 0.31 | 14.86 | **R** |
| 6 | MK-SSAC *vs* MK-AAE (Makhzani et al., 2015) | 1.41 | 0.17 | 18.55 | **R** | 1.81 | 0.38 | 10.65 | **R** |
| 7 | MK-SSAC *vs* MK-ladder network (Rasmus et al., 2015) | 0.42 | 0.23 | 4.08 | **R** | 0.46 | 0.25 | 4.11 | **R** |

our model achieved the highest accuracy, specificity and AUC and second highest sensitivity. It reveals that our previously proposed MV-KBC model achieved a relatively good performance, and replacing the ResNet-50 used in MV-KBC with the proposed SSAC model, which can be trained by using both labeled and unlabeled data in a semi-supervised way, yields further improved classification accuracy, specificity and AUC. Furthermore, it suggests that the end-to-end deep models (i.e., Method B and MK-SSAC) outperform the two-stage methods (i.e., Method A).

### 5.4. Statistical analysis

In this part, the paired $t$-test with a significant level $\alpha$ of 0.01 was adopted to determine whether the accuracy / AUC gain by the proposed MK-SSAC model over any one of the seven compared methods is statistically significant. We assumed that the image classification accuracies / AUC values of our MK-SSAC model and each compared method are random variables $X_1$ and $X_2$, respectively, each following a Gaussian distribution, i.e. $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$. The difference between $X_1$ and $X_2$ is defined as $\Delta = X_1 - X_2$. Thus, the hypotheses to be tested are $H_0$: $\mu_\Delta \leq 0$ versus $H_1$: $\mu_\Delta > 0$. Given $\alpha = 0.01$, $t_\alpha(n-1) = t_\alpha(5-1) = 3.747$ and the rejection domain is $W = t > 3.747$, where $t = \frac{\sqrt{n}\mu_\Delta}{\sigma_\Delta}$ and $n$ is the number of independent trials.

Table 5 gives the results of paired $t$-test. It shows that all $t$ values belong to the rejection domain $W$. Thus we rejected the hypothesis $H_0$ and accepted $H_1$ when compared with each method. It suggests that our MK-SSAC model achieved more accurate classification than each of seven compared methods, and the performance improvement is statistically significant. The results also prove that our MK-SSAC model outperforms significantly the fully supervised model (i.e. MV-KBC) and three SSL models (i.e. Method 5–7).

## 6. Discussion

### 6.1. Ablation studies

The contribution of the T layers that bridge the reconstruction network and classification network has been illustrated in the first experiment. Besides T layers, the reconstruction network itself and the adversarial learning used in it play a pivotal role in the proposed MK-SSAC model. To demonstrate the contributions of these two modules, we conducted ablation studies via constructing the MK-C model and MK-SSGC model. In MK-C, each of 27 submodels contains only the classification network (i.e. the reconstruction network was removed). In MK-SSGC model, the reconstruction network in each of 27 submodels is merely an autoencoder (i.e. the discriminator was removed).

The performance of our MK-SSAC model and these two variants was compared in Table 6. It shows that our model achieved higher accuracy, sensitivity, specificity, and AUC than MK-C and MK-SSGC. This result indicates that both the reconstruction network itself and the adversarial learning used in it contribute to the improvement of performance. Specifically, the performance gain obtain from the semi-supervised generator (MK-C vs MK-SSGC), adversarial training (MK-SSGC vs MK-SSAC) are 0.60% and 0.31% in accuracy, 1.35% and 0.30% in sensitivity, 0.31% and 0.21% in specificity, and 0.27% and 0.26% in AUC, respectively. Particularly, the semi-supervised generator can achieve a 1.35% improvement in sensitivity, which suggests that using unlabeled data to facilitate the training of deep models is not only effective, but also critical to make our model more suitable for lung nodule screening.

**Table 6**
Performance of the MK-SSAC model with different configurations.

| Methods | Nodules | Results (%) | | | |
|---|---|---|---|---|---|
| | L/UL | Accuracy | Sensitivity | Specificity | AUC |
| MK-C | 1945/0 | 91.62 | 83.29 | 95.76 | 95.28 |
| MK-SSGC | 1945/1839 | 92.22 | 84.64 | 95.97 | 95.55 |
| Our MK-SSAC | 1945/1839 | **92.53** | **84.94** | **96.28** | **95.81** |

We also qualitatively compared our MK-SSAC model to the MK-C and MK-SSGC models via visualizing the 54-dimensional features produced by the penultimate FC layer of each model. Fig. 5 shows the proximity and between boundary points computed using the t-distributed stochastic neighborhood embedding (t-SNE) (van der Maaten and Hinton, 2008). It reveals that jointly using the semi-supervised generator and adversarial training distinctively improves the separability between benign and malignant lung nodules.

