Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network

Article *in* The Oncologist · April 2019

DOI: 10.1634/theoncologist.2018-0908

CITATIONS
56

READS
1,145

24 authors, including:

Chao Zhang
Guangdong Academy of Medical Sciences and General Hospital
56 PUBLICATIONS **724** CITATIONS
SEE PROFILE

Xing Sun
The University of Hong Kong
69 PUBLICATIONS **1,004** CITATIONS
SEE PROFILE

Kang Dang
Nanyang Technological University
22 PUBLICATIONS **137** CITATIONS
SEE PROFILE

Qing Zhou
Guangdong lung cancer institute, Guangdong Academy of Medical Sciences and G…
229 PUBLICATIONS **5,340** CITATIONS
SEE PROFILE

Published Ahead of Print on April 17, 2019 as 10.1634/theoncologist.2018-0908.

The Oncologist®

Cancer Diagnostics and Molecular Pathology

Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network

Chao Zhang,[a,†] Xing Sun,[d,†] Kang Dang,[d] Ke Li,[d] Xiao-wei Guo,[d] Jia Chang,[e] Zong-qiao Yu,[d] Fei-yue Huang,[d] Yun-sheng Wu,[d] Zhu Liang,[d] Zai-yi Liu,[b] Xue-gong Zhang,[f] Xing-lin Gao,[c] Shao-hong Huang,[g] Jie Qin,[g] Wei-neng Feng,[h] Tao Zhou,[h] Yan-bin Zhang,[i] Wei-jun Fang,[i] Ming-fang Zhao,[j] Xue-ning Yang,[a] Qing Zhou,[a] Yi-long Wu,[a] Wen-zhao Zhong[a]

[a]Guangdong Lung Cancer Institute, Guangdong Provincial Key Laboratory of Translational Medicine in Lung Cancer, [b]Department of Radiology, and [c]Department of Respiration, Guangdong Provincial People's Hospital & Guangdong Academy of Medical Sciences, Guangzhou, People's Republic of China; [d]Tencent Youtu Lab, Shanghai, People's Republic of China; [e]Tencent, Shenzhen, People's Republic of China; [f]MOR Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & System Biology, Department of Automation, Tsinghua University, Beijing, People's Republic of China; [g]The Third Affiliated Hospital of Sun Yat-Sen University, Guangzhou, People's Republic of China; [h]First People's Hospital of Foshan, Foshan, People's Republic of China; [i]Guangzhou Chest Hospital, Guangzhou, People's Republic of China; [j]Department of Medical Oncology, The First Hospital of China Medical University, Shenyang, People's Republic of China

[†]Contributed equally

Disclosures of potential conflicts of interest may be found at the end of this article.

Key Words. Lung cancer • Convolutional neural network • Pulmonary nodule • Diagnostics

Abstract

Background. Computed tomography (CT) is essential for pulmonary nodule detection in diagnosing lung cancer. As deep learning algorithms have recently been regarded as a promising technique in medical fields, we attempt to integrate a well-trained deep learning algorithm to detect and classify pulmonary nodules derived from clinical CT images.

Materials and Methods. Open-source data sets and multicenter data sets have been used in this study. A three-dimensional convolutional neural network (CNN) was designed to detect pulmonary nodules and classify them into malignant or benign diseases based on pathologically and laboratory proven results.

Results. The sensitivity and specificity of this well-trained model were found to be 84.4% (95% confidence interval [CI], 80.5%–88.3%) and 83.0% (95% CI, 79.5%–86.5%), respectively. Subgroup analysis of smaller nodules (<10 mm) have demonstrated remarkable sensitivity and specificity, similar to that of larger nodules (10–30 mm). Additional model validation was implemented by comparing manual assessments done by different ranks of doctors with those performed by three-dimensional CNN. The results show that the performance of the CNN model was superior to manual assessment.

Conclusion. Under the companion diagnostics, the three-dimensional CNN with a deep learning algorithm may assist radiologists in the future by providing accurate and timely information for diagnosing pulmonary nodules in regular clinical practices. ***The Oncologist*** 2019;24:1–7

Implications for Practice: The three-dimensional convolutional neural network described in this article demonstrated both high sensitivity and high specificity in classifying pulmonary nodules regardless of diameters as well as superiority compared with manual assessment. Although it still warrants further improvement and validation in larger screening cohorts, its clinical application could definitely facilitate and assist doctors in clinical practice.

Introduction

Lung cancer is the leading cause of cancer death worldwide, accounting for 1.6 million deaths annually [1]. Despite the availability of multimodality treatment, the 5-year survival rate for advanced lung cancer remains low, varying from 4% to 17% [2].

