COMPUTED TOMOGRAPHY

A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images

Qianqian Ni[1] · Zhi Yuan Sun[1] · Li Qi[1] · Wen Chen[2] · Yi Yang[3] · Li Wang[1] · Xinyuan Zhang[1] · Liu Yang[1] · Yi Fang[1] · Zijian Xing[4] · Zhen Zhou[5] · Yizhou Yu[6] · Guang Ming Lu[1] · Long Jiang Zhang[1,7]

Abstract

Objectives To utilize a deep learning model for automatic detection of abnormalities in chest CT images from COVID-19 patients and compare its quantitative determination performance with radiological residents.

Methods A deep learning algorithm consisted of lesion detection, segmentation, and location was trained and validated in 14,435 participants with chest CT images and definite pathogen diagnosis. The algorithm was tested in a non-overlapping dataset of 96 confirmed COVID-19 patients in three hospitals across China during the outbreak. Quantitative detection performance of the model was compared with three radiological residents with two experienced radiologists' reading reports as reference standard by assessing the accuracy, sensitivity, specificity, and F1 score.

Results Of 96 patients, 88 had pneumonia lesions on CT images and 8 had no abnormities on CT images. For per-patient basis, the algorithm showed superior sensitivity of 1.00 (95% confidence interval (CI) 0.95, 1.00) and F1 score of 0.97 in detecting lesions from CT images of COVID-19 pneumonia patients. While for per-lung lobe basis, the algorithm achieved a sensitivity of 0.96 (95% CI 0.94, 0.98) and a slightly inferior F1 score of 0.86. The median volume of lesions calculated by algorithm was 40.10 cm^3. An average running speed of 20.3 s ± 5.8 per case demonstrated the algorithm was much faster than the residents in assessing CT images (all $p < 0.017$). The deep learning algorithm can also assist radiologists make quicker diagnosis (all $p < 0.0001$) with superior diagnostic performance.

Conclusions The algorithm showed excellent performance in detecting COVID-19 pneumonia on chest CT images compared with resident radiologists.

Key Points

- *The higher sensitivity of deep learning model in detecting COVID-19 pneumonia were found compared with radiological residents on a per-lobe and per-patient basis.*
- *The deep learning model improves diagnosis efficiency by shortening processing time.*
- *The deep learning model can automatically calculate the volume of the lesions and whole lung.*

Keywords COVID-19 · Deep learning · Multidetector computed tomography · Diagnosis · Pneumonia

---

Qianqian Ni and Zhi Yuan Sun contributed equally to this work.

✉ Long Jiang Zhang
   kevinzhlj@163.com

1 Department of Medical Imaging, Jinling Hospital, Medical School of Nanjing University, Nanjing 210002, Jiangsu, China

2 Department of Medical Imaging, Taihe Hospital, Shiyan 442008, Hubei, China

3 Department of Medical Imaging, Wuhan First Hospital, Wuhan 430022, Hubei, China

4 Deepwise AI Lab, Beijing 100080, China

5 School of Electronics Engineering and Computer Science, Peking University, Beijing 10080, China

6 Department of Computer Science, The University of Hong Kong, Pok Fu Lam, Hong Kong

7 Department of Medical Imaging, Medical Imaging Center, Nanjing Clinical School, Southern Medical University, 305 Zhongshan East Road, Xuanwu District, Nanjing 210002, Jiangsu, China

Abbreviations

COVID-19	2019 coronavirus disease
CT	Computed tomographic
GGO	Ground-glass opacity
MERS	Middle East respiratory syndrome
NPV	Negative prediction value
PHEIC	Public health emergency of international concern
PPV	Positive prediction value
RT-PCR	Reverse transcription-polymerase chain reaction
SARS	Severe acute respiratory syndrome
SARS-CoV	Severe acute respiratory syndrome coronavirus
STD	Standard deviation
MERS-CoV	Middle East respiratory syndrome coronavirus
WHO	World Health Organization