### 6.2. Deep analysis for adversarial training

In this section, we further analyzed the reason why adversarial training can improve the classification performance. The adversarial training was used in the reconstruction network to jointly evaluate the discrepancy between the distributions of input images and reconstructed ones, aiming to avoid focusing only on minimizing the mean squared error (MSE), which is a pixel-wise measurement and sensitive to geometrical distortions. The reconstruction performance was assessed by measuring the MSE and peak-signal-to-noise-ratio (PSNR) of the images generated by 27 reconstruction networks. The results show that incorporating the adversarial training into the reconstruction process reduced the MSE from 0.09 to 0.08 and increased the PSNR from 15.06 to 17.61.

Meanwhile, we also randomly selected 10 patches from single view ROIs of lung nodules and visualized the corresponding reconstructed patches, which were generated by the MK-SSAC model with/without adversarial training. The comparison given in Fig. 6 shows that using adversarial training can generally improve the quality of reconstructed ROIs, and hence makes the reconstruction network have a stronger ability to characterize lung nodules. In this case, transferring the learned image representation ability of the reconstruction network to the classification network leads to a better classification performance than the model without using adversarial training.

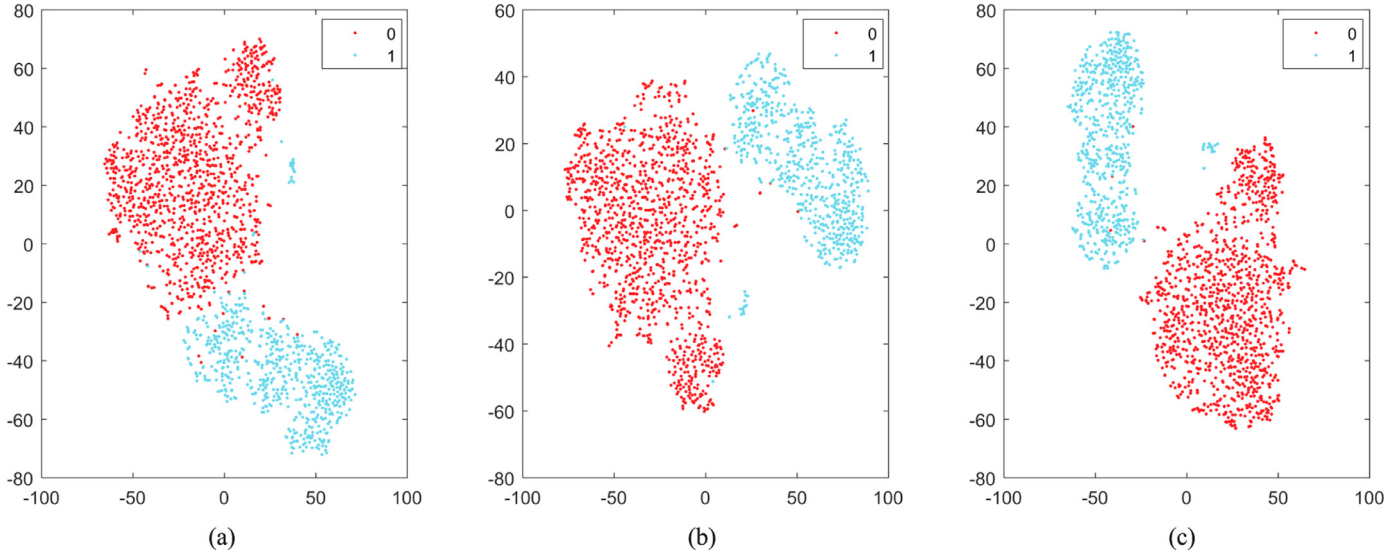


**Fig. 5.** T-SNE visualization of the 54-dimensional features obtained by the penultimate FC layer of (a) MK-C model, (b) MK-SSGC model, and (c) Our MK-SSAC model. Improved separability between the features from benign (red) and malignant (blue) nodules can be observed in (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

