

Classification of Lung Nodule Malignancy Risk on Computed Tomography Images Using Convolutional Neural Network: A Comparison Between 2D and 3D Strategies

Xingjian Yan^{1,3}, Jianing Pang^{1,2}, Hang Qi^{1,3}, Yixin Zhu^{1,3}, Chunxue Bai⁴, Xin Geng⁵, Mina Liu⁶, Demetri Terzopoulos^{1,3}, and Xiaowei Ding^{1,3}(✉)

¹ VoxelCloud Inc., Los Angeles, CA 90012, USA
xding@voxelcloud.io

² Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

³ University of California, Los Angeles, Los Angeles, CA 90095, USA

⁴ Shanghai Zhongshan Hospital, Shanghai, China

⁵ Department of Cardiothoracic Surgery,

Huashan Hospital of Fudan University, Shanghai, China

⁶ Shanghai Chest Hospital of Shanghai Jiao Tong University, Shanghai, China

Abstract. Computed tomography (CT) is the preferred method for non-invasive lung cancer screening. Early detection of potentially malignant lung nodules will greatly improve patient outcome, where an effective computer-aided diagnosis (CAD) system may play an important role. Two-dimensional convolutional neural network (CNN) based CAD methods have been proposed and well-studied to extract hierarchical and discriminative features for classifying lung nodules. It is often questioned if the transition to 3D will be a key to major step forward in performance. In this paper, we propose a novel 3D CNN on the 1018-patient Lung Image Database Consortium collection (LIDC-IDRI). To the best of our knowledge, this is the first time to directly compare three different strategies: slice-level 2D CNN, nodule-level 2D CNN and nodule-level 3D CNN. Using comparable network architectures, we achieved nodule malignancy risk classification accuracies of 86.7%, 87.3% and 87.4% against the personal opinion of four radiologists, respectively. In the experiments, our results and analyses demonstrates that the nodule-level 2D CNN can better capture the z-direction features of lung nodule than a slice-level 2D approach, whereas nodule-level 3D CNN can further integrate nodule-level features as well as context features from all three directions in a 3D patch in a limited extent, resulting in a slightly better performance than the other two strategies.

1 Introduction

Lung cancer is the deadliest type of cancer worldwide. It is estimated that lung cancer caused 158,040 deaths in the US in 2015, which accounts for nearly 40% of all cancer deaths in the country [1]. Worldwide, it caused 1.69 million deaths

in 2012, nearly 20% of the total cancer deaths in the world [2]. The prognosis of lung cancer depends critically on the stage at which it is diagnosed. The five-year survival rate of early-stage disease is over 50%, whereas that of advanced stage is less than 5%. Currently, more than half of the diagnosed lung cancer cases are in advanced stage [1]. Therefore, effective screening of lung cancer is crucial for detecting the disease in an early and more treatable stage, consequently improving patient survival rates.

Computed tomography (CT) is the current preferred method for non-invasive lung cancer screening due to its high sensitivity. The National Lung Screening Trial, which enrolled more than 50,000 high-risk subjects, demonstrated that low dose CT screening reduced lung cancer mortality by more than 20% compared with chest radiography screening [3]. Despite its promise, the current lung cancer screening method by CT bears several limitations:

1. Interpretation of the CT images requires analyzing hundreds of images at a time, considerably increasing the workload of radiologists;
2. Significant interobserver and intraobserver variations make the screening result subjective and less reliable [4];
3. The false positive rate remains high, limiting the utility of CT as a early screening modality [5].

In this work, we implemented a self-contained artificial neural network for classifying malignant and benign lung nodules, and tested three strategies for feeding the 3D nodule volume into the network: independent 2D slices with nodule-level voting, simultaneous multi-slice input, and full 3D volumetric input. All nodules were extracted from the annotated images in the publicly available Lung Image Database Consortium (LIDC) [6].

