Pattern Recognition 61 (2017) 663–673

Contents lists available at ScienceDirect

Pattern Recognition

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

Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification

Wei Shen [a,b], Mu Zhou [c], Feng Yang [d,*], Dongdong Yu [a,b], Di Dong [a,b], Caiyun Yang [a,b], Yali Zang [a,b], Jie Tian [a,b,**]

[a] Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] Beijing Key Laboratory of Molecular Imaging, Beijing 100190, China
[c] Stanford Center for Biomedical Informatics Research, Stanford University, CA 94305, USA
[d] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

ARTICLE INFO

Article history:
Received 28 January 2016
Received in revised form
29 April 2016
Accepted 24 May 2016
Available online 26 May 2016

Keywords:
Lung nodule
Malignancy suspiciousness
Convolutional neural network
Multi-crop pooling

ABSTRACT

We investigate the problem of lung nodule malignancy suspiciousness (the likelihood of nodule malignancy) classification using thoracic Computed Tomography (CT) images. Unlike traditional studies primarily relying on cautious nodule segmentation and time-consuming feature extraction, we tackle a more challenging task on directly modeling raw nodule patches and building an end-to-end machine-learning architecture for classifying lung nodule malignancy suspiciousness. We present a Multi-crop Convolutional Neural Network (MC-CNN) to automatically extract nodule salient information by employing a novel multi-crop pooling strategy which crops different regions from convolutional feature maps and then applies max-pooling different times. Extensive experimental results show that the proposed method not only achieves state-of-the-art nodule suspiciousness classification performance, but also effectively characterizes nodule semantic attributes (subtlety and margin) and nodule diameter which are potentially helpful in modeling nodule malignancy.

1. Introduction

Lung cancer is an aggressive disease carrying a dismal prognosis with a 5-year survival rate at 18% [1]. Despite the development of multi-modality treatments over the past decade, lung cancer remains the leading death of cancer and accounts for approximately 27% of all cancer deaths [2]. Technological advances in Computed Tomography (CT) have been routinely used in lung cancer detection, risk assessment, and clinical management. In particular, the increasing quantity of CT image assays has created a unique avenue for data-driven analysis to capture underlying cancer characteristics at a macroscopic level, allowing identification of prognostic imaging biomarkers [3].

In this study, we investigate the problem of automatic lung nodule malignancy suspiciousness classification using CT imaging data. The annotation of nodule malignancy suspiciousness has permitted a chance to evaluate consensus assessments from different experienced radiologists. Specifically, the automatic classification of malignancy suspiciousness on CT studies is a worthy task, because it would facilitate radiologists to assess early risk factors which is essential in lung cancer research [4,5]. A typical implication of such analysis is to provide useful cues for subsequent therapeutic plannings and holds promise for improving individualized patient management. For example, distinct malignancy likelihood derived from imaging can be used to recommend follow-up treatments including CT surveillance (e.g. low likelihood score) or biopsy test and surgical resection (e.g. high likelihood score) [6]. Despite different approaches were proposed for lung nodule diagnosis, novel data-driven techniques are required to advance the predictive power with CT imaging, especially for the prediction on malignancy suspiciousness.

Image-based techniques for analyzing lesions are normally performed with detection [7,8], segmentation [9–12], hand-crafted feature engineering [13,14], and category labelling [15–18]. Zinovev et al. [19] adopted a belief decision tree approach to predict nodule semantic attributes. Chen et al. [20] proposed to use a neural network ensemble scheme to distinguish probably benign, uncertain and probably malignant lung nodules. Han et al. [16] used a 3-D image-based texture feature analysis for nodule diagnosis. More recently, Balagurunathan et al. [14] and Aerts et al. [13] extracted a number of nodule image features to investigate their prognostic

* Corresponding author.
** Corresponding author at: Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
E-mail addresses: wei.shen@ia.ac.cn (W. Shen), fengyang@bjtu.edu.cn (F. Yang), tian@ieee.org (J. Tian).
URL: http://www.3dmed.net (J. Tian).

http://dx.doi.org/10.1016/j.patcog.2016.05.029
0031-3203/

Table 1
Some classification results on LIDC-IDRI dataset from literatures. "NA" denotes "nodule attributes" and "MS" denotes "malignancy suspiciousness".

Related work	Label	Accuracy	AUC	Sample size
Zinovev et al. [19]	NA	54.32%	–	914
Chen et al. [20]	MS	78.70%	–	47
Han et al. [16]	MS	–	0.927	1356

power. Related studies on the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset [21] are shown in Table 1. However, all these methods rely on nodule segmentation as a prerequisite. Notably, automatic nodule segmentation may affect classification results since methods such as region growing and level set typically depend on initialization. Working on these segmented regions may yield inaccurate features that lead to erroneous outputs. To derive a suspiciousness-sensitive descriptor in CT imaging, we need to overcome at least two major obstacles: the difficulty of nodule delineation caused by a large range of nodule morphology variation, and the challenge posed by the nodule radiological heterogeneity for computational models to capture quantitative characteristics.

Image patch-based approaches provide an alternative way for the region of interest (ROI) definition [22,23]. Researchers are seeking visual feature descriptors, such as Local Binary Patterns (LBP) [24] and Histogram of Oriented Gradients (HOG) [25], to refine measurement on lung cancer imaging. Nevertheless, the yielded textural features are largely determined by the parameter setting. Thus, using them to accurately describe the variability of lung nodules is difficult.

In response to these challenges, we utilize the Convolutional Neural Network (CNN) [26–28] to build an end-to-end computational architecture which is robust in lung nodule image feature extraction and malignancy suspiciousness classification. We propose a computational architecture—the Multi-crop Convolutional Neural Network (MC-CNN)—to learn high-level suspiciousness-specific features for lung nodule classification. As outlined in Fig. 1, our approach automatically classifies nodule malignancy suspiciousness by extracting a set of highly compact features. It is an end-to-end architecture which embeds nodule feature extraction into a hierarchical network. The proposed method simplifies conventional lung nodule malignancy suspiciousness classification by removing nodule segmentation and hand-crafted feature (e.g., texture and shape compactness) engineering work. Our main contributions can be summarized as follows:

1. We demonstrate that even without nodule segmentation and hand-crafted feature engineering which are time-consuming and subjective, the high-level features extracted by our MC-CNN from detected nodule patches are able to present high inter-class variations related to nodule malignancy suspiciousness (Fig. 2), bridging the gap between the raw nodule image and the malignancy suspiciousness.
2. We propose a multi-crop pooling operation which is a specialized pooling strategy for producing multi-scale features to surrogate the conventional max-pooling operation. Without using multiple networks to produce multi-scale features, the proposed approach applying on a single network is effective in computational complexity (Section 4.2).
3. Beyond nodule malignancy suspiciousness classification, we extend the proposed approach to quantify nodule semantic labels as well as to estimate nodule diameter that may potentially assist researchers in evaluating malignancy uncertainty (Section 4.5). Our results showed the possible applications of the proposed method in other lung nodule-relevant analysis that may potentially assist researchers in evaluating malignancy uncertainty.

Applying a supervised learning scheme in deep feature extraction, our approach is in contrast with an auto-encoder approach [30] that applied an unsupervised learning method without prior labeling information. The proposed method also differs from our previous work based on the multi-scale CNN model [31] which utilized multiple CNNs in parallel with different scales of nodule images. In [31], a resampling strategy was used to uniformly represent nodule patches. However, multiple networks become the main burden for training CNNs efficiently since they involve more computational costs, especially when dealing with high-resolution images. As opposed to the design of multiple CNNs [31], the proposed model simplified the training process by replacing multiple CNNs with the multi-crop pooling architecture that is specially tailored to lung nodule malignancy suspiciousness classification. Furthermore, our model underscored the knowledge extraction from feature space rather than image space. In other words, the computation is specified on the intermediate convolutional features (i.e., feature space), rather than different scales of raw input signals (i.e., image space).

