Abbreviate title page

Full title:

Enhanced Pulmonary Nodule Detection Using Fully Automated Deep Learning: A Multifactor Investigation

Commented [WH1]: Perhaps reconsider the title: Evaluating a Fully Automated Pulmonary Nodule Detection Approach and its Impact on Radiologist Performance

Article type:

original research article

Keywords:

lung cancer screening, pulmonary nodule detection, deep learning, computer-aided detection

Abbreviations:

CT: computed tomography

DL: deep learning

CAD: computer-aided detection

CNN: convolutional neural network

RPN: regional proposal network

ROI: region of interest

FROC: free-response receiver operative characteristic

LROC: localization receiver operative characteristic

AUC: area under LROC curve

TPR: true positive rate

FDR: false discovery rate

MIP: maximum intensity projection

Implications for Patient Care:

- Deep learning can be used to assist in the d~~D~~etection of pulmonary nodules is, ~~enhanced by using the deep learning model as assistance,~~ which could benefit nodule management.
- When screening ~~for~~ the elderly population, scans should be more carefully interpreted since manual detection sensitivity could be negatively affected by older age.

Summary statement:

~~The~~ Our DL model ~~showed~~ showed ~~elevated~~ improved sensitivity ~~than~~ over manual ~~detection~~ identification of pulmonary nodules and ~~was~~ was insensitive to radiation dose, patient age, or ~~CT manufacture~~device manufacturer. ~~It can also~~Our model also enhanced manual ~~detection~~ review ~~via~~ by increasing the sensitivity and reducing the reading time.

Commented [WH2]: In what way?

Commented [WH3]: This statement is vague and requires further investigation as to why detection performance is lower in this age group than what the paper provides.

Abstract

Purpose:

This study was to compare the detection sensitivity of lung nodules between the deep learning (DL) model and radiologists from various aspects. It ~~was~~ also ~~to verify~~verified whether the radiologists' detection performance could be enhanced when using the DL model as assistance.

Materials and Methods:

A total of 12754 thin-section chest computed tomography (CT) scans dated from January 2012 to June 2017 were retrospectively collected for the DL model training, validation and testing. Pulmonary nodules of this data were categorized into 6 types according to solidity and size. The testing dataset was divided into 3 cohorts based on radiation dose, patient age, and CT manufacturer. Detection performance of the DL model was analyzed using free-response receiver operative characteristic (FROC). Sensitivities of the DL model and radiologists were compared using exploratory data analysis. False discovery rates (FDR) of the DL model were also compared within each cohort. Detection performance of the same radiologist without and with using the DL model were compared using nodule-wise sensitivity and patient-wise localization receiver operative characteristic (LROC) curves.

Results:

~~For all cohorts t~~The DL model showed elevated overall sensitivity than manual detection of pulmonary nodules. No significant dependence on radiation dose and scanner manufacturer were observed ~~for all readings~~. Though the DL model was ~~also~~ insensitive to patient age, significant dependence on age was observed for the less-experienced radiologist. By using the DL model as assistance, radiologists' performance was improved and reading time was shortened.

Conclusion:

DL is promising to enhance pulmonary nodule detection and benefit its management.

Commented [WH4]: Do you mean false POSITIVE rates?

Commented [WH5]: I would use the phrase "manual review" or "manual identification"

Commented [WH6]: Should this be a separate sentence or is this somehow related to the DL model?

Introduction

Lung cancer continues to have the highest incidence and mortality rates worldwide in 2018.[1] Because of its aggressive and heterogeneous nature, detection and intervention at early stage where the cancer manifests as pulmonary nodules are vital to improve the survival rate.[2] Currently low-dose computed tomography (CT) is widely used for early stage lung cancer screening, as extensive studies have shown that the mortality rate could be significantly reduced.[3–6]

Although detection of pulmonary nodules has been improved using new generation CT scanners, certain nodules may still be oversighted due to the nodule appearance, image quality or perception error of the radiologist, which could be caused by inappropriate reading conditions, fatigue or distraction.[7,8] In a frequently used dataset for research, the original manual miss-detection rate was 76% (38 out of 50 nodules were missed).[9] All missed nodules were later proven to be cancerous and they could be repeatedly miss-detected for up to 3 years. Computer-aided detection (CAD) systems have been developed to improve the nodule detection rate.[9–12] However, based on conventional image processing techniques, these systems typically require complicated image processing steps and may not be robust across various data sources and nodule types.

Taking advantage of the most recent development of artificial intelligence, deep learning (DL) technique using convolution neural networks (CNN) has been shown to be a promising approach to assist lung nodule detection and management.[13–16] Fundamentally different from conventional CAD systems, the DL model can be easily optimized and readily applied to read a large amount of data. However, ~~till now~~a ~~and sensitive~~ nodule detection with high sensitiity, which would be the precondition for ~~proper~~ reliable nodule management~~, such as follow-up screening, immediate investigation or intervention,~~ remains challenging.

