A Novel Deep Learning Model to Distinguish Malignant Versus Benign Solid Lung Nodules

Shuwen Wang
Leilei Zhou
Xiaoran Li
Jie Tang
Jing Wu
Xindao Yin
Yu-Chen Chen
Lingquan Lu

Corresponding Authors: Yu-Chen Chen, e-mail: chenyuchen1989@126.com, Lingquan Lu, e-mail: njllq333@sina.com

Financial support: This work was supported by the 333 High-level Talents Training Project of Jiangsu Province (No. BRA2019122)

Conflict of interest: None declared

Background: In this study we aimed to establish a new transfer learning model based on noncontrast and thin-layer computed tomography (CT) scans to distinguish between malignant and benign solid lung nodules.

Material/Methods: CT images from 202 patients with 210 lesions (malignant: 127, benign: 83) manifesting as solid lung nodules from January 2016 to December 2020 from 3 institutions were retrospectively collected, and each nodule was histopathologically confirmed. Two experienced thoracic radiologists reviewed all images and determined the regions of interest (ROIs) in the three-dimensional (3D) images layer-by-layer. We divided the lesions and images into training and testing sets at a ratio of 7:3. The Inception V3 model was pretrained by the training dataset. Five-fold cross-validation was used to choose the optimal model. Receiver operator characteristic curves (ROC curves) for methods to evaluate the performance of the models were drafted.

Results: In the validation set, the AUC, accuracy, sensitivity, and specificity of Inception V3 model (lesion-level) were 0.999, 0.989, 0.983, and 1.0, respectively, which is higher than the image-level (0.997, 0.933, 0.922, and 0.948, respectively). The Inception V3 model (lesion-level) performed better than the image-level but there was no significant difference between the models (P>0.05). The ResNet50 model based on image level achieved AUC, accuracy, sensitivity, and specificity of 0.963, 0.926, 0.916, and 0.944, respectively, which is lower than that of Inception V3.

Conclusions: Our study developed a novel deep learning model based on noncontrast and thin-layer CT scans to classify benign vs malignant lung nodules, and the Inception V3 model greatly improved the differentiation accuracy and specificity.

Keywords: Deep Learning • Neural Networks, Computer • Solitary Pulmonary Nodule

Full-text PDF: https://www.medscimonit.com/abstract/index/idArt/936830
Background

Lung cancer has long been a leading cause of death. Early and accurate diagnosis plays an important role in survival since only 15% of lung cancers are diagnosed at an early stage [1,2]. CT is widely used as a tool for physical examinations due to its high performance in lung nodule screening. According to the Lung CT Screening Reporting and Data System (Lung RADS version 1.1) [3], solid nodules scoring more than 4 points, which means the size is more than 8 mm, are suspicious for malignancy, and additional diagnostic testing is recommended. Nodule management guidelines [4] also recommend that the follow-up interval is a range rather than a precise time point. Apart from size, radiologists also make the propensity diagnosis relying on radiographic features such as spiculation, lobulation, and pleural reaction, among others. However, overdiagnosis and a high false-positive rate (FPR) remain [5,6] due to the misclassification of these lesions, which is related to their similar appearance on CT images and scan parameters in terms of slice thickness and contrast enhancement of various structures within the image [7,8]. Therefore, clinicians still rely on histological testing, which is an invasive procedure, as the criterion standard, and this sometimes results in technical challenges or complications.

Radiomics and deep learning (DL) models have been widely proposed and developed for application in medical fields. The former consists of a series of processes, including lesion segmentation, radiomics feature extraction and selection, and model establishment and evaluation. Hawkins et al [9] extracted 23 features for predicting malignant nodules (AUC=0.83). Chen et al [10] obtained an accuracy of 0.84 in nodule classification by developing a support vector machine (SVM) algorithm using 72 cases. Unlike radiomics, deep learning classifiers contain training, testing, and performance evaluation [11], and there are many techniques used, such as convolutional neural networks (CNNs) [12], unsupervised learning [13], and transfer learning [14]. Transfer learning allows for pretraining on a very large dataset of images and then tuning the resulting model using specific samples, which is useful for classification to train a stable, unbiased, and non-overfitting deep learning architecture from the very beginning [15]. Inception and ResNet models are the most frequently used for medical studies, such as the Inception V3 model, which performed well in pathological classification of NSCLC [16], breast cancer [17], and knee injury for MRI images [18] and the ResNet50 model, which is widely applied in diagnosis of brain diseases [19], classification of skin lesions [20], and coronary artery calcium detection [21].