Introduction

On December 31, 2019, a novel coronavirus probably originated from Wuhan, Hubei Province, China, was reported to the World Health Organization (WHO) [1–3]. Subsequently, WHO declared the novel coronavirus Public Health Emergency of International Concern (PHEIC) on January 30, 2020. At a situation report on March 4, 2020, this largest and most widespread outbreak of the novel coronavirus resulted in a total of 90,870 confirmed patients diagnosed as coronavirus disease-19 (COVID-19) and 3112 deaths. Global spread of this new coronavirus has resulted in 10,566 confirmed cases cross 72 countries with 166 deaths [4, 5]. To date, COVID-19 has led to more deaths than combination of SARS and MERS despite the relatively lower mortality rate [6]. Rapid spread of COVID-19 resulted from human-to-human transmission [7, 8]. During any outbreak, prompt recognition and patient quarantine play a vital role in containment of the threat. As this is a newly discovered virus, the spectrum of the effective diagnostic tools remains narrow. Instead of real-time polymerase chain reaction (RT-PCR), which is partially restricted by insufficient testing kits, delayed testing cycle, and questionable extraction technique, CT is expected to be applied in initial screening of suspected patients to accelerate the definite diagnosis especially with the emergence of artificial intelligence (AI) techniques [9–12].

In the last few years, AI in health care has been widely suggested as an important tool to guide the disease detection and clinical decisions [13, 14]. It is notable that AI is emphasized to work effectively in current epidemic for prediction of outbreaks as a Canadian company (Blue Dot) successfully reported the location of this outbreak in late December 2019. In addition, AI is also used to aid in the development of image checking in order to distinguish COVID-19 pneumonia with other benign respiratory illness [15]. Also, some ongoing works based on AI are attempting to find new ways to control the spread of COVID-19 and eliminate or reduce the threats from the epidemic. Despite current success in outbreak prediction and COVID-19 recognition, no exploritation of AI in accurate assessment of COVID-19 pneumonia has been reported officially. Thus, the purpose of our study is to automatically detect and quantitatively analyze the pneumonia lesions in chest CT images from patients diagnosed as COVID-19.

Methods

Study population

We used a commercially available deep learning algorithm (Deepwise & League of PhD Technology Co. Ltd.) [16] which was previously trained and validated in 19,291 CT scans from 14,435 patients collected from seven hospitals in China (mean age 40.9 ± 0.9; 51% male, 49% female) with the inclusion criteria of (1) CT images with slice thickness ≤ 2 mm and (2) patients diagnosed as pneumonia or healthy participants, and the exclusion criteria of (1) patients had history of pulmonary surgery; (2) CT images diagnosed as infection but not pneumonia, such as pulmonary tuberculosis; and (3) CT images with poor quality, e.g., heavy breathing artifacts and metal artifacts. Among all the 14,435 collected patients, 2154 patients were diagnosed as COVID-19 by pathogenic test, while 5874 patients were diagnosed as other pulmonary pneumonia (bacterial pneumonia, fungal pneumonia, and other viral pneumonia).

The algorithm was tested in this non-overlapping set of 96 consecutive patients in 3 hospitals from January 20, 2020, to February 10, 2020, who were diagnosed with COVID-19 by RC-PCR test using respiratory secretions extracted from nasopharyngeal or oropharyngeal swabs (84 patients from Taihe Hospital, Shiyan, Hubei; 11 from Wuhan First Hospital, Wuhan, Hubei; 1 from Jinling Hospital, Nanjing, Jiangsu). All patients involved underwent chest thin-slice CT. Patients who had (1) incomplete CT imaging data, (2) chest radiograph only, and (3) other CT examinations were excluded. The study was approved by the institutional review boards of all 3 hospitals with all written informed consents waived. The mean age of enrolled patients was 44 years. Forty-six of them are male with the age of 45 years ± 17, and 50 of them are female with the age of 43 years ± 13.

CT protocols

All patients included underwent non-contrast CT scans using the following multidetector CT scanners (Somatom definition

AS, Somatom definition flash, Siemens Healthcare; Optima CT540, Optima 680, GE Healthcare). Each CT scan was performed during end stage of inspiration with supine position, ranging from lung apex to diaphragm. The detailed CT parameters were listed as follows: (1) voltage 120 kVp, (2) reference tube current 110–250 mAs, (3) detector collimation 16–320×0.5–0.625 mm, (4) slice thickness 1.0–1.25 mm, (5) slice interval of 0.9–1.25 mm, and (6) pitch of 1–1.375.