In this study, we developed a fully automated DL model using cutting-edge CNN architectures. The DL model performance was compared with the radiologists with varying data attributes including radiation dose, patient age, and scanner manufacturers. Next, to assess whether manual detection could be enhanced

Commented [WH7]: Explicitly name the dataset

Commented [WH8]: Specify

when using the DL model as ~~assistance~~a first-pass reader, the radiologist's detection accuracy~~ies~~ with and without ~~and with~~ using the DL model were compared, while the mimicking real-world clinical reading ~~situations~~environment.

Materials and Methods

Data preparation and categorization

___ provided software and hardware support. Authors who are not affiliated with ___ had control of data and information. Institutional review board (IRB) approvals from all hospitals were received and written informed consents were waived since the study had minimal risk and would not adversely affect the subjects' rights and welfare. A total of 13159 thin-section chest CT scans dated from January 2012 to June 2017 from multiple hospitals in China were retrospectively collected by convenience sampling. The data were composed of both screening and in-patient scans, and patient age (≥18) was the single eligibility criterion. Scans ~~would b~~were excluded from the study if: (1) all lung lobes ~~cannot~~ were not be fully ~~seen~~ visible in the field of view, (2) the image had motion artifacts, (3) the image ~~does~~ did not comply with the DICOM standard, and (4) the radiologists who were making the ground truth labeling ~~cannot~~ were unable to annotate the images confidently. After the selection procedure shown in Fig. 1, 12754 scans were included in our study. ~~, where~~ 11625 scans (91.1%, 11625/12754) ~~mainly~~ from 3 top-tier hospitals were ~~used~~ selected for model training and validation, and 1129 scans (8.9%, 1129/12754) from over 10 other hospitals were ~~preserved~~ used for testing. The split ratio between training and validation scans was ~9:1, and the model was tuned based on the fixed validation set. For the training/validation dataset, 5777 (~49.7%) were male with age 54 ± 15, and 5848 (~ 50.3%) were female with age 55 ± 15. For the testing dataset, the patient age was 57 ± 20. All acquired axial images had the matrix size of 512×512, and the slice thickness ranged from 0.8 to 2 mm.

To generate the ground truth for the entire dataset, two radiologists (Radiologist A and B) each having ~10 years' experience of ~~chest CT~~reviewing chest CTs ~~first~~ independently reviewed all 12754 scans with the original radiology report. ~~The DICOM data were~~ The studies were reviewed using RadiAnt DICOM Viewer

Version 4.2.0 (Poznan, Poland). The window level and width were typically set as -600 and 1500 Hounsfield Unit (HU). To guarantee the best reading, the radiologists ~~could~~ were able to make ~~preferable~~ preferential adjustments based on scan-specific properties and ~~had~~ were allotted unlimited reading time. The detected nodule was marked by a square bounding box~~,~~ with the nodule at the center. Based on the National Comprehensive Cancer Network (NCCN) guidelines for lung cancer screening (version 2.2019)[17], nodules in our dataset were categorized into 6 types~~: , including the~~ solid nodule (SN, ≤6 or >6 mm), ~~the~~ subsolid nodule (SSN, ≤5 or >5 mm), ~~the~~ calcified nodule (CN), and ~~the~~ pleural nodule (PN). The size standard for SN and SSN was different because they had different follow-up management. These ground truth nodules types were later used to assess differences in~~the~~ detection ~~variance~~ rates across all types.

Commented [WH9]: There are only four types listed

There was a significant overlap between the two radiologists' annotation and the nodule size was determined by taking the average of their measurements. Samples that were differently annotated by Radiologists A and B were checked by a third radiologist (Radiologist C) having ~15 years' experience and consensus was reached within the 3 radiologists. For the entire dataset, 65821 nodules were annotated with an average occurrence of 5.2 nodules per patient, and the distribution was shown in Table 1. Since the focus of this study was nodule detection, to reduce manual annotation cost, different types of ground glass nodules (GGN) were generally categorized as SSN and no further diagnosis or pathological details about nodules were studied.

Commented [WH10]: Would it be possible to provide interrater agreement score?

DL model development

The DL model in our study consists of two CNN ~~structures~~models: a Faster R-CNN-based model as the detection architecture and a DenseNet as the feature map extractor.[18,19] ~~Originally,~~ The original implementation of Faster R-CNN takes only one image as input, then feeds the extracted features into a region proposal network (RPN) to propose potential regions of interest (ROI), which are further processed to generate potential objects' classification and their bounding boxes. ~~However,~~Given that the CT scan is a 3D image volume~~, . To utilize information from the extra dimension,~~ we modified the Faster R-CNN network to take successive slices as input, forming a multi-channel 2.5D CNN[20]. Here, 2.5D simply means

the model could take successive slices as input but does not use 3D convolution, since the 3rd dimension (axial) was not continuous and resolution was not consistent.