1.1 Related Work

Considerable efforts have been devoted to developing efficient, observer-independent computer aided diagnosis (CAD) methods for differentiating malignant from benign nodules [7–14]. The general scheme usually includes first designing and extracting features (e.g. geometry, texture, opacity, etc.) from image patches, and then training a classifier (e.g. linear discriminant analysis, support vector machine, artificial neural network, etc.) to categorize nodules. The performance is usually evaluated by the receiver operating characteristics (ROC).

In contrast to handcrafted features, deep learning methods learn a representation of data via training end-to-end and are capable of automatically extracting features specific to the learning task at hand. The image data is fed directly into a multi-layer convolutional neural network (CNN) that includes convolutional, pooling, and fully connected layers. Leveraging the availability of large training datasets, advances in training algorithms, and increasingly accessible computational power, such methods have achieved impressive performance in various tasks in computer vision (e.g. [15, 16]). Several deep learning based approaches

for CT lung nodule classification have also been proposed, with differences in network configuration, nodule extraction strategies, and whether the network is self-contained or requires a separate classifier (i.e. network is used for feature extraction only) [17–19]. The major limitation of these works is the use of individual 2D patches as the network input and the subsequent patch-level classifications. Certain information from the nodule-level, e.g. texture and context features in the z-direction, are ignored, which is a sub-optimal setup as the nodules are intrinsically 3D objects.

1.2 Contributions

Our work made the following three contributions:

1. We proposed a 3D convolutional neural network method for lung nodule malignancy risk classification.
2. We compared three neural network input strategies: 2D slice level CNN, 2D nodule level CNN, and 3D nodule level CNN.
3. Using 3D CNN approach, we achieved the best classification accuracy reported (87.4%) on LIDC-IDRI dataset.

2 Data and Method

We implemented three convolutional neural network (CNN) input strategies (slice-level 2D CNN, nodule-level 2D CNN, and 3D CNN) and evaluated their performance on lung nodule malignancy risk classification. All nodules were extracted from the annotated images in the publicly available Lung Image Database Consortium (LIDC) [6] and the reference standard of nodule malignancy risk were obtained using a reader opinion voting procedure described below.

2.1 Data

Both locations and malignancy risk scores of the lung nodules in each CT scan are annotated in 1018 CT images from the LIDC database. The scores range from 1 (not suspicious to be malignant) to 5 (highly suspicious to be malignant) and are given by panels of up to four radiologists. To extract a single score from the panels’ readings, majority voting was conducted to account for the subjectivity of each expert’s experience. a score of less than 3 was considered one vote for benign, and one above 3 was one vote for malignant, while a score of 3 was discarded. The final classification of a nodule was then determined by the majority vote. If there were equal votes for both classes, the nodule would be removed from the study. A similar voting approach was also used in [17]. Note that the binary label for each nodule was based on the subjective opinions of the readers, therefore was not equivalent to a biopsy or outcome proved ground truth. In the 1882 nodules used in our experiment, the number of nodules voted benign was roughly twice that of nodules voted as malignant. Simple data augmentations were performed (e.g. rotation) to increase the size of the training set and also balance the number of benign and malignant nodules.

2.2 Slice-Level 2D Convolutional Neural Network

A typical self-contained CNN image classifier is a function that takes an image as an input and outputs a $c \times 1$ vector, where c is the number of classes and the i -th element of the vector is the probability of the input image belonging to the i -th class.

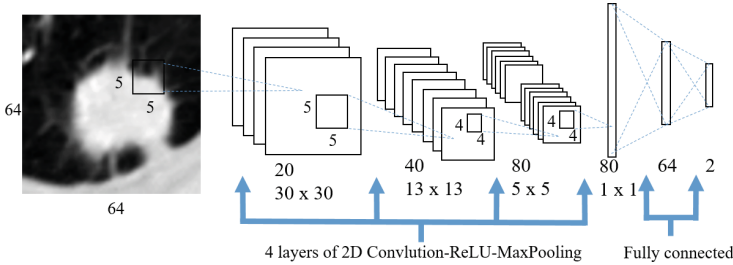


Fig. 1. The network architecture of slice-level 2D CNN.