For early-stage lung cancer, successful surgical dissection can be curative: The 5-year survival rate for patients undergoing non-small cell lung cancer (NSCLC) resection is 75%–100% for stage IA NSCLC but only 25% for stage IIIA NSCLC [3]. Yet, it is

Correspondence: Wen-zhao Zhong, M.D., or Yi-long Wu, M.D., Guangdong Lung Cancer Institute, Guangdong Provincial People's Hospital and Guangdong Academy of Medical Science, 106 Zhongshan 2nd Rd., Yuexiu Qu, Guangzhou 510080, Guangdong, People's Republic of China. Telephone: 86-20-8387785; e-mail: 13609777314@163.com or syylwu@live.cn Received December 24, 2018; accepted for publication March 6, 2019. http://dx.doi.org/10.1634/theoncologist.2018-0908

Published Ahead of Print on April 17, 2019 as 10.1634/theoncologist.2018-0908.

difficult to confirm its pathological status by biopsy, especially for small pulmonary nodules in early stage. The limitation may subsequently impair clinical decision making and management.

The extensive application of computed tomography (CT) has enabled more early-stage NSCLC to be diagnosed, making complete surgical dissection possible. The National Lung Screening Trial (NLST) compared a large screening population separated into two groups of patients in whom low-dose computed tomography (LDCT) and chest radiography were used, respectively, and revealed a significant 20% reduction in lung cancer-specific mortality in the LDCT group [4]. Conventional CT analysis requires radiologist assessment and is highly laborious, and conventional CT-based lung cancer screening often produces false-positive testing results [4, 5]. Thus, interest in deep convolutional neural networks (CNNs) based pulmonary nodules detection and classification has grown rapidly in recent years [6], owing to the fact that CNNs have demonstrated high accuracy in many other computer vision tasks and less manual intervention [7].

In this study, a substantial amount of open-source image data was applied to pretrain a CNN model for the detection and classification of pulmonary nodules. Subsequently, the model was further trained and validated using newly collected images derived from multiple clinical centers across China. Finally, a further effect analysis was performed to illustrate the potential use of this deep learning algorithm in clinical practice.

Materials and Methods

Data Sets

We collected three sets of data for different purposes. First, CT images derived from the Lung Nodule Analysis 2016 Challenge (LUNA16) data set [8] and Kaggle data set were used to pretrain the CNN model. These data sets contain both diagnostic results and thoracic CT scans from lung cancer screening. Second, thoracic CT images contributed by Guangdong Provincial People's Hospital, The Third Affiliated Hospital of Sun Yat-Sen University, Foshan First People's Hospital, and Guangzhou Chest Hospital from May 2015 to October 2016 were used for training and validating the algorithm. Third, data from 50 patients, who underwent surgical dissection and had preoperative CT images in Guangdong Lung Cancer Institute since January 2017, were prospectively collected for final assessment of our algorithm. Half of these images contained pathologically confirmed malignant nodules; the other half were associated with benign diseases. Ethics review and informed patient consent were obtained from each participating hospital.

Grading

Images from the LUNA16 data set record the results of a two-phase image annotation process performed by four experienced thoracic surgeons with marked-up annotated lesions. For the Kaggle data set, only image-level diagnosis was provided, so we applied our algorithm to obtain the nodule-level detection results, which were subsequently reviewed and adjusted by two specialized radiologists.

Images derived from different centers were graded by up to eight radiologists for the presence of pulmonary nodules, including lesion status and diameter. Patients with malignant disease and benign tumors underwent surgical dissection with confirmed pathological results (supplemental online Fig. 1) based on the latest International Association for the Study of Lung Cancer pathological classification, whereas patients diagnosed with tuberculosis were authenticated based on laboratory and microbiological examination. Malignant disease was divided into four groups, including adenocarcinoma, squamous cell carcinoma, lymphoepithelioma-like carcinoma (LELC) and others. Benign disease was divided into four groups: tuberculosis; hamartoma; inflammatory pseudotumor; and any other pathogens that may cause pulmonary nodules. Moreover, nodule diameter was remeasured and subdivided by three radiologists for further analysis.

Validation

LUNA16 and Kaggle data sets were used to pretrain the CNN model. After pretraining, K-fold cross-validation was performed using the data set obtained from the participating centers. Subgroup analysis was implemented for different diameters and pathological subtype to validate efficacy in these specific parameters. Moreover, 25 licensed physicians, including radiologists, thoracic surgeons, and respiratory doctors with more than 5 years of attending doctor work experience, were invited to grade 50 prospectively collected thoracic CT images. Their grading was directly compared with the trained deep learning algorithm. The grading results have been measured in terms of accuracy, specificity, and specialty in average values. In addition, to examine the competitiveness of these results, we applied the current top-ranked algorithm [9] from Kaggle by using its publicly available code in our training and validation cohort. To ensure fairness, we have made a head-to-head comparison between our model and the top-ranked Kaggle algorithm [9] trained on identical public data sets.