~~Additionally,~~ DenseNet was used for feature extraction and back propagation in our model. Different from regular CNN, where feature maps are mostly connected once, in DenseNet all maps are directly linked thus forming a densely connected network. Such network could reduce the number of layers, maintain feature density during propagation and improve overall expressive power of the model. Detailed model structure is shown in Fig. 2.

DL model training and testing

For the training phase, nearly all inputs to the model were nodule-positive slices. Each training step consisted of 9 successive slices, which typically covered an entire nodule given our thin-section scans. For every 100 nodule-positive slices, 1 nodule-negative slice randomly selected from all the lung regions was injected for model training to avoid bias. The training process was monitored using the validation dataset.

Once training was finished, the DL model was ~~tested~~ evaluated on the testing dataset using all axial slices from each patient's scan. Similar as the training process, 9 successive slices were loaded into the DL model for each computation step (i.e., slice 1 to 9 for the first step, slice 2 to10 for the second step, etc.). The CT images were ~~and we did~~ not downsize ~~d~~the CT images. Output of the model was the detection marked with a square bounding box and the confidence, the nodule type can be automatically reported as well.

Testing data differentiation

Radiation dose, patient age, and scanner manufacturer were ~~aspects~~ investigated in our study. ~~A recent article~~Ohno et al concluded that there was no significant difference between radiologists' detection using low-dose and standard-dose CT.,[21] They did not compare a ~~however~~ DL model's detection performance on scans using different dose ~~has not been compared~~. Based on the national lung cancer screening guidelines ~~with low-dose CT of~~ used in China,[22] ~~the~~ a scan was labeled~~testing data scan was registered~~ as low-dose if

the X-ray tube current was below 60 mAs; otherwise, the scan ~~, otherwise it~~ was ~~registered as~~considered as conventional-dose. For all scans, the peak voltage was not differentiated and ~~and it~~ was typically120 kVp.

Since the pulmonary structure and texture are age dependent,[~~23,24~~] which may affect lung nodule detection[~~23,24~~], the ~~testing~~ test data was empirically ~~separated~~ stratified into 3 age groups: ~~in terms of age, with patient age~~ below 30 years, 31-60 and over 61. Though smaller age interval or even regression with age could be used, such analysis was not performed in our study considering the clinical necessity and the relatively slow change of pulmonary structures with age.

The third variation was made with regard to CT scanner manufacturer, ~~who may~~which may employ different image acquisition techniques ~~adopt different instrumentation engineering~~ and image reconstruction algorithms that ~~may~~ affect the DL model's detection performance. ~~Manufactures~~ Our dataset included ~~in our study were Canon~~scans from devices manufactured by Canon Medical Systems, GE Healthcare, Philips, and Siemens.

Experimental design and data analysis

Performance of the DL model was first demonstrated using the free-response receiver operative characteristic (FROC) curve, where the sensitivity was plotted versus the false positive detections per scan. To compare the sensitivity between the DL model and radiologists, the testing data were also independently examined by two testing radiologists (Radiologist 1 and 2, with ~5 and 10 years' experience respectively, note that they were not the radiologists determining the ground truth) and exploratory data analysis was conducted.

~~Similar reading conditions and instruction as the radiologists establishing the ground truth were given to~~ Radiologist 1 and 2 were given similar instructions and reading environment as the radiologists who established the reference set. H~~, h~~owever, they ~~had no~~did not have access to the original radiology reports as reference, and the nodule type was not required to be reported. For both the DL model and the testing radiologists, nodule types from the ground truth were used to assess their detection variance across all types.

The detection sensitivity of each nodule type was cross-tabulated with the dose level, age range and the manufacture, and a chi-squared independence test was conducted to examine the dependence of nodule detection on these factors. Within each cohort, the overall detection sensitivity was compared between the DL model and the testing radiologists. Dependence of the DL model's overall false discovery rates (FDR) on the 3 factors were tested as well (the false positive rate was not used because the "true negative" detections cannot realistically reflect the model's performance). Note that since there were 3 variations, Bonferroni correction was used for the critical significance level, *i.e.*, $\alpha = 0.05/3 \approx 0.0167$.

Second~~ly~~, to verify whether the DL model could enhance manual detection in clinical situations, two smaller sized data batches (Batch 1 and 2) containing 123 and 148 scans were ~~both~~ examined by ~~another~~ two additional~~2~~ radiologists (Radiologist 3 and 4, having ~10 years' experience respectively)~~, without or with~~with and without ~~using~~ the DL model. Batch 1 was used to test the nodule-wise detection enhancement for the radiologists, while Batch 2 was used to test the patient-wise detection enhancement. The radiologists first read the scans alone without using the DL model, then they would use the DL model as assistance during their second reading. A washout period of one-week was used between the two readings, and the scans within each batch were shuffled. Such data amounts were selected approximating the radiologists' two days' clinical workload, and their reading time was limited for each scan (up to ~20 minutes, a typical reading period for radiologists at a top-tier hospital). The nodule type was required to report for each detection, together with a confidence level ranging from 0 to 1 with a step of 0.1. Considering the clinical significance, only solid nodules >3 mm were included in this analysis.

