# Knowledge-based Collaborative Deep Learning for Benign-Malignant Lung Nodule Classification on Chest CT

Yutong Xie, Yong Xia, *Member, IEEE*, Jianpeng Zhang, Yang Song, *Member, IEEE*, Dagan Feng, *Fellow, IEEE*, Michael Fulham, and Weidong Cai, *Member, IEEE*

***Abstract*— The accurate identification of malignant lung nodules on chest CT is critical for the early detection of lung cancer, which also offers patients the best chance of cure. Deep learning methods have recently been successfully introduced to computer vision problems, although substantial challenges remain in the detection of malignant nodules due to the lack of large training data sets. In this paper, we propose a multi-view knowledge-based collaborative (MV-KBC) deep model to separate malignant from benign nodules using limited chest CT data. Our model learns 3-D lung nodule characteristics by decomposing a 3-D nodule into nine fixed views. For each view, we construct a knowledge-based collaborative (KBC) submodel, where three types of image patches are designed to fine-tune three pre-trained ResNet-50 networks that characterize the nodules' overall appearance, voxel, and shape heterogeneity, respectively. We jointly use the nine KBC submodels to classify lung nodules with an adaptive weighting scheme learned during the error back propagation, which enables the MV-KBC model to be trained in an end-to-end manner. The penalty loss function is used for better reduction of the false negative rate with a minimal effect on the overall performance of the MV-KBC model. We tested our method on the benchmark LIDC-IDRI data set and compared it to the five state-of-the-art classification approaches. Our results show that the MV-KBC model achieved an accuracy of 91.60% for lung nodule classification with an AUC of 95.70%. These results are markedly superior to the state-of-the-art approaches.**

***Index Terms*— Lung nodule classification, deep learning, collaborative learning, computed tomography (CT).**

## I. Introduction

THE 2015 Global Cancer Statistics show that lung cancer accounts for approximately 13% of 14.1 million new cancer cases and 19.5% of cancer-related deaths each year [1]. The 5-year survival for patients who present 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 before it has spread [2]. Early lung cancer detection therefore offers the best chance for cure.

The National Lung Screening Trial [2], [3] shows that screening with CT results in a 20% reduction in lung cancer deaths through the identification of early disease. A "spot on the lung" on a chest CT is defined as a lung nodule, and it can be benign or malignant [4]. 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 requires a high degree of skill and concentration, and is time-consuming, expensive, and prone to operator bias. Although computer-aided diagnosis systems (CADs) have been employed to assist radiologists in reading chest CT scans, automated identification of benign and malignant nodules on chest CTs remains problematic due to at least two reasons: the difficulty of lung nodule delineation caused by a large range of nodule shape and texture variation and the visual similarities shared by malignant and benign nodules. As a results, non-professionals can have difficulty in separating them.

Lung CADs typically: (1) segment nodules from the background, (2) extract features from each segmented nodule and, (3) use the features to train a classifier to characterize the nodule as benign or potentially malignant. Traditional lung nodule segmentation methods involve lung segmentation and the detection and segmentation of a region of interest (ROI) that includes the nodule. These methods can be generally categorized as morphologic [5], [6], [7], region growing [8], energy optimization [9], [10] and statistical learning based methods [11], [12]. To address the difficulty in nodule

Manuscript received August 10, 2018; revised October 3, 2018; accepted October 8, 2018. Date of publication October 17, 2018; date of current version April 2, 2019. This work was supported in part by the National Natural Science Foundation of China under Grants 61771397 and 61471297 and in part by the Australian Research Council (ARC) Grants. *(Corresponding author: Yong Xia).*

Y. Xie, Y. Xia, and J. Zhang are with the Shaanxi Key Lab of Speech and Image Information Processing, Centre for Multidisciplinary Convergence Computing, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: xuyongxie@mail.nwpu.edu.cn; yxia@nwpu.edu.cn; james.zhang@mail.nwpu.edu.cn.)

Y. Song, D. Feng, and W. Cai are with the Biomedical and Multimedia Information Technology Research Group, School of Information Technologies, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: yang.song@sydney.edu.au; dagan.feng@sydney.edu.au; tom.cai@sydney.edu.au).

M. Fulham is with the Department of Molecular Imaging, Royal Prince Alfred Hospital, NSW 2050, Australia, also with the Sydney Medical School, The University of Sydney, Sydney, NSW 2006, Australia, and also with the Centre for Multidisciplinary Convergence Computing, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: michael.fulham@sydney.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2018.2876510

segmentation caused by attachments between nodules and other lung structures, Diciotti *et al.* [5] applied an automated correction method, which is based on a local shape analysis using 3D geodesic distance map representations, to an initial rough segmentation of nodules. Song *et al.* [8] proposed a novel toboggan based growing automatic segmentation approach (TBGA), which included automatic initial seed point selection, multi-constraints 3D lesion extraction, and lesion refinement. Farag *et al.* [9] fused the image intensity statistical information in a variational level set framework for lung nodule segmentation. Wu *et al.* [11] used a conditional random field (CRF) model that incorporates texture, gray-level intensities, shape, and edge cues to improve the segmentation of nodule boundaries. Tao *et al.* [12] presented a multi-level statistical learning-based framework for automated detection and segmentation of ground glass nodules (GGN). After segmentation, feature extraction translates the lung nodule into a feature vector. Most commonly used features include texture descriptors, such as the gray level co-occurrence matrix (GLCM)-based features [13], [14], [15], local binary pattern (LBP) [16] and histogram of oriented gradients (HOG) [13], and shape descriptors such as the Fourier shape descriptor [17], [18] and spherical harmonics [19]. The extracted visual features can then be sued to train a classifier, such as the support vector machine (SVM) [20], K-nearest neighbor (KNN) [21], back propagation neural network (BPNN) [22], [23], and random forest [24]. Despite their prevalence, these lung CADs reply heavily on handcrafted features and classifiers.

Recently, deep learning techniques have achieved profound success in computer vision, since they provide a uniform feature extraction-classification framework to free users from troublesome handcrafted feature extraction [25], [26], [45], [52]–[56]. This success has prompted many investigators to employ deep convolutional neural networks (CNNs) in medical image analysis. For image segmentation, the fully convolutional network (FCN), which involves up-sampling layers to make the size of output match that of the input image, provides a new direction. Recently, Ronneberger *et al.* [27] reported a new FCN called U-Net for biomedical image segmentation with promising results. For lung nodule classification, Hua *et al.* [28] applied the deep CNN and deep belief network (DBN) to separate benign from malignant lung nodules and reported that deep learning achieved better discrimination. Kumar *et al.* [57] used auto-encoders and CNNs to classify lung nodules as malignant or benign, with an accuracy of 77.52%. Shafiee *et al.* [58] leveraged stochastic sequencers, consisting of three stochastically-formed convolutional layers, to obtain an accuracy of 84.49%. Hussein *et al.* [29] used an end-to-end trainable multi-view CNN (MV-CNN) for lung nodule characterization. Shen *et al.* [59] proposed a multi-scale CNN that captures lung nodule heterogeneity via extracting discriminative features from alternatingly stacked layers. They further extended this model to a multi-crop CNN [30] that is able to automatically extract salient nodule information via cropping different regions from convolutional feature maps and applying max-pooling at varying times. Hussein *et al.* [31] proposed 3D CNN multi-task learning for lung nodule characterization.

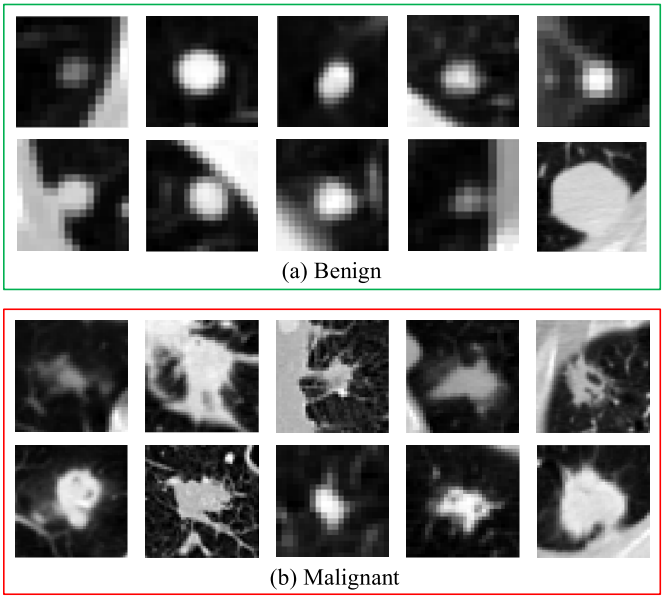
Although these deep learning techniques are more accurate than handcrafted features-based methods, they have not achieved the same performance on routine lung nodule classification as they have done in the ImageNet Challenge. The suboptimal performance is attributed mainly to the overfitting of deep models caused by inadequate training data, as there is usually a small dataset in medical image analysis and this relates to the work required in acquiring the image data and then in image annotation.

There are many attempts in the deep learning community to address the small data issue. First, it has been reported that the image representation ability learned from large-scale datasets, such as the ImageNet, can be transferred to generic visual recognition tasks, which have limited training data [32]. Hu *et al.* [33] proposed a deep transfer metric learning method to transfer discriminative knowledge from a labeled source domain to an unlabeled target domain to overcome this limitation. Shen *et al.* [34] used insufficient lung nodule data and formulated a domain-adaptation framework that learns transferable DCNN-based features for patient-level prediction of malignant lung nodules.

Second, although it is straightforward to design 3D CNN for medical image analysis [31], [35], [36], extending the use of 2D CNN to the analysis of volumetric medical images on a slice-by-slice basis, together with data augmentation, enables us to have more training samples [29], [30], [37], [38]. Volumetric data are firstly decomposed into fixed tri-planar views (sagittal, coronal, and axial planes). Thereafter, two strategies can be performed. First, all multi-view patches can be fed into a 2D CNN [29], [30]. Second, as suggested by Setio *et al.* [37], a 3D lung nodule can be decomposed into nine fixed view planes and be processed, using a multi-view architecture, in which each 2D CNN is trained with the image patches extracted on each plane, and the outputs of all CNNs are combined using the late-fusion strategy, i.e. performing fusion in a richer feature level.