Image readings and definition of reference standard

Clinical readings were independently performed by three cardiothoracic resident radiologists (L.Q., L.W., and X.Y.Z. with 6, 5, and 2 years of experiences in chest imaging interpretation, respectively) who were blinded to clinical data and previous imaging results. All 3 readers are first required to record the presence or absence of COVID-19 and the number and location (lobe and segment) of the lesions if present.

Then, abnormal features of chest CT images were recorded including (1) ground-glass opacity (GGO) presented as an area of increased attenuation with no obscuration of bronchial and vessels [9]; (2) pulmonary consolidation; (3) crazy paving pattern; (4) diffused, central, or peripheral distribution of lesions defined based on one previous publication [17]; (5) thoracic lymphadenopathy with the short-axis diameter of lymph nodes ≥ 10 mm; as well as (6) other pulmonary illness such as emphysema or fibrosis. The number of abnormal lobes was also recorded. In addition, CT severity score was calculated according to chest CT images. The scoring of each lung lobe was identified as follows: 0 normal and 1 abnormal (any lesions detected regardless of their opacities and extent). Accordingly, the maximum score was recorded as a cumulative of 5 with all the 5 lobes involved. CT severity score in this study is expressed as $(n)/5 \times 100\%$ (n = the number of involved lung lobes). CT severity was categorized into the following classes: (a) mild ($\leq 20\%$), (b) moderate (20–50%), and (c) severe ($> 50\%$).

The reference standard for the presence of COVID-19 and imaging features on chest CT was defined by two well-experienced senior radiologists (G.M.L. and Z.Y.S. with 37 and 18 years of experiences in chest radiology) who made the final decision in consensus combining the patients' clinical, laboratory, and chest CT imaging data.

Deep learning algorithm development

An automatic AI pneumonia detection and evaluation system was used to extract CT features and quantitatively estimate the pulmonary involvement of abnormalities. This system is built based on deep neural networks, where three major steps are designed to ensure the final accuracy which will be available to detect the patients with COVID-19 pneumonia, including (1) abnormality detection, (2) voxel segmentation, and (3) pulmonary lobe segmentation. All the processes were performed by AI system automatically without any interaction of human.

Abnormality detection and segmentation

In this study, COVID-19 pneumonia-based lung lesions included consolidation, GGO, nodules and others such as fibrosis. A convolutional MVP-Net [18] is exploited to achieve automatic detection of the lesions. Domain knowledge is incorporated in clinical practice during the model design. Considering that radiologists tending to inspect multiple windows to obtain accurate diagnosis, we achieved this idea by using a multi-view feature pyramid network, where multi-view features were extracted from images rendered with varied window widths and window levels. To effectively combine this multi-view information, a channel-wise attention module is employed to capture complementary information across different views. The overall architecture of the network is shown in our previous published work [18]. A three-pathway architecture is built to extract the most prominent features from each representative view, followed by a classifier and regressor to classify and localize the potential abnormal regions in CT images. Afterwards, 3D U-Net [19] was introduced to classify voxels that represented the abnormality in the detected regions. Thus, we could acquire the extracted voxel-wise regions of abnormality. As a natural result based on the output of the abovementioned methods, a number of metrics, such as the volume and CT value of the lesions, could be calculated and output.

Pulmonary lobe segmentation

In order to provide the localization information of lesions in the lung, pulmonary lobe segmentation was necessary. To this end, a 3D U-Net is adopted as the basic segmentation network. Besides, a smooth margin loss is proposed to mine the most informative samples for training. To guarantee a desired result, two effective metrics which leverage anatomical priors were used to help select the best model during training [20].

All the CT data in this study has never been used before, and there is no overlapping among the patient identities among all datasets. After analyzing the CT images with this system, the presence or absence of COVID-19 pneumonia was recorded on a per-patient and per-lobe basis.

Deep learning algorithm training, validation, and testing

A total of 19,291 pulmonary CT scans from 14,435 individuals were used for the deep learning algorithm training and validating, among which 3854 scans were derived from 2154 COVID-19 patients, 6871 scans were collected from 5847

patients with patients diagnosed as other pneumonia (bacterial pneumonia, fungal pneumonia, and other viral pneumonia), and the rest 8566 scans were taken from 6434 healthy people. All the 96 CT scans were enrolled in validation set without overlap between training set and validation set (Fig. 1).