The configuration of our slice-level 2D CNN is illustrated in Fig. 1. The images are first fed into four sets of 2D convolution-ReLU-pooling layers with 20, 40, 80, and 80 filters of size 5×5 , 5×5 , 4×4 , and 4×4 respectively. The activation function is set as rectified linear unit (ReLU) [20]. The output feature map from each convolution filter all feed into max pooling layers with kernel size 2×2 . To maximize the complexity of the model on such input with small size in each direction (64×64), the kernel of the three consecutive convolution-activation-maxpooling layers is set such that the dimension of the last output feature map is 1×1 . Finally, we compose a fully connected layer with 64 neurons (with batch-normalization and 50% dropout [21]) and a softmax layer with two outputs. The entire network, as well as for the following two networks, is trained from scratch with randomly initialized weights. To accommodate the large dataset size, the optimization method was chosen to be stochastic gradient descent with 100 nodules per batch (same of the following two models), following the heuristic proposed in [16] to control the learning rate.

The nodule patches are automatically extracted using the labeled center-of-mass, contour, and diameter information. As shown in Fig. 2, the distribution of lung nodule sizes varies from 1.5 to 35 mm, with a mean of 8 mm and standard deviation of 6 mm. Since the neural network requires a fixed size of input patches, for each nodule, we cropped out 5 patches of of size 64×64 pixels in the $x - y$ plane (approximately 40 mm field of view), with the middle patch’s center being the center of mass of the nodule. During testing, we choose the majority vote of the output from the network of the five patches to be the final result.

2.3 Nodule-Level 2D Convolutional Neural Network

The slice-level strategy treats all 2D slices as independent examples. Essentially, it “shuffles” nodules in the training examples and discards all information

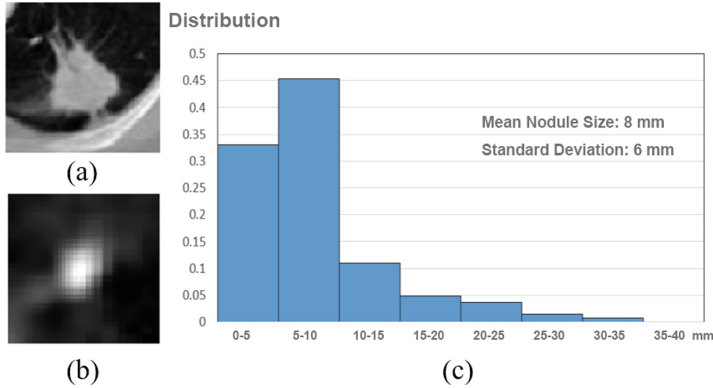


Fig. 2. (a) and (b) are examples of axial slices of two different lung nodules with size of 20 mm and 3 mm. (c) shows the distribution of the nodules’ size in the x direction. The horizontal axis is the size range, and the vertical axis is the probability of the total nodules in different size intervals.

along z direction, including texture features and correlations between slices that belonged to the same nodule. To address these shortcomings, we proposed a modified configuration that classified the nodules using multiple 2D slices simultaneously.

The simultaneous multi-slice network took a 3D patch and internally interpreted it as an image with multiple channels. The network has exactly the same architecture as that of the Slice-level 2D CNN as shown in Fig. 1.

The pipeline takes a 3D patch with size $64 \times 64 \times 5$, obtained the same way as in the Slice-level case, and consider it as one image with five channels. The nodule level 2D CNN architecture allows the network to be trained and tested on a nodule by nodule basis and eliminates the need for voting in the slice-level classification approach described above.