Third, the prior domain knowledge can be incorporated into the solution to regularize the deep model. For example, there is a high correspondence between a nodule's malignancy and its heterogeneity (see Fig. 1) [39]. In our previous work [17], we used the GLCM-based texture descriptors and Fourier shape descriptor to explore the nodule's heterogeneity in voxel values (HVV) and heterogeneity in shapes (HS), respectively, and combined both descriptors with the information learned by a nine-layer DCNN for lung nodule classification at the decision level. Although we reported improved accuracy, this method still used hand-crafted features to characterize the heterogeneity of nodules, and they are less effective.

In this paper, we propose a multi-view knowledge-based collaborative (MV-KBC) deep neural network model for benign-malignant lung nodule classification on chest CT. We firstly decompose each 3D lung nodule into nine fixed views (sagittal, coronal, axial and six diagonal planes) to learn 3D nodule characteristics. Then, for each view, we construct a knowledge-based collaborative (KBC) submodel, where three types of image patches are designed to fine-tune three pre-trained ResNet-50 networks, aiming to transfer the image representation abilities of those ResNet-50 networks to characterizing the overall appearance (OA), HVV and HS



(a) Benign

(b) Malignant

Fig. 1. Examples of CT lung nodules in the axial plane. It shows that there is a high correspondence between a nodule's malignancy and its heterogeneity in voxel values (HVV) and heterogeneity in shapes (HS). (a) Benign. (b) Malignant.

of lung nodules, respectively. Finally, nine KBC submodels are used jointly to classify nodules with an adaptive weighting scheme learned during the error back propagation, thus enabling the MV-KBC model to be trained in an end-to-end manner. Furthermore, we also introduce the penalty loss function to manipulate the tradeoff between false positive rate and false negative rate of the MV-KBC model.

The contribution of the proposed MV-KBC model is threefold: (1) To the best of our knowledge, this work is one of the first to incorporate domain knowledge into a deep learning model for benign-malignant lung nodule classification. (2) Fusing multi-view (i.e. transverse, sagittal, coronal and six diagonal planes) / multi-appearance (i.e. OA, HS, and HVV) submodels at the decision level enables the entire model to be trained in an end-to-end manner, which avoids the troublesome setting of weighting coefficients and improves the classification accuracy. (3) The results suggest that our model provides a substantial performance improvement, and the fast online testing suggests that our model could be used in a routine clinical workflow.

A preliminary version of this work was presented in MICCAI 2017 [40]. In this paper, we have substantially revised and extended the original paper. The main extension includes decomposing each 3D lung nodule onto nine fixed view planes, using the patches extracted on each view plane to train a KBC submodel, hierarchical ensemble of 27 ResNet-50 networks and replacing the cross-entropy loss with the penalty loss function.

## II. Dataset

The LIDC-IDRI database [41]–[43] in the Cancer Imaging Archive (TCIA) contains 1018 clinical chest CT scans with lung nodules obtained from seven institutions. There is an associated XML file that details the locations of nodules on each 512×512 slice. The nodule diameters range from 3mm to 30mm. Each suspicious lesion is categorized as a non-nodule, a nodule <3 mm, or a nodule 3 mm diameter in the long axis. For this study, we only considered nodules 3 mm in diameter, since nodules <3 mm were not considered to be clinically relevant by current screening protocols [14], [15], [29], [30], [37], [44]. The malignancy of each nodule was evaluated with a 5-point scale, from benign to malignant, by up to four experienced thoracic radiologists. Following the procedures used in previous studies [14], [15], [29], [30], [44], we selected those nodules which were annotated by at least one radiologist for this study, calculated the median malignancy level (MML) of each nodule, and annotated a nodule whose MML <3 as benign, a nodule whose MML = 3 as uncertain, and a nodule whose MML > 3 as malignant. Thus there are 1301 benign, 612 uncertain, and 644 malignant nodules. To reduce the impact of uncertain evaluation of nodule malignancy, we excluded all 'uncertain lung nodules' from our experiments. The distribution of nodules over their MML and annotation is shown in TABLE I.

TABLE I
Median Malignancy Level (MML) in LIDC-IDRI Dataset

| Dataset | Benign | | Uncertain | Malignant | |
|---|---|---|---|---|---|
| MML | 1 | 2 | 3 | 4 | 5 |
| # of Nodules | 358 | 943 | 612 | 474 | 170 |

## III. Methods

The proposed MV-KBC algorithm consists of four major steps: (1) extracting 2D nodule slices from nine views of planes, (2) extracting the OA, HVV and HS patches on 2D nodule slices, (3) constructing nine KBC submodels and training each of them using the patches extracted on each view of planes, and constructing and training the MV-KBC model for lung nodule classification. A diagram that summarizes this algorithm was shown in Fig. 2.

### A. Multi-View Slice Extraction

Since chest CT scans have variable spatial resolution, we resampled them to a unified voxel size of 1.0×1.0×1.0 $mm^3$ using the spline interpolation [30]. We assumed that lung nodules had been detected, and hence limited the scope of this study solely to benign-malignant nodule classification. To avoid the inaccuracy caused by nodule detection, we defined the location of a nodule as the middle of the nodule's centers given by radiologists. For each lung nodule, we first cropped a 64×64×64 cube that is centered on its location such that the nodule is always contained completely in the cube. Then, we extracted nine 2D slices on the transverse, sagittal, coronal and six diagonal planes, respectively, where each diagonal plane cuts two opposite faces of the cube in diagonal and has two opposite edges of the cube and four vertices (see Fig. 2 (a)). Thus, we obtained nine views of slices for each nodule.

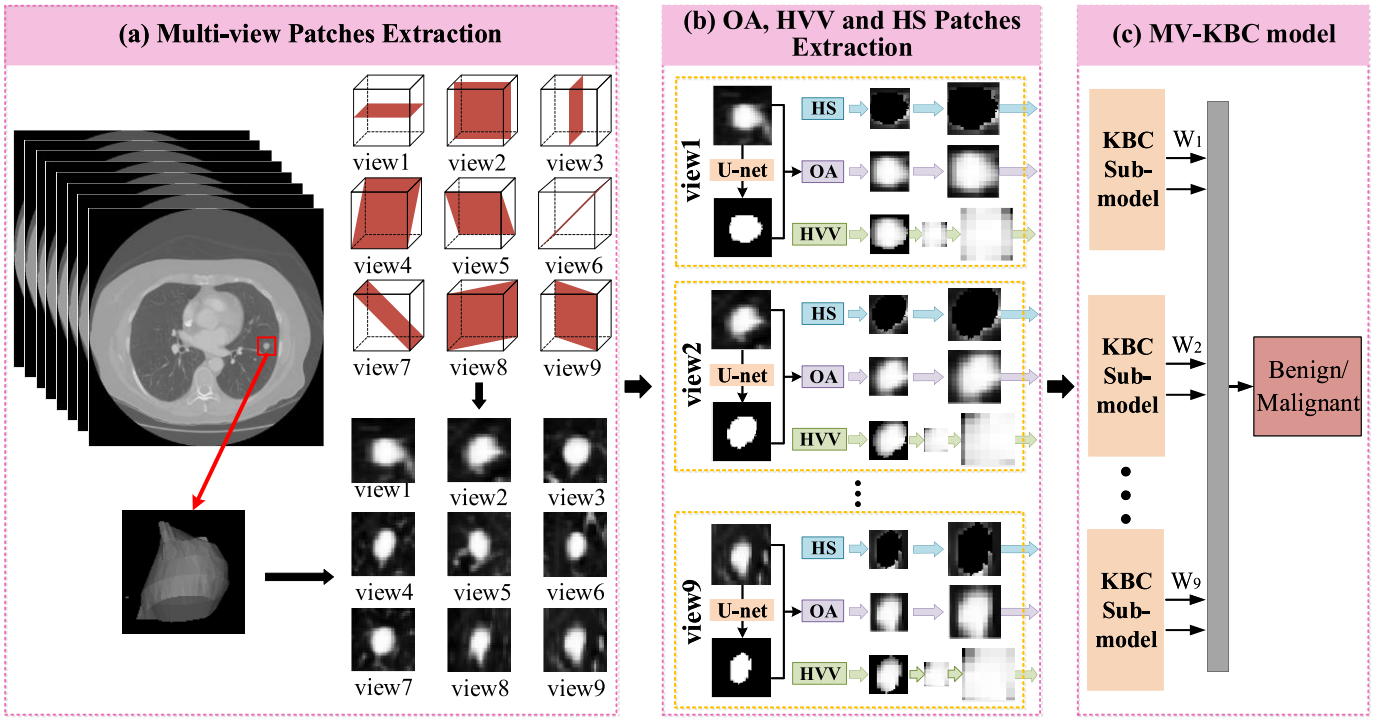



Fig. 2. Framework of our proposed MV-KBC algorithm.

## B. OA, HVV and HS Patch Extraction

The extraction of OA, HVV and HS patches is based on the segmentation of lung nodules on each slice. We adopted the U-Net network [27], a fully convolutional network (FCN) model, to segment the lung nodule. It has a contracting path and an expansive path (see Fig. 3). The contracting path follows the typical architecture of a convolutional neural network, in which there is the repeated application of two 3×3 padded convolutional layers, each followed by the ReLU function. Four 2×2 max pooling layers with a stride of 2 are used to downsample the obtained feature maps. Every step in the expansive path has an upsampling of the feature maps followed by a 2×2 convolutional layer, a concatenation with the corresponding feature map from the contracting path and two 3×3 convolutional layers, each followed by the ReLU function. The last layer is a 1×1 convolutional layer, which maps each 32-component feature vector to the desired number of classes.