The rest of the paper is organized as follows. Section 2 introduces the proposed multi-crop CNN architecture. Section 3 presents the detail of the dataset and data augmentation. Section 4 describes the experimental setup and results. Section 5 is the discussion and Section 6 concludes the paper.

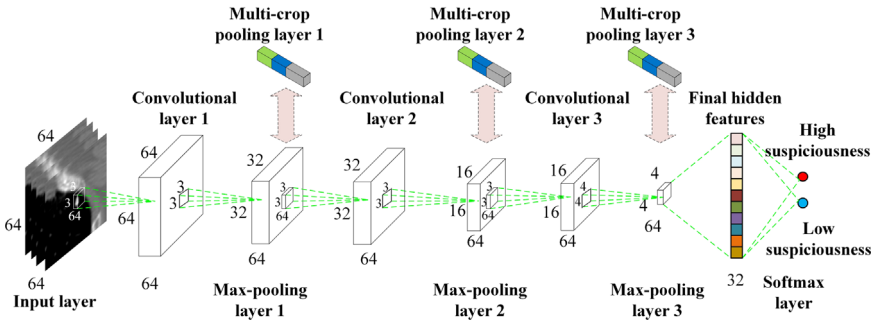

Fig. 1. The proposed MC-CNN architecture for lung nodule malignancy suspiciousness classification. The numbers along each side of the cuboid indicate the dimensions of the feature maps. The inside cuboid represents the 3-D convolution kernel and the inside square stands for the associated 2-D pooling region. The dimension of the final hidden feature layer is marked at the bottom. The output layer is a softmax layer that predicts the probability of the class of nodule malignancy suspiciousness, i.e., low malignancy suspiciousness and high malignancy suspiciousness. The pink arrow indicates a multi-crop pooling layer that serves as a surrogate of a max-pooling layer for improving classification performance. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

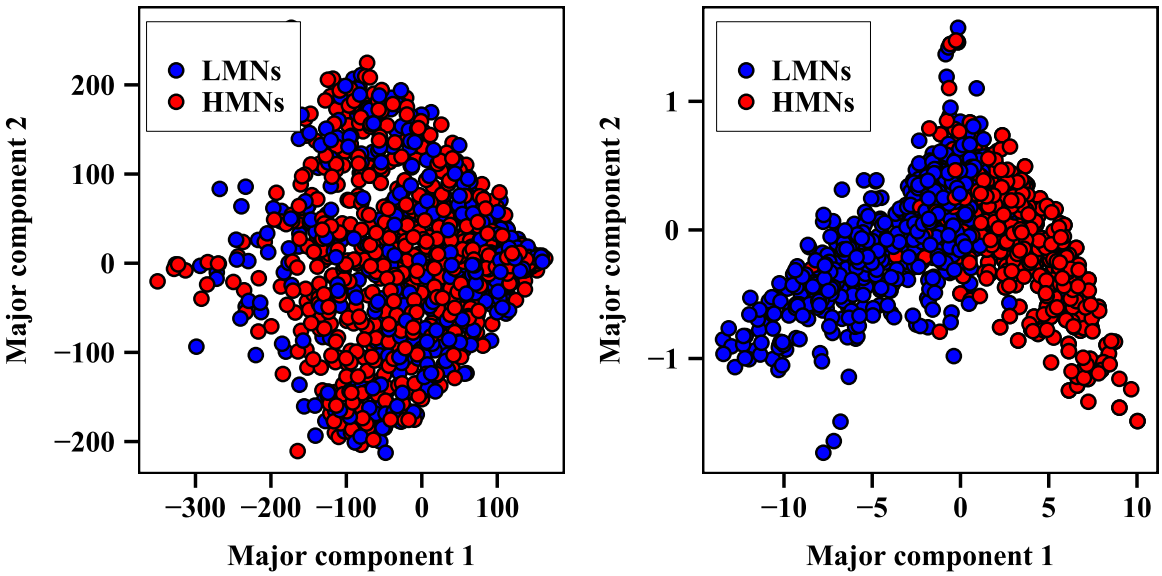

Fig. 2. Feature visualization. "LMNs" indicate low malignancy-suspicious nodules and "HMNs" represent high malignancy-suspicious nodules. Two major components were computed and projected in a 2-D space by the Principal Component Analysis (PCA) [29]. *Left*: features from original pixel-based nodule patches; *right*: deep features from the proposed method. Visualization indicates that the MC-CNN is effective in yielding highly-discriminative features.

2. Methods

In recent studies [26,32,33], the CNN architecture has been brought to the forefront in the image processing field. The core computation seeks a feature representation, also known as the activation of the final hidden layer in the network, that is transformed from high-dimensional features in $R^{M\times N}$ and remains well separated in a low-dimensional $R^{P\times 1}$ space ($P < M \times N$). Specifically, two computational units including *convolutional layers* and *pooling layers* are used to quantify the mechanism. The network defines a feature-extraction cascade consisting of concatenated convolutional layers and pooling layers (i.e., "Conv+Pool"). Thus, the formed hierarchical network can learn high-level compact features from signal activations of high layers. As shown in Fig. 1, our MC-CNN also consists of "Conv+Pool" layers. However, a proposed multi-crop pooling layer (Section 2.2) is used to surrogate the max-pooling layer to extract multi-scale features.

Given the lung nodule CT images, our goal is to discover a set of discriminative features from the proposed hierarchical neural networks and thus to capture the essence of suspiciousness-specific nodule information. The challenge is that the image space is heterogeneous including both healthy tissues and nodules at different visual scales. Compared to the conventional feature extraction [13,14,34], we propose an integrated computational deep learning architecture. The major components which form the basis of our multi-crop CNN are presented from Sections 2.1 to 2.3.

2.1. Conv+pool layer design

The CNN starts from a convolutional layer where we adopt the Randomized Leaky Rectified Linear Units (RReLU) [35,36] as a non-linear transformation. Formally, the convolution operation is defined by

$$y^j = \mathrm{RReLU}\left(\sum_i c^{ij} * h^i + b^j\right), \tag{1}$$

where h^i and y^j are the ith input map and the jth output map, respectively. We define c^{ij} as the convolution kernel between the ith input map and the jth output map ($*$ denotes the 2-D convolution). b^j is the bias of the jth output map. h^i, y^j and c^{ij} are all 2-D. The entire input, output and convolution kernel of a convolutional layer are a stack of h^i, y^j, and c^{ij}. As seen in Fig. 1, there are 64 CT slices (h^i) in the input layer and the convolutional layer outputs 64 convolutional feature maps (y^j). Accordingly, the number of convolution kernels is 64 with dimension of $3 \times 3 \times 64$ voxels. Both c^{ij} and b^j are continuously learned in the network training process. The non-linear transformation function RReLU(x) [35] is expressed as

$$\mathrm{RReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ \dfrac{x}{a} & \text{if } x < 0,\ a \sim U(b_l, b_u) \end{cases} \tag{2}$$

where a is a random factor sampled from a uniform distribution $U(b_l, b_u)$, and b_l and b_u are the lower and higher bounds of the distribution respectively. RReLU allows for a small, non-zero gradient initialization for unit activation that has been proven to be less prone to overfit the dataset [35,36] than conventional ReLU [26] in a classification task, especially when the training samples are limited.