Our proposed system consists of three modules as follows.

Preprocessing Module (Supplemental Online Fig. 2)

The first step of the preprocessing module is to isolate image regions containing lung tissues from the rest of the CT slices. We remove irrelevant content by setting a predetermined threshold value, that is, 0 HU, such that bone and soft tissues outside the lung regions could be excluded. To further improve the segmentation accuracy in lung region, we apply image segmentation with adaptive thresholds (which means no fixed cutoff is applied during the process) to segment the lung tissues out, followed by the operation of three-dimensional (3D) dilation and erosion to correct small segmentation errors. The second step is to reconstruct the segmented 3D lung images to new images of standardized slice thickness and resolutions to avoid issues with different types of CT equipment. The final step is to enhance the image contrast to highlight information of the lung tissues. The enhanced 3D lung images are used as inputs to the next stage of our algorithm.

Nodule Cancer Diagnosis Network (Supplemental Online Fig. 3)

A 3D pulmonary nodule detection network is built to obtain 3D features from the lung images. It employs a multiresolution mechanism such that the network can identify nodules of both

Published Ahead of Print on April 17, 2019 as 10.1634/theoncologist.2018-0908.

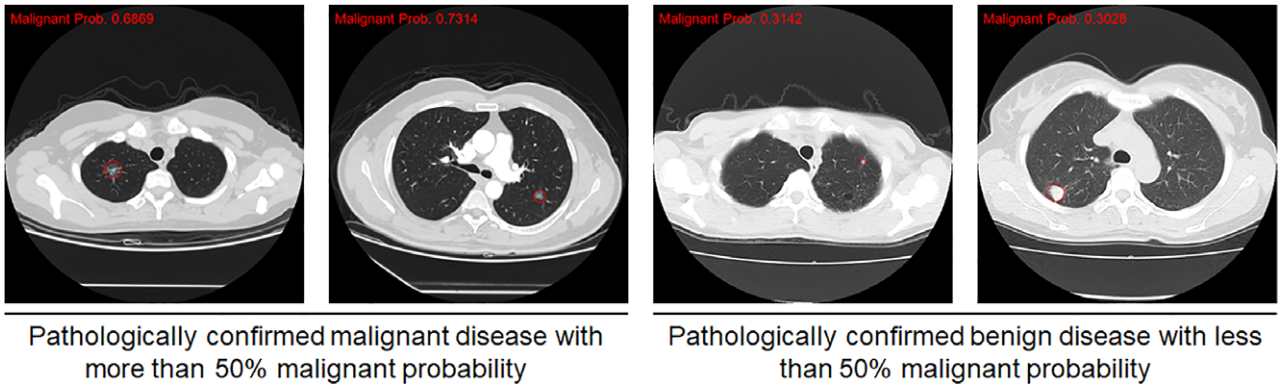

Figure 1. Final output performance from our proposed system. Four computed tomography images containing pathologically confirmed malignant or benign disease are presented. The target nodules will be automatically circled out and given probability value of malignance. If value was more than 50%, it was considered malignant disease, whereas it was the opposite for those less than 50%.

large and small sizes. By adding another network branch containing two fully connected layers to the nodule detection network, a nodule cancer diagnostic network is obtained. The nodule cancer diagnostic network simultaneously identifies suspicious pulmonary nodules and calculates the probability of detected nodules being malignant. It combines pulmonary nodule detection and classification into one unified process, which is more efficient and effective compared with applying separated nodule detection and classification processes.

Our CNN model is implemented on the Pytorch platform [10]. We employ a two-stage training strategy to increase the stability of CNN learning. In the first stage, a nodule detection network is trained with input images and the corresponding annotated nodule locations. In the second stage, the nodule cancer diagnostic network is initialized according to the detection network parameters from the first stage, and then fined tuned with input images and the associated diagnosis results. During network training, only representative non-nodular samples are selected in each training epoch to circumvent the impact from unbalanced nodular and non-nodular samples. These representative non-nodular samples are selected with online hard negative mining algorithm [11]: Images that do not contain nodules but are falsely detected by our model will be regarded as representative non-nodular samples for model adjustment during the next round of learning. In addition, batch normalization [12] and dropout [13] are respectively used in the network to improve the training effectiveness and avoid over-fitting [14]. Stochastic gradient descent is used as the optimization algorithm for parameter updating during the model training.