We applied the U-Net to the LIDC-IDRI dataset with the 10-fold cross validation. Each of the first nine folds has 195 nodules, and the tenth fold has 190 nodules. Each time, one fold of nodules was used for testing, and others were used for training the U-Net. Hence, the testing set has never been used for U-Net training. All training images and their segmentation maps, defined as the intersection of the areas marked by radiologists, were used to train the network in an end-to-end manner to minimize the cross entropy loss. The mini-batch stochastic gradient descent algorithm with a batch size of 32 was adopted as the optimizer. The maximum epoch number was set to 100 and the learning rate was set to 0.001. Moreover, we randomly chose 10% of the training patches to form a validation set and terminate the training process even before reaching the maximum epoch number, if the error on the other 90% of training patches continues to decline but

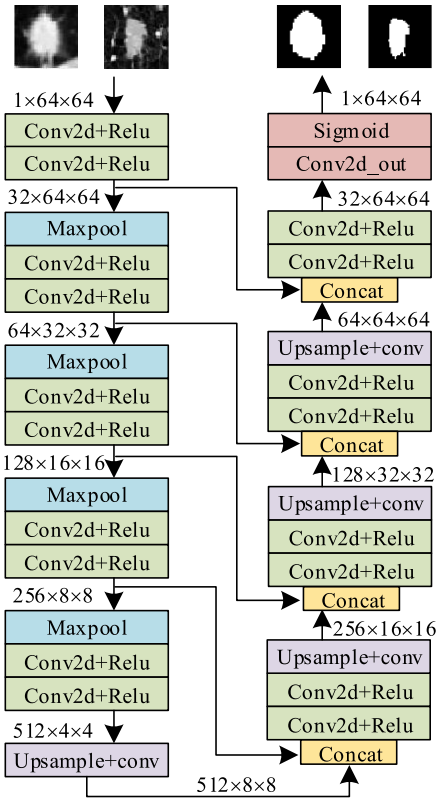



Fig. 3. Architecture of the U-Net used for lung nodule segmentation.

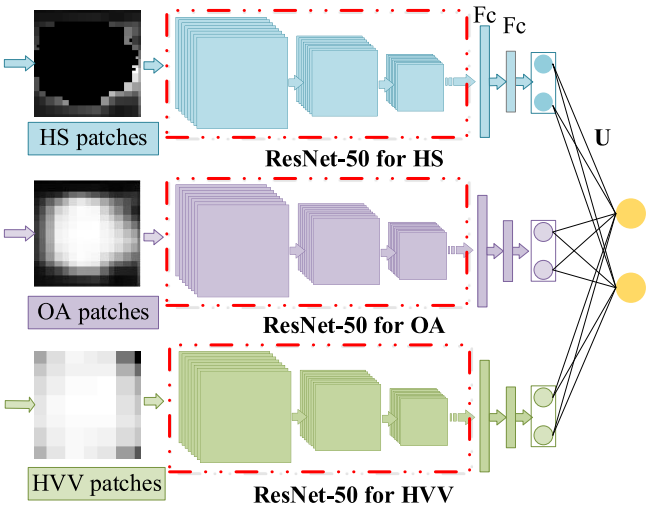



Fig. 4. Architecture of our proposed KBC submodel for a specific view.

the error on the validation set stops decreasing. At the testing stage, nodule segmentation was performed on a slice-by-slice basis by using the trained U-Net.

Based on nodule segmentation, a square ROI encapsulating the nodule on each slice was identified as an OA patch to represent the lung nodule's overall appearance. The OA patches obtained on different slices have variable sizes. 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 16×16, the average size of all OA patches, a 16×16 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.

Since data augmentation alleviates the overfitting of deep learning models by adding variants to the dataset [37], we generated four augmented data for each training patch using random image translation, rotation and horizontal or vertical flip [30]. The translation step was selected from [0, 6] voxels, and the rotation angle was randomly selected from {90°, 180°, 270°}. Then, all OA, HVV and HS patches were resized to 224×224.

### C. KBC Submodel

The OA, HVV or HS patches extracted on each of nine views of planes, together with the augmented data, were used to train a KBC submodel, which contains three pre-trained ResNet-50 networks [45] (see Fig. 4). The ResNet-50 network used for this study contains 50 learnable layers, including consequently a 7×7 convolutional layer that produces 64 feature maps, a 3×3 max pooling layer, four bottleneck architectures, an average pooling layer and a FC layer with 1000 neurons. Each bottleneck architecture consists of three convolutional layers with a map size of 1×1, 3×3, and 1×1, respectively (see TABLE II). The feature map channel increases for the 1st to 4th bottleneck layers, whereas the feature map size (i. e. output size) gradually decreases as the layer goes deeper.

To transfer the image representation ability learned on large scale image databases to characterizing lung nodules,

TABLE II
FOUR BOTTLENECK ARCHITECTURES OF THE RESNET-50

| Bottleneck layer | Replication | Three convolutional layers (1×1, 3×3, 1×1) | |
|---|---|---|---|
| | | Channels | Output Size |
| 1st | 3 | 64, 64, 256 | 56×56, 56×56, 56×56 |
| 2nd | 4 | 128, 128, 512 | 28×28, 28×28, 28×28 |
| 3rd | 6 | 256, 256, 1024 | 14×14, 14×14, 14×14 |
| 4th | 3 | 512, 512, 2048 | 7×7, 7×7, 7×7 |

the parameters used to initialize each ResNet-50 network have been converged by training with the ImageNet dataset [45], [46]. To adapt the ResNet-50 network to our benign-malignant nodule classification problem, we removed its last fully connected layer, and then added three fully connected layers with 2048, 1024 and 2 neurons, respectively. The weights of these three fully connected layers were randomly initialized by using Xaiver algorithm, and the activation function in the last layer was set to the sigmoid function. The modified ResNet-50 network was then fine-tuned in a layer-wise manner, starting with tuning only the last layer and finally tuning all layers.

For the $k$-th KBC submodel, let the $n$-th input patch triplet be denoted by $\left\{\boldsymbol{X}_{nk}^{(OA)}, \boldsymbol{X}_{nk}^{(HVV)}, \boldsymbol{X}_{nk}^{(HS)}\right\}$, and the corresponding output of three ResNet-50 networks be denoted by $\left\{\boldsymbol{O}_{nk}^{(OA)}, \boldsymbol{O}_{nk}^{(HVV)}, \boldsymbol{O}_{nk}^{(HS)}\right\}$, where $\boldsymbol{O}_{nk}^{(\#)} \in R^2$. Then, the output of the $k$-th KBC submodel can be calculated as

$$M_{nkj} = f\left(\sum\nolimits_{\#}\sum\nolimits_{i=1}^{2} U_{kij}^{(\#)} O_{nki}^{(\#)}\right) \tag{1}$$

where $\left\{U_{kij}^{(\#)}: \# \in \{OA, HVV, HS\}\right\}$ is the assemble of weights between the output layer of each ResNet-50 network and the output layer of the $k$-th KBC submodel. The parameter $i \in \{1, 2\}$ indicates the $i$-th neuron of the output layer in each ResNet-50 network. The summation over $i$ means the weighted sum of the outputs in each ResNet-50 network. The summation over # means the weighted sum of the outputs of three ResNet-50 networks. The parameter $j \in \{1, 2\}$ indicates the $j$-th neuron of the output layer in each KBC submodel. The function $f(\cdot)$ is a softmax activation function.

### D. MV-KBC Model

The proposed MV-KBC model consists of nine KBC submodels (see Fig. 2 (c)). The two-neuron output layer of each KBC submodel is connected to the same one-neuron classification layer followed by the sigmoid function. The output of this classification layer is the prediction made by the MV-KBC model, which can be formulated as

$$P_n = f\left(\sum\nolimits_{k=1}^{9}\sum\nolimits_{j=1}^{2} W_{kj} M_{nkj}\right) \tag{2}$$

where $\left\{W_{kj}: k = 1, 2, \cdots, 9; j \in \{1, 2\}\right\}$ is the assemble of weights between the output layer of each KBC submodel and the classification layer. The summation over $j$ means

the weighted sum of the output in each KBC submodel. The summation over $k$ means the weighted sum of the output of nine KBC submodels. The function $f(\cdot)$ is the sigmoid activation.

The cross-entropy loss is usually insensitive to the identity of the assigned class in case of misclassification [47]. However, misclassifying a malignant nodule as 'benign' (false negative) may be costlier than misclassifying a benign as 'malignant' (false positive), since the clinical practice may be falsely reassured that the nodule is 'benign' thus missing the opportunity to effectively treat an early lung tumor. To address this issue, we propose the following penalty cross-entropy loss that provides the means to distinguish between false negative and false positive nodules by penalizing each error differently.

$$l(y_n, P_n) = -\delta_n[y_n \log(P_n) + (1 - y_n)\log(1 - P_n)] \quad (3)$$

where the penalty factor

$$\delta_n = \begin{cases} C, & y_n - P_n > 0.5 \\ 1, & otherwise \end{cases} \quad (4)$$

For this study, we empirically set $C = 2$ to give a larger penalty to false negative cases.

Both groups of weights $\left\{U_{kij}^{(\#)}\right\}$ and $\{W_{kj}\}$ can be updated during the error back propagation process, and hence the MV-KBC model can be trained in an end-to-end way. The change of weight $\{W_{kj}\}$ is in proportion to descend along the min-batch gradient [47], [48], shown as follows:

$$\Delta W_{kj} = \frac{\eta}{b}\sum_{n=1}^{b} \delta_n(y_n - P_n)M_{nkj} \quad (5)$$

where $b$ is the batch size, and $\eta$ is learning rate.

We chose the min-batch stochastic gradient descent as the optimizer and set the maximum epoch number to 100. Since our training dataset is small, we employed the following variable learning rate scheme

$$\eta(t) = \frac{\eta(0)}{1 + 10^{-4}t} \quad (6)$$

where $t$ is the index of iterations, and the initial learning rate $\eta(0)$ is set to 0.0001. Moreover, we randomly choose 10% of the training patches to form a validation set and terminate the training process even before reaching the maximum epoch number, if the error on the other 90% of training patches continues to decline but the error on the validation set stops decreasing. The steps of training the proposed MV-KBC model was summarized in Algorithm I.

It should be noted that, although we used the ResNet-50 network for this study, our MV-KBC model allows a DCNN of any arbitrary structure to be embedded.