Following the convolutional layer, a max-pooling layer is commonly introduced to select feature subsets. It is defined as

$$y^i_{(j,k)} = \max_{0\leq m,n<s} \{h^i_{(j\cdot s+m,k\cdot s+n)}\}, \tag{3}$$

where s is the size of pooling region. $y^i_{(j,k)}$ represents the neuron at position (j,k) in the ith output map. $h^i_{(j\cdot s+m,k\cdot s+n)}$ denotes the neuron at position $(j\cdot s + m, k\cdot s + n)$ in the ith input map where m and n are the offsets of the position. The advantage of using the max-pooling layer is its translation invariability even when different nodule images are not well-aligned. In the following section, we introduce our multi-crop pooling strategy which can surrogate the traditional max-pooling operation.

2.2. Multi-crop pooling strategy

We extend the traditional max-pooling layer into our multi-crop pooling layer which allows the capture of nodule-centric visual features. Traditional max-pooling layers in the network play a role of selecting feature subsets and reducing the size of the feature maps. However, the max-pooling operation performs

uniformly on each feature map, and thus the max-pooling layer is a single-level feature reduction operation. Such a setting hinders it from capturing accurate object-specific information when the size of objects varied largely in the images. As seen in Fig. 3, a large variability of nodule sizes steers us to pursue an alternative strategy for capturing nodule sensitive information.

We proposed a multi-crop layer to surrogate the conventional max-pooling layer. It is a strategy with repetitive pooling features, enabling a multi-scale feature extraction from the input feature maps on nodule samples. Given a stack of feature maps from a previous convolutional layer, a multi-crop strategy is designed to fully capture nodule-centric features. As shown in Fig. 4, the concatenated nodule-centric feature $f = [f_0, f_1, f_2]$ is formed from three nodule-centric feature patches R_0, R_1, R_2 respectively. Specifically, let the size of R_0 be $l \times l \times n$, where $l \times l$ is the dimension of the feature map and n is the number of feature maps:

$$f_i = \max - \text{pool}^{(2-i)}\{R_i\},\ i = \{0, 1, 2\}, \tag{4}$$

where R_1, R_2 are two center regions with size of $(l/2) \times (l/2) \times n$ and $(l/4) \times (l/4) \times n$. The superscript of "max-pool" indicates the frequency of the utilized max-pooling operation on R_i. In Fig. 4, the input of multi-crop pooling operation is the convolutional features R_0 obtained from either the original image or the pooled features. R_1 is the center region cropped from R_0 and R_2 is the center region cropped from R_1. Then, R_0 is max-pooled twice and R_1 is max-pooled once to generate pooled feature maps f_0 and f_1. R_2 serves as f_2 without any pooling operation. The final multi-crop feature is made up with the concatenation of f_0, f_1, and f_2. Specifically, the strategy on targeting nodule-specific patches allows us to feed multi-scale nodule sensitive information into the following convolutional layers. The functionality of the multi-crop pooling layer is similar to that of the max-pooling layer since they both pool the input feature maps. Thus, it can surrogate any max-pooling layers for the purpose of extracting multi-scale features. The effectiveness of multi-crop features is discussed in Section 4.2.

The objective of multi-crop pooling strategy is to extract multi-scale features from a single network. The strategy draws inspiration from spatial pyramid pooling network (SPPNet) [37] which concatenated the feature pyramid as the final feature vector. Although both SPPNet and the proposed method share a similarity in extracting features at different scales, several remarkable differences are recognized: (1) the pooling frequency of multi-crop pooling relies on feature location in the feature map, while spatial pyramid pooling strategy pools features at different location equally; (2) the output features of our multi-crop pooling layer at different scales have the same dimension while the feature dimensions from spatial pyramid pooling are determined by their pooling levels; (3) a third and more important distinction is that the output of our multi-crop pooling at each scale can be concatenated to feed into the following convolutional layer, while the output of spatial pyramid pooling can only be placed at the top of a CNN.

Speaking of computational complexity, unlike conventional multi-scale features [31] with multiple parallel networks in image space, the multi-crop pooling layer could generate multi-scale features from a singular MC-CNN pipeline, which greatly simplifies the training process and shortens the training time without sacrificing the classification accuracy (Section 4.2).

2.3. Loss function

In general, multiple pairs of concatenated Conv+Pool layers consist of a major network architecture and the last pooling layer is usually connected to a fully-connected layer. The output layer of the entire network is a 2-way softmax layer (see Fig. 1) predicting the probability distribution over low malignancy suspiciousness and high malignancy suspiciousness:

$$p_j = \frac{\exp(y'_j)}{\exp(y'_0) + \exp(y'_1)},\ j = \{0, 1\}, \tag{5}$$

where $y'_j = \sum_{i=1}^{32} h'_i \cdot w_{i,j} + b'_j$ is the linear combination of the input h'_i (the activations of the final hidden features in Fig. 1). $w_{i,j}$ is the weight and b'_j is the scalar.

The network is learned by minimizing the cross entropy loss, which can be expressed as

$$LOSS = -(q \log p_1 + (1 - q) \log p_0), \tag{6}$$

where q indicates suspiciousness label with the value 1 or 0 corresponding to being high suspiciousness or low suspiciousness respectively. The network is trained using Stochastic Gradient Descent (SGD) with a standard backprop [38,39].

2.4. Prediction modeling and model evaluation

In addition to classify nodule malignancy suspiciousness category, we also predict nodule attributes including nodule subtlety, margin, and diameter (Section 4.5), which could potentially be used to model nodule malignancy uncertainty. For malignancy suspiciousness, subtlety and margin, we model them as a binary classification problem and predict whether the nodule belongs to the high score category or the low score category. For nodule diameter estimation, we modify our MC-CNN to be a regression model by replacing the last softmax layer with a single neuron which predicts the estimated diameter in a real value. Balanced datasets, obtained by sample selection, are prepared for the classification tasks. For diameter estimation model, the entire dataset without any balancing process is used. More detailed discussions on the validation setting are given in Section 3.

In order to do model selection for predicting outcomes, we split the dataset into the training set, validation set and test set. Each network model was trained for 5000 iterations, and we saved the trained model at every 100 iterations. After the entire training process, the associated validation scores obtained from the validation set were sorted in a descending order. We then selected the top 3 models as the final trained models and the prediction outcome of a test patch was the average of the ensemble probability scores.

The performance of classifying malignancy suspiciousness (Section 4.2), subtlety and margin (Section 4.5.1), and of estimating nodule diameter (Section 4.5.2) were evaluated via five-fold cross validation. In each experiment, three folds were used as the training set. One fold was used as the validation set and the rest one as the test set. We reported the classification performance by averaging the classification accuracies and the area under the curve (AUC) scores across 30 times tests.

3. Dataset description

3.1. Dataset

The dataset used in this work is the LIDC-IDRI dataset [21], consisting of 1010 patients with lung cancer thoracic CT scans as well as mark-up annotated lesions. We included nodules along with their annotated centers from the nodule collection report.[1] The diameters of the nodules range from 3 mm to 30 mm. Since the resolution of the images varied, we resampled images using

[1] http://www.via.cornell.edu/lidc.

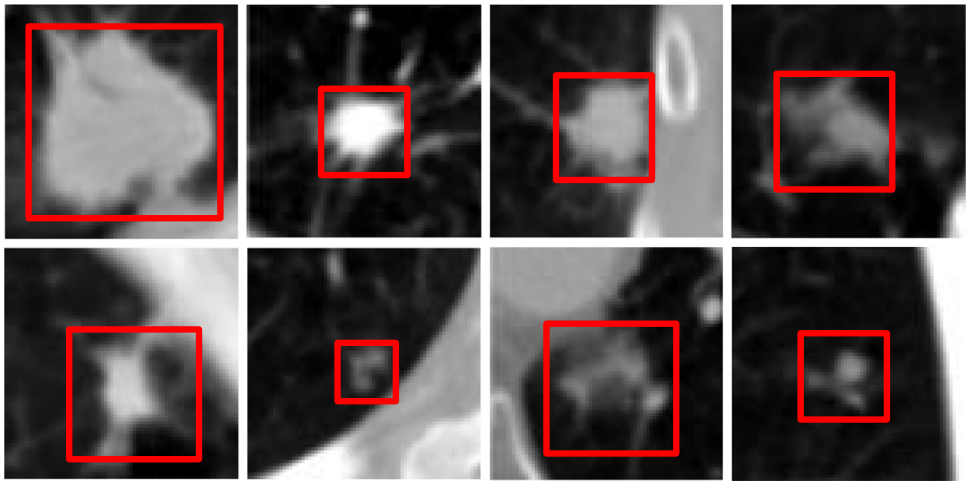

Fig. 3. Nodule sample images. We illustrate that both high malignancy suspicious cases (first row) and low malignancy suspicious (second row) cases have a large diameter range (3–30 mm).