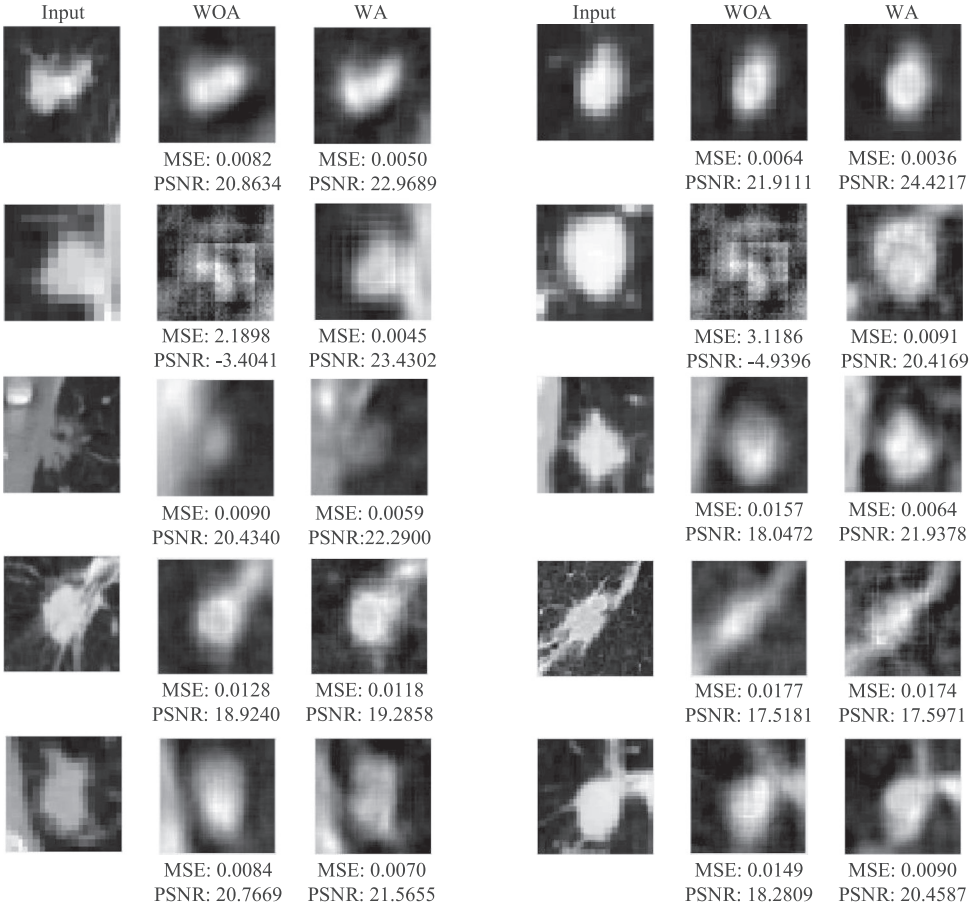


**Fig. 6.** 10 patches from single view ROIs of lung nodules (1st and 4th column) and the corresponding reconstructed patches generated by the MK-SSAC model without adversarial training (WOA) or with adversarial training (WA).

### *6.3. Trade-off between labeled and unlabeled data*

We suggested that jointly using labeled and unlabeled data in a semi-supervised way can alleviate the over-fitting of deep models trained with only the small labeled dataset, and thus improves the classification performance. To evaluate if this strategy still works on classification problems with an even smaller labeled dataset, we kept the testing dataset unchanged and randomly selected 80%, 60%, 40%, and 20% training nodules, respectively, from each fold of the labeled training data to train the proposed MK-SSAC model and the supervised MK-C model. The results in Fig. 7 show that the classification accuracy and AUC of both models improve with

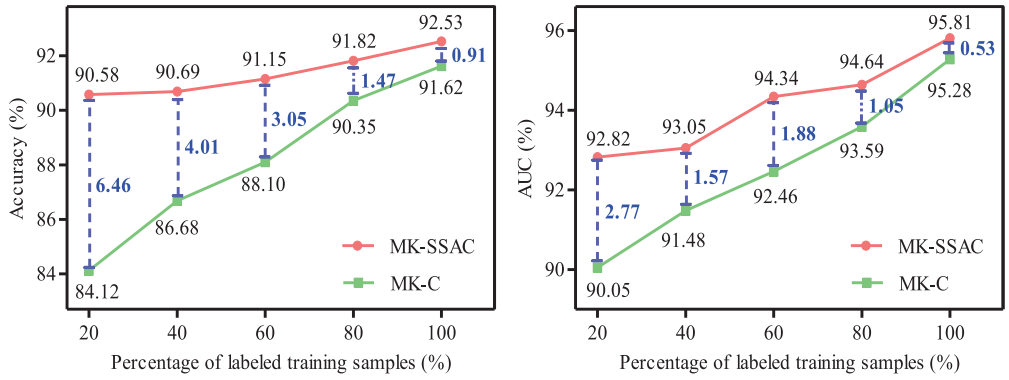


**Fig. 7.** Classification accuracy and AUC of the MK-C model and our MK-SSAC model when using different percentages of labeled training data.

the increasing percentage of labeled training data. Furthermore, with the unlabeled data, our semi-supervised model outperforms the fully-supervised one steadily no matter how many percentages of labeled training images are used. We also find that reducing the number of labeled training data to 20% leads the supervised MK-C model to more severe over-fitting, which results in the substantial drop of classification accuracy (from 91.62% to 84.12%) and AUC (from 95.28% to 90.05%). However, with the help of unlabeled data, our MK-SSAC model is relatively robust to the variation of the number of labeled training data, as its accuracy only drops moderately from 92.53% to 90.58% and its AUC drops from 95.81% to 92.82%. Meanwhile, it is interesting that the accuracy and AUC gain achieved by our MK-SSAC model increases from 0.91% to 6.46% and from 0.53% to 2.77%, respectively, when the number of labeled training data decreases from 100% to 20%. It suggests that the less labeled training data we have, the greater role the unlabeled data plays, and the more significant the improvement of our method is. These results demonstrate that using unlabeled data can largely compensate the loss of image representation power caused by the small-data issue.