2.4 3D Convolutional Neural Network

Given the two configurations above, the intuitive next step is to extract features from all three dimensions at the same time, i.e. building a 3D CNN models that takes the advantages of all 3D information provided by the images. Our implementation is based on [22]. Each convolutional layer contains a number of 3D filters, where each has a size of $w \times h \times z$, where w , h , and z are the width, height, and depth, respectively. Such 3D filters are more powerful than 2D filters [22, 23] in the sense that they not only capture spatial relationships between different axial slices, but also are capable to detect the volumetric differences between nodules.

An illustrated in Fig. 3, the 3D CNN network has a similar structure to nodule-level 2D CNN. The network has four sets of 3D convolution-ReLU-pooling layers, followed by two fully connected layers with 50% dropout. The

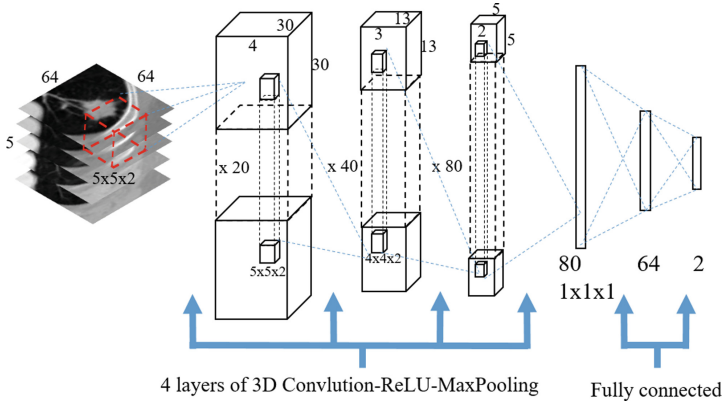


Fig. 3. Illustration of network architecture for 3D CNN.

last layer of the network is a softmax fully connected layer. Each convolutional layer consists of 20, 40, 80, and 80 filters and kernels of size $5 \times 5 \times 2$, $5 \times 5 \times 2$, $4 \times 4 \times 2$, and $4 \times 4 \times 2$, respectively. 2×2 max-pooling in the x and y dimension is applied in the pooling layers. The two fully connected layers have 64 and 2 nodes, respectively (with batch normalization as well as 50% dropout). The convolution kernels of each layer of the 3D network has one times more weights than the above 2D models. With increased complexity, the 3D network should theoretically produce at least as good result as the 2D networks. Stochastic gradient descent was used to train the 3D network. For comparison, the same $64 \times 64 \times 5$ patch is used to train the 3D CNN.

3 Experiment Results

3.1 Implementation Details

First, 300 nodules were randomly selected as the testing set with balanced number of benign and malignant nodules, while the remaining over 1500 nodules were selected as training set. Note that the number of benign nodules in the original train set was almost double that of malignant nodules. Thus, for the purpose of balancing the dataset, we doubled the number of malignant nodules in the training set by adding a copy of them with small random translation. Then, we augmented the training set by rotating the nodules 90° four times along the z axis with respect to the center of each patch and flipping to generate over 25000 nodules. Such a setting helped capture a range of translation and rotation invariant features. We split the training set equally into 5 sets and used them to perform a 5-fold cross validation for evaluating classification performance of the three types of CNN we have trained. In each fold, there were 5000 nodules and the number of benign and malignant nodules were very close due to augmentation and shuffling as described above. Each fold of cross validation of the

three models was trained for 20 epochs, and the loss generally converges after 10 epochs. The CNN implementation used in this work was the deep learning toolkit Torch [24].

3.2 Results

The three models each with five set of weights produced from cross validation is tested with the testing set of nodules. The five individual accuracies for each network in the 5-fold cross-validation phase are very close (within 3%). The averaged classification accuracy on the testing set achieved by the Slice-level 2D CNN, Nodule-level 2D CNN, and 3D CNN are 86.7%, 87.3% and 87.4% respectively. Table 1 shows the performance metrics, including accuracy, sensitivity, and specificity, averaged over five outcomes of cross validation.