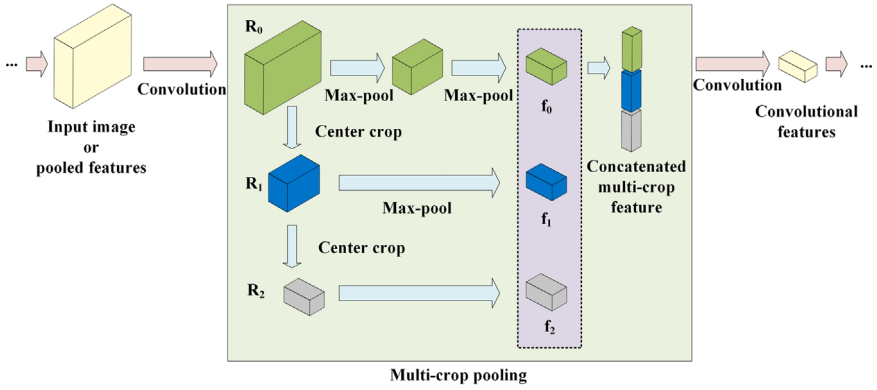

Fig. 4. Illustration of the multi-crop pooling operation. The input of multi-crop pooling operation is the convolutional features R_0 obtained from either the original image or the pooled features. R_1 is the center region cropped from R_0 and R_2 is the center region cropped from R_1. Then, R_0 is max-pooled twice and R_1 is max-pooled once to generate pooled feature maps f_0 and f_1. R_2 serves as f_2 without any pooling operation. The final multi-crop feature is made up with the concatenation of f_0, f_1, and f_2.

spline interpolation to have a fixed resolution with 0.5 mm/voxel along all three axes. Each nodule patch was cropped from the resampled CT image based on the annotated nodule center.

The malignancy suspiciousness of each nodule is rated from 1 to 5 by four experienced thoracic radiologists, indicating an increasing degree of malignancy suspiciousness. We chose the averaged malignancy rating for each nodule as [40,31,16]: for those with an average score lower than 3, we labelled them as low malignancy-suspicious nodules (LMNs); for those with an average score higher than 3, we labelled them as high malignancy-suspicious nodules (HMNs). We removed nodule samples with ambiguous IDs. Overall, there were 880 LMN and 495 HMN cases included for performance evaluation. For nodules with an average rating of 3, we followed the study in [16] by conducting two additional experiments of excluding them from the evaluation and including them in another experiment, respectively. The number of nodules with an average rating of 3 was 1243 in total. These nodules will be referred to as uncertain nodules (UNs) in the following sections since they do not fall to any distinct categories.

Similarly, for nodule subtlety and margin attributes analysis in Section 4.5, we selected equal numbers of positive nodule samples (average attribute rating >3) and negative nodule samples (average attribute rating <3) from the LIDC-IDRI dataset, resulting in 756 nodules for subtlety classification and 658 nodules for margin classification. For nodule diameter estimation in Section 4.5, we used the entire dataset of 2618 nodules for the regression modeling.

3.2. Data augmentation

We sought to train the MC-CNN model with augmented training samples that complemented the learning process given limited training samples. We augmented nodules by random image translation, rotation and flip operations as in [26,41,42]. The translation was in a range of $[-6, 6]$ voxels; the rotation was done by first swapping the three axes in 3-D followed by a 2-D rotation of [90°, 180°, 270°]. Nodule patches were augmented when fed into the input layer. Such augmentation helped the MC-CNN capture nodule features invariant to image-level translation, rotation and flip operations.

4. Experiments and results

In this section, we evaluate our MC-CNN performance by measuring the classification accuracy and the AUC score of computed deep features. We perform a systematic evaluation against different parameters including the number of convolution kernels,

the position of multi-crop pooling layer in the whole architecture, and dataset sample sizes. We first describe the experimental setup. Then, we report results of our MC-CNN and compare it with the state-of-the-art approaches. We also conduct an exploratory analysis on modeling nodule malignancy uncertainty. Finally, we demonstrate the effectiveness of our MC-CNN in nodule subtlety prediction, nodule margin prediction and nodule diameter estimation.

4.1. Experimental setup

To observe performance with regard to the network configuration, we investigated different configurations including the number of convolution kernels of each convolution layer and the position of a multi-crop pooling layer. The MC-CNN which had n_{ker} convolution kernels for each convolutional layer and a multi-crop pooling layer surrogating the ith max-pooling layer (the position of the multi-crop pooling layer) was named as $\mathrm{MC-CNN}_i^{n_{ker}}$, where $n_{ker} = \{16, 32, 64\}$ and $i = \{1, 2, 3\}$. Thus, there were 9 different configurations in total. For simplicity, $\mathrm{MC-CNN}_i$ was used to indicate a general MC-CNN with an ith max-pooling layer surrogated by a multi-crop pooling layer but with an arbitrary n_{ker}. The number of neurons n_h in the final hidden feature layer was fixed to 32. All results of these parameter settings are discussed in Section 4.2.

To capture a majority of nodule morphology, the input nodule patch size was set to $64 \times 64 \times 64$ voxels. We set the learning rate to be 1.0×10^{-3}. In order to relieve the risk of overfitting, we added an L-2 norm weight decay during the training process and the weight decay coefficiency was 5×10^{-4}. We let $b_l = 3$ and $b_u = 8$ in Eq. (2) as [35]. The pooling region size s was 3 with pooling stride of 2 in the first two pooling layers, while s was 4 with a stride of 4 in the third pooling layer to decrease the feature dimension. The size of convolution kernel was 3×3. These chosen parameters were commonly used as discussed in [26,43].

Our MC-CNN implementation was based on CAFFE [44]. HOG and LBP descriptors, which were implemented in the scikit-image package [45], were compared for the performance evaluation of segmentation-free classification methods with our method (Section 4.3). The classifier used was the Support Vector Machine (SVM) classifier from the scikit-learn package [46].

4.2. MC-CNN classification performance

In this section, we perform a systematic evaluation against different parameters including the number of convolution kernels, the position of multi-crop pooling layer, and the dataset sample size. During each round of the five-fold cross validation, there were originally 825 nodules (528 LMNs and 297 HMNs) in the training set and 275 nodules (176 LMNs and 99 HMNs) in either of the validation set and test set. We oversampled HMN samples to approximately balance the training set.