Radar plots and localization receiver operative characteristic (LROC) curves were used to show the results of the nodule-wise and patient-wise analyses respectively.[25,26] For the patient-wise analysis, the true-positives (TPs), false-positives (FPs), true-negatives (TNs) and false-negatives (FNs) were defined as follow: at a certain confidence threshold, only if all nodules were correctly detected (location and type), such patient was counted as TP; and only if no detections were made on patient with no nodules, the patient was counted as TN. On the other hand, if the scan was partially correctly annotated (nodules may be miss-

detected, wrongly-categorized or wrongly-located), the patient was counted as FN. Finally, if detections were made on patient having no nodules, the patient was counted as FP. The true positive rate (TPR) and false positive rate (FPR) were then calculated accordingly to generate the LROC data points. Areas under the patient-wise LROC curves (AUC) were calculated for both radiologists as well.

Results

Detection performance of the DL model

Nodule detection performance of the DL model was demonstrated using the FROC metric. On average, when there was 1 false positive detection per scan, the sensitivity was 0.74. The sensitivity improved at a cost of specificity and reached a maximum of 0.86 when there were 8 false positives per scan. The FROC curve was shown in Fig 3.

Next, we ~~will~~ show the performance of the DL model across the radiation dose, patient age, and scanner manufacturer. Since such information may be lost for certain scans, these scans were excluded when the data were separated into the 3 cohorts and the nodule numbers may not be consistent.

Commented [WH11]: This sentence is not understandable.

Effect of radiation dose

Across all nodule types, for the sensitivity of the DL model, no dependence on radiation dose level was observed ($X^2 = 1.1036$, $p = 0.9538$). The same result was observed for the radiologists, which was consistent as reported in Ref 21 (Radiologist 1: $X^2 = 1.6562$, $p = 0.8944$; Radiologist 2: $X^2 = 1.5293$, $p = 0.9097$). Yet, for this testing dataset, the more experienced radiologist (Radiologist 2) showed higher overall sensitivity than the less experienced, and the DL model showed advanced overall sensitivity on both dose levels than both radiologists. The results were summarized in Table 2.

Commented [WH12]: Superscript?

Effect of patient age

The chi-squared test showed different patient age dependence of the DL model than the radiologists. While detection sensitivity of the DL model was independent of the patient's age ($X^2 = 6.1676$, $p = 0.8010$), the ~~the~~ less-experienced radiologist showed ~~significantly~~a statistically significant association ~~age-dependent~~

~~sensitivity~~ ($X^2 = 46.0263$, $p < 0.0001$). The more experienced radiologist showed no significant dependence ($X^2 = 20.6033$, $p = 0.0240$), ~~dependence~~ but the p-value was ~~slightly~~ only slightly above the corrected critical level. Again, the DL model showed higher overall sensitivity at each age range; ~~and~~ the results ~~awere~~ re summarized in Table 3.

Effect of scanner manufacture

For the scanner manufacturers, as expected, the results showed no ~~dependence~~ association ~~for both~~between the DL model and the radiologists (DL: $X^2 = 10.5136$, $p = 0.7862$; Radiologist 1: $X^2 = 9.0240$, $p = 0.8763$; Radiologist 2: $X^2 = 14.6075$, $p = 0.4800$). The less experienced radiologist consistently had the lowest overall detection sensitivity across all 3 factors and the results were summarized in Table 4.

False positive nodule detections

Besides sensitivity, false positives of the DL model were also counted for the 3 aspects and were reported in the tables. Since "true negative nodules" were not available, we calculated the false discovery rate (FDR), which was defined as: false positives/(false positives + true positives). Chi-squared independence test was performed for the FDR of the DL model, and no dependence on the 3 factors was observed (dose: $X^2 = 0.5640$, $p = 0.4527$; age: $X^2 = 0.4734$, $p = 0.7892$; manufacture: $X^2 = 3.7270$, $p = 0.2925$).

~~Detection enhancement using~~Radiologist performance using the DL model

~~Results verifying d~~Detection ~~enhancement~~ performance for Radiologists 3 and 4 when using the DL model ~~were~~ are shown in Fig. 4. From ~~data~~ Batch 1, the radiologists' ~~nodule-wise~~ detection sensitivity ~~was~~ improved across all ~~types~~nodule types. For Batch 2, patient-~~wise~~ level detection ~~from data Batch 2, the LROC curves were also~~also improved, with AUC ~~increased~~ increasing from 0.67 to 0.77 for Radiologist 3, and from 0.65 to 0.78 for Radiologist 4. ~~For b~~Both radiologists experienced, ~~when using the DL model,~~ shorter reading time than without using the model were reported, from ~15 minutes to ~5-10 minutes per patient.