We picked weights of the three models that performed best in testing and drew their testing results as ROC curve as shown in Fig. 4. The area under the ROC curves (ROC AUC) are listed in Table 1. The overall performance of three classifiers suggested that our method can achieve promising results. The 3D approach slightly outperformed the other two in global accuracy and Sensitivity. The advantages of 3D approach in classification performance is limited. It can be ascribed to a factor that although 3D convolutional neural networks can produce useful dimensional reduction without losing the information from the third dimension that is very helpful for lung nodule classification tasks, the low spacial resolution of CT images in z direction (compared to the resolution in x-y directions) limited this capability of 3D CNN.

Table 1. Accuracy measure of three different models

Models	Accuracy	Sensitivity	Specificity	ROC AUC
2D CNN slice-level	86.7%	78.6%	91.2%	0.926 ± 0.014
2D CNN nodule-level	87.3%	88.5%	86.0%	0.937 ± 0.014
3D CNN	87.4%	89.4%	85.2%	0.947 ± 0.014

4 Discussion

Several previous works also utilized the LIDC dataset and proposed CNN based methods for lung nodule malignancy risk classification. Shen et al. [17] proposed to use a multi-scale CNN for feature extraction from center 2D nodule patches, and support vector machine or random forest for classification. The highest achieved accuracy was 86.8%. Kumar et al. [18] proposed to use a five-layer autoencoder to extract features from 2D patches and a decision tree for classification. The mean achieved accuracy was 75.01%. Hua et al. [19] proposed to use a deep belief network (DBN) or CNN for both feature extraction and classification from independent 2D patches. The achieved sensitivity and specificity for the DBN/CNN approach were 73.4/73.3% and 82.2/78.7%, respectively.

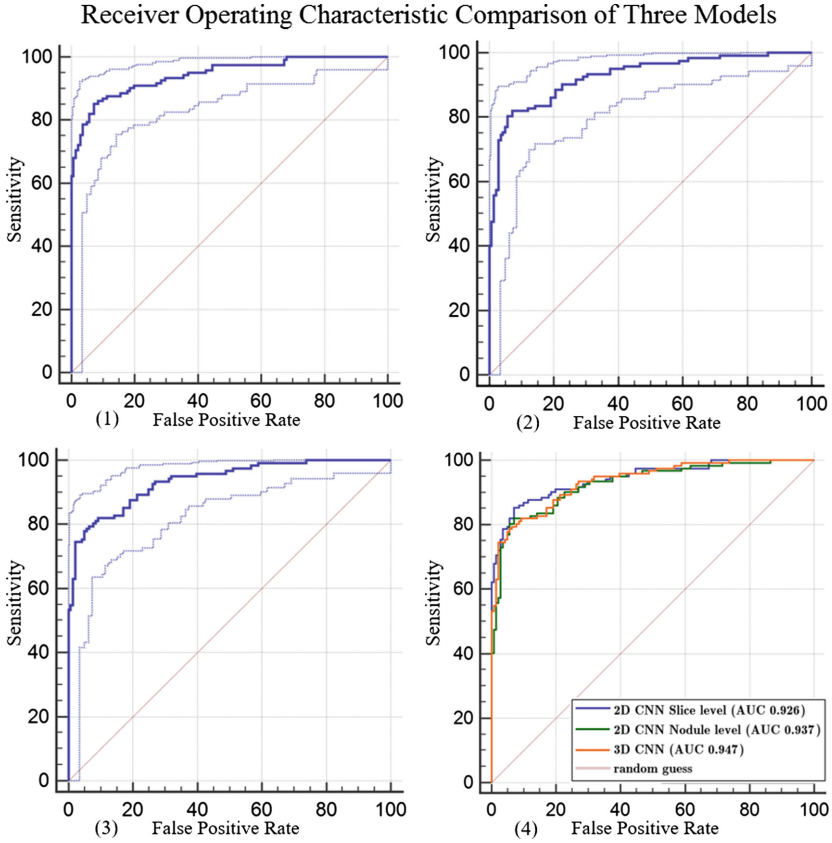


Fig. 4. Receiver operating characteristic comparison of three different models. Plot (1), (2), and (3) are ROC responses of 3D CNN, 2D CNN Nodule Level, and 2D CNN Slice Level respectively. The dashed lines indicate the 95% confidence interval. Plot (4) is a comparison of the ROC responses.