Comparison of deep learning algorithm and radiologists

The dataset of 96 COVID-19 patients with chest CT images was used for the comparison of diagnostic performance of three independent radiologists (resident 1, 6 years; resident 2, 5 years; resident 3, 2 years of experiences in chest imaging interpretation) and deep learning algorithm. Pneumonia lesions detected per-patient or per-lobe basis were used for the evaluation of diagnostic performance.

We also investigated the impact of deep learning algorithm on guiding the diagnosis of the three radiologists. To avoid the potential memorization bias, residents were requested to make a diagnosis with the assistance of AI system after 2 weeks of initial test. Abnormality detection, voxel segmentation, and pulmonary lobe segmentation were processed by AI system automatically. During the second round of reading the same CT images, AI system will present the labeled lesions in CT slices and provide its diagnosis of lesion detection of each lobe. Residents were requested to make final diagnosis with the assistance of AI system and compared the diagnostic performance with residents' initial reports.

The reference standard was defined by two well-experienced senior radiologists (G.M.L. and Z.Y.S. with 37 and 18 years of experiences, respectively, in chest radiology) who made the final decision in consensus combining the patients' clinical, laboratory, and chest CT imaging data.

Statistical analysis

We performed statistical analysis using commercially available statistical software SPSS (V23.0, IBM SPSS Inc.). Categorical variables were presented as numbers and percentages. Continuous data was presented as mean ± standard deviation (std) or median (interquartile range), as appropriate. On a per-patient and per-lobe basis, the accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and 95% confidence intervals of three resident radiologists' evaluations and deep learning algorithm were assessed. Sensitivities and specificities of deep learning algorithm were compared with three residents by chi-square test with two experienced radiologists' reading reports as reference standard. A p value cutoff of 0.017 was used based on Bonferroni correction for three comparisons. Reading time per graphic assessment unit between deep learning model and radiological residents were compared with unpaired t test. F1 score was calculated as harmonic mean of recall and

Fig. 1 Flow diagram shows the overview of deep learning algorithm and participant selection
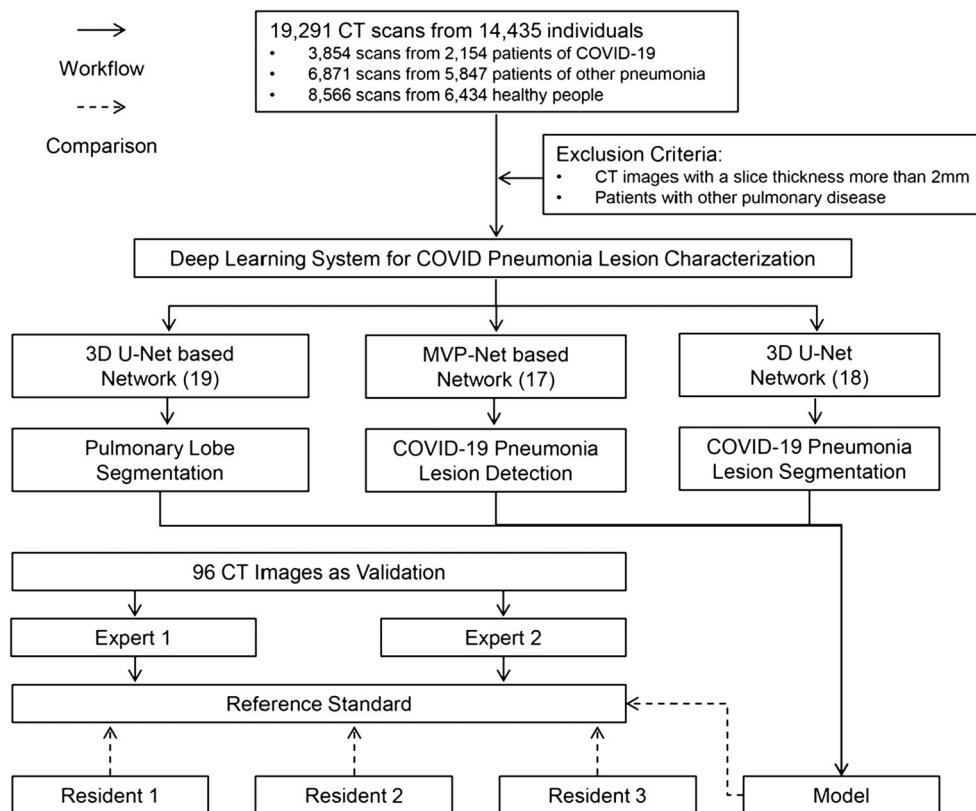
precision. F1 scores and confusion matrix were calculated using scikit-learn 0.19 (scikit-learn.org). p value < 0.05 was regarded as the significant threshold, not corrected for multiple comparisons.