4.2.1. Results with different network configurations

Following the description in Section 4.1, there were 9 different network configurations in total with respect to the number of convolution kernels ($n_{ker} = \{16, 32, 64\}$) and the ith position ($i = \{1, 2, 3\}$) of the multi-crop layer. As shown in Fig. 5, our MC-CNN was stable to different configurations with all being above 86% in accuracy with a maximum standard variation of 0.27% and above 0.90 for the AUC score with a maximum standard variation of 0.0016. The highest classification accuracy obtained was 87.14% from $\mathrm{MC-CNN}_1^{64}$, and the highest AUC score was 0.93 from $\mathrm{MC-CNN}_1^{16}$. Besides n_{ker} and i, we also evaluated the effect of the number of neurons (n_h) in the hidden feature layer. We trained $\mathrm{MC-CNN}_1$ with different n_h from $\{16, 32, 64\}$. Classification accuracy comparison is shown in Fig. 6. It was obvious that the performance was quite stable and the variation was less than 0.3% for a certain n_{ker}. Thus, we chose to fix n_h as 32 in all the remaining experiments due to its relative stability. The encouraging results of the MC-CNN can be ascribed to that the hierarchical learning network selects high-level discriminative features through the multi-crop pooling strategy. And the stable outcomes can be explained that the weight-decay term (Section 4.1) regularizes the weights during the learning process, making results less sensitive to different network capacities.

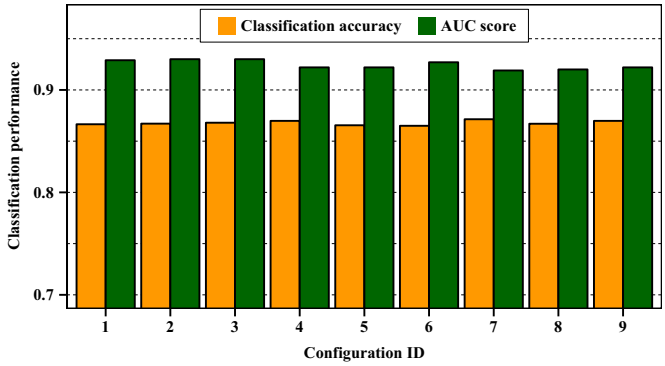

Fig. 5. The classification accuracies and AUC scores of our MC-CNNs using 9 different configurations. The final hidden feature dimension is fixed to 32 for simplicity. Each configuration is assigned to a unique ID for display convenience.

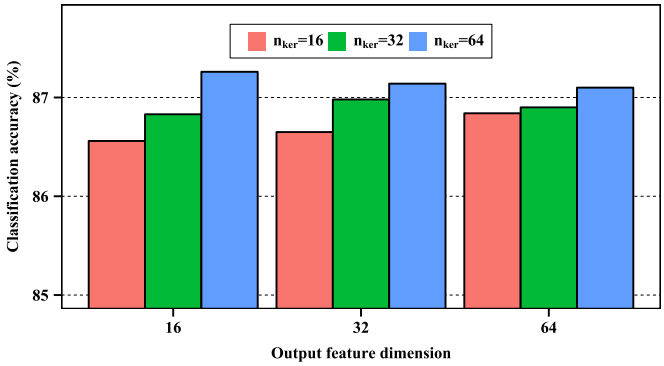

Fig. 6. Classification accuracies from MC-CNNs with different numbers of final hidden feature nodes (n_h). The variation is less than 0.3% for a certain n_{ker} indicating n_h is not a crucial parameter to the performance of our MC-CNN.

4.2.2. Effectiveness of multi-crop pooling features

We justify the effectiveness of the multi-crop pooling features by comparing our MC-CNNs with three other networks without applying multi-crop pooling operation in the feature space. First, since multi-crop pooling on feature maps helped achieve high classification accuracy, we also applied multi-crop pooling operation directly to the image space (i.e., the input nodule patches). Instead of using max-pooling inside the multi-crop pooling layer, we used average-pooling which simulated the image

Table 2
Classification accuracies of a multi-crop CNN with mean-pooling (MC-CNN-MP), a Multi-scale CNN (MCNN), a single CNN (CNN-S) and our $\mathrm{MC-CNN}_1^{64}$.

Network	MC-CNN-MP	MCNN	CNN-S	$\mathrm{MC-CNN}_1^{64}$
Accuracy (%)	86.24	86.53	86.32	**87.14**

Table 3
Classification accuracies on different dataset sizes.

Dataset size	340	1030	1375
Accuracy (%)	83.09	86.36	87.14

downsampling process (MC-CNN-MP). Second, in order to show that the proposed method simplified the training process without sacrificing classification accuracy compared to the traditional multi-scale CNN (MCNN) pipeline, we trained an MCNN [31] using the same input patch size and the same number of layers with our $MC-CNN_1^{64}$. Finally, a traditional single scale CNN was also tested to serve as a baseline. Again, the input patch size and the number of layers were the same with our $MC-CNN_1^{64}$. The results are shown in Table 2. The classification accuracies of all these three networks were lower than that of our $MC-CNN_1^{64}$. The results confirmed three aspects of feature learning for lung nodule classification. First, multi-cop pooling applied on input image patches (MC-CNN-MP) lowered the image resolution leading to an information loss and decreased the classification accuracy. Second, multi-scale features learned in a single network had comparable or even better representative capability than those learned from multiple networks (MCNN). Third, the improvement over CNN-S could be explained that nodule-centric features from nodules with different sizes were consistently persevered in the MC-CNN, while the conventional CNN extracted single scale features from both small and large nodules. Furthermore, speaking of time complexity, the training time of our MC-CNN was nearly one-third of that of the MCNN which indicated the efficiency of the computation.

4.2.3. Performance with varying data samples

We evaluate the performance of $MC-CNN_1^{64}$ on datasets with different sizes by randomly sampling different numbers of nodules from the original dataset including three sub-datasets: a quarter, a half and the entire dataset. The classification accuracies are shown in Table 3. Adding more training data improved the model performance from 83.09% to 87.14%, leading to a performance increase by around 4%. Although Table 3 demonstrated empirical success of the MC-CNN with regard to different sizes of samples, we would expect to collect more nodule samples to further improve and validate the stability of the proposed approach.

4.3. Competing with state-of-the-art approaches

We compare our method with both segmentation-free and segmentation-dependent classification methods in this section. Segmentation-free methods included LBP and HOG descriptors working on nodule patches. Segmentation-dependent methods relied on nodule image segmentation for feature engineering.

Table 4
Classification accuracies using HOG descriptor with different s_w and LBP descriptor with different n_{pt}.

Descriptor	Parameter	32 (%)	64 (%)	96 (%)
HOG	$s_w=8$	**74.18**	66.69	64.07
	$s_w=16$	63.27	66.40	65.16
	$s_w=32$	49.82	56.15	56.58
LBP	$n_{pt}=8$	64.58	49.24	36.00
	$n_{pt}=16$	66.40	59.93	52.22
	$n_{pt}=24$	**67.35**	59.20	54.84

Table 5
Classification performance comparison on the same dataset size.

Method	Autoencoder [30]	Massive-feat [13]	Our method
Accuracy (%)	80.29	83.21	87.14
AUC score	0.86	0.89	0.93
Sensitivity	0.73	0.87	0.77
Specificity	0.85	0.78	0.93

4.3.1. Comparison with HOG and LBP based classification

We first compared our results with commonly used descriptors including HOG and LBP descriptors. For HOG descriptor, we used different cell window sizes, $s_w = \{8, 16, 32\}$ with the number of orientations $n_o=8$. For LBP descriptor, the uniform LBP descriptor was computed with different neighbourhood points $n_{pt} = \{8, 16, 24\}$. The SVM classifier was used for classification. We extracted HOG descriptors and LBP descriptors with three scales on nodule patches, i.e., $32 \times 32 \times 32$ voxels, $64 \times 64 \times 64$ voxels, and $96 \times 96 \times 96$ voxels. Accuracies of HOG and LBP descriptors were shown in Table 4. We found that HOG descriptor was quite sensitive to the size of the cell window (s_w). For LBP descriptor, we observed that the number of neighborhood points (n_{pt}) was positively related to the performance probably because sophisticated neighborhood structures led to improved results. However, when competing with the best results among these two descriptors, our method outperformed them by 12.96% and 19.79% respectively. Overall, our observation confirmed that parametric textural descriptors were sensitive to various parameters.