Commented [WH13]: It was stated in the Implications on Patient Care that a significant difference existed when reviewing images from elderly patients. Is that supported here? Elaborate more in the discussion as to why this would be.

Commented [WH14]: It would be preferred if the 95% confidence intervals could be reported and test of statistical significance could be performed.

Discussion

Exploratory result analysis showed that the DL model could detect most of the nodules when using a relatively low specificity, and its sensitivity generally outperformed the manual detection. Success of this model relied on the combination of two CNN structures. Considering the non-homogenous features of pulmonary nodules, DenseNet played a critical role to sufficiently extract the features and maintain their density through model propagation. Meanwhile, capability of the Faster R-CNN to yield nodule location makes LROC available for more reliable model performance assessment. Besides, robustness of the model was also considered by constructing the data from multiple hospitals. The testing data were never exposed to the model till the training was finished, and decent sensitivity could still be achieved.

Contingency test using all listed 6 nodule types showed that the DL model's detection performance (both sensitivity and FDR) does not depend on the radiation dose level, patient age or CT scanner manufacture, which indicated that the DL model can be broadly applied under different imaging conditions with no restriction. In addition, as the less-experienced radiologist may be more sensitive to patient-specific reading tasks, the model might also be a training tool for the junior radiologist to accumulate experience and improve their detection performance.

Though the model was insensitive to the investigated factors, it showed different sensitivities across the nodule types. The model had relatively higher sensitivity for the solid nodule >6 mm and the calcified nodule, and lower sensitivity for the smaller nodules. Such results were consistent with expectations: larger nodules had more abundant features and the calcified nodules typically had higher signal intensity on the CT images. Adjusting the detection layers' resolution of the model may improve the detection for smaller nodules.

It was also interesting to note that when using the DL model as assistance, the patient-wise LROC curve pattern was largely different between Radiologist 3 and 4. The difference could be caused by how the radiologists interpreted the DL model's detection. With using the model, Radiologist 3 may annotate some

nodules with 100% confidence, which were indeed true positives therefore the curve started above the (0, 0) point, where the confidence threshold was the highest. However, since the DL model could have low specificity, over-relying on the model may cost his/her specificity as well: at the high specificity region (close to the 0 point), the curve of with using the DL model was right shifted with no benefit for sensitivity. On the other hand, Radiologist 4 may have cautiously referred to the model's detection and his/her sensitivity was steadily improved without the cost of specificity.

Limitations of the study also exist. First is the relatively high FDR of the model, which is ~49% for the entire testing data. Though the model was trained with emphasis on sensitivity considering more severe consequence of miss-detection than false-positive detection, such high FDR may still mislead and add burden to radiologists (like Radiologist 3 in our study). To reduce the FDR in the future, we may inject more nodule-negative slices for training. Another approach may be using maximum intensity projection (MIP) image volumes for model training and testing. Since MIP has been demonstrated able to improve manual detection[27], it may help to achieve similar effect for the DL model.

Another limitation lies in the data collection. Though the patient age was examined, patients' smoking history (pack-year) was not investigated. This was because the smoking history was separately stored from the DICOM information and cannot be accessed. Besides, nodule biopsy information was not collected since this was an image and report based retrospective study using numerous samples, and no diagnosis or grading was involved. However, this would be critical to further verify the efficacy of the DL model. Biopsy confirmation of the nodules may be performed in the future using certain testing samples.

At last, for the patient-wise LROC analysis, the sensitivity and specificity may appear not quite satisfactory. This maybe because we chose a highly strict nodule size cutoff, where solid nodules >3 mm and all subsolid nodules were considered in the analysis. However, for baseline screening, nodules ≤6 mm usually do not require immediate investigations. Extra LROC plots using adjusted cutoff size were shown in the supplemental figure, where detection enhancement was still observed.

In conclusion, the automatic DL model achieved decent pulmonary nodule detection sensitivity with high robustness. The model's performance did not depend on multiple external factors and can be used with no restrictions. It could also enhance radiologists' detection, especially for the larger nodules and could reduce the reading time when used as assistance. The model's performance may be improved by fine tuning of the model and using different data curation in the future.

Acknowledgements

(blinded for review).

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel R, Torre L, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;0:1-31. doi:10.3322/caac.21492.