However, deep learning algorithms require intensive computational training and more expertise for tuning. To achieve better performance, many investigators have attempted to optimize the models by adjusting the input images. For instance, a multi-CNN model [22] reached an accuracy of 0.87 for binary nodule classification based on multiple down-sampling. Juan et al [23] proposed use of ML-xResNet to classify different types of lung nodule malignancies and achieved 92.19% accuracy. Nasrullah et al [24] obtained a sensitivity of 94% when combining ML-xResNet with clinical factors. A 3D fully-convolutional neural network for reduction of false-positive rate in lung nodule classification was created as well [25]. Additionally, there is still room for improvements in the performance of image acquisition parameters. Some researchers have proven that radiomic features are most affected by varying acquisition parameters [26-28]. Dilger et al [29] improved pulmonary nodule classification by combining the radiomics of intra- and perinodular regions (AUC=0.80). He et al [28] showed that thin-slice and noncontrast CT can provide more accurate information on radiomics features. Dou et al [30] showed that combining the radiomics of perinodular regions can be used to predict lymph node metastasis. Many investigators [31-33] have attempted to distinguish granulomas from malignancies using quantitative radiomics or computerized feature-based analysis as the main concept in DL approaches, which is to expose the network to every possible combination of imaging acquisition parameters.

It is imperative to look further into computer-aided diagnosis to develop novel, accurate approaches to aid in monitoring individuals with pulmonary nodules and to allow for safe and cost-effective diagnosis while preventing unnecessary procedures in cases of benign growths. In this study, 2 modern deep learning models that take advantage of transfer learning and use thin slices and noncontrast CT images were designed for nodule classification. Thus, the model could potentially provide a better prediction of the clinical outcomes of lung nodules, even for very small nodules, intending to considerably improve patient prognosis.

Material and Methods

Patients

First, we retrospectively reviewed the electronic pathological records of patients with solid nodules in the lung between January 2016 and December 2020. The inclusion criteria were: 1) histopathology-confirmed, 2) solid nodule in the lung, 3) lesion size ≥6 mm in axial CT images, and 4) without treatment. Then, 359 patients with 367 nodules were included. The exclusion criteria were: 1) CT images with thin slices measuring more than 3 mm and 2) CT images taken with contrast medium. Finally, a total of 202 patients (125 males and 77 females, age range: 24-83 years) were enrolled, including 127 malignant (104 adenocarcinoma and 23 squamous cell carcinoma) and 83 benign (3 organizing pneumonia, 41 tuberculosis, 13 pulmonary cryptococcosis, 6 inflammatory pseudotumors, and 20 granulomas).
nODULES AS THE SOURCES OF NODULE DATASETS, WE COLLECTED DATA ON 73 PATIENTS (49 MALES, 24 FEMALES) WITH 78 NODULES FROM NANJING FIRST HOSPITAL. IN THIS COHORT, 55 MALIGNANT AND 23 BENIGN NODULES WERE INCLUDED. THEN, WE COLLECTED DATA ON 98 PATIENTS (69 MALES, 29 FEMALES) WITH 98 NODULES FROM GAOCUN PEOPLE HOSPITAL, WHICH INCLUDED 72 MALIGNANT AND 26 BENIGN NODULES. FINALLY, WE INCLUDED 31 PATIENTS (3 MALES AND 28 FEMALES) WITH 34 BENIGN NODULES FROM NANJING SECOND HOSPITAL. CATEGORICAL CLINICAL CHARACTERISTICS INCLUDED SEX (MALE OR FEMALE), AGE (MEAN SD), NODULE SIZE (0.6-1.0 CM, 1.0-3.0 CM, AND >3.0 CM), LOCATION, LOBULATION (ABSENT/PRESENT), AND SPICULATION (ABSENT/PRESENT). THE WORKFLOW IS SHOWN IN FIGURE 1.