### 6.4. Analysis of the MV-KBC architecture

In our MK-SSAC model, we decomposed each 3D lung nodule into nine fixed planar views to learn the 3D characteristics of nodules, following the decomposition scheme proposed in Xie et al. (2018a) and Setio et al. (2016). We also employed three SSAC models to characterize each nodule's OA, HVV and HS (Xie et al., 2018a, 2017a) on each planar view, respectively, based on the prior knowledge that there is a high correspondence between a nodule's malignancy and its heterogeneity in shape and voxel values.

To demonstrate that extending the SSAC model to the MV-SSAC model is effective, three baseline models were implemented for comparison. The first baseline is the single-view SSAC without using the prior knowledge (SWO-SSAC) model, which characterizes only a nodule's OA in the transaxial plane. The second baseline is the single-view knowledge-based collaborative SSAC (SK-SSAC) model, which characterizes a nodule's OA, HVV and HS in the transaxial plane. The third baseline is the multi-view SSAC without knowledge (MWO-SSAC) model, which characterizes a nodule's OA from nine planar views. Table 7 gives the performance of three baseline models and the proposed MK-SSAC model. It shows that embedding SSAC model into the MV-KBC architecture can exploit more information of the nodule and achieve the best classification performance when compared to the single-view or without-prior-knowledge architecture.

**Table 7**
Performance of the MK-SSAC model and three baselines.

| Methods | Nodules | Results (%) | | | |
|---|---|---|---|---|---|
| | L/UL | Accuracy | Sensitivity | Specificity | AUC |
| SWO-SSAC | 1945/1839 | 90.42 | 81.98 | 94.60 | 94.49 |
| SK-SSAC | 1945/1839 | 91.33 | 84.55 | 94.69 | 95.15 |
| MWO-SSAC | 1945/1839 | 91.45 | 84.04 | 95.12 | 95.36 |
| MK-SSAC | 1945/1839 | **92.53** | **84.94** | **96.28** | **95.81** |

### 6.5. Robustness to LUNGx challenge dataset

The LUNGx challenge dataset (Armato III et al., 2015a) in the Cancer Imaging Archive (TCIA) contains 83 lung nodules, including 10 nodules (5 benign and 5 malignant nodules) for training and 73 nodules (37 benign and 36 malignant nodules) for testing. We performed two experiments on this dataset: (1) directly testing the trained MK-SSAC model on the 73 test nodules without fine-tuning, and (2) fine-tuning the trained MK-SSAC model using the 10 training nodules and then testing it on 73 nodules. Both experiments were performed 5 times independently. The obtained mean and standard deviation of the accuracy, sensitivity, specificity and AUC were given in Table 8. It shows that, comparing to the MV-KBC method, the trained MK-SSAC model improves the average accuracy by 0.54% (from 75.62% to 76.16%) and average AUC by 0.74% (from 76.85% to 77.59%). It not only suggests that our model trained by both labeled and unlabeled data is superior over the fully supervised method, but also proves that our model has a strong generalization ability. By further fine-tuning the trained model with the LUNGx training dataset, our method achieves the highest average accuracy of 77.26% and average AUC of 78.83%, substantially higher than the average AUC of 12 best-performed methods.

### 6.6. Robustness to ISIC-2017 challenge dataset

The dataset of 2017 International Skin Imaging Collaboration (ISIC) skin lesion classification challenge (Codella et al., 2018) consists of 2000 training, 150 validation, and 600 test dermoscopy images. Each image is paired with the gold standard (definitive) lesion diagnosis, including the melanoma, nevus, and seborrheic keratosis. We also collected 1320 additional dermoscopy images from the ISIC Archive[2] as our unlabeled dataset.

To adapt our SSAC model to the skin lesion classification task, we redesigned the classification network $C$ and the reconstruction network $R$. Following the advanced skin lesion classi-

[2] https://isic-archive.com/.

**Table 8**
Performance of our MK-SSAC model and 12 best-performed methods on the LUNGX challenge testing dataset. The results of Methods 1 to 11 were adopted from Armato III et al. (2015a).

| No. | Nodule segmentation | Classifier | AUC (%) |
|---|---|---|---|
| 1 | Voxel-intensity-based segmentation | SVM | 50 ± 6.8 |
| 2 | Region growing | WEKA | 50 ± 5.6 |
| 3 | None required | Rules based on histogram-equalized pixel frequencies | 54 ± 6.7 |
| 4 | Bidirectional region growing | Uses tumor perfusion surrogate | 54 ± 6.6 |
| 5 | Region growing | WEKA | 55 ± 6.7 |
| 6 | Graph-cut-based surface detection | Random forest | 56 ± 5.4 |
| 7 | Gray-level thresholding + morphological operations | SVM | 59 ± 6.6 |
| 8 | None required | Convolutional neural network | 59 ± 5.3 |
| 9 | GrowCut region growing | SVM | 61 ± 5.4 |
| 10 | Radiologist-provided | Discriminant function | 66 ± 6.3 |
| 11 | Semi-automated thresholding | Support vector regressor | 68 ± 6.2 |

| 12 | MV-KBC (Xie et al., 2018a) | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC (%) |
|---|---|---|---|---|---|
| | | 75.62 ± 1.15 | 87.22 ± 7.24 | 64.32 ± 7.00 | 76.85 ± 0.17 |
| - | MK-SSAC | 76.16 ± 0.75 | 86.11 ± 7.61 | 66.49 ± 8.46 | 77.59 ± 0.39 |
| - | Fine-tuned MK-SSAC | **77.26** ± 0.75 | **87.22** ± 7.25 | **67.57** ± 6.34 | **78.83** ± 0.75 |