4.3.2. Comparison with segmentation-dependent classification

We have reported related results on the LIDC-IDRI dataset in the literature in Table 1. Although we noticed that different number of samples were used which made the fair comparison difficult, the results from our MC-CNN were still quite competitive in terms of both classification accuracy and the AUC score. In this section, we included two more metrics: sensitivity and specificity. The sensitivity and the specificity of our method were 0.77 and

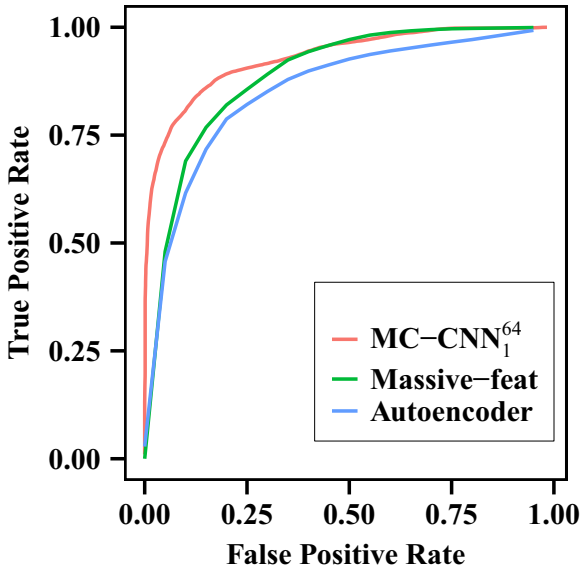

Fig. 7. The receiver operating characteristic curve (ROC curve) of our $MC-CNN_1^{64}$, the Massive-feat method and the Autoencoder method. It can be seen that the ROC of our $MC-CNN_1^{64}$ (AUC=0.93) is very competitive compared to the Autoencoder method (AUC=0.86) and the Massive-feat method (AUC=0.89).

0.93 respectively. We implemented two approaches in the literature for additional comparison with the same number of nodules (see Table 5 and Fig. 7). The first one is the autoencoder-based method (Autoencoder) [30] in an unsupervised learning scheme. We tested it on the same dataset and achieved the classification accuracy of 80.29% with an AUC score of 0.86. The sensitivity and the specificity were 0.73 and 0.85 respectively. The lack of prior label information in the unsupervised learning may lead to a sub-optimal feature learning, causing its outcomes lower than that of the MC-CNN. Second, we implemented a massive feature mining pipeline (Massive-feat) following the strategy in [13]. Four types of nodule image features were included: first order statistics, shape and size features, textural features, and wavelet features. After feature selection using the mRMR score [47], the top 54 features were chosen from the extracted 319 dimensional features. Applying the SVM classifier led to the best accuracy of 83.21% with an AUC score of 0.89. The sensitivity and specificity were 0.87 and 0.78 respectively. Both the classification accuracies and the AUC scores from two implemented approaches were shown inferior to the proposed MC-CNN. Though the sensitivity of our method was lower than that of the Massive-feat method, the specificity of our MC-CNN was higher compared to those of the other two methods. More importantly, the reporting figures of our approach here were not meant to lead a significant improvement over the current literature. Instead, we sought to demonstrate an alternative feature extraction pipeline that can complement state-of-the-art architectures for lung nodule analysis.

4.4. Exploratory analysis on modeling nodule malignancy uncertainty

Since our prior results were based on a binary setting of malignancy suspiciousness classification, in this section, we extended to estimate nodule malignancy uncertainty by taking into account nodules with a moderate score of 3. The uncertainty estimation on nodule malignancy suspiciousness is challenging because of ambiguous assessment from human experts. We provided exploratory evidence to model uncertain nodules by analyzing inclination of uncertain samples to the distinct group of LMNs or HMNs, so that we may be able to gain insight into better patient sub-group stratification.

Two tasks were designed to quantify nodule malignancy uncertainty by applying the model of MC – CNN_1^{64}. First, the uncertain nodules with a score 3 were either categorized into LMNs or HMNs respectively. Second, we additionally treated them as an independent category and performed classification on three groups. Table 6 showed the classification results with uncertain nodules included. Comparing the first row in Table 6 with our best results, we found that uncertain nodules made the model slightly inferior to that trained without uncertain nodules, probably since inclusion of uncertain nodules introduced variation within each category. Comparing first and second columns, we found that incorporating uncertain cases into LMNs led to better results than incorporating them into HMNs.

Table 6
Classification accuracies including uncertain nodules. "UNs" indicate uncertain nodules. "IC" indicates an independent category.

Settings	UNs as LMNs (%)	UNs as HMNs (%)	UNs as IC (%)
UNs in training set	86.12	85.60	–
UNs in both training and test sets	87.29	72.57	62.46

The observation indicated that uncertain cases shared more similarities with LMNs. The finding presented that radiologists seemed to have a biased scoring towards classifying some LMS cases into uncertain nodules. Our observation was consistent with the study [16]. Also, the dropped accuracy was observed when the uncertain nodules were regarded as an independent category. The result is not surprising since nodules with a moderate score present heterogeneous characteristics in nature, leading to deteriorate performance in classification during both training and test phases. The evidence data here suggests that a more sophisticated comparison will be needed to investigate between subtle sub-groups in the future.

4.5. Nodule semantic prediction and diameter estimation

Beyond nodule malignancy suspiciousness classification, we quantify nodule semantic prediction including subtlety and margin and nodule diameter estimation using the MC – CNN_1^{64}.

4.5.1. Nodule subtlety and margin prediction

We performed semantic label prediction including two attributes: subtlety and margin. Subtlety indicates the difficulty in nodule detection which refers to the contrast between the lung nodule and its surroundings, while margin describes how well-defined the margins of the defining nodule [48]. The model we used here is the classification model of MC – CNN_1^{64} and the evaluation method is also the five-fold cross validation. The classification accuracy of subtlety is 74.32% and that of margin is 76.99%.

4.5.2. Nodule diameter estimation

The diameter of a solitary pulmonary nodule (SPN) can be a useful predictor of malignancy—larger diameter indicates the increasing suspiciousness of nodule malignancy [49]. In this experiment, given the five-fold cross validation, we used the regression version of MC – CNN_1^{64} (Section 2.4). The metric used to evaluate the estimation performance was the relative estimation error

$$E_r = \frac{|d_{est} - d_{truth}|}{d_{truth}}, \tag{7}$$

where d_{est} is the estimated diameter and d_{truth} indicates the ground truth diameter. The distribution of E_r is shown in Fig. 8. The population of E_r less than 0.2, 0.3 and 0.4 respectively occupy 73.78%, 84.54% and 90.15% of the entire dataset. The results suggest an alternative way of estimating the nodule diameter, indicating a strong correlation of the learned deep features with lung nodule diameter distribution.

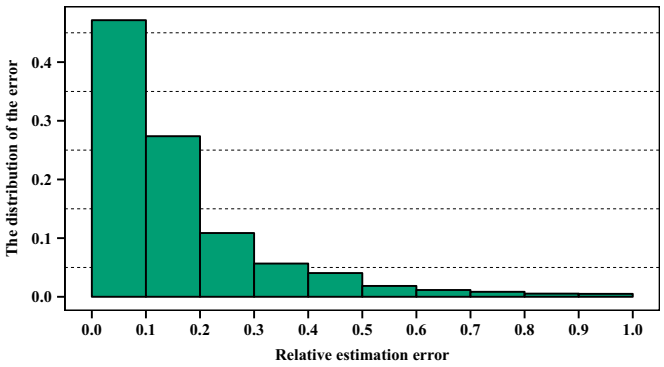

Fig. 8. The distribution of the relative estimation error E_r. The population of E_r within 0.2, 0.3 and 0.4 occupy 73.78%, 84.54% and 90.15% of the entire dataset.