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin*. 2018;68(00):7-30. doi:10.3322/caac.21442

3. van Klaveren RJ, Oudkerk M, Prokop M, et al. Management of lung nodules detected by volume CT scanning. *N Engl J Med*. 2009;361(23):2221-2229. doi:10.1056/NEJMoa0906085

4. Team TNLSTR. Reduced lung cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395-409. doi:10.1056/NEJMoa1414264

5. National T, Screening L. Results of Initial Low-Dose Computed Tomographic Screening for Lung Cancer. *N Engl J Med*. 2013;368(21):1980-1991. doi:10.1056/NEJMoa1209120

6. Diederich S, Wormanns D, Semik M, et al. Screening for early lung cancer with low-dose spiral CT: prevalence in 817 asymptomatic smokers. *Radiology*. 2002;222(3):773-781. doi:10.1148/radiol.2223010490

7. Manning DJ, Ethell SC, Donovan T. Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *Br J Radiol*. 2004;77(915):231-235. doi:10.1259/bjr/28883951

8. Hossain R, Wu CC, de Groot PM, Carter BW, Gilman MD, Abbott GF. Missed lung cancer. *Radiol Clin North Am*. 2018;56(3):365-375. doi:10.1016/j.rcl.2018.01.004

9. Armato 3rd SG, Li F, Giger ML, MacMahon H, Sone S, Doi K. Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program. *Radiology*. 2002;225(3):685-692. doi:10.1148/radiol.2253011376

10. Arimura H, Katsuragawa S, Suzuki K, et al. Computerized scheme for automated detection of lung nodules in low-dose computed tomography images for lung cancer screening. *Acad Radiol*. 2004;11(6):617-629. doi:10.1016/j.acra.2004.02.009

11. Liang M, Tang W, Xu DM, et al. Low-Dose CT Screening for Lung Cancer: Computer-aided Detection of Missed Lung Cancers. *Radiology*. 2016;281(1):279-288. doi:10.1148/radiol.2016150063

12. Li F, Arimura H, Suzuki K, et al. Computer-aided Detection of Peripheral Lung Cancers Missed at CT: ROC Analyses without and with Localization. *Radiology*. 2005;237(2):684-690. doi:10.1148/radiol.2372041555

13. Shen W, Zhou M, B FY, Yang C, B JT. Multi-scale Convolutional Neural Networks for Lung Nodule Classification. In: *International Conference on Information Processing in Medical Imaging*. Vol 9123. Cham: Springer International Publishing; 2015:588-599. doi:10.1007/978-3-319-19992-4

14. Ciompi F, Chung K, Van Riel SJ, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci Rep*. 2017;7(March):1-11. doi:10.1038/srep46479

15. Causey JL, Zhang J, Ma S, et al. Highly accurate model for prediction of lung nodule malignancy with CT scans. *Sci Rep*. 2018;8(1):1-12. doi:10.1038/s41598-018-27569-w

16. Yu-Jen Chen Y-J, Hua K-L, Hsu C-H, Cheng W-H, Hidayati SC. Computer-aided classification of

lung nodules on computed tomography images via deep learning technique. *Onco Targets Ther*. 2015:2015. doi:10.2147/OTT.S80733

17. NCCN. Lung cancer screening. www.nccn.org/patients. Published 2018.

18. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(6):1137-1149. doi:10.1109/TPAMI.2016.2577031

19. Huang G, Liu Z, Maaten L Van Der, Weinberger KQ. Densely connected convolutional networks. *CVPR*. 2017;1(2):4700-4708. doi:10.1109/CVPR.2017.243

20. Roth HR, Lu L, Seff A, et al. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland P, Hata N, Barillot C, Hornegger J, Howe R, eds. *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2014*. Cham: Springer International Publishing; 2014:520-527.

21. Ohno Y, Koyama H, Yoshikawa T, et al. Standard-, Reduced-, and No-Dose Thin-Section Radiologic Examinations: Comparison of Capability for Nodule Detection and Nodule Type Assessment in Patients Suspected of Having Pulmonary Nodules. *Radiology*. 2017;284(2):562-573. doi:10.1148/radiol.2017161037

22. Zhou Q, Fan Y, Wang Y, et al. China national lung cancer screening guideline with low-dose computed tomography (2018 version). *Chinese J Lung Cancer*. 2018;21(2):67-75. doi:10.3779/j.issn.1009-3419.2018.02.01

23. Turner JM, Mead J, Wohl ME. Elasticity of human lungs in relation to age. *J Appl Physiol*. 1968;25(6):664-671.

24. Gillooly M, Lamb D. Airspace size in lungs of lifelong non-smokers : effect of age and sex. *Thorax*. 1993;48:39-43.

25. Park SH, Goo JM, Jo C-H. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J Radiol*. 2004;5(March):11-18.

26. Gifford HC, King MA, Wells RG, Hawkins WG, Narayanan M V, Pretorius PH. LROC analysis of detector-response compensation in SPECT. *IEEE Trans Med Imaging*. 2000;19(5):463-473.

27. Gruden JF, Ouanounou S, Tigges S, Norris SD, Klausner TS. Incremental benefit of maximum-intensity-projection images on observer detection of small pulmonary nodules revealed by multidetector CT. *Am J Roentgenol*. 2002;179(July):149-157.