Considering these previous attempts, the major contributions of our work included: (1) implementation of a voting procedure to integrate classification results from multiple 2D patches to provide a nodule-level prediction; (2) implementation of a 3D input strategy to utilize the additional information from the additional dimension; and (3) direct comparison between the 3D approach and 2D approaches for lung nodule malignancy risk classification. Intuitively, lung nodules are 3D objects, therefore it is suboptimal to only extract 2D features from independent slices since all texture and context features in the third dimension are ignored. By constructing the CNN to take 3D volumes directly, it is now possible to extract the object-level features and improve the diagnostic performance over the 2D approaches. For example, for a 2D model, it is not possible to classify a patch that does not intersect or just intersect by a very small area with the nodule itself, while for a 3D implementation such “empty”

patches may be included for extracting context features in the third dimension. Although “empty” patches can be avoided on a research dataset where nodules are carefully segmented and boundaries are perfectly defined, in most real world situations only weak labels, e.g. a rough nodule bounding box, are available, and a full 3D model may be superior than the 2D approaches in handling this situation.

Future efforts are warranted to address the limitations of the presented work to further improve and validate the performance. First, our networks were trained and evaluated using the malignancy risk scores given by a panel of radiologists. Such reference standard is subjective and depends on the reader’s training and experience. Therefore, it is desirable to use a more objective reference standard such as a biopsy-based malignancy rating. A direct comparison between the performance of the proposed method and a human reader will also become possible. Second, a network with larger feature maps and more layers may further improve the classification performance, as suggested by previous experiences in the general image classification tasks. A larger training set will also be beneficial. Lastly, the locations and sizes of the nodules are given in the LIDC dataset, which is not necessarily provided in a real-world scenario. Hence, it is desirable to integrate a nodule detection module to the current work flow in order to provide a complete solution towards clinical feasibility.

5 Conclusion

In this work, we implemented a 3D CNN based method for lung nodule malignancy classification, trained and tested using the LIDC images. We compared three strategies: 2D slice-level, 2D nodule-level, and full 3D nodule-level. Using the malignancy risk score from a panel of readers as the reference labels, the accuracies of the three input methods were 86.7%, 87.3% and 87.4%, respectively. With comparable network architectures, it was found that incorporating 3D context can only slightly improve the risk prediction performance of the CNN based classifier. But 3D CNN model may be superior than the 2D approaches in handling the situation that only weakly labeled or rough nodule region is available.

References

1. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2016. *CA Cancer J. Clin.* **66**, 7–30 (2016)
2. Stewart, B., Wild, C.P., et al.: World cancer report 2014. World (2015)
3. Team, N., et al.: Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**, 395 (2011)
4. Erasmus, J.J., Gladish, G.W., Broemeling, L., Sabloff, B.S., Truong, M.T., Herbst, R.S., Munden, R.F.: Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. *J. Clin. Oncol.* **21**, 2574–2582 (2003)