Results

CT image findings

As shown in Table 1, pneumonia was detected in 88 patients (91.7%) on chest CT images from all the 96 patients involved. Most of these 88 patients (74, 77.1%) diagnosed as COVID-19 pneumonia had multiple lesions in initial CT images.

Table 1 Overview of CT imaging features in 96 patients

Variables	All, $n = 96$
Positive CT findings, n (%)	88 (91.7%)
Numbers of lesions	
Solitary, n (%)	14 (14.6%)
Multiple, n (%)	74 (77.1%)
Number of lobes affected	
≥ 2 lobes affected, n (%)	66 (68.8%)
≥ 3 lobes affected, n (%)	54 (56.3%)
≥ 4 lobes affected, n (%)	43 (44.8%)
Locations	
Right lung, n (%)	75 (78.1%)
Upper lobe, n (%)	55 (57.3%)
Middle lobe, n (%)	50 (52.1%)
Lower lobe, n (%)	57 (59.4%)
Left lung, n (%)	73 (76.0%)
Upper lobe, n (%)	53 (55.2%)
Lower lobe, n (%)	65 (67.7%)
Both lungs, n (%)	75 (78.1%)
CT severity score	
Mean score, medium (quartile)	0.25 (0.0875–0.60)
Mild ≤ 20%, n (%)	30 (31.3%)
Moderate 20–50%, n (%)	12 (12.5%)
Severe > 50%, n (%)	54 (56.3%)
CT imaging features	
GGO, n (%)	10 (10.4%)
Consolidation, n (%)	3 (3.1%)
GGO + consolidation, n (%)	75 (78.1%)
Rounded morphology, n (%)	69 (71.9%)
Other morphology, n (%)	67 (69.8%)
Crazy paving pattern, n (%)	32 (33.3%)
Interstitial changes, n (%)	50 (52.1%)
Subpleural distribution, n (%)	82 (85.4%)
Diffuse distribution, n (%)	5 (5.2%)

Sixty-six patients (68.8%) were found to have more than 2 lobes involve. For all the lesions identified by experienced radiologists, 75 patients had abnormalities in the right lung and 73 patients in the left lung. Seventy-five patients (78.1%) presented as bilateral lung involvement. All the 88 patients had GGOs (10, 10.4%), consolidation (3, 3.1%), or the integration of GGOs and consolidation (75, 78.1%). Crazy paving pattern was observed in 32 patients (33.3%), and interstitial abnormalities were found in 50 (52.1%) patients. Eighty-two patients (85.4%) had subpleurally distributed diseases. Typical CT features are listed in Table 1. As defined above, CT severity score ≤ 20% (mild) was seen in 30 patients, 20–50% (moderate) in 12 patients, and > 50% (severe) in 54 patients.