CT IMAGE ACQUISITION AND IMAGE ANNOTATION

CT IMAGE ACQUISITION PARAMETERS

This was a multi-center study. CT examinations without contrast medium were acquired from Nanjing First Hospital and Nanjing Second Hospital using Philips Ingenuity CT Core 128 scanners (Philips Medical Systems, Haifa, Matam) and Philips Brilliance 64 CT scanners (Philips Medical Systems, Jiangsu, China), while images taken with GE LightSpeed RT16 CT scanners (GE Medical Systems, Milwaukee, WI, USA) were acquired from Gaocun People’s Hospital. All CT scans were performed with a fixed-tube voltage of 120 kVp and X-ray tube current exposure time of 30-300 mAs. The pixel spacing of the CT image ranged from 0.625 to 0.843 mm depending on patient size, and the reconstruction slice thickness was 1 mm or 3 mm. The slice images were reconstructed with a matrix of 512×512 pixels.

Two thoracic radiologists (Li and Wu, with 5 and 10 years of experience in chest image interpretation, respectively) who were blinded to the results independently reviewed all CT images on the Picture Archiving and Communication Systems (PACS) and reached a consensus by discussion in case of disagreement. Lesion size, spiculation, and lobulation were evaluated as CT morphological features.

CNN INPUT IMAGE ANNOTATION

The thin-slice and noncontrast images were downloaded from PACS and stored in DICOM format, and the window width and
level were set at 750 HU and -500 HU, respectively. The images were subsequently imported into ITK-SNAP software (version 3.8.0 http://fssf.org/), and ROIs for the 3D images were determined layer-by-layer by 2 thoracic radiologists. The dataset of lesions and handcrafted annotation CT images is shown in Table 1.

Establishment of a Model

Two deep learning models (Inception V3 and ResNet50) were trained and optimized to distinguish between benign and malignant lung nodules. We set 2 nodes (benign/malignant) in the SoftMax layer, and the other structures of the models remained the same. All models were pretrained on ImageNet [34], which includes 300 000 000 natural images and trained with the combined set of training images from both datasets. Five-fold cross-validation was used to choose the optimal model. A total of 147 lesions were randomized into 5 equal subgroups, and 63 lesions were included in the test group. Five different combinations of subsets were analyzed. Finally, the result of each subset and the mean result of all subsets were reported. CT images with discordant interpretations made by the deep learning models and radiologists are shown in Figure 2.

Statistical Analysis

SPSS (v.25.0; IBM) software was used to analyze the clinical characteristic data. The \( \chi^2 \) test was used to analyze the statistical significance of these qualitative data, including sex and lesion location, lobulation, and spiculation. An independent \( t \) test was used to compare the CT clinical data, including age and size. A \( P \) value of less than 0.05 was considered statistically significant.

The performance of the proposed models for nodule classification was evaluated according to various statistical measures.

Figure 2. CT images with discordant interpretations between deep learning models and radiologists. (A) Model False-Positives: a 63-year-old male patient with histologically-confirmed benign nodule when the model predicts a malignant nodule. (B) Model False-Negatives: a 47-year-old male patient with histological-confirmed malignant nodule when the model predicts a benign nodule.

Wang S. et al: A novel deep learning model for distinguishing solid lung nodules © Med Sci Monit, 2022; 28: e936830

DATABASE ANALYSIS

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]
Table 2. Characteristics of solid nodules in the lung.