Output Module (Fig. 1)

After the nodule cancer diagnosis network produces the malignancy scores for all detected nodules, we fuse them to obtain the final image-level malignancy score. Under the independence assumption of malignancy of each nodule, the fused image-level malignancy score can be calculated by the formula $P_{f\ used} = 1 - (1 - P_1)(1 - P_2) \ldots (1 - P_N)$, where P stands for the probability of nodule malignancy and N refers to the number of nodules. Examples of system output are shown in Figure 1.

Statistical Analysis and Performance Comparison

Sample evaluation was performed using PASS 11 software (PASS 11; NCSS, LLC, Kaysville, UT, www.ncss.com). The alpha was set at 0.05 with two-sided confidence interval (CI). Based on this study, the values of sensitivity and specificity were both hypothesized as 0.8. Exact Clopper-Pearson CIs have been applied for calculations, and the evaluated sample size was 715, providing adequate sample enrollment in this study. The 95% CIs for the sensitivity and specificity of the algorithm at the one operating point were calculated as exact Clopper Pearson CIs. Based on the results of the validation cohort and performance comparison, the receiver operating characteristic curve and tables were generated to characterize the sensitivity and specificity of the algorithm.

Results

Details of the CT images in development and validation sets are summarized in Figure 2 and Table 1. In total, 888 CT images and 1,397 CT images were extracted from the LUNA16 data set and Kaggle data set, respectively. These data were used to pretrain the CNN model. Among 757 eligible CT images retrospectively collected from multiple centers

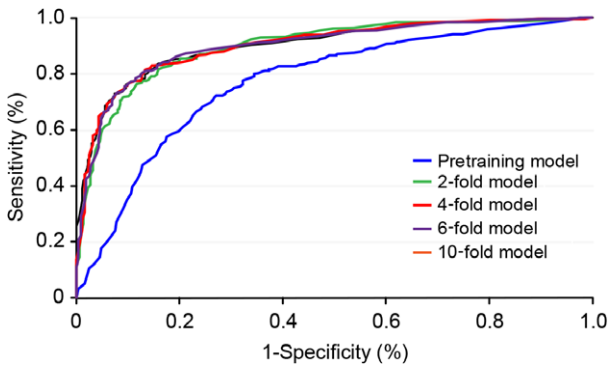

Figure 2. Receiver operating characteristic curve of proposed deep learning algorithms on multicenter images. Each curve represents the model based on different folds compared with the pretraining model, with an area under curve of 0.799. The 2-fold, 4-fold, 6-fold, and 10-fold models respectively achieved an area under curve of 0.898, 0.900, 0.900, and 0.901.

Published Ahead of Print on April 17, 2019 as 10.1634/theoncologist.2018-0908.

Table 1. Data set characteristics

Characteristics	Pretraining sets		Training and validation sets
	LUNA16	Kaggle	
No. of images	888	1,397	757
No. of detectable nodules	1,186	1,173	855
No. of radiologists	4	Not mentioned	9
Patient demographics			
No. of patients	NA	NA	757
Age, mean (SD), years	NA	NA	59.8 (±11.2)
Female, *n*/total *n* (%)	NA	NA	340/757 (44.9)
Image quality distribution			
Thickness of CT images, mm (%)	≤2.5 (100)	Not mentioned	<2.5 (26.0) ≥2.5 (74.0)
Pulmonary nodule information			
Malignant disease ratio, %	37.4	25.9	43.7
Small nodules (≤30 mm), %	99.8		82.2

Summary of image characteristics and detailed information in pretraining, training, and validation set. Patient demographics of the pretraining set could not be acquired from the open-source data.
Abbreviations: CT, computed tomography; NA, not applicable.

to train and validate the model, 197 (26.02%) were classified as thin sections (slice thickness not greater than 2.5 mm), whereas 560 (73.98%) were considered thick sections (slice thickness greater than 2.5 mm).

The sensitivity of our proposed deep learning algorithm trained by the multicenter images achieved 84.4% (95% CI, 80.5%–88.3%), based on 83.0% (95% CI, 79.5%–86.5%) specificity in 10-fold cross-validation (Fig. 3; Table 2). The area under the curve of this algorithm was 0.855 (95% CI, 0.830–0.880). Among different folds that were tested, sensitivity increased with the increases in the number of the training images (Table 2, 10-fold vs. 2-fold training, etc.). The comparison of our deep learning algorithm with the first-placed algorithm from the Kaggle competition determined a sensitivity of 0.752 for our deep learning model and 0.661 for the first-placed algorithm, based on a specificity of 0.757. Our deep learning algorithm had an area under the curve of 0.803 (95% CI, 0.775–0.831), whereas the value was 0.767 (95% CI, 0.737–0.797) for the first-placed algorithm (Fig. 4A).