**Table 9**
Performance of our SSAC model and other methods on the ISIC-2017 testing dataset. "M" and "SK" represent melanoma and seborrheic keratosis, respectively.

| Methods | Nodules | M Classification | | | | SK Classification | | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| | L/UL | Acc | Sen | Spe | AUC | Acc | Sen | Spe | AUC | AUC (%) |
| Ours | 2000/1320 | 83.5 | 55.6 | 90.3 | 87.3 | 91.2 | 88.9 | 91.6 | 95.9 | 91.6 |
| Baseline (He et al., 2016) | 2000/0 | 83.0 | 53.8 | 90.1 | 82.7 | 88.8 | 82.2 | 90.0 | 94.3 | 88.5 |
| ARL-CNN (Zhang et al., 2019) | 2000/0 | 83.7 | 59.0 | 89.6 | 85.9 | 90.8 | 77.8 | 93.1 | 95.1 | 90.5 |
| #5 (Yang et al., 2017) | 2000/0 | 83.0 | 43.6 | 92.5 | 83.0 | 91.7 | 70.0 | 99.5 | 94.2 | 88.6 |
| ARL-CNN (Zhang et al., 2019) | 3320/0 | 85.0 | 65.8 | 89.6 | 87.5 | 86.8 | 87.8 | 86.7 | 95.8 | 91.7 |
| #1 (Matsunaga et al., 2017) | 3444/0 | 82.8 | 73.5 | 85.1 | 86.8 | 80.3 | 97.8 | 77.3 | 95.3 | 91.1 |
| #2 (Díaz, 2017) | 2900/0 | 82.3 | 10.3 | 99.8 | 85.6 | 87.5 | 17.8 | 99.8 | 96.5 | 91.0 |
| #3 (Menegola et al., 2017) | 9544/0 | 87.2 | 54.7 | 95.0 | 87.4 | 89.5 | 35.6 | 99.0 | 94.3 | 90.8 |
| #4 (Bi et al., 2017) | 3600/0 | 85.8 | 42.7 | 96.3 | 87.0 | 91.8 | 58.9 | 97.6 | 92.1 | 89.6 |

fication method (Zhang et al., 2019), we employed the ResNet-50 network (He et al., 2016) pre-trained on the ImageNet dataset (Deng et al., 2009) as the backbone of C for a better initialization. To adapt ResNet-50 to our classification task, we kept only three output neurons in the last FC layer, and the weights of the modified FC layer were randomly initialized. The reconstruction network $R$ consists of a generator $G$ (i.e. an encoder + a decoder) and a discriminator $D$. The encoder has the same architecture as the feature learning part (i.e. before the GAP layer) of $C$. The decoder (see Fig. 8) consists of four upsampling layers, eight conv-bn-relu blocks and a conv-sigmoid block. The backbone of discriminator $D$ is the pre-trained ResNet-50, but has only two output neurons in the last FC layer. We designed 17 learnable T layers to bridge $R$ and $C$ by transferring the feature maps obtained by the conv-bn-relu block, four conv blocks, and 12 identity blocks of $R$ to the corresponding blocks of $C$.

We used the same settings, including the optimizer, learning rate, weighting coefficient, and batch size as described in Section 3.1 to train the SSAC model. To alleviate the overfitting of deep models, we employed the same data argumentation method as Zhang et al. (2019) to enlarge the training dataset. We utilized the ISIC-2017 validation set to monitor the performance of SSAC model and terminated the training process when the network falls into overfitting. Four performance metrics (i.e. accuracy, sensitivity, specificity, and AUC) were used to assess the results separately for melanoma classification task (i.e. melanoma vs. others) and seborrheic keratosis classification task (i.e. seborrheic keratosis vs. others). Note that the ISIC-2017 challenge ranks the performance of each method on the testing dataset according to the average AUC across both categories.