### E. Evaluation

We applied the MV-KBC model to the LIDC-IDRI dataset 5 times independently, with the 10-fold cross validation. The performance was assessed by the mean and standard deviation of obtained accuracy, sensitivity/recall, specificity, precision with the cut-off value of 0.5, F-score metric and area under the receiver operator curve (AUC) [49]. Accuracy shows the

**Algorithm 1** Batch-Based Learning Process of the Penalty Loss and $W_{kj}$

**Input:** Batch size: b = 32
True labels: $\boldsymbol{Y} = (y_1, y_2, \ldots, y_b)$
Prediction vector: $\boldsymbol{P} = (P_1, P_2, \ldots, P_b)$
Learning rate: $\eta$

**For** $n$=1 : $b$
penalty factor: $\delta_n = \begin{cases} C, & y_n - P_n > 0.5 \\ 1, & otherwise \end{cases}$
penalty loss: $l(y_n, P_n) = -\delta_n[y_n \log(P_n) + (1 - y_n)\log(1 - P_n)]$
**end**
Batch penalty loss: $L(\boldsymbol{Y}, \boldsymbol{P}) = \frac{\sum_{n=1}^{b}(l(y_n, P_n))}{b}$
Update $W_{kj} : k = 1, \ldots, 9;\ j \in \{1, 2\}$
$W_{kj}^{'} = W_{kj} + \Delta W_{kj}$

$$\Delta W_{kj} = -\eta\frac{\partial L(\boldsymbol{Y}, \boldsymbol{P})}{\partial W_{kj}} = -\eta\frac{\partial\frac{\sum_{n=1}^{b}(l(y_n, P_n))}{b}}{\partial W_{kj}}$$
$$= -\frac{\eta}{b}\sum_{n=1}^{b}\left(\frac{\partial\{-\delta_n[y_n\log(P_n) + (1-y_n)\log(1-P_n)]\}}{\partial W_{kj}}\right)$$
$$= -\frac{\eta}{b}\sum_{n=1}^{b}\left\{-\delta_n\frac{\partial P_n}{\partial W_{kj}}\left[\frac{y_n}{P_n} - \frac{(1-y_n)}{(1-P_n)}\right]\right\}$$
$$= \frac{\eta}{b}\sum_{n=1}^{b}\left\{\delta_n P_n(1 - P_n)M_{nkj}\left[\frac{y_n}{P_n} - \frac{(1-y_n)}{(1-P_n)}\right]\right\}$$
$$= \frac{\eta}{b}\sum_{n=1}^{b}\delta_n(y_n - P_n)M_{nkj}$$

performance of our model in classifying nodules as malignant or benign. Sensitivity and specificity measure the proportion of malignant and benign nodules that are correctly identified, respectively. Precision is the fraction of retrieved true positive instances among the retrieved positive instances. The F-score is a measure of a test's accuracy and considers precision and recall. The AUC is sensitive to imbalance among the classes.

We used the entire LIDC-IDRI dataset (i.e. 1301 benign and 644 malignant nodules) and evaluated our MV-KBC model against six lung nodule classification methods, which were abbreviated as method A, B, C, D, E and F have been described in introduction section. For the method C, D, E and F, we repeated the codes and tested them on our dataset 5 times independently using the 10-fold cross validation. For other compared methods, we did not have the code, and hence adopted its performance reported in published paper.

## IV. Results

### A. Comparisons in Benign-Malignant Classification

TABLE III shows the mean and standard deviations of the accuracy, sensitivity/recall, specificity, precision, F-score, and AUC of the proposed MV-KBC model and six other lung nodule classification methods. For methods A and B, we adopted its performance reported in the published papers. Though these lung nodules are from the same LIDC-IDRI dataset, the images used to train the models are different and method B used a larger training set than method A. Our MV-KBC model used the LIDC-IDRI dataset (i.e. 1301 benign and 644 malignant nodules) and achieved a best performance compared with method A and B.

We also compared our model with method C, D, E and F in our dataset. Method C only uses the 3D GLCM-based texture

TABLE III
PERFORMANCE OF SEVEN LUNG NODULE CLASSIFICATION METHODS. 'B' AND 'M' ARE THE NUMBER OF BENIGN AND MALIGNANT LUNG NODULES

| | Methods | Number | | Results (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | B | M | Accuracy | Sensitivity / Recall | Specificity | AUC | Precision | F score |
| A | Shen et al., 2017 [30] (Multi-crop CNN) | 528 | 297 | 87.14 | 77.00 | 93.00 | 93.00 | Not given | Not given |
| B | Hussein et al., 2017 [31] (3D CNN) | 635 | 509 | 91.26 | Not given | Not given | Not given | Not given | Not given |
| C | Han et al., 2015 [14] (3D GLCM feature+SVM) | 1301 | 644 | 85.38±0.10 | 70.20±0.15 | 92.80±0.20 | 88.19±0.16 | 82.85±0.38 | 75.99±0.10 |
| D | Dhara et al., 2016 [15] (Multi-visual features) | 1301 | 644 | 87.90±0.17 | 84.50±0.19 | 89.09±0.25 | 93.77±0.15 | 79.31±0.37 | 81.82±0.21 |
| E | Xie et al., 2018 [17] (Deep + visual features) | 1301 | 644 | 88.73±0.15 | 84.40±0.20 | 90.88±0.13 | 94.02±0.20 | 82.09±0.24 | 83.23±0.21 |
| F | Xie et al., 2017 [40] (TMME with Resnet-50) | 1301 | 644 | 91.01±0.10 | 83.83±0.15 | **94.56**±0.13 | 95.35±0.15 | **88.40**±0.24 | 86.07±0.15 |
| - | **Proposed MV-KBC** (mean±standard deviation) | 1301 | 644 | **91.60**±0.15 | **86.52**±0.25 | 94.00±0.30 | **95.70**±0.24 | 87.75±0.52 | **87.13**±0.16 |

features to describe nodule appearance, and hence achieved the lowest accuracy in this table. Method D performed massive mining of shape, margin sharpness and GLCM-based texture features for better representing nodules, and thus has a higher accuracy. Method E is our pervious method, which combined traditional visual features with deep features learned by a CNN and further improved accuracy. Our recent method F proposed the transferable multi-model ensemble (TMME) model which used three pre-trained and fine-tuned ResNet-50 networks to characterizing the OA, HVV and HS of lung nodule and avoided the adverse impact of insufficient training dataset and improper handcrafted features. Thus it can improve the performance of lung nodule classification compared to the methods C, D and E. Though there is a little lower specificity, our MV-KBC model can achieve a higher accuracy, sensitivity and AUC than Method F. It indicates that the MV-KBC model can characterize the OA, HVV and HS of lung nodules more effectively than method F by using the multi-view patches and penalty cross-entropy loss, instead of only axial plane patches and traditional cross-entropy loss. Furthermore, three methods (i.e., methods E, F and MV-KBC) using deep-learning-based features of lung nodules outperform the approaches (i.e., methods C and D) using hand-crafted features. It also reveals that integrating feature extraction and classifier training into a unified framework (as our MV-KBC model) can boost the performance of classification.

We also show 12 examples of the classification results produced by our MV-KBC model in Fig. 5. For each classification result, we provided its classification confidence value ($Conf$) under the image patch, which can be calculated as:

$$Conf_n = \begin{cases} 1 - P_n & benign \\ P_n & malignant \end{cases} \quad (7)$$

where $P_n$ is the prediction probability given by our model, as defined in Section III-$D$. The value of $Conf$ ranges from 0 (most insecure) to 1 (most confident). It shows in Fig. 5 that our model has a high confidence for the benign and malignant nodules in the left three columns, which have large visual differences, and a low confidence for the examples in the right three columns, which share many visual similarities. Nevertheless, it reveals that, although some benign nodules look similar to malignant ones, our MV-KBC model is still able to separate them. These results clearly demonstrate that our MV-KBC model has superior ability to differentiate malignant from benign nodules.

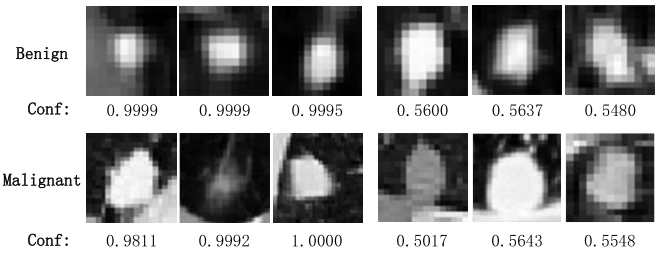



Fig. 5. Visualization of 12 examples of the classification results produced by our MV-KBC model, with the classification confidence value *Conf* being given beneath each example. Top row: 6 benign nodules; bottom row: 6 malignant nodules.

TABLE IV
FOUR PARTITIONS OF THE LIDC-IDRI DATASET

| Partition | True Samples | | False Samples | |
|---|---|---|---|---|
| | MML | # of Nodules | MML | # of Nodules |
| P1 | 1 | 358 | 2, 4, 5 | 1587 |
| P2 | 2 | 933 | 1, 4, 5 | 1002 |
| P3 | 4 | 474 | 1, 2, 5 | 1471 |
| P4 | 5 | 170 | 1, 2, 4 | 1775 |

## B. Comparisons in Subgroups

This experiment aims to evaluate the performance of our MV-KBC model in differentiating the nodules from each MML subgroup. To this end, we created four partitions of the LIDC-IDRI dataset, in each of which the nodules with a specific MML are annotated as true samples, and other nodules are annotated as false samples (see TABLE IV). We compared our MV-KBC model to the methods C, D, E and F using each of those data partitions and presented the results in TABLE V. It shows that our MV-KBC model achieved the highest accuracy, sensitivity, specificity and AUC on all those partitions. It further demonstrates that the MV-KBC model,