Comparison of performance between deep learning model and radiological residents

The performances of deep learning model and radiological residents in detecting abnormalities from chest CT images are listed on Table 2 based on per-patient and per-lung lobe analysis, respectively. For lesion detection-based per-patient lobe level, the algorithm had a sensitivity of 1.00 (95% CI 0.96, 1.00) in the identification of patients with abnormal CT images. The reading reports from three residents showed sensitivities of 0.94 (95% CI 0.87, 0.98), 0.93 (95% CI 0.86, 0.97), and 0.89 (95% CI 0.80, 0.94)), respectively. The specificity of algorithm was 0.25 (95% CI 0.03, 0.65), while the specificities of residents were 1.00 (95% CI 0.63, 1.00), 0.75 (95% CI 0.35, 0.97), and 1.00 (95% CI 0.63, 1.00). F1 score of the algorithm was 0.97, which was higher than those of the resident 2 and resident 3 (0.95 and 0.94, respectively) and slightly lower than that of resident 1 (0.97). Accordingly, the sensitivity of algorithm was superior to residents in detecting abnormal CT images. Considering the trade-off effect, the specificity of algorithm is inevitably lower, while F1 score of the algorithm is comparable with that of 3 residents. For lesion detection-based per-lung lobe level, accuracy, sensitivity, and specificity of the algorithm were 0.82 (95% CI 0.79, 0.86), 0.96 (95% CI 0.94, 0.98), and 0.63 (95% CI 0.55, 0.69). The accuracy and sensitivity of the algorithm are superior or similar to those of residents, and the specificity of algorithm is slightly inferior to residents. F1 score of the algorithm was 0.86, which was slightly lower than residents (0.89, 0.89, and 0.89, respectively). Overall, the sensitivity of the algorithm is significantly higher than residents, but the specificity is inferior to residents (all p values < 0.017). Figure 2 shows the representative CT images of confirmed COVID-19 patients and the corresponding outputs of the deep learning algorithm.

The comparisons of diagnostic sensitivity between the algorithm and residents based on lung lobe are shown in Table 3. In terms of F1 score, the algorithm was slightly

Table 2 Performance of deep learning model versus radiology residents

Variables	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score	p value sensitivity	p value specificity
Patient level								
Model	0.94 (0.87, 0.98)	1.00 (0.96, 1.00)	0.25 (0.03, 0.65)	0.94 (0.91, 0.96)	1.00 (0.63, 1.00)	0.97		
Resident 1	0.95 (0.88, 0.98)	0.94 (0.87, 0.98)	1.00 (0.63, 1.00)	1.00 (0.95, 1.00)	0.62 (0.41, 0.79)	0.97	0.0233	0.0019
Resident 2	0.92 (0.84, 0.96)	0.93 (0.86, 0.97)	0.75 (0.35, 0.97)	0.98 (0.93, 0.99)	0.50 (0.30, 0.70)	0.95	0.0126	0.0455
Resident 3	0.90 (0.82, 0.95)	0.89 (0.80, 0.94)	1.00 (0.63, 1.00)	1.00 (0.94, 1.00)	0.44 (0.31, 0.60)	0.94	0.0011	0.0019
Lung lobe level								
Model	0.82 (0.79, 0.86)	0.96 (0.94, 0.98)	0.63 (0.55, 0.69)	0.78 (0.75, 0.81)	0.93 (0.87, 0.96)	0.86		
Resident 1	0.88 (0.85, 0.91)	0.87 (0.82, 0.91)	0.90 (0.84, 0.93)	0.92 (0.89, 0.95)	0.83 (0.78, 0.87)	0.89	< 0.0001	< 0.0001
Resident 2	0.88 (0.84, 0.90)	0.84 (0.79, 0.88)	0.93 (0.88, 0.96)	0.94 (0.91, 0.96)	0.80 (0.76, 0.84)	0.89	< 0.0001	< 0.0001
Resident 3	0.88 (0.85, 0.91)	0.82 (0.77, 0.86)	0.96 (0.92, 0.98)	0.97 (0.94, 0.98)	0.79 (0.75, 0.83)	0.89	< 0.0001	< 0.0001

Dates in parentheses are 95% confidence intervals. *p* value for sensitivity represents the *p* value of sensitivity in comparison between algorithm and residents, *p* value for specificity represents the *p* value of specificity in comparison between algorithm and residents. *PPV* positive predictive value, *NPV* negative predictive value

inferior to residents in detection of lesions located on the right upper and right lower lung lobes (F1 score 0.87 vs. 0.92, 0.94, and 0.93, 0.84 vs. 0.88, 0.88, and 0.87). While for right middle, left upper, and left lower lobes, the algorithm was similar to residents with F1 scores of 0.84, 0.87, and 0.90, respectively. The utilization of algorithm enabled high sensitivity of 0.96 (95% CI 0.87, 1.00), 0.94 (95% CI 0.83, 0.99), 0.98 (95% CI 0.91, 1.00), 0.96 (95% CI 0.87, 1.00), and 0.97 (95% CI 0.89, 1.00) in all the five lung lobes, respectively. The sensitivity of the algorithm is slightly higher than residents, demonstrating the distinct advantage of the algorithm in detecting abnormalities from CT images of patients confirmed of COVID-19.