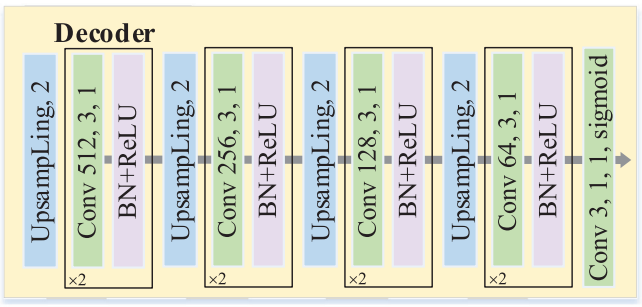


**Fig. 8.** Detailed architecture of the decoder.

We compared our SSAC model to the baseline ResNet50 model (He et al., 2016), the advanced ARL-CNN model (Zhang et al., 2019) and five top-ranking methods in the ISIC-2017 skin lesion classification challenge leader board (Matsunaga et al., 2017; Díaz, 2017; Menegola et al., 2017; Bi et al., 2017; Yang et al., 2017). The obtained performance metrics were compared in Table 9. It shows that our SSAC model that jointly uses 2000 labeled and 1320 unlabeled data attained an average AUC of 91.6%, noticeably higher than the AUC of the baseline, ARL-CNN and #5 method, which were trained with the same labeled dataset. Meanwhile, compared to the methods using external labeled data, our model achieved the second highest average AUC, only 0.1% less than average AUC obtained by the ARL-CNN model, which was trained with 3320 labeled data.

## 7. Conclusion

We presented the MK-SSAC model to differentiate malignant lung nodules from benign ones on chest CT by proposing a novel semi-supervised strategy to effectively use the unlabeled nodules and taking into account the nodule's heterogeneity in shape and

voxel values on nine planar views. Experimental results on the LIDC-IDRI dataset demonstrate the effectiveness of our MK-SSAC model for improving the state-of-the-art nodule classification systems. Although our model is built upon the specific application of lung nodule classification, our semi-supervised strategy itself is generic and can be applied to other deep learning-based medical image classification tasks to reduce the requirement of a large number of annotated medical images or to further improve the classification performance. In our future work, we will focus on applying the reinforcement learning to optimize automatically the architecture of the reconstruction network and classification network and the settings of hyper-parameters, such as the weighting factor $\lambda$, batch size, and learning rate, aiming to make the proposed MK-SSAC model more accurate and more efficient. Moreover, it would be necessary to study the incorporation of other information about the subjects (e.g. pathological data, biomarker, and diagnostic reports) into the deep learning model for more accurate benign–malignant lung nodule classification.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61771397, in part by the Science and Technology Innovation Committee of Shenzhen Municipality, China, under Grants JCYJ20180306171334997, in part by Synergy Innovation Foundation of the University and Enterprise for Graduate Students in Northwestern Polytechnical University under Grants XQ201911, and in part by the Project for Graduate Innovation team of Northwestern Polytechnical University. We acknowledged the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC-IDRI Database used in this work.

## Conflict of interest

There are no potential conflicts of interest to report.

## References

Alilou, M., Orooji, M., Madabhushi, A., 2017. Intra-perinodular textural transition (IPRIS): a 3D descriptor for nodule diagnosis on lung ct. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer International Publishing, Cham, pp. 647–655. doi:10.1007/978-3-319-66179-7_74.

Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Beek, E.J.R., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P.-Y., Roberts, R.Y., Smith, A.R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Vande Casteele, A., Gupte, S., Sallam, M., Heath, M.D., Kuhn, M.H., Dharaiya, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y., Clarke, L.P., 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on ct scans. Med. Phys. 38 (2), 915–931. doi:10.1118/1.3528204.

Armato III, S.G., Hadjiiski, L., Tourassi, G.D., Drukker, K., Giger, M.L., Li, F., Redmond, G., Farahani, K., Kirby, J.S., Clarke, L.P., 2015a. Lungx challenge for computerized lung nodule classification: reflections and lessons learned. J. Med. Imaging 2 (2).

Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Clarke, L. P., 2015b. Data from LIDC-IDRI. The cancer imaging archive.

Bach, P.B., Mirkin, J.N., Oliver, T.K., 2012. Benefits and harms of CT screening for lung cancer: a systematic review. JAMA 307 (22), 2418–2429. doi:10.1001/jama.2012.5521.

Baur, C., Albarqouni, S., Navab, N., 2017. Semi-supervised deep learning for fully convolutional networks. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer International Publishing, Cham, pp. 311–319. doi:10.1007/978-3-319-66179-7_36.

Bi, L., Kim, J., Ahn, E., Feng, D., 2017. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv:1703.04197.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. doi:10.1023/A:1010933404324.

Buty, M., Xu, Z., Gao, M., Bagci, U., Wu, A., Mollura, D.J., 2016. Characterization of lung nodule malignancy using hybrid shape and appearance features. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer International Publishing, Cham, pp. 662–670. doi:10.1007/978-3-319-46720-7_77.

Chapelle, O., Scholkopf, B., A. Zien, E., 2009. Semi-supervised learning. IEEE Trans. Neural Netw. 20 (3). doi:10.1109/TNN.2009.2015974. 542–542

Cheng, Y., Zhao, X., Cai, R., Li, Z., Huang, K., Rui, Y., 2016. Semi-supervised multimodal deep learning for RGB-D object recognition. In: Proc. of the 25th International Joint Conference on Artificial Intelligence (IJCAI), pp. 3345–3351.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F., 2013. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging 26 (6), 1045–1057. doi:10.1007/s10278-013-9622-7.

Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al., 2018. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 168–172.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. doi:10.1007/BF00994018.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. doi:10.1109/CVPR.2009.5206848.

Dey, R., Lu, Z., Hong, Y., 2018. Diagnostic classification of lung nodules using 3D neural networks. In: IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 774–778. doi:10.1109/ISBI.2018.8363687.

Dhara, A.K., Mukhopadhyay, S., Dutta, A., Garg, M., Khandelwal, N., 2016. A combination of shape and texture features for classification of pulmonary nodules in lung ct images. J. Digit. Imaging 29 (4), 466–475. doi:10.1007/s10278-015-9857-6.

Díaz, I. G., 2017. Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. arXiv:1703.01976.

Dou, Q., Chen, H., Jin, Y., Lin, H., Qin, J., Heng, P.-A., 2017a. Automated pulmonary nodule detection via 3D convnets with online sample filtering and hybrid-loss residual learning. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer International Publishing, Cham, pp. 630–638. doi:10.1007/978-3-319-66179-7_72.

Dou, Q., Chen, H., Yu, L., Qin, J., Heng, P., 2017b. Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection. IEEE Trans. Biomed. Eng. 64 (7), 1558–1567. doi:10.1109/TBME.2016.2613502.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.-A., 2017c. 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. 41, 40–54. doi:10.1016/j.media.2017.05.001.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 2672–2680.

Haeusser, P., Mordvintsev, A., Cremers, D., 2017. Learning by association a versatile semi-supervised training method for neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 626–635. doi:10.1109/CVPR.2017.74.

Han, F., Wang, H., Zhang, G., Han, H., Song, B., Li, L., Moore, W., Lu, H., Zhao, H., Liang, Z., 2015. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. J. Digit. Imaging 28 (1), 99–115. doi:10.1007/s10278-014-9718-8.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Hecht-Nielsen, R., 1992. Theory of the backpropagation neural network. In: Neural Networks for Perception. Academic Press, p. 6593. doi:10.1016/B978-0-12-741252-8.50010-8.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507. doi:10.1126/science.1127647.

Hua, K.-L., Hsu, C.-H., Hidayati, S.C., Cheng, W.-H., Chen, Y.-J., 2015. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets Ther. 8, 20152022. doi:10.2147/OTT.S80733.

Hussein, S., Cao, K., Song, Q., Bagci, U., 2017a. Risk stratification of lung nodules using 3D CNN-based multi-task learning. In: Information Processing in Medical Imaging (IPMI). Springer International Publishing, Cham, pp. 249–260. doi:10.1007/978-3-319-59050-9_20.

Hussein, S., Gillies, R., Cao, K., Song, Q., Bagci, U., 2017b. Tumornet: lung nodule characterization using multi-view convolutional neural network with gaussian process. In: IEEE 14th International Symposium on Biomedical Imaging (ISBI), pp. 1007–1010. doi:10.1109/ISBI.2017.7950686.

Jia, H., Xia, Y., Song, Y., Cai, W., Fulham, M., Feng, D.D., 2018. Atlas registration and ensemble deep convolutional neural network-based prostate segmentation using magnetic resonance imaging. Neurocomputing 275, 1358–1369. doi:10.1016/j.neucom.2017.09.084.

Kingma, D. P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980.

Li, X., Dou, Q., Chen, H., Fu, C.-W., Qi, X., Belav, D.L., Armbrecht, G., Felsenberg, D., Zheng, G., Heng, P.-A., 2018. 3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images. Med. Image Anal. 45, 41–54. doi:10.1016/j.media.2018.01.004.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Snchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. doi:10.1016/j.media.2017.07.005.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605. doi: 10.1.1.457.7213.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B., 2015. Adversarial autoencoders. arXiv:1511.05644.

Matsunaga, K., Hamada, A., Minagawa, A., Koga, H., 2017. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv:1703.03108.

Menegola, A., Tavares, J., Fornaciali, M., Li, L. T., Avila, S., Valle, E., 2017. Recod titans at ISIC challenge 2017. arXiv:1703.04819.

Metz, S., Ganter, C., Lorenzen, S., van Marwick, S., Holzapfel, K., Herrmann, K., Rummeny, E.J., Wester, H.-J., Schwaiger, M., Nekolla, S.G., Beer, A.J., 2015. Multiparametric mr and pet imaging of intratumoral biological heterogeneity in patients with metastatic lung cancer using voxel-by-voxel analysis. PLoS One 10 (7), 1–14. doi:10.1371/journal.pone.0132386.

Qin, Y., Zheng, H., Huang, X., Yang, J., Zhu, Y.M., 2019. Pulmonary nodule segmentation with ct sample synthesis using adversarial networks. Med. Phys. 46 (3), 1218–1229.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434.