TABLE V
PERFORMANCE OF FIVE LUNG NODULE CLASSIFICATION METHODS ON EACH OF FOUR PARTITIONS OF THE LIDC-IDRI DATASET. ESPECIALLY, 'AC', 'SE', AND 'SP' ARE ACCURACY, SENSITIVITY AND SPECIFICITY

| | | Methods | Results (%) | | | |
|---|---|---|---|---|---|---|
| | | | Ac | Se | Sp | AUC |
| P1 | C | 3D GLCM+SVM [14] | 86.83 | 49.58 | 95.24 | 85.17 |
| | D | Multi-visual features [15] | 88.46 | 62.73 | 94.28 | 87.61 |
| | E | Deep + visual features [17] | 89.62 | 69.69 | 94.14 | 88.94 |
| | F | TMME (ResNet-50) [40] | 91.09 | 71.70 | 95.47 | 90.34 |
| | - | **Proposed MV-KBC** | **92.95** | **73.26** | **97.38** | **91.57** |
| P2 | C | 3D GLCM+SVM [14] | 74.23 | 72.5 | 75.85 | 77.03 |
| | D | Multi-visual features [15] | 76.13 | 78.53 | 73.88 | 78.08 |
| | E | Deep + visual features [17] | 78.71 | 83.37 | 74.33 | 81.03 |
| | F | TMME (ResNet-50) [40] | 80.82 | 85.32 | 76.58 | 83.44 |
| | - | **Proposed MV-KBC** | **84.12** | **89.82** | **78.76** | **87.76** |
| P3 | C | 3D GLCM+SVM [14] | 75.75 | 22.11 | 93.03 | 70.99 |
| | D | Multi-visual features [15] | 77.52 | 35.03 | 91.21 | 75.49 |
| | E | Deep + visual features [17] | 80.63 | 46.48 | 91.64 | 80.04 |
| | F | TMME (ResNet-50) [40] | 82.84 | 49.79 | 93.48 | 83.03 |
| | - | **Proposed MV-KBC** | **84.53** | **54.19** | **94.30** | **89.08** |
| P4 | C | 3D GLCM+SVM [14] | 88.14 | 31.76 | 93.54 | 73.65 |
| | D | Multi-visual features [15] | 88.31 | 45.18 | 92.44 | 80.73 |
| | E | Deep + visual features [17] | 90.96 | 55.23 | 94.38 | 87.30 |
| | F | TMME (ResNet-50) [40] | 92.59 | 60.18 | 95.69 | 90.32 |
| | - | **Proposed MV-KBC** | **93.72** | **67.23** | **96.26** | **94.86** |

TABLE VI
PERFORMANCE OF THE PROPOSED MV-KBC MODEL ON THE DATASETS, WHERE UNCERTAIN LUNG NODULES (ULNS) WERE CATEGORIZED AS BENIGN ONES AND MALIGNANT ONES, RESPECTIVELY

| | Results (%) | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC |
| ULNs as benign nodules | 89.81 | 75.63 | 94.74 | 93.23 |
| ULNs as malignant nodules | 74.05 | 66.66 | 82.06 | 80.12 |

which characterizes the OA, HVV and HS of lung nodules on multi-view patches using ResNet-50, is more effective in classifying malignant and benign nodules than using either hand-crafted features (methods C, D and E) or only axial plane patches (method F).

### C. Exploratory Analysis for Uncertain Nodules

For this study, we excluded 612 uncertain lung nodules from our experiments, each of which has a median malignancy level of 3, an ambiguous assessment, from experienced thoracic radiologists. Following the work done by Han *et al.* [14], Shen *et al.* [30], and Dhara *et al.* [15], we designed two experiments, in which those uncertain nodules were categorized as benign ones and malignant ones, respectively, to evaluate the impact of nodules with malignancy suspiciousness on the performance of the proposed MV-KBC model. The obtained classification accuracy was given in TABLE VI. It reveals that grouping uncertain lung nodules into the benign category leads to higher classification accuracy than grouping them into the malignant category, which indicates that those uncertain nodules share more similarities with benign nodules. This finding is consistent with those reported in [14], [15], and [30].

The underlying reason could be that, since a false negative case may lead to much higher cost than a false positive case in lung nodule screening, the thoracic radiologists who annotated the LIDC-IDRI dataset are prone to give "safer" annotations, i.e. ranking more benign nodules (malignancy level 2) than malignant nodules (malignancy level 4) as uncertain ones (malignancy level 3). As a result, most of those uncertain nodules are in fact benign. Therefore, categorizing those nodules as benign ones leads to higher classification accuracy than categorizing them as malignant ones. We are not sure if such operator-related bias is general or just a specific phenomenon for this lung nodule dataset. However, the classification results given in TABLE VI indicate that the performance of a medical image analysis method relies heavily on the quality of image annotation. Besides paying more attention to the quality of data acquisition, we would further investigate semi-supervised learning to use the nodules with uncertain annotations.

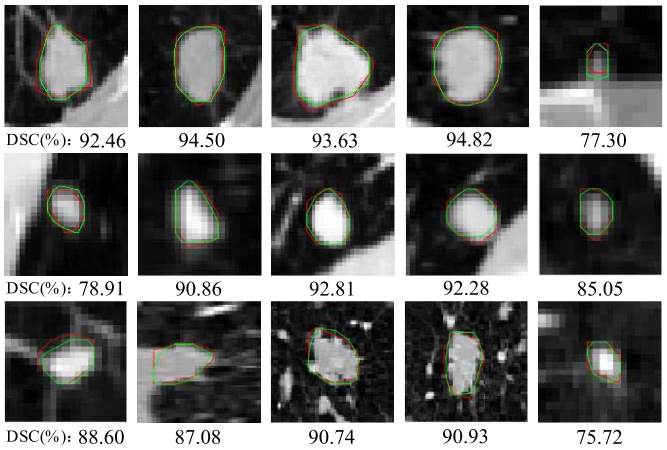



Fig. 6. 15 lung nodules selected randomly from the testing set and the segmentation results of the U-Net; obtained boundaries are highlighted in green; the ground truth is outlined in red.

### D. Impact of Nodule Segmentation

When using the proposed MV-KBC model to classify a lung nodule, it requires to segment the nodule in each slice such that the OA, HVV and HS patches can be extracted as the input of the model. For this study, we employ the U-Net for nodule segmentation. Let the intersection of the nodule areas marked manually by four radiologists be the ground truth, and the U-Net yields an average dice similarity coefficient (DSC) of 80.23% and a sensitivity of 90.04%. We visualized the segmentation results of 15 lung nodules randomly selected from the testing set in Fig. 6. It shows that U-Net can segment the major nodule area well and the discrepancy between the segmentation results and ground truth is small.

Since both the HS patches and HVV patches were extracted based on the nodule boundaries on each slice, the segmentation inaccuracy, though small, may have some impact on

TABLE VII
PERFORMANCE OF THE PROPOSED MV-KBC MODEL WHEN USING EITHER THE U-NET-BASED NODULE SEGMENTATION OR GROUND TRUTH (GT)

| Method | Results (%) | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC |
| MV-KBC | 91.60 | 86.52 | 94.00 | 95.70 |
| MV-KBC + GT | **92.10** | **87.15** | **94.45** | **96.13** |

TABLE VIII
PERFORMANCE OF OUR MV-KBC MODEL TRAINED ON THE PATCHES EXTRACTED ON DIFFERENT COMBINATION OF VIEWS

| Views used in MV-KBC | Results (%) | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC |
| view1 | 89.06 | 77.10 | 95.02 | 92.32 |
| view1 ~ 3 | 90.01 | 83.41 | 93.12 | 94.17 |
| view4&7&9 | 90.12 | 84.31 | 93.05 | 94.42 |
| view1 ~ 6 | 90.86 | 81.37 | **95.52** | 94.83 |
| view4 ~ 9 | 90.71 | 83.63 | 94.30 | 95.12 |
| view1 ~ 9 | **91.60** | **86.52** | 94.00 | **95.70** |

the performance of our MV-KBC model. To evaluate such impact quantitatively, we repeated the nodule classification experiments by using the ground truth as the segmentation results. The performance given in TABLE VII shows that using the nodule segmentation ground truth resulted in a slightly higher AUC, accuracy, sensitivity and specificity when compared to using the nodule segmentation produced by U-Net. It indicates that U-Net is a relatively good choice for nodule segmentation in the proposed MV-KBC model, though more accurate nodule segmentation may further but slightly improve the performance of our model.

### E. Analysis for Multi-View

In our MV-KBC model, we decompose each 3D lung nodule into nine fixed views and construct a KBC submodel to characterize the lung nodule patches extracted on each view from three perspectives, i.e. OA, HVV and HS. To demonstrate that multi-view ensemble learning is effective, we tested each of nine KBC submodels and plotted the ROC curves in Fig. 7. As we expected, multi-view learning outperforms each single-view learning, since it uses more information of nodules. Furthermore, it also shows that using view 4, 7, or 9 achieves higher AUC than using other views.

To validate if these three views comprise the most discriminative information for nodule classification, we jointly used these three views and compared this combination to other combination of views, such as using the first one, three, six, and nine views and using all diagonal views. The classification performance given in TABLE VIII shows that jointly using the view 4, 7, and 9 performs better than using the first one or three views, but worse than using other combinations. Therefore, these three views are not adequate to construct an effective

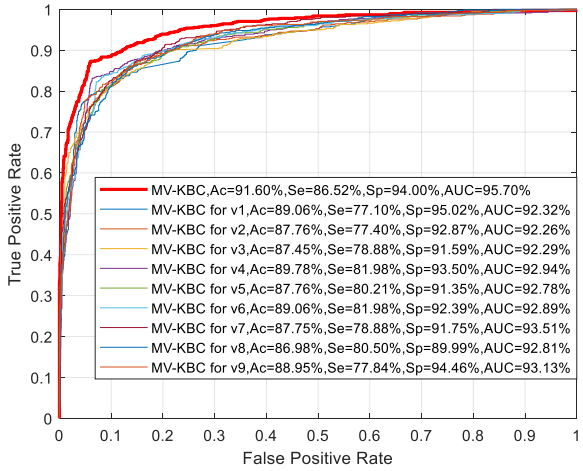



Fig. 7. ROC curves of the proposed MV-KBC model and nine models, each being trained on the patches extracted on one view of plane.

nodule classifier. Meanwhile, it also shows in this table that the more the views we used, the higher the classification accuracy and AUC we obtained. The result is not surprising, since more and more information of nodules can be exploited with the increase of views used in our model.

### F. Different Ensemble Strategies

In our study, the ensemble is performed hierarchically at two decision levels. First, we connect the two-neuron output layer of each of three ResNet-50 networks to a two-neuron layer followed by the softmax function. Such an ensemble of three ResNet-50 networks is called a KBC submodel. Second, we further connect the two-neuron output layer of each KBC submodel to a one-neuron layer followed by the sigmoid function. This ensemble is our proposed MV-KBC model.