Additional sensitivity analyses using the multicenter data set were conducted for two subcategories—diameter and pathological result. Diameters were divided into three subgroups: 0–10 mm, 10–20 mm, and 20–30 mm. No statistically significant differences in sensitivity or specificity were detected between these three subgroups (Fig. 4B). In the 0–10-mm group, the sensitivity and specificity were 85.7% (95% CI, 70.8%–100.0%) and 91.1% (95% CI, 86.8%–95.2%), respectively; in the 10–20-mm group, 85.7% (95% CI, 77.1%–94.3%) and 90.1% (95% CI, 84.8%–95.4%), respectively; and in the 20–30-mm group, 78.9% (95% CI, 66.0%–91.8%) and 91.3% (95% CI, 83.2%–99.4%), respectively. Combined with pathological status, the algorithm-based prediction for adenocarcinoma had the highest accuracy of 85.7%, compared with 65.0% for squamous cell carcinoma and 75% for LELC.

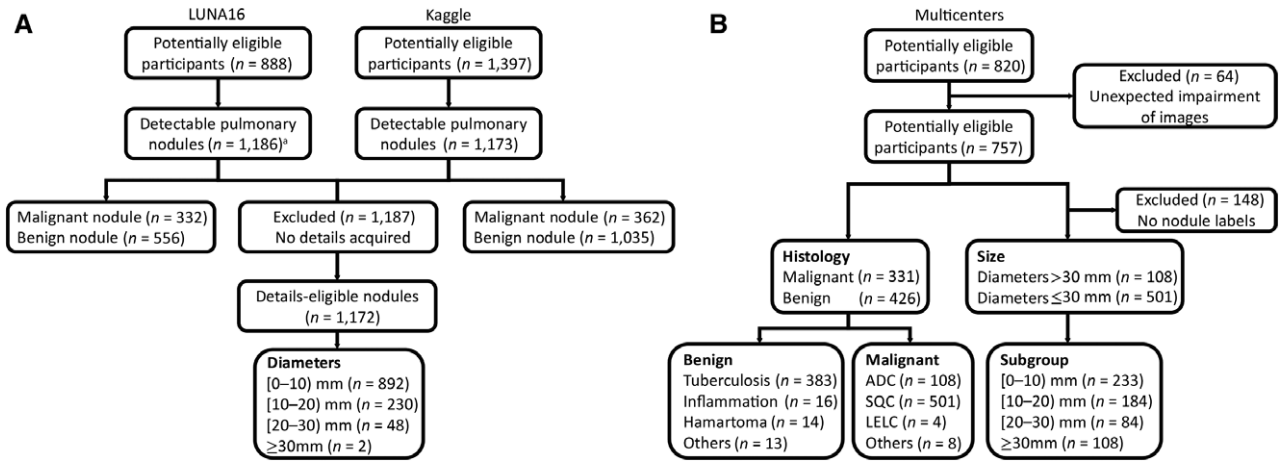

Figure 3. Flow charts of open-source and multicenter data sets. **(A):** Data sets derived from LUNA16 and Kaggle. **(B):** Multicenter data sets with well-labeled and histology-confirmed results.
Abbreviations: ADC, adenocarcinoma; LELC, lymphoepithelioma-like carcinoma; LUNA16, Lung Nodule Analysis 2016 challenge; SQC, squamous carcinoma.

Published Ahead of Print on April 17, 2019 as 10.1634/theoncologist.2018-0908.

Table 2. Performance of proposed deep learning algorithms on multicenter data sets (training and validation through multicenter data sets)

Model	Accuracy, %	Sensitivity, %	Specificity, %
Pretraining	75.0	75.8	74.3
2-fold training	82.8	82.3	83.2
4-fold training	83.3	83.1	83.4
6-fold training	83.7	83.4	83.9
10-fold training	83.7	84.4	83.0

Table 3. Comparisons between baseline approaches and proposed algorithm on the 50-image evaluation set (performance comparison of 50 images)

Model	Accuracy, %	Sensitivity, %	Specificity, %
Doctors	79.6	81.3	77.9
Pretraining model	80.0	84.0	76.0
Trained model	92.0	96.0	88.0
Kaggle Top 1	71.2	70.3	71.8

As a final evaluation set, we constructed a 50-image set where the patients underwent surgical dissection and had preoperative CT images prospectively collected. This evaluation set was used to compare the performance of 25 licensed physicians and our proposed algorithm (Table 3). Doctor manual assessment had an average accuracy of 79.6%, with 81.3% (95% CI, 66.0%–96.6%) sensitivity and 77.9% (95% CI, 61.6%–94.1%) specificity. The performance of the trained algorithm on the same data set had an accuracy of 92.0%, with 96.0% (95% CI, 88.3%–100.0%) sensitivity and 88.0% (95% CI, 76.0%–100.0%) specificity. In supplemental online Figure 8, we display some challenging examples, for example, certain benign nodules being visually similar to malignant ones. For example, the first image of Figure 8 shows a nodule with irregular boundaries like malignant ones; however, it is a benign nodule, as the CT intensities within are quite uniform. Our algorithm can correctly classify such challenging examples.