Ranzato, M.A., Szummer, M., 2008. Semi-supervised learning of compact document representations with deep networks. In: Proceedings of the 25th International Conference on Machine Learning (ICML). ACM, pp. 792–799. doi:10.1145/1390156.1390256.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T., 2015. Semi-supervised learning with ladder networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 3546–3554.

Rojas, R., 1996. The Backpropagation Algorithm. Springer Berlin Heidelberg, pp. 149–182. doi:10.1007/978-3-642-61068-4_7.

Sakamoto, M., Nakano, H., Zhao, K., Sekiyama, T., 2018. Lung nodule classification by the combination of fusion classifier and cascaded convolutional neural networks. In: IEEE 15th International Symposium on Biomedical Imaging (ISBI), pp. 822–825. doi:10.1109/ISBI.2018.8363698.

Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S.J., Wille, M.M.W., Naqibullah, M., Snchez, C.I., van Ginneken, B., 2016. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. IEEE Trans. Med. Imaging 35 (5), 1160–1169. doi:10.1109/TMI.2016.2536809.

Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J., 2015. Multi-scale convolutional neural networks for lung nodule classification. In: Information Processing in Medical Imaging (IPMI). Springer International Publishing, Cham, pp. 588–599. doi:10.1007/978-3-319-19992-4_46.

Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., Zang, Y., Tian, J., 2017. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognit. 61, 663–673. doi:10.1016/j.patcog.2016.05.029.

Slatore, C.G., Wiener, R.S., Laing, A.D., 2016. What is a lung nodule? Am. J. Respir. Crit. Care Med. 193 (7), 11–12. doi:10.1164/rccm.1937P11.

Springenberg, J. T., 2015. Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv:1511.06390.

Torre, L.A., Siegel, R.L., Jemal, A., 2016. Lung Cancer Statistics. Springer International Publishing, Cham, pp. 1–19. doi:10.1007/978-3-319-24223-1_1.

Vigneault, D.M., Xie, W., Ho, C.Y., Bluemke, D.A., Noble, J.A., 2018. Omega-net: fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks. Med. Image Anal. 48, 95–106. doi:10.1016/j.media.2018.05.008.

Wang, S., Zhou, M., Liu, Z., Liu, Z., Gu, D., Zang, Y., Dong, D., Gevaert, O., Tian, J., 2017. Central focused convolutional neural networks: developing a data-driven model for lung nodule segmentation. Med. Image Anal. 40, 172–183. doi:10.1016/j.media.2017.06.014.

Wu, G.X., Raz, D.J., 2016. Lung Cancer Screening. Springer International Publishing, Cham, pp. 1–23. doi:10.1007/978-3-319-40389-2_1.

Xie, Y., Xia, Y., Zhang, J., Feng, D.D., Fulham, M., Cai, W., 2017a. Transferable multi-model ensemble for benign-malignant lung nodule classification on chest ct. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer International Publishing, Cham, pp. 656–664. doi:10.1007/978-3-319-66179-7_75.

Xie, Y., Xia, Y., Zhang, J., Song, Y., Feng, D., Fulham, M., Cai, W., 2018a. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct. IEEE Trans. Med. Imaging doi:10.1109/TMI.2018.2876510. 1–1

Xie, Y., Zhang, J., Liu, S., Cai, W., Xia, Y., 2017b. Lung nodule classification by jointly using visual descriptors and deep features. In: Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging. Springer International Publishing, Cham, pp. 116–125. doi:10.1007/978-3-319-61188-4_11.

Xie, Y., Zhang, J., Xia, Y., Fulham, M., Zhang, Y., 2018b. Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. Inf. Fusion 42, 102–110. doi:10.1016/j.inffus.2017.10.005.

Yan, X., Pang, J., Qi, H., Zhu, Y., Bai, C., Geng, X., Liu, M., Terzopoulos, D., Ding, X., 2017. Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: a comparison between 2D and 3D strategies. In: ACCV 2016 Workshops. Springer International Publishing, Cham, pp. 91–101. doi:10.1007/978-3-319-54526-4_7.

Yang, X., Zeng, Z., Yeo, S. Y., Tan, C., Tey, H. L., Su, Y., 2017. A novel multi-task deep learning model for skin lesion segmentation and classification. arXiv:1703.01025.

Zhang, J., Xia, Y., Xie, Y., Fulham, M., Feng, D.D., 2018. Classification of medical images in the biomedical literature by jointly using deep and handcrafted visual features. IEEE J. Biomed. Health Inform. 22 (5), 1521–1530. doi:10.1109/JBHI.2017.2775662.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2019. Attention residual learning for skin lesion classification. IEEE Trans. Med. Imaging.

Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z., 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer International Publishing, Cham, pp. 408–416. doi:10.1007/978-3-319-66179-7_47.

Zhu, X., 2006. Semi-Supervised Learning Literature Survey, 2. Computer Science, University of Wisconsin-Madison, p. 4.