To demonstrate the effectiveness of this ensemble strategy, we compared our MV-KBC model to the method in [30], where all multi-view patches are fed into a network (Strategy-I), and the late-fusion method in [37], which performs the fusion at a richer feature level (Strategy-II). To implement the Strategy-I, we fine-tuned three pre-trained ResNet-50 networks using multi-view OA, HVV, and HS patches and combined them at the decision level. To implement the Strategy-II, we concatenated the outputs of the first fully connected (FC) layer (containing 2048 neurons) of each ResNet-50 in each KBC submodel and connected them to a new FC layer with 1024 neurons, followed by a classification layer, which contains one neuron with the sigmoid function. The results in TABLE IX show that, by using the hierarchical ensemble at two decision levels, our MV-KBC model achieves the best performance.

### G. Other Pre-Trained DCNNs

Although we used ResNet-50 as each DCNN component for this study, any DCNN, such as GoogLeNet [52] and VGGNet-19 [53], can be embedded in the proposed MV-KBC model. TABLE X gives the accuracy, sensitivity, specificity

TABLE IX
PERFORMANCE OF DIFFERENT ENSEMBLE STRATEGIES

| Ensemble Strategies | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC (%) |
|---|---|---|---|---|
| Strategy-I | 90.58 | 84.42 | 93.72 | 93.95 |
| Strategy-II | 90.40 | 85.55 | 92.89 | 93.82 |
| MV-KBC | **91.60** | **86.52** | **94.00** | **95.70** |

TABLE X
PERFORMANCE OF OUR MV-KBC MODEL WHEN USING GOOGLENET, VGGNET-19 AND RESNET-50 AS EACH DCNN COMPONENT, RESPECTIVELY

| DCNN used in MV-KBC | Results (%) | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC |
| GoogLeNet (22 layers) | 90.76 | 84.02 | **94.10** | 94.82 |
| VGGNet-19 (19 layers) | 91.24 | 82.97 | 95.35 | 95.48 |
| ResNet-50 (50 layers) | **91.60** | **86.52** | 94.00 | **95.70** |

TABLE XI
PERFORMANCE AND TIME COST OF THE KBC SUBMODEL WHEN USING RESNET-50, RESNET-101 AND RESNET-152 AS EACH DCNN COMPONENT, RESPECTIVELY

| DCNN used in KBC for view1 | Results (%) | | | | Time (h) |
|---|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC | |
| ResNet-50 | 89.06 | 77.10 | 95.02 | 92.32 | **2.2** |
| ResNet-101 | 89.24 | **77.83** | 94.92 | 92.63 | 3.7 |
| ResNet-152 | **89.39** | 77.74 | **95.15** | **92.84** | 4.5 |

and AUC of our MV-KBC model when each DCNN component is a pre-trained GoogLeNet, VGGNet-19 and ResNet-50, respectively. It shows that using ResNet-50 achieves the best performance, particularly a significantly improved sensitivity value (i.e., nearly 3.6% higher than using VGGNet-19 and nearly 2.5% higher than using GoogLeNet). Meanwhile, due to the use of $1 \times 1$ convolutions, ResNet-50 has 25.5 million parameters and is computationally more efficient than VGGNet-19, which has about 144 million parameters.

Moreover, the results TABLE X also suggest that a deeper DCNN seems to lead a higher classification accuracy. To validate this finding, we evaluated the performance of ResNet-101 or ResNet-152 against ResNet-50 on the patches extracted on axial planes (view 1). The results in TABLE XI show that using ResNet-101 or ResNet-152 can further but slightly improve the classification accuracy. However, since there are 27 DCNNs in our model, we chose to use ResNet-50 due to the consideration of the spatial and computational complexity of ResNet-101, ResNet-152, and other deeper networks.

## V. DISCUSSION

### A. Rationale of Designing KBC Submodel

The design of KBC submodel (see Fig. 4) for a specific view is based on the assumption that three fine-tuned ResNet-50 networks can characterize lung nodules from different aspects, i.e. the OA, HVV and HS, and hence yield complementary features. We randomly selected four single view ROIs of lung nodule and visualized the corresponding patches and learned feature maps in Fig. 8. In each subfigure, the top row shows the OA patch, HVV patch and HS patch generated from the single view ROI, and the bottom row shows the normalized sum of 2048 feature maps produced by the last convolutional layer of the corresponding ResNet-50 network. It reveals that the feature maps learned from three types of input image patches highlight different areas and may mutually complement each other.

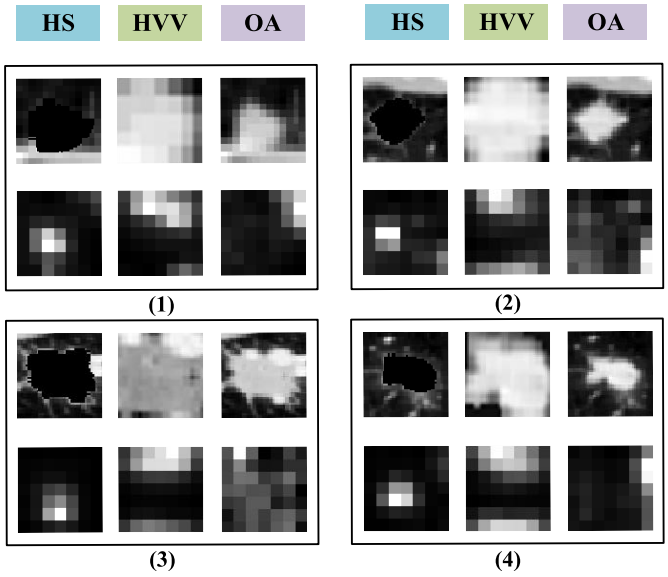



Fig. 8. Three types of image patches generated from four example nodule ROIs and the sum of corresponding feature maps learned by each fine-tuned ResNet-50 network.

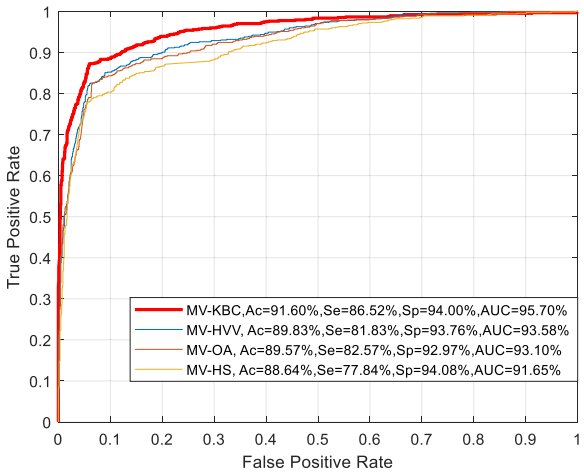



Fig. 9. ROC curves of the proposed MV-KBC model and three models, which only characterize 3D lung nodules from one perspective by inputting multi-view OA, HVV and HS patches, respectively. Especially, 'Ac', 'Se' and 'Sp' are accuracy, sensitivity and specificity.

To quantify our findings, we compared the performance of our MV-KBC model to that of three models (MV-OA, MV-HVV and MV-HS model). Each of them uses the multi-view OA, HVV or HS patches to fine-tune nine pre-trained ResNet-50 networks, and thus only characterizes 3D lung nodules from one of three perspectives. The receiver operator curve (ROC) curves of these four models were plotted

TABLE XII
PERFORMANCE OF MV-KBC MODEL WITH ONE AND THREE PRE-TRAINED RESNET-50 NETWORKS IN THE KBC SUBMODEL

| Methods | Results (%) | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC |
| MV-KBC (1 ResNet-50 in KBC) | 91.05 | 86.06 | 93.57 | 94.97 |
| MV-KBC (3 ResNet-50 in KBC) | **91.60** | **86.52** | **94.00** | **95.70** |

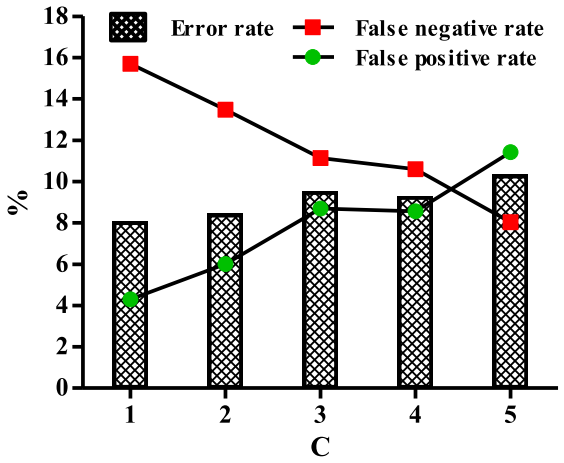



Fig. 10. Variation of accuracy, sensitivity and specificity of the proposed MV-KBC model over the increase of the penalty factor *C*.

in Fig. 9. It shows that using HS patches resulted in the lowest performance, using OA patches and HVV patches produced similar ROC curves, and jointly using three types of patches led to a further performance gain. Moreover, comparing the performance of our MV-KBC model to that of the MV-OA, MV-HVV and MV-HS models, the classification accuracy improves at least 1.67% and the AUC improves at least 2.12%. This experiment demonstrates that the image representation learned from each type of patches has complementary discriminative power and using a combination of them can result in higher classification accuracy than using each of them alone.

To prove the advantage of having three ResNet-50 networks in KBC, we compared the performance of our MV-KBC model with the baseline which concatenated OA, HVV and HS patches to form a 3-dimensional tensor in each KBC. The tensor can be considered as a 2D image with 3 channels and used to fine-tune a pre-trained ResNet-50 network. The results in TABLE XII show that our method can lead to a notable performance gain compared to the baseline. Hence, designing three ResNet-50 networks in KBC is a better choice than concatenating the patches.