5. Swensen, S.J., Jett, J.R., Hartman, T.E., Midthun, D.E., Mandrekar, S.J., Hillman, S.L., Sykes, A.M., Aughenbaugh, G.L., Bungum, A.O., Allen, K.L.: CT screening for lung cancer: five-year prospective experience 1. *Radiology* **235**, 259–265 (2005)
6. Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., 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**, 915–931 (2011)
7. Rubin, G.D., Lyo, J.K., Paik, D.S., Sherbondy, A.J., Chow, L.C., Leung, A.N., Mindelzun, R., Schraedley-Desmond, P.K., Zinck, S.E., Naidich, D.P., et al.: Pulmonary nodules on multi-detector row CT scans: performance comparison of radiologists and computer-aided detection 1. *Radiology* **234**, 274–283 (2005)
8. Furuya, K., Murayama, S., Soeda, H., Murakami, J., Ichinose, Y., Yauuchi, H., Katsuda, Y., Koga, M., Masuda, K.: New classification of small pulmonary nodules by margin characteristics on highresolution CT. *Acta Radiol.* **40**, 496–504 (1999)
9. Gurney, J.W., Swensen, S.J.: Solitary pulmonary nodules: determining the likelihood of malignancy with neural network analysis. *Radiology* **196**, 823–829 (1995)
10. Kawata, Y., Niki, N., Ohmatsu, H., Kusumoto, M., Kakinuma, R., Mori, K., Nishiyama, H., Eguchi, K., Kaneko, M., Moriyama, N.: Computerized analysis of 3-d pulmonary nodule images in surrounding and internal structure feature spaces. In: *Proceedings of 2001 International Conference on Image Processing*, vol. 2, pp. 889–892. IEEE (2001)
11. Kido, S., Kuriyama, K., Higashiyama, M., Kasugai, T., Kuroda, C.: Fractal analysis of internal and peripheral textures of small peripheral bronchogenic carcinomas in thin-section computed tomography: comparison of bronchioloalveolar cell carcinomas with nonbronchioloalveolar cell carcinomas. *J. Comput. Assist. Tomogr.* **27**, 56–61 (2003)
12. Shiraishi, J., Abe, H., Engelmann, R., Aoyama, M., MacMahon, H., Doi, K.: Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: ROC analysis of radiologists' performance - initial experience 1. *Radiology* **227**, 469–474 (2003)
13. Armato, S.G., Altman, M.B., Wilkie, J., Sone, S., Li, F., Doi, K., Roy, A.S.: Automated lung nodule classification following automated nodule detection on CT: a serial approach. *Med. Phys.* **30**, 1188–1197 (2003)
14. Mori, K., Niki, N., Kondo, T., Kamiyama, Y., Kodama, T., Kawada, Y., Moriyama, N.: Development of a novel computer-aided diagnosis system for automatic discrimination of malignant from benign solitary pulmonary nodules on thin-section dynamic computed tomography. *J. Comput. Assist. Tomogr.* **29**, 215–222 (2005)
15. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: *IEEE Conference on Computer Vision and Pattern Recognition 2009, CVPR 2009*, pp. 1794–1801. IEEE (2009)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
17. Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multi-scale convolutional neural networks for lung nodule classification. In: *Ourselin, S., Alexander, D.C., Westin, C.-F., Cardoso, M.J. (eds.) IPMI 2015. LNCS*, vol. 9123, pp. 588–599. Springer, Heidelberg (2015). doi:[10.1007/978-3-319-19992-4_46](https://doi.org/10.1007/978-3-319-19992-4_46)
18. Kumar, D., Wong, A., Clausi, D.A.: Lung nodule classification using deep features in CT images. In: *2015 12th Conference on Computer and Robot Vision (CRV)*, pp. 133–138. IEEE (2015)

19. Hua, K.L., Hsu, C.H., Hidayati, S.C., Cheng, W.H., Chen, Y.J.: Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *Onco Target Ther.* **8**, 2015–2022 (2015)
20. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 807–814 (2010)
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014)
22. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: C3D: generic features for video analysis. *CoRR*, abs/1412.0767 **2** 7 (2014)
23. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 221–231 (2013)
24. Collobert, R., Kavukcuoglu, K., Farabet, C.: Torch7: a matlab-like environment for machine learning (2011)