Tables

Table 1. Categorization and number of retrospectively detected pulmonary nodules.

Types	Training set		Testing set	
	no. of nodules	no. of patients	no. of nodules	no. of patients
Solid nodules (SN)	18554	7636	4734	1036
≤ 6mm	15225	5456	4406	848
> 6mm	3329	2180	328	188
Subsolid nodules (SSN)	31275	6680	1716	583
≤ 5mm	17850	3487	1252	343
> 5mm	13425	3193	464	240
Calcified nodules (CN)	6262	991	496	256
Pleural nodules (PN)	1987	1050	797	355
Total	58078	11625	7743	1129

Table 2. Dose-related detection sensitivity of the DL model and radiologists

Dose and nodule types		Gold standard	Detected nodules (sensitivity, %)		
			DL model	Radiologist 1	Radiologist 2
Low dose	SN ≤6 mm	719	517 (71.9)	300 (41.7)	358 (49.8)
	SN >6 mm	44	39 (88.6)	41 (93.2)	36 (81.8)
	SSN ≤5 mm	333	204 (61.3)	75 (22.5)	187 (56.2)
	SSN >5 mm	61	52 (85.2)	41 (67.2)	50 (82.0)
	CN	59	51 (86.4)	28 (47.5)	39 (66.1)
	PN	223	168 (75.3)	137 (61.4)	162 (71.7)
	Overall TP	**1439**	**1031 (71.6)**	**622 (43.2)**	**832 (57.8)**
	FP (FDR, %)		**653 (38.8)**	**-**	**-**
Conventional dose	SN ≤6 mm	2680	1727 (64.4)	968 (36.1)	1347 (50.3)
	SN >6 mm	215	189 (87.9)	166 (77.2)	149 (69.3)
	SSN ≤5 mm	993	676 (68.1)	260 (26.2)	565 (56.9)
	SSN >5 mm	371	301 (81.1)	216 (58.2)	316 (85.2)
	CN	265	244 (92.1)	127 (47.9)	147 (55.5)
	PN	400	313 (78.3)	203 (50.8)	261 (65.3)
	Overall TP	**4924**	**3450 (70.1)**	**1940 (39.4)**	**2785 (56.6)**
	FP (FDR, %)		**3241 (48.4)**	**-**	**-**

(SN: solid nodule, SSN: subsolid nodule, CN: calcified nodule, PN: pleural nodule, TP: true positive, FP: false positive, FDR: false discovery rate)

Table 3. Age-related detection sensitivity of the DL model and radiologists

Age groups and nodule types		Gold standard	Detected nodules (sensitivity, %)		
			DL Model	Radiologist 1	Radiologist 2
Group A (below 30)	SN ≤6 mm	340	218 (64.1)	141 (41.5)	181 (53.2)
	SN >6 mm	30	28 (93.3)	23 (76.7)	23 (76.7)
	SSN ≤5 mm	24	13 (54.2)	15 (62.5)	18 (75.0)
	SSN >5 mm	12	11 (91.7)	11 (91.7)	12 (100)
	CN	15	12 (80.0)	11 (73.3)	12 (80.0)
	PN	39	33 (84.6)	12 (30.8)	16 (41.0)
	Overall TP	**460**	**315 (68.5)**	**213 (46.3)**	**262 (57.0)**
	FP (FDR, %)		**238 (43.0)**	**-**	**-**
Group B (31-60)	SN ≤6 mm	1706	1146 (67.2)	645 (37.8)	879 (51.5)
	SN >6 mm	130	114 (87.7)	112 (86.2)	104 (80.0)
	SSN ≤5 mm	650	456 (70.2)	206 (31.7)	355 (54.6)
	SSN >5 mm	247	221 (89.5)	166 (67.2)	206 (83.4)
	CN	154	143 (92.9)	72 (46.8)	89 (57.8)
	PN	297	241 (81.1)	158 (53.2)	197 (66.3)
	Overall TP	**3184**	**2321 (72.9)**	**1359 (42.7)**	**1830 (57.5)**
	FP (FDR, %)		**1921 (45.3)**	**-**	**-**
Group C (over 60)	SN ≤6 mm	1310	855 (65.3)	511 (39.0)	679 (51.8)
	SN >6 mm	99	86 (86.9)	82 (82.3)	74 (74.7)
	SSN ≤5 mm	510	329 (64.5)	119 (23.3)	304 (59.6)
	SSN >5 mm	159	111 (69.8)	60 (37.7)	118 (74.2)
	CN	142	127 (89.4)	71 (50.0)	78 (54.9)
	PN	259	190 (73.4)	140 (54.1)	189 (73.0)
	Overall TP	**2479**	**1698 (68.5)**	**983 (39.7)**	**1442 (58.2)**
	FP (FDR, %)		**1693 (50.1)**	**-**	**-**