|                      | Malignant nodules (n=127) | Benign nodules (n=83) | P value |
|----------------------|---------------------------|-----------------------|---------|
| Gender n (%)         |                           |                       | 0.078   |
| Male                 | 67 (52.8)                 | 54 (65.1)             |         |
| Female               | 60 (47.2)                 | 29 (34.9)             |         |
| Age (mean)           | 65.17±8.09                | 62.31±11.6            | 0.053   |
| Location n (%)       |                           |                       | 0.172   |
| Left upper lobe      | 30 (23.6)                 | 25 (30.1)             |         |
| Left lower lobe      | 19 (15.0)                 | 11 (13.3)             |         |
| Left mixed           | 2 (1.6)                   | 0                     |         |
| Right upper lobe     | 36 (28.3)                 | 26 (31.3)             |         |
| Right middle lobe    | 9 (7.1)                   | 0                     |         |
| Right lower lobe     | 27 (21.3)                 | 21 (25.3)             |         |
| Right mixed          | 4 (3.1)                   | 0                     |         |
| Lobulation n (%)     |                           |                       | <0.001* |
|Absent               | 29 (22.8)                 | 47 (56.6)             |         |
|Present              | 98 (77.2)                 | 36 (43.4)             |         |
| Spiculation n (%)    |                           |                       | <0.001* |
|Absent               | 15 (11.8)                 | 43 (51.8)             |         |
|Present              | 112 (88.2)                | 40 (48.2)             |         |
| Size (mm)            |                           |                       | 0.904   |
| 6-10                 | 11 (8.7)                  | 8 (9.6)               |         |
| 11-30                | 75 (59.1)                 | 52 (62.6)             |         |
| >30                  | 41 (32.3)                 | 23 (27.7)             |         |
| Median size (range)  | 26.65 (7-100)             | 24.36 (6-67)          |         |
| Pathology n (%)      |                           |                       | /       |
|ADC                   | 104 (81.9)                |                       |         |
|SCC                   | 23 (18.1)                 |                       |         |
|OP                    | 3 (3.6)                   |                       |         |
|TB                    | 41 (49.4)                 |                       |         |
|PCP                   | 13 (15.6)                 |                       |         |
|IPT                   | 6 (7.2)                   |                       |         |
|Granuloma             | 20 (24.1)                 |                       |         |

ADC – adenoacarcinoma; SCC – squamous cell carcinoma; OP – organizing pneumonia; TB – tuberculosis; PCP – pulmonary cryptococcosis; IPT – inflammatory pseudotumor; Granulomas – histopathology-confirmed granulomas but unknown to the specific disease. * P<0.05.
such as sensitivity, specificity, accuracy, and area under the receiver operating curve (AUC). AUC values ranged from 0.5 to 1. The Delong test was used to compare the differences. A P value less than 0.05 was considered statistically significant.

A receiver operator characteristic (ROC) curve, a graphical technique for describing and comparing the accuracy of diagnostic tests, was used to evaluate the sensitivity and specificity of the models. The matplotlib was used to generate the ROC curves.

**Results**

**Clinical Characteristics and Subjective CT Features of Nodules**

In this study, 28.3% of the malignant nodules and 31.3% of the benign nodules were located in the right upper lobe. A high proportion of nodules (59.1% of malignant and 62.6% of benign) measured between 11 cm and 30 cm. Spiculation was found in 48.2% of benign nodules and lobulation was found in 43.4%. There was a significant difference between the malignant and benign nodules in terms of subjective CT features, including spiculation and lobulation (P<0.001), whereas age and sex showed no difference (P>0.05). The patient characteristics are presented in **Table 2**.

**Evaluating Performance of the Models**

We separately developed the Inception V3 model based on image level and lesion level. The performance is shown in **Table 3**. There was no significance among the 2 models in the training group and the test group (P>0.05). The ResNet50 model was built based on image level, and we compared the results between the Inception V3 model and ResNet50 model (**Table 4**). The ROC curve of the 2 models exhibited a high prediction ability when the true positive rate was close to 1 (**Figure 3**). There was no significant difference among the 2 models in the training group and the test group (P>0.05). In addition, the diagnostic accuracy of the Inception V3 and ResNet50 models were very high and Fisher’s exact test showed no significant difference (P>0.05).

**Discussion**

Herein, we developed and validated a new transfer learning model based on thinner slice and noncontrast CT images, which is a noninvasive diagnostic tool for differentiating between malignant nodules and benign lung nodules.