Discussion

Our comprehensive experiments demonstrate the feasibility of applying a deep learning algorithm to clinical practice for lung cancer screening and diagnosis. Compared with doctor assessments, the deep learning algorithm exhibited significantly better performance in detecting and classifying pulmonary nodules. Further comparisons with the first-place algorithm from the Kaggle competition also revealed better nodule classification using our proposed CNN model.

Previous studies applying deep learning algorithms in various therapeutic areas such as skin cancer and diabetic retinopathy reported marked success [7, 15]. In lung cancer, several studies previously explored the detection and classification of pulmonary nodules. Ciompi et al. [16] introduced a T-distributed stochastic neighbor embedding algorithm to obtain insight into the types of features learned by the network, and achieved better accuracy in classifying nodule subtypes. Hua et al. [17] first reported the application of a deep learning algorithm to nodule classification. Both deep belief network and CNN models revealed encouraging results for nodule classification. A recent study focusing on the use of deep CNNs to detect pulmonary nodules achieved remarkable results, with 91.7% sensitivity and 88.6% specificity [18]. However, no deep residual architecture was leveraged to perform automatic nodule detection, and no further comparison of performance was made. Furthermore, the lack of validation based on real-world data or pathological confirmation may have confounded the results. Most importantly, the lack of high-quality labeled-nodule data in these studies created a major hindrance to the algorithm development, rendering it incapable of surpassing the accuracy by doctor manual assessment.

In contrast to some previous deep learning algorithms applied to pulmonary nodule classification, our deep learning algorithm-based model has a unified rather than separated nodule detection and classification processes (supplemental online Fig. 4). Previous approaches apply a serial/sequential model, meaning a linear stack of layers. This process may result in error accumulation during calculation, and it is less effective and efficient compared with our approach.

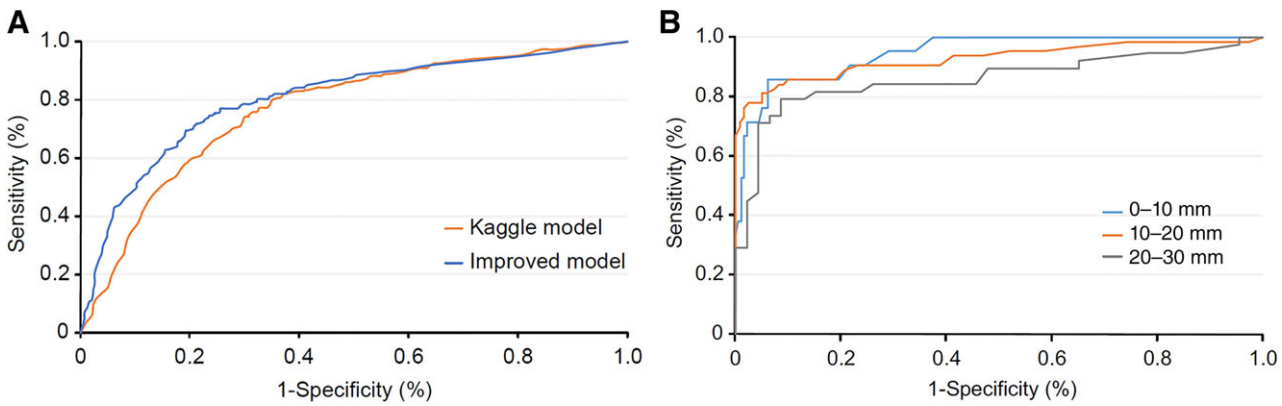

Figure 4. Analysis of superiority with our deep learning algorithms and subgroup analysis based on nodule diameters. **(A):** The improved deep learning algorithm demonstrated excessive superiority, with an area under curve of 0.803, compared with 0.767 for the first-place algorithm (Kaggle competition). **(B):** Sensitivity and specificity were 85.7% and 91.0% for the 0–10-mm-diameter subgroup; 85.7% and 90.1% for the 10–20-mm subgroup; and 78.9% and 91.3% for the 20–30-mm subgroup.