Table 7
Processing time of each method. Training time is measured for one round in the five-fold cross validation and the test time is averaged for one single nodule.

Methods	Segmentation	Training time		Test time
		Feature extraction	Classifier	
Massive-feat [13]	Manual	10h15min25s	0.09 s	32.76 s
Autoencoder [30]	Manual	7.29 s	0.14 s	0.01 s
MC – CNN_1^{64}	–	47min01s		0.23 s

5. Discussion

In this paper, we proposed a deep learning computational architecture, called MC-CNN, to classify nodule malignancy suspiciousness using CT images. It is an end-to-end architecture which embeds nodule feature extraction into a hierarchical network and simplifies conventional lung nodule malignancy suspiciousness classification by removing nodule segmentation and hand-crafted feature engineering. Providing early suspiciousness estimation from imaging allowed a strategy of non-invasively identifying patient sub-groups before treatments of needle biopsy or surgical resection. Thus, it has the potential to facilitate radiologists to discern the underlying risk factors for better individualized patient management. Experimental results demonstrated that our proposed method achieved promising results in both classification accuracy (87.14%) and the AUC score (0.93). Additional semantic prediction and diameter estimation reaffirmed the strength of the proposed approach in characterizing nodule-related information. To further assimilate diagnostic values of the proposed approach, we would expect to validate the deep learning architecture by incorporating additional lung cancer imaging studies with both radiologists' opinions and follow-up pathologic scores. As vast quantity of clinical imaging sequences are becoming increasingly available, our data-driven model holds promise for early diagnosis with more rapid clinical translation.

Although accurate processing time comparison of different methods is difficult, we list the time consumed by the Massive-feat method [13], the Autoencoder method [30] and our MC – CNN_1^{64} in Table 7. All the methods run on the same machine with 12GB memory and a 6-core Intel Xeon CPU. Nvidia Tesla K40 GPU was enabled for our MC – CNN_1^{64} model and the Autoencoder method. We did not migrate the feature extraction code in the Massive-feat method to GPU. The manual segmentation time for the Massive-feat method and the Autoencoder method were not included because the segmentation was provided in the dataset. We found that hand-crafted feature extraction in the Massive-feat method was very time-consuming which took more than 10 h and the test time for one single nodule was also much longer than that of the other two methods. The input data of the Autoencoder method was the 2-D nodule slice while that of our method was 3-D which brought much more computational cost. This could explain why our method took more time than the Autoencoder method. It was also obvious that our method could simplify the traditional nodule analysis pipeline by removing nodule segmentation and feature extraction.

The rationale for seeking "deep features" is that deep learning networks would make mostly correct assumptions about the nature of images by varying the depth and breadth of the network capacity [26]. As the results in [50], through hierarchical networks, the CNN can produce a dimensional reduction that is particularly helpful for image-related classification. Our MC-CNN prioritized a repetitive pooling strategy for nodule-centric feature extraction. By considering different regions of the feature maps independently, the strategy preserved more details of the salient region of nodules. Thus, features from small nodules were also well kept and forwarded to the following layers, indicating that our MC-CNN was able to capture a variety of nodule dynamic structures.

Dropout [51] is a known strategy to prevent the CNN from over-fitting. However, we did not observe much test difference between networks with or without dropout on the LIDC-IDRI dataset. The reason may be ascribed to that the LIDC-IDRI dataset is quite different from generic image datasets (e.g., the imagenet dataset) having thousands of categories, where the learning models are prone to overfit decision boundaries easily. The use of 3-D augmentation created augmented training samples that preserve intra-class variation to minimize the potential over-fitting issue. Additionally, results on semantic prediction and diameter estimation revealed the generalized performance of the proposed method.

The proposed study complemented the traditional approaches. The only prerequisite of our method is the identification of the nodule central location, which is a substitute of conventional nodule image segmentation. Both multi-crop pooling and max-pooling can tolerate a small amount of shift of a nodule center point, thus the process actually does not require an accurate localization of nodule centers. This suggests an appealing strategy of approach initialization. However, when dealing with a growing number of clinical imaging sequences, an extra automatic nodule detection process might be needed for both our method and the traditional nodule classification methods in order to speed up the diagnosis process.

6. Conclusion

Deep learning architecture is a rising computational paradigm in developing predictive models of diseases. In this paper, we introduced a deep learning model of MC-CNN to tackle the challenging problem of lung nodule malignancy suspiciousness classification. We demonstrated that the learned deep features were able to capture nodule salient information by the multi-crop pooling strategy. The encouraging results on nodule malignancy suspiciousness classification showed the effectiveness of our MC-CNN. Additional experiments on nodule semantic prediction and nodule diameter estimation revealed that the proposed method could be potentially helpful to other aspects of nodule-relevant characteristics analysis. In general, the extracted deep features can be considered to be integrated with conventional image features to further improve the precision performance for lung cancer patients.

Acknowledgment

This paper is supported by the Chinese Academy of Sciences Key Deployment Program under Grant No. KGZD-EW-T03, the National Natural Science Foundation of China under Grant Nos. 81227901, 81527805, 61231004, 81370035, 81230030, 61301002, 61302025, 81301346, and 81501616, the Beijing Natural Science Foundation under Grant No. 4132080, the Fundamental Research Funds for the Central Universities under Grant No. 2013JBZ014, 2016JBM018, the Scientific Research and Equipment Development Project of Chinese Academy of Sciences under Grant No. YZ201457. The authors also acknowledge 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 study. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2015. CA: Cancer J. Clin. 65 (1) (2015) 5–29.

[2] A.C. Society, Cancer Facts and Figures 2015, 2015.

[3] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G. van Stiphout, P. Granton, C.M. Zegers, R. Gillies, R. Boellard, A. Dekker, et al., Radiomics: extracting more information from medical images using advanced feature analysis, Eur. J. Cancer 48 (4) (2012) 441–446.

[4] M.K. Gould, L. Ananth, P.G. Barnett, A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules, CHEST J. 131 (2) (2007) 383–388.

[5] M.M. Wahidi, J.A. Govert, R.K. Goudar, M.K. Gould, D.C. McCrory, Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer?: Accp evidence-based clinical practice guidelines, CHEST J 132 (3_suppl) (2007) 94S–107S.

[6] M.K. Gould, J. Donington, W.R. Lynch, P.J. Mazzone, D.E. Midthun, D.P. Naidich, R.S. Wiener, Evaluation of individuals with pulmonary nodules: when is it lung cancer?: diagnosis and management of lung cancer: American college of chest physicians evidence-based clinical practice guidelines, CHEST J. 143 (5_suppl) (2013) e93S–e120S.

[7] Valente IRS, Cortez PC, Neto EC, Soares JM, de Albuquerque VHC, Tavares JMR, Automatic 3d pulmonary nodule detection in ct images: a survey, Comput. Methods Programs Biomed. 2016;124:91–107.

[8] M.N. Gurcan, B. Sahiner, N. Petrick, H.-P. Chan, E.A. Kazerooni, P.N. Cascade, L. Hadjiiski, Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system, Med. Phys. 29 (11) (2002) 2552–2558.

[9] Y. Gu, V. Kumar, L.O. Hall, D.B. Goldgof, C.-Y. Li, R. Korn, C. Bendtsen, E. R. Velazquez, A. Dekker, H. Aerts, et al., Automated delineation of lung tumors from ct images using a single click ensemble segmentation approach, Pattern Recognit. 46 (3) (2013) 692–702.