(SN: solid nodule, SSN: subsolid nodule, CN: calcified nodule, PN: pleural nodule, TP: true positive, FP: false positive, FDR: false discovery rate)

Table 4. Manufacture-related detection sensitivity of the DL model and radiologists

Manufactures and nodule types		Gold standard	Detected nodules (sensitivity, %)		
			DL Model	Radiologist 1	Radiologist 2
Manufacture A	SN ≤6 mm	321	194 (60.4)	119 (37.1)	194 (60.4)
	SN >6 mm	39	33 (84.6)	32 (82.1)	33 (84.6)
	SSN ≤5 mm	146	60 (41.1)	42 (28.8)	92 (63.0)
	SSN >5 mm	82	53 (64.6)	57 (69.5)	62 (75.6)
	CN	42	36 (85.7)	27 (64.3)	30 (71.4)
	PN	45	32 (71.1)	19 (42.2)	26 (57.8)
	Overall TP	**675**	**408 (60.4)**	**296 (43.9)**	**437 (64.7)**
	FP (FDR, %)		**545 (57.2)**	**-**	**-**
Manufacture B	SN ≤6 mm	1214	890 (73.3)	505 (41.6)	477 (39.3)
	SN >6 mm	56	51 (91.1)	44 (78.6)	38 (67.9)
	SSN ≤5 mm	603	433 (71.8)	176 (29.2)	292 (48.4)
	SSN >5 mm	125	114 (91.2)	80 (64.0)	95 (76.0)
	CN	80	75 (93.8)	46 (57.5)	51 (67.5)
	PN	284	235 (82.7)	165 (58.1)	184 (64.8)
	Overall TP	**2362**	**1798 (76.1)**	**1016 (43.0)**	**1137 (48.1)**
	FP (FDR, %)		**1461 (44.8)**	**-**	**-**
Manufacture C	SN ≤6 mm	1311	786 (60.0)	554 (42.3)	775 (59.1)
	SN >6 mm	105	92 (87.6)	90 (85.7)	83 (79.0)
	SSN ≤5 mm	245	176 (71.8)	83 (33.9)	141 (57.6)
	SSN >5 mm	102	92 (90.2)	63 (61.8)	93 (91.2)
	CN	145	129 (89.0)	63 (43.4)	75 (51.7)
	PN	195	138 (70.8)	105 (53.8)	147 (75.4)
	Overall TP	**2092**	**1413 (67.5)**	**958 (45.8)**	**1314 (62.8)**
	FP (FDR, %)		**1115 (44.1)**	**-**	**-**
Manufacture D	SN ≤6 mm	380	254 (66.8)	146 (38.4)	246 (64.7)
	SN >6 mm	34	31 (91.2)	30 (88.2)	30 (88.2)
	SSN ≤5 mm	137	82 (59.9)	42 (30.7)	90 (65.7)
	SSN >5 mm	93	69 (74.2)	59 (63.4)	82 (88.2)
	CN	34	32 (94.1)	20 (58.8)	23 (67.6)
	PN	75	56 (74.7)	37 (49.3)	47 (62.7)
	Overall TP	**753**	**524 (69.6)**	**334 (44.4)**	**518 (68.8)**
	FP (FDR, %)		**605 (53.6)**	**-**	**-**

(SN: solid nodule, SSN: subsolid nodule, CN: calcified nodule, PN: pleural nodule, TP: true positive, FP: false positive, FDR: false discovery rate)

Figure legends

Figure 1. Schematic of training/validation (A) and testing (B) datasets preparation procedure. Qualified profiles were selected based on 4 steps: (1) profiles with no clinical reports were excluded; (2) post-operative scans were excluded; (3) profiles indicating diffuse pulmonary nodules were excluded; (4) patients with other lung diseases, such as pneumonia and tuberculosis, were excluded. For the testing dataset, it was further divided into different cohorts based on the factors to be investigated.

Figure 2. Framework of the proposed model. N successive slices before and after the center slice are collected together as the input. Convolution is performed on each of the images and feature maps are extracted using DenseNet. The features are fed into regional proposal network (RPN) to obtain potential regions first, then features inside the proposed regions are further processed to obtain both the nodule classification and location.

Figure 3. FROC of the DL model's detection performance.

Figure 4. (A&B) Nodule-wise detection sensitivity comparison for Radiologist 3 and 4 between without or with using the DL model as assistance. (C&D) Patient-wise detection LROC comparison for Radiologist 3 and 4 between without or with using the DL model as assistance.

Supplemental Figure 1. Patient-wise detection LROC curves of Radiologists 3 and 4 with nodule size cutoff at 6 mm. (A) For Radiologist 3, the AUC without using the DL model was 0.877, and it was 0.953 with using the model. (B) For Radiologist 4, the AUC without using the DL model was 0.908, and it was 0.954 with using the model. For both radiologists, detection enhancement was seen with using the DL model.