Published Ahead of Print on April 17, 2019 as 10.1634/theoncologist.2018-0908.

Table 4. Comparison between deep learning-based model and CAD-based model previously reported

	Deep learning-based model		CAD-based mean diameter and volume model				
Study	Multicenter data	Testing cohort	Mathilde et al.	Tammemagi et al. (diameter model)	Tammemagi et al. (volume model)	Tammemagi et al. (diameter model)	Tammemagi et al. (volume model)
Detected nodules (validation)	855	50	1,152	1,711	1,711	3,680	3,680
Population	Non-intent to screen	Non-intent to screen	DLCST data	PanCan data (development data)	PanCan data (development data)	NLST data (validation data)	NLST data (validation data)
Malignancy ratio	43.7%	50.0%	5.7%	6.5%	6.5%	12.0%	12.0%
Methods	CNN	CNN	PanCan risk prediction model	PanCan risk prediction model	PanCan risk prediction model	PanCan risk prediction model	PanCan risk prediction model
AUC	0.90	0.94	0.87	0.95	0.95	0.81	0.82
Sensitivity	83.4	96.0	NA	88.0	86.3	75.0[a]	75.0[a]
Specificity	83.9	88.0	NA	89.7	90.2	75.0[a]	75.0[a]

[a]Both sensitivity and specificity values were adapted from receiver operating characteristic curves.
Abbreviations: AUC, area under the curve; CAD, computer-aided detection; CNN, convolutional neural network; DLCST, Danish Lung Cancer Screening Trial; NA, not applicable; NLST, National Lung Screening Trial.

In this study, application of a deep learning-based model was optimized and extended for a medical setting, using improved deep neural networks and large data sets with matched pathologically confirmed labels. This generated algorithm achieved 84.4% sensitivity and 83.0% specificity, minimizing both false-positive and false-negative results. Furthermore, subgroup analysis showed there was high efficacy for the detection of small (<10 mm) pulmonary nodules, similar to that for larger nodules (10–30 mm). Computer-aided detection (CAD) [19] tools, a well-known system for nodules measurements and risk prediction, have been previously reported and validated through specific data sets [20, 21]. More comparable and even superior results were observed using the CNN model compared with the CAD-based model (Table 4). Moreover, the CNN model has the potential for improvement, as relatively few data were currently input.

Beyond the superiority of the CNN model, there were some limitations with this algorithm. As a large subset of the validation data was derived from a multicenter data set, heterogeneity in image quality was inevitable, which may have affected the ability for cancer prediction. The study only included a limited number of ground glass nodules (GGNs) representing early-stage disease [22], which was not intended for screening; thus, the model should be further refined for GGN detection. Another major limitation involved the nature of deep neural networks. Neural networks only directly connect the image with the eventual result, with no opportunity to gain insight into the process by which the result is derived. Despite the application of multitask learning and multiattribute loss to help the model learn features such as lobulation and malignancy, it remains difficult to comprehensively illustrate all the features that the model has learned. Moreover, as all images for training and validation came from a single scan, information reflecting the micro-alteration in CT scans during imaging surveillance might have been missed. The feasibility of applying deep learning algorithms to imaging surveillance should be discussed in the future.

Conclusion

This study developed a deep learning algorithm with both high sensitivity and high specificity for pulmonary nodule detection and classification, which is superior to radiologist assessment under certain circumstances requiring efficiency. Indeed, a deep learning algorithm can never replace physicians considering the unavoidable errors it may make. However, with a comparable performance with the radiologist, deep learning algorithms can be a good helper in CT assessment. Moreover, inspired by the NLST, a further clinical question is whether such algorithms may one day improve patient outcomes and reduce lung cancer-specific mortality. A prospective, multicenter clinical trial generating large data sets should investigate this question to ultimately figure out whether algorithm-based models will be a clinically viable approach for nodule classification and cancer diagnosis.

Acknowledgments

We thank all the doctors who provided assistance in the performance comparison from the participating centers in China listed in supplemental online Appendix 1. We also acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health for their critical role in creating the publicly available, free LUNA16 Database used in this study. This study was supported by National Key R&D Program of China (Grant 2016YFC1303800, to Q.Z.); Special Fund of Public Interest by National Health and Family Control Committee (Grant 201402031, to Y.-l.W.); Project of National Natural Science Foundation (Grant 81872510, to W.-z.Z); and Research Fund from Guangzhou Science and Technology Bureau (Grant 201704020161).

Author Contributions

Conception/design: Chao Zhang, Xing Sun, Xiao-wei Guo, Xue-gong Zhang, Xue-ning Yang, Yi-long Wu, Wen-zhao Zhong

Published Ahead of Print on April 17, 2019 as 10.1634/theoncologist.2018-0908.