### B. Parameter of Penalty Factor C

The proposed MV-KBC model replaces the conventional binary cross-entropy loss with the penalty loss, which provides a way to control the trade-off between the false negative rate and false positive rate by penalizing the false negative with a factor C. Fig. 10 depicts the false negative rate, false positive rate and classification error rate obtained when setting the factor C to different values. It shows that, when the factor C increased from 1 to 5, the false negative rate decreased from 15.70% to 8.03%, the false positive rate increased from 4.30% to 11.42%, and the classification error rate only increased from 8.04% to 10.28%. Therefore, when applying the proposed model to lung nodule screening, we can set the penalty factor C to a large value to reduce the false negative rate. In order not to affect the overall performance of the classifier, we the optimal penalty factor C as 2, since the false negative rate significantly decreased by 2.22% and the classification error rate only increased by 0.36%.

TABLE XIII
PERFORMANCE OF THE 3D-KBC MODEL AND OUR MV-KBC MODEL

| Methods | Results (%) | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC |
| 3D-KBC (scratch 3D ResNet-50) | **90.07** | 81.05 | **94.98** | 92.56 |
| MV-KBC (scratch ResNet-50) | 89.96 | **83.06** | 93.50 | **93.39** |

### C. Multi-View Architecture vs. 3D Network

We also extended our MV-KBC model to the 3D-KBC model, in which the multi-view ResNet-50 network is replaced with the 3D ResNet-50 network, and compared their performance. Similarly, training the 3D-KBC model has 4 steps. Step 1, we cropped a 64×64×64 volume centered on the location of each nodule from the chest CT data and segmented it using the 3D U-Net [60]. Step 2, based on the segmentation result, we defined the cubic volume of interest (VOI) that encapsulates the nodule as an OA volume, set nodule voxels inside the OA volume to 0 to form a HS volume, and set non-nodule voxels inside the OA volume to 0 and, if the OA volume is larger than 16×16×16, extracted a 16×16×16 cube that contains the maximum nodule voxels to form a HVV volume. Step 3, we resized all OA, HVV, and HS volumes to 64×64×64 and applied the same data augmentation method that is used in the MV-KBC model to them. Step 4, all OA, HVV, and HS volumes, together with their augmented versions, were used to train the 3D-KBC model, which consists of three 3D ResNet-50 networks designed to learn the nodules' overall appearance, heterogeneity in voxel values, and heterogeneity in shape, respectively. Each 3D ResNet-50 network contains 50 learnable layers, including consequently a 7×7×7 convolutional layer that produces 64 feature maps, a 3×3×3 max pooling layer, four bottleneck architectures, an average pooling layer, and a FC layer with two neurons. Each bottleneck architecture consists of three convolutional layers with a map size of 1×1×1, 3×3×3, and 1×1×1, respectively. The two-neuron output layers of these three networks are connected to the same one-neuron layer, followed by the sigmoid function for final classification. We used the same loss and optimizer to train the 3D-KBC model. Both models were trained from scratch for a fair comparison.

The results in TABLE XIII show that although 3D-KBC model obtains a marginally higher accuracy (with a 0.11% improvement) and specificity (1.48% improvement), our

TABLE XIV
PERFORMANCE OF OUR MV-KBC MODEL WITH SCRATCH AND PRE-TRAINED RESNET-50 NETWORK

| Methods | Results (%) | | | |
|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | AUC |
| MV-KBC (scratch ResNet-50) | 89.96 | 83.06 | 93.5 | 93.39 |
| MV-KBC (pre-trained ResNet-50) | **91.60** | **86.52** | **94.00** | **95.70** |

TABLE XV
PERFORMANCE OF FOUR METHODS FOR REGRESSING LUNG NODULE MALIGNANCY SCORES

| Methods | RMSE | MAD |
|---|---|---|
| Three-layer CNN, 2017 [51] | 0.8940 | Not Given |
| Deep+visual features, 2017 [13] | Not Given | 0.9200 |
| 3D CNN MTL, 2017 [31] | Not Given | 0.4593 |
| Proposed method | **0.6213** | **0.4374** |

MV-KBC model achieves an improved sensitivity (2.01% improvement) and AUC (0.83% improvement). Since a higher sensitivity indicates a lower false negative rate, our MV-KBC model, which uses the multi-view learning, is more suitable for lung nodule screening and potentially more useful in clinical practice than the 3D-KBC model.

### D. Impact of Transfer Learning

To demonstrate that transfer learning can improve the performance of lung nodule classification, we compared the performance of our MV-KBC model based on pre-trained ResNet-50 to a deep model that has the same architecture but is based on the ResNet-50 networks trained from scratch. As shown in TABLE XIV, although the model with scratch ResNet-50 networks achieves a relatively good performance, transferring the image representation abilities of pre-trained ResNet-50 networks helps to better characterize lung nodules and brings a further performance gain.

### E. Regressing Malignancy Scores

Our MV-KBC model can be extended to regress the malignancy score (from 1-5) of nodules by simply replacing the penalty loss with the mean square error loss. We compared this extended method to a three-layer CNN [51], combining deep and hand-crafted visual features [13], and 3D CNN multi-task learning (MTL) [31]. The performance of malignancy regression was assessed by the root mean square error (RMSE) and mean absolute score error (MASE) between the predicted score and the true score. The results given in TABLE XV show that our method achieves a RMSE of 0.6213 and a MASE of 0.4374, which are lower than those achieved by other three methods. Therefore, the proposed MV-KBC model is also a good choice for the regression of nodule malignancy score.

### F. LUNGx Challenge Dataset

The LUNGx challenge dataset [61] in the Cancer Imaging Archive (TCIA) contains 70 clinical chest CT scans with 83 lung nodules. The nodule diameters range from 3mm to 45mm. A set of 10 calibration scans was made available as a training set. Five of the 10 calibration scans contained a single confirmed benign nodule (2 confirmed based on nodule stability for at least 2 years; 2 confirmed based on nodule resolution; 1 confirmed based on pathological assessment), and the other five scans contained a single pathologically-confirmed malignant nodule (2 small cell carcinomas, 1 poorly- and 1 moderately-differentiated adenocarcinomas, and 1 non-small cell carcinoma, not otherwise specified). The other 60 scans with a total of 73 nodules were considered as a test set, which contained 37 benign nodules (including 13 confirmed based on nodule stability for at least 2 years, 19 confirmed based on nodule resolution, and 5 confirmed based on pathologic assessment) and 36 malignant nodules (including 15 adenocarcinomas, 9 non-small cell carcinomas not otherwise specified, 7 small cell carcinomas, 2 carcinoid tumors, 1 squamous cell carcinoma, and 2 nodules suspicious for malignancy). We applied our MV-KBC model to the LUNGx challenge dataset 5 times independently and assessed the mean and standard deviation of the accuracy, sensitivity/recall, specificity with the cut-off value of 0.5, precision, F-score, and AUC. The results shown in TABLE XVI show that our MV-KBC model achieved the highest AUC compared to the 11 best-performing methods listed in the challenge leaderboard [62]. These results show that the proposed model also has superior ability to classify malignant from benign nodules on this dataset.

### G. Robustness to Noise Corruption

To demonstrate the robustness of our MV-KBC model against noisy images, we further trained and tested the model on images with additive Gaussian noise, whose mean is zero and standard deviation is 1, 5, and 10, respectively. The performance of our model on noise-free and noisy images was shown in TABLE XVII. It reveals that the performance of our model decreases with the increase of the noise level. However, the accuracy of our model only drops slightly from 91.60% to 90.15% when the standard deviation of the Gaussian noise added to the images increases from 0 to 10. The robustness of our model against noisy images may be ascribed to (1) the powerful ResNet-50 network, (2) joint use of each nodule's OA, HS, HVV as its representation, and (3) the ensemble of submodels constructed using the patches extracted on each of nine views of planes.

### H. Model Complexity

All DCNN models were fine-tuned using the open source Keras and Tensorflow software packages. Since there are 27 ResNet-50 models embedded in it, the proposed MV-KBC model has a relatively high computational complexity during training. In our experiments, it takes about 20 hours to train the model and less than 0.5 second to apply it to classify each lung nodule on a server with 8 NVIDIA GTX Titan XP GPUs and 512GB Memory. Although training the model is time-consuming, it can be done offline. The fast online testing

TABLE XVI
PERFORMANCE OF OUR MV-KBC MODEL AND 11 BEST-PERFORMING METHODS [62] ON THE LUNGx CHALLENGE DATASET

| Methods | Nodule segmentation | Classifier | AUC (%) |
|---|---|---|---|
| 1 | Voxel-intensity-based segmentation | SVM | 50.00±6.80 |
| 2 | Region growing | WEKA | 50.00±5.60 |
| 3 | None required | Rules based on histogram-equalized pixel frequencies | 54.00±6.70 |
| 4 | Bidirectional region growing | Uses tumor perfusion surrogate | 54.00±6.60 |
| 5 | Region growing | WEKA | 55.00±6.70 |
| 6 | Graph-cut-based surface detection | Random forest | 56.00±5.40 |
| 7 | Manual initialization, gray-level thresholding, morphological operations | SVM | 59.00±6.60 |
| 8 | None required | Convolutional neural network | 59.00±5.30 |
| 9 | GrowCut region growing with automated initial label points | SVM | 61.00±5.40 |
| 10 | Radiologist-provided nodule semantic ratings | Discriminant function | 66.00±6.30 |
| 11 | Semi-automated thresholding | Support vector regressor | 68.00±6.20 |

| - | **Proposed MV-KBC** (mean±standard deviation) | Accuracy (%) | Sensitivity / Recall (%) | Specificity (%) | Precision (%) | F-score (%) | AUC (%) |
|---|---|---|---|---|---|---|---|
| | | 75.62±1.15 | 87.22±7.24 | 64.32±7.00 | 70.63±2.61 | 77.84±1.77 | **76.85**±0.17 |

TABLE XVII
PERFORMANCE OF OUR MV-KBC MODEL ON NOISE-FREE AND NOISE-CORRUPTED LIDC-IDRI DATASETS. THE STANDARD DEVIATION OF GAUSSIAN NOISE IS DENOTED BY $\sigma$

| σ | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|
| 0 | **91.60** | **86.52** | **94.00** | **95.70** |
| 1 | 91.02 | 85.73 | 93.66 | 95.18 |
| 5 | 90.54 | 84.66 | 93.49 | 94.78 |
| 10 | 90.15 | 84.40 | 93.17 | 94.25 |

suggests that our approach could be used in a routine clinical workflow.