[10] J. Song, C. Yang, L. Fan, K. Wang, F. Yang, S. Liu, J. Tian, Lung lesion extraction using a toboggan based growing automatic segmentation approach, IEEE Trans. Med. Imaging 99 (2015) 1, http://dx.doi.org/10.1109/TMI.2015.2474119.

[11] T. Messay, R.C. Hardie, T.R. Tuinstra, Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset, Med. Image Anal. 22 (1) (2015) 48–62.

[12] A. Qu, J. Chen, L. Wang, J. Yuan, F. Yang, Q. Xiang, N. Maskey, G. Yang, J. Liu, Y. Li, Segmentation of hematoxylin-eosin stained breast cancer histopathological images based on pixel-wise svm classifier, Science China Inf. Sci. (2015) 1–13.

[13] H.J. Aerts, E.R. Velazquez, R.T. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nature Commun. 2014;5.

[14] Y. Balagurunathan, Y. Gu, H. Wang, V. Kumar, O. Grove, S. Hawkins, J. Kim, D. B. Goldgof, L.O. Hall, R.A. Gatenby, R.J. Gillies, Reproducibility and prognosis of quantitative features extracted from ct images, Transl. Oncol. 7 (1) (2014) 72–87.

[15] T.W. Way, L.M. Hadjiiski, B. Sahiner, H.-P. Chan, E.A. Kazerooni, P.N. Cascade, N. Bogot, C. Zhou, Computer-aided diagnosis of pulmonary nodules on ct scans: segmentation and classification using 3d active contours, Med. Phys. 33 (7) (2006) 2323–2337.

[16] F. Han, H. Wang, G. Zhang, H. Han, B. Song, L. Li, W. Moore, H. Lu, H. Zhao, Z. Liang, Texture feature analysis for computer-aided diagnosis on pulmonary nodules, J. Digit. Imaging 28 (1) (2015) 99–115.

[17] F. Ciompi, B. de Hoop, S.J. van Riel, K. Chung, E.T. Scholten, M. Oudkerk, P.A. de Jong, M. Prokop, B. van Ginneken, Automatic classification of pulmonary perifissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box, Med. Image Anal. 26 (1) (2015) 195–202.

[18] C. Dong, Y. Yin, X. Yang, Detecting malignant patients via modified boosted tree, Sci. China Inf. Sci. 53 (7) (2010) 1369–1378.

[19] D. Zinovev, J. Feigenbaum, J. Furst, D. Raicu, Probabilistic lung nodule classification with belief decision trees, in: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, IEEE, Boston, MA, USA, 2011, pp. 4493–4498.

[20] H. Chen, W. Wu, H. Xia, J. Du, M. Yang, B. Ma, Classification of pulmonary nodules using neural network ensemble, in: Advances in Neural Networks—ISNN 2011, Springer, Guilin, China, 2011, pp. 460–466.

[21] S.G. Armato III, G. McLennan, L. Bidaut, M.F. McNitt-Gray, C.R. Meyer, A. P. Reeves, B. Zhao, D.R. Aberle, C.I. Henschke, E.A. Hoffman, et al., The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans, Med. Phys. 38 (2) (2011) 915–931.

[22] Y. Song, W. Cai, Y. Zhou, D.D. Feng, Feature-based image patch approximation for lung tissue classification, IEEE Trans. Med. Imaging 32 (4) (2013) 797–808.

[23] F. Zhang, Y. Song, W. Cai, M.-Z. Lee, Y. Zhou, H. Huang, S. Shan, M.J. Fulham, D. D. Feng, Lung nodule classification with multilevel patch-based context analysis, IEEE Trans. Biomed. Eng. 61 (4) (2014) 1155–1166.

[24] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[25] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1, 2005, pp. 886–893.

[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[27] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Columbus, Ohio, USA, 2014, pp. 1891–1898.

[28] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Columbus, Ohio, USA, 2014, pp. 580–587.

[29] I. Jolliffe, Principal Component Analysis, Wiley Online Library, 2005, http://dx.doi.org/10.1002/9781118445112.stat06472.

[33] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[34] A.P. Reeves, Y. Xie, A. Jirapatnakul, Automated pulmonary nodule ct image characterization in lung cancer screening, Int. J. Comput. Assist. Radiol. Surg. (2015) 1–16.

[35] B. Xu, N. Wang, T. Chen, M. Li, Empirical Evaluation of Rectified Activations in Convolutional Network, arXiv preprint arXiv:1505.00853.

[36] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proceedings of the ICML, vol. 30, 2013.

[37] K. He, X. Zhang, S. Ren, J. Sun, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, arXiv preprint arXiv:1406.4729.

[38] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.

[39] D.R.G.H.R. Williams, G. Hinton, Learning representations by back-propagating errors, Nature (1986) 323–533.

[40] F. Han, G. Zhang, H. Wang, B. Song, H. Lu, D. Zhao, H. Zhao, Z. Liang, A texture feature analysis for diagnosis of pulmonary nodules using lidc-idri database, in: IEEE International Conference on Medical Imaging Physics and Engineering, 2013, pp. 14–18.

[41] H.R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, R. M. Summers, A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, Springer, 2014, pp. 520–527.

[42] H.R. Roth, J. Yao, L. Lu, J. Stieger, J.E. Burns, R.M. Summers, Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications, in: Recent Advances in Computational Methods and Clinical Applications for Spine Imaging, Springer, Boston, MA, USA, 2015, pp. 3–12.

[43] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv preprint arXiv:1409.1556.

[44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the ACM International Conference on Multimedia, ACM, New York, NY, USA, 2014, pp. 675–678.

[45] S. van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, N. Yager, E. Gouillart, T. Yu, Scikit-Image: Image Processing in Python, Technical Report, PeerJ PrePrints, 2014.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[47] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[48] M.F. McNitt-Gray, S.G. Armato, C.R. Meyer, A.P. Reeves, G. McLennan, R.C. Pais, J. Freymann, M.S. Brown, R.M. Engelmann, P.H. Bland, et al., The lung image database consortium (lidc) data collection process for nodule detection and annotation, Acad. Radiology. 14 (12) (2007) 1464–1474.

[49] H.T. Winer-Muram, The solitary pulmonary nodule 1, Radiology 239 (1) (2006) 34–49.

[50] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1735–1742.

[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

Wei Shen is a Ph.D. candidate in Pattern Recognition and Intelligent Systems, Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include medical image processing and pattern recognition.

Mu Zhou is a postdoctoral fellow at Stanford Center for Biomedical Informatics Research, Stanford University, CA, USA. He earned his Ph.D. degree in computer engineering in 2015 at University of South Florida, Tampa, USA. His research interests include medical image analysis, deep learning, data mining, and multi-scale biomedical data integration.

Feng Yang is an associate professor from School of Computer Science and Information Technology in Beijing Jiaotong University. Her current research interests include medical image processing, cardiac fiber architecture reconstruction from DTI, and deep learning.

Dongdong Yu is a Ph.D. candidate in Pattern Recognition and Intelligent Systems, Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include medical image registration, optimization and pattern recognition.

Di Dong received his Ph.D. degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences, in 2013. He is an assistant professor at the Institute of Automation, Chinese Academy of Sciences. His current research interests include radiomics on lung cancer and glioma.

Caiyun Yang received the Ph.D. degree from The University of Tokyo in 2012. She is an assistant professor at the Institute of Automation, Chinese Academy of Sciences. Her research interests include image processing and image modeling.

Yali Zang received the Ph.D. degree from University of Chinese Academy of Sciences, in 2013. She is an assistant professor at the Institute of Automation, Chinese Academy of Sciences. Her research interests include image processing, biometrics, radiomics of lung cancer and GBM.

Jie Tian received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1993. He is a professor at the Institute of Automation, Chinese Academy of Sciences. His research interests include medical image processing and analysis and pattern recognition.