Detection and characterization of pulmonary nodules is an important issue since it is the first step in early lung cancer diagnosis. In our study, the clinical characteristics and radiographic features (lobulation and spiculation) were more common in the lung cancer group. Our results agree with several previous studies reporting that lung cancer more commonly affects the upper lobes, especially the right upper lobe [35]. Nodule size is also a primary determinant of malignancy [36], and the guidelines specify that a suspicious nodule with a diameter larger than 6 mm has a risk of malignancy [4]. Therefore, the minimum size of nodules in our study was 7 mm, which was smaller than in previous studies. Additionally, our dataset included many subtypes of benign nodules, such as organizing pneumonia, pulmonary cryptococcosis, and inflammatory pseudotumors, which are hard to distinguish from lung cancer in terms of morphology.
DL has shown strong performance in the medical field since it can be trained end-to-end in a supervised method while learning highly discriminative image features [37,38]. There is also a growing body of research on predicting pulmonary nodules [39-42], even lymph metastasis, and it has shown great potential in evaluating the survival rate [30,43]. Our one-classifier construction was Inception [15], which is being studied more often because the multiscale features of input have different CT features. The other model was ResNet [44], which can effectively help to compensate for the loss of small-nodule details. Therefore, we built the models at the image level (5008 images) instead of at the lesion-level only (210 lesions), which avoided having too few training examples to learn a full deep representation. Meanwhile, unlike the radiomics classifier [45], the CNN model can automatically extract high-throughput features instead of features calculated from segmented objects as input information, thus avoiding the complicated process of artificial feature extraction [46].

In our study, the InceptionV3 model based on lesion level exhibited good performance, with AUCs of 0.998 in the training set and 0.999 in the validation set, and the accuracy was 0.984 and 0.989, which is higher than in previous studies [10,29,47], whose AUCs were 0.86-0.94 and diagnostic accuracies were 0.84-0.96. The AUCs and diagnostic accuracy of the InceptionV3 model (image-level) was 0.977 and 0.933, which is higher than that of the ResNet50 model (image-level). PET/CT is also good at discriminating nodules. Lai [48] built a model based on PET/CT and achieved a lower accuracy of 0.79. In addition, PET/CT is expensive and time consuming. Thus, we prefer to build the models based on CT, which has better performance and is more beneficial than conventional CT diagnostic methods for patients who cannot tolerate contrast enhancement agents due to allergies or renal failure.

Lesion segmentation is the most challenging aspect of model evaluation [49,50], and we used expert image annotation, which can be an effective way to optimize the model. Referring to earlier lung nodule classification studies, our model takes the effect of CT scanning parameters, including image thickness and contrast medium injection, into consideration because biological heterogeneity within the tumor can be detected and described with plain-phase CT images. However, this phenomenon may be confused by the intravenous contrast agents of existing intratumor contrast agents. Therefore, the present study both simplifies the process and greatly improves the accuracy and specificity.

Our study has some limitations. First, this was a retrospective study and thus had inherent selection bias. Second, the sample we selected was small because of strict inclusion criteria, which could easily have caused a data imbalance. To reduce the bias risk caused by samples with more patients with malignant nodules than those with benign nodules, during the training process, the benign dataset was dynamically oversampled at the lesion level and image level [51]. Second, the study was also limited to discrimination between benign and malignant nodules and did not thoroughly identify benign nodules such as OP and TB. Third, use of different CT scanners may have affected the evaluation of some CT results caused by a partial-volume effect.
Conclusions

In conclusion, we built a novel transfer learning model with high accuracy in distinguishing malignant vs benign lung nodules. It can provide added diagnostic value to differentiate lung nodules and reduce the need for invasive diagnostic procedures, and it may assist clinicians in creating personalized treatment strategies and choosing the optimal intervention.

References:

1. Jemal A, Bray F, Center MM, et al. Global cancer statistics. Cancer J Clin. 2011;61(2):69-90
2. Valerie IR, Cortez PC, Neto EC, et al. Automatic 3D pulmonary nodule detection in CT images: A survey. Comput Methods Programs Biomed. 2016;124:91-107
3. Lung-RADS® ACRCoal. Lung-RADS Assessment Categories version 1.1
4. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: From the Fleischner Society 2017. Radiology. 2017;284(1):228-43
5. Carter S, Barratt A. What is overdiagnosis and why should we take it seriously in cancer screening? Public Health Res Pract. 2017;27(3):2731722
6. Patel E, Pinsky P, Gattozzi C, et al. Overdiagnosis in low-dose computed tomography screening for lung cancer. JAMA Intern Med. 2014;174(2):269-74
7. Yuan X, Zhang J, Quan C, et al. Differentiation of malignant and benign pulmonary nodules with first-pass dual-input perfusion CT. Eur Radiol. 2013;23(9):2469-74
8. Shan F, Zhang Z, Xing W, et al. Differentiation between malignant and benign solitary pulmonary nodules using early phase perfusion and combined with routine computed tomography. Eur J Radiol. 2012;81(11):3598-605
9. Hawkins S, Wang H, Liu Y, et al. Predicting malignant nodules from screening CT scans. J Thorac Oncol. 2016;11(12):2120-28
10. Chen CH, Chang CK, Tu CY, et al. Radiomic features analysis in computed tomography images of lung nodule classification. PLoS One. 2018;13(2):e0192002
11. Meng F, Guo Y, Li M, et al. Radiomics nomogram: A noninvasive tool for preoperative evaluation of the invasiveness of pulmonary adenocarcinomas manifesting as ground-glass nodules. Transl Oncol. 2021;14(1):100936
12. Budi S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med Image Anal. 2021;71:102062
13. Liu J, Gong M, He H. Deep associative neural network for associative memory based on unsupervised representation learning. Neural Netw. 2019;113:41-53
14. Liujen G, Kool T, Bejnordi B, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88
15. Wang X, Mao K, Wang L, et al. An appraisal of lung nodules automatic classification algorithms for CT images. Sensors (Basel). 2019;19(1):194
16. Coudray N, Ocampo P, Sakellaroopouls T, et al. Classification and muta tion prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24(10):1559-67
17. Zhang P, Ma Z, Zhang Y, Chen X, Wang GIG. Improved Inception V3 method and its effect on Radiologists' performance of tumor classification with automated breast ultrasound system. Gland Surg. 2021;10(7):2232-45
18. Sridhar S, Anutharaj J, Valsalan P, et al. A Torn ACL mapping in knee MRI procedures, and it may assist clinicians in creating personalized treatment strategies and choosing the optimal intervention.

Declaration of Figures’ Authenticity

All figures submitted have been created by the authors, and confirm that the images are original with no duplication and have not been previously published in whole or in part.
42. Hu X, Gong J, Zhou W, et al. Computer-aided diagnosis of ground glass pulmonary nodule by fusing deep learning and radiomics features. Phys Med Biol. 2021;66(6):065015
43. He B, Dong D, She Y, et al. Predicting response to immunotherapy in advanced non-small-cell lung cancer using tumor mutational burden radiomic biomarker. J Immunother Cancer. 2020;8(2):e000550
44. Hong Y, Pan H, Jia Y, et al. ResDNet: Efficient dense multi-scale representations with residual learning for high-level vision tasks. IEEE Trans Neural Netw Learn Syst. 2022;PP [Online ahead of print]
45. Wang J, Liu X, Dong D, et al. Prediction of malignant and benign of lung tumor using a quantitative radiomic method. Annu Int Conf IEEE Eng Med Biol Soc. 2016;2016:1272-75
46. Sarıgül M, Ozyildirim B, Avci M. Differential convolutional neural network. Neural Netw. 2019;116:279-87
47. Arevalo J, Gonzalez F, Ramos-Pollan R, et al. Convolutional neural networks for mammography mass lesion classification. Annu Int Conf IEEE Eng Med Biol Soc. 2015;2015:797-800
48. Lai Y, Wu K, Tseng N, et al. Differentiation between malignant and benign pulmonary nodules by using automated three-dimensional high-resolution representation learning with fluorodeoxyglucose positron emission tomography-computed tomography. Front Med (Lausanne) 2022;9:773041
49. Gillies R, Kinahan P, Hricak HR. Radiomics: Images are more than pictures, they are data. Radiology. 2016;278(2):563-77
50. Majkowska A, Mittal S, Steiner DF, et al. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology. 2020;294(2):421-31
51. Zhou L, Zhang Z, Chen YC, et al. A deep learning-based radiomics model for differentiating benign and malignant renal tumors. Transl Oncol. 2019;12(2):292-300