## VI. CONCLUSION

We present the MV-KBC model to separate benign from malignant lung nodules on chest CT by taking into account the nodule appearance on nine view planes and the nodule heterogeneity and by applying an adaptive weighting scheme so that our model can be trained in an end-to-end manner. The results show that our model is more accurate than current state-of-the-art approaches on the LIDC-IDRI dataset. In future work, we will extend the proposed model to a semi-supervised learning framework, such that we can use the nodules with an uncertain level of malignancy and unlabeled nodules as training samples to reduce the need for data annotation. Meanwhile, we will investigate the compression of the DCNN structure used in our model, with the aim of making the training of the model computationally more efficient. Moreover, it will also be necessary to investigate the incorporation of other pathological information into the deep model for a more accurate benign-malignant lung nodule classification.

## ACKNOWLEDGMENT

We acknowledged the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC-IDRI Database used in this work.

## REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA, Cancer J. Clin.*, vol. 63, no. 1, pp. 5–29, 2015.
[2] G. X. Wu and D. J. Raz, "Lung cancer screening," in *Cancer Treatment and Research*, vol. 170. Cham, Switzerland: Springer, 2016, pp. 1–23.
[3] G. X. Wu, D. J. Raz, L. Brown, and V. Sun, "Psychological burden associated with lung cancer screening: A systematic review," *Clin. Lung Cancer*, vol. 17, no. 5, pp. 315–324, 2016.
[4] American Thoracic Society, "What is a lung nodule?" *Amer. J. Respiratory Crit. Care Med.*, vol. 193, no. 7, pp. 11–12, 2016.
[5] S. Diciotti, S. Lombardo, M. Falchini, G. Picozzi, and M. Mascalchi, "Automated segmentation refinement of small lung nodules in CT scans by local shape analysis," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 12, pp. 3418–3428, Dec. 2011.
[6] T. Messay, R. C. Hardie, and S. K. Rogers, "A new computationally efficient CAD system for pulmonary nodule detection in CT imagery," *Med. Image Anal.*, vol. 14, no. 3, pp. 390–406, Jun. 2010.
[7] A. Soliman *et al.*, "Accurate lungs segmentation on CT chest images by adaptive appearance-guided shape modeling," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 263–276, Jan. 2017.
[8] J. Song *et al.*, "Lung lesion extraction using a toboggan based growing automatic segmentation approach," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 337–353, Jan. 2016.
[9] A. A. Farag, H. E. A. El Munim, J. H. Graham, and A. A. Farag, "A novel approach for lung nodules segmentation in chest CT using level sets," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5202–5213, Dec. 2013.
[10] B. C. Lassen, C. Jacobs, J.-M. Kuhnigk, B. van Ginneken, and E. M. van Rikxoort, "Robust semi-automatic segmentation of pulmonary subsolid nodules in chest computed tomography scans," *Phys. Med. Biol.*, vol. 60, no. 3, pp. 1307–1323, Jan. 2015.
[11] D. Wu *et al.*, "Stratified learning of local anatomical context for lung nodules in CT images," in *Proc. IEEE CVPR*, Jun. 2010, pp. 2791–2798.
[12] Y. Tao *et al.*, "Multi-level ground glass nodule detection and segmentation in CT lung images," in *Proc. MICCAI*, 2009, pp. 715–723.
[13] S. Chen *et al.*, "Automatic scoring of multiple semantic attributes with multi-task feature leverage: A study on pulmonary nodules in CT images," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 802–814, Mar. 2017.
[14] F. Han *et al.*, "Texture feature analysis for computer-aided diagnosis on pulmonary nodules," *J. Digit. Imag.*, vol. 28, no. 1, pp. 99–115, Feb. 2015.
[15] A. K. Dhara, S. Mukhopadhyay, A. Dutta, M. Garg, and N. Khandelwal, "A combination of shape and texture features for classification of pulmonary nodules in lung CT images," *J. Digit. Imag.*, vol. 29, no. 4, pp. 466–475, Aug. 2016.
[16] L. Sorensen, S. B. Shaker, and M. de Bruijne, "Quantitative analysis of pulmonary emphysema using local binary patterns," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 559–569, Feb. 2010.

[17] Y. Xie, Y. Xia, J. Zhang, M. Fulham, and Y. Zhang, "Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT," *Inf. Fusion*, vol. 42, pp. 102–110, Jul. 2018.

[18] Y. Xie, J. Zhang, S. Liu, W. Cai, and Y. Xia, "Lung nodule classification by jointly using visual descriptors and deep features," in *Proc. MICCAI Workshops MCV BAMBI*, 2017, pp. 116–125.

[19] A. El-Baz, M. Nitzken, E. Vanbogaert, G. Gimel'farb, R. Falk, and M. A. El-Ghar, "A novel shape-based diagnostic approach for early diagnosis of lung nodules," in *Proc. IEEE ISBI*, Mar./Apr. 2011, pp. 137–140.

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[21] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Jul./Aug. 1985.

[22] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Proc. IEEE IJCNN*, Jun. 1989, pp. 593–605.

[23] H. Chen, J. Zhang, Y. Xu, B. Chen, and K. Zhang, "Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans," *Expert Syst. Appl.*, vol. 39, no. 13, pp. 11503–11509, 2012.

[24] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[25] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[28] K.-L. Hua, H. Che-Hao, S. C. Hidayati, W.-H. Cheng, and Y.-J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," *Onco Targets Therapy*, vol. 8, pp. 2015–2022, Aug. 2015.

[29] S. Hussein, R. Gillies, K. Cao, Q. Song, and U. Bagci, "TumorNet: Lung nodule characterization using multi-view convolutional neural network with Gaussian process," in *Proc. ISBI*, 2017, pp. 1007–1010.

[30] W. Shen *et al.*, "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification," *Pattern Recognit.*, vol. 61, pp. 663–673, Jan. 2017.

[31] S. Hussein, K. Cao, Q. Song, and U. Bagci, "Risk stratification of lung nodules using 3D CNN-based multi-task learning," in *Proc. IPMI*, vol. 10265, 2017, pp. 249–260.

[32] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. CVPR*, 2014, pp. 1717–1724.

[33] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *Proc. CVPR*, 2015, pp. 325–333.

[34] W. Shen *et al.*, "Learning from experts: Developing transferable deep features for patient-level lung cancer prediction," in *Proc. MICCAI*, 2016, pp. 124–131.

[35] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.

[36] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.

[37] A. A. A. Setio *et al.*, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.

[38] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *Proc. MICCAI*, 2013, pp. 246–253.

[39] S. Metz *et al.*, "Multiparametric MR and PET imaging of intratumoral biological heterogeneity in patients with metastatic lung cancer using voxel-by-voxel analysis," *PLoS ONE*, vol. 10, no. 7, p. e0132386, Jul. 2015. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0132386

[40] 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 *Proc. MICCAI*, 2017, pp. 656–664.

[41] S. G. Armato, III, *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, 2011.

[42] S. G. Armato, III, G. Mclennan, M. F. Mcnittgray, C. R. Meyer, and A. P. Reeves. (2015). *Data From LIDC-IDRI. The Cancer Imaging Archive*. [Online]. Available: http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX

[43] K. Clark *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013.

[44] F. Han *et al.*, "A texture feature analysis for diagnosis of pulmonary nodules using LIDC-IDRI database," in *Proc. IEEE ICMIPE*, Oct. 2013, pp. 14–18.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Jun. 2016, pp. 770–778.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, Jun. 2009, pp. 248–255.

[47] R. Rojas, "The backpropagation algorithm," in *Neural Networks*. Berlin, Germany: Springer, 1996.

[48] M. Nielsen. (Dec. 2017). *How the Backpropagation Algorithm Works*. [Online]. Available: http://neuralnetworksanddeeplearning.com/chap2.html

[49] Y. S. Resheff, A. Mandelbaum, and D. Weinshall. (Apr. 2017). "Every untrue label is untrue in its own way: Controlling error type with the log bilinear loss." [Online]. Available: https://arxiv.org/abs/1704.06062

[50] 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," *BioMed. Eng. OnLine*, vol. 15, no. 1, pp. 1–17, Jan. 2016. [Online]. Available: https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-015-0120-7

[51] X. Li, Y. Kao, W. Shen, X. Li, and G. Xie, "Lung nodule malignancy prediction using multi-task convolutional neural network," *Proc. SPIE*, vol. 10134, p. 1013424, Mar. 2017.

[52] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE CVPR*, Jun. 2016, pp. 1–9.

[53] K. Simonyan and A. Zisserman. (Apr. 2015). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[54] J. Zhang, Y. Xia, Y. Xie, M. Fulham, and D. D. Feng, "Classification of medical images in the biomedical literature by jointly using deep and handcrafted visual features," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1521–1530, Sep. 2018.

[55] J. Zhang, Y. Xia, H. Cui, and Y. Zhang, "Pulmonary nodule detection in medical images: A survey," *Biomed. Signal Process. Control*, vol. 43, pp. 138–147, May 2018.

[56] J. Zhang, Y. Xia, H. Zeng, and Y. Zhang, "NODULe: Combining constrained multi-scale LoG filters with densely dilated 3D deep convolutional neural network for pulmonary nodule detection," *Neurocomputing*, vol. 317, pp. 159–167, Nov. 2018.

[57] D. Kumar, A. Wong, and D. A. Clausi, "Lung nodule classification using deep features in CT images," in *Proc. CRV*, 2015, pp. 133–138.

[58] M. J. Shafiee, A. G. Chung, D. Kumar, F. Khalvati, M. Haider, and A. Wong. (2015). "Discovery radiomics via stochasticnet sequencers for cancer detection." [Online]. Available: https://arxiv.org/abs/1511.03361

[59] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *Proc. IPMI*, 2015, pp. 588–599.

[60] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*, 2016, pp. 424–432.

[61] S. G. Armato, III, *et al.* (2015). *Lung Nodule Classification Challenge Dataset. The Cancer Imaging Archive*. [Online]. Available: http://doi.org/10.7937/K9/TCIA.2015.UZLSU3FL

[62] S. G. Armato *et al.*, "LUNGx Challenge for computerized lung nodule classification," *J. Med. Imag.*, vol. 3, no. 4, p. 044506, 2016.