Provision of study material or patients: Zai-yi Liu, Xing-lin Gao, Shao-hong Huang, Jie Qin, Wei-neng Feng, Tao Zhou, Yan-bin Zhang, Wei-jun Fang
Collection and/or assembly of data: Chao Zhang, Zai-yi Liu, Xing-lin Gao, Jie Qin, Tao Zhou, Wei-jun Fang, Wen-zhao Zhong
Data analysis and interpretation: Xiao-wei Guo, Jia Chang, Zong-qiao Yu, Fei-yue Huang, Yun-sheng Wu, Zhu Liang
Manuscript writing: Chao Zhang, Xing Sun, Kang Dang, Ke Li, Yi-long Wu, Wen-zhao Zhong
Final approval of manuscript: Chao Zhang, Xing Sun, Kang Dang, Ke Li, Xiao-wei Guo, Jia Chang, Zong-qiao Yu, Fei-yue Huang, Yun-sheng Wu, Zhu Liang, Zai-yi Liu, Xue-gong Zhang, Xing-lin Gao, Shao-hong Huang, Jie Qin, Wei-neng Feng, Tao Zhou, Yan-bin Zhang, Wei-jun Fang, Ming-fang Zhao, Xue-ning Yang, Qing Zhou, Yi-long Wu, Wen-zhao Zhong

Disclosures

Yi-long Wu: AstraZeneca (C/A), Roche (RF), AstraZeneca, Roche, Eli Lilly, Pfizer, Sanofi (H). The other authors indicated no financial relationships.

(C/A) Consulting/advisory relationship; (RF) Research funding; (E) Employment; (ET) Expert testimony; (H) Honoraria received; (OI) Ownership interests; (IP) Intellectual property rights/inventor/patent holder; (SAB) Scientific advisory board

References

1. Ferlay J, Soerjomataram I, Dikshit R et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:E359–386.

2. American Cancer Society. Cancer facts & figures 2015. Available at http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-044552.pdf. Accessed April 23, 2018.

3. Le Chevalier T. Adjuvant chemotherapy for resectable non-small-cell lung cancer: Where is it going? Ann Oncol 2010;21(suppl 7):vii196–198.

4. National Lung Screening Trial Research Team, Aberle DR, Adams AM et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395–409.

5. Chiolero A, Paccaud F, Aujesky D et al. How to prevent overdiagnosis. Swiss Med Wkly 2015; 145:w14060.

6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–444.

7. Esteva A, Kuprel B, Novoa RA et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–118.

8. Setio AAA, Traverso A, de Bel T et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 2017;42:1–13.

9. Liao F, Liang M, Li Z et al. Evaluate the malignancy of pulmonary nodules using the 3D Deep Leaky Noisy-or Network. 2017. arXiv preprint arXiv:1711.08324.

10. Paszke A, Gross S, Chintala S et al. Automatic differentiation in PyTorch. 2017.

11. Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016;761–769.

12. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 2015. arXiv preprint arXiv: 1502.03167.

13. Srivastava N, Hinton G, Krizhevsky A et al. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014; 15:1929–1958.

14. Bottou L. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Heidelberg, Germany: Physica-Verlag, 2010:177–186.

15. Gulshan V, Peng L, Coram M et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–2410.

16. Ciompi F, Chung K, van Riel SJ et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Sci Rep 2017; 7:46479.

17. Hua KL, Hsu CH, Hidayati SC et al. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther 2015;8:2015–2022.

18. Nibali A, He Z, Wollersheim D. Pulmonary nodule classification with deep residual networks. Int J Comput Assist Radiol Surg 2017;12: 1799–1808.

19. Horeweg N, van der Aalst CM, Vliegenthart R et al. Volumetric computed tomography screening for lung cancer: Three rounds of the NELSON trial. Eur Respir J 2013;42:1659–1667.

20. Tammemagi M, Ritchie AJ, Atkar-Khattra S et al. Predicting malignancy risk of screen-detected lung nodules-mean diameter or volume. J Thorac Oncol 2019;14:203–211.

21. Winkler Wille MM, van Riel SJ, Saghir Z et al. Predictive accuracy of the PanCan Lung Cancer Risk Prediction Model - External validation based on CT from the Danish Lung Cancer Screening Trial. Eur Radiol 2015;25:3093–3099.

22. Kakinuma R, Noguchi M, Ashizawa K et al. Natural history of pulmonary subsolid nodules: A prospective multicenter study. J Thorac Oncol 2016;11:1012–1028.

See http://www.TheOncologist.com for supplemental material available online.