Performance of deep learning model in CT severity scoring

As defined in methods, the accuracy for grading CT severity on a scale was 0.66 (95% CI 0.55, 0.75) for the algorithm, and 0.79 (95% CI 0.70, 0.87), 0.77 (95% CI 0.67, 0.85), and 0.84 (95% CI 0.76, 0.91) for residents. The confusion matrix in Fig. 3 demonstrated the grading discrepancies of the algorithm and residents. The algorithm showed superiority in grading severe CT images, and the algorithm was similar to residents in grading moderate CT images. For mild CT images, the algorithm was inferior to the residents.

Fig. 2 Representative cases. Panels **a** and **b**. Chest CT images of a 53-year-old female diagnosed with COVID-19 pneumonia. **a** Axial unenhanced chest CT image and (**b**) corresponding output with deep learning algorithm show the multifocal subpleurally distributed GGOs with consolidation in the upper lobe of left lung (arrow). Panels **c** and **d**. Chest CT images of a 56-year-old female diagnosed with COVID-19 pneumonia. **c** Axial unenhanced chest CT image and (**d**) corresponding output with deep learning algorithm show the multifocal subpleurally distributed GGOs with crazy paving sign in both lungs (arrows)
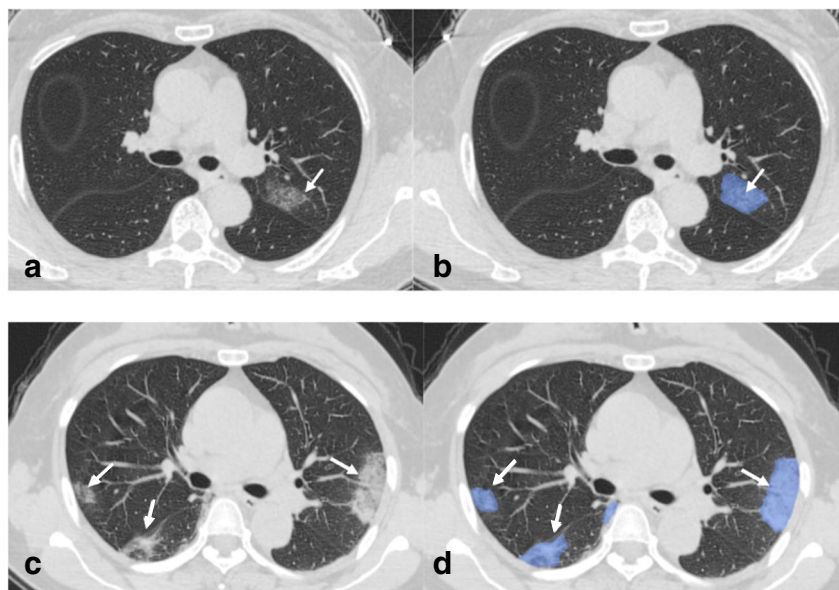
Table 3 Performance of deep learning model versus radiology residents based on anatomical structure

Variables	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score	p value sensitivity	p value specificity
Right upper lobe								
Model	0.83 (0.74, 0.90)	0.96 (0.87, 0.99)	0.66 (0.49, 0.79)	0.79 (0.71, 0.85)	0.93 (0.77, 0.98)	0.87		
Resident 1	0.91 (0.83, 0.96)	0.91 (0.80, 0.97)	0.90 (0.77, 0.97)	0.93 (0.83, 0.97)	0.88 (0.76, 0.95)	0.92	0.2413	0.0076
Resident 2	0.93 (0.86, 0.97)	0.93 (0.82, 0.98)	0.93 (0.80, 0.98)	0.94 (0.85, 0.98)	0.90 (0.79, 0.96)	0.94	0.4011	0.0027
Resident 3	0.93 (0.86, 0.97)	0.91 (0.80, 0.97)	0.95 (0.83, 0.99)	0.96 (0.87, 0.99)	0.89 (0.77, 0.95)	0.93	0.2413	0.0008
Right middle lobe								
Model	0.81 (0.72, 0.88)	0.94 (0.83, 0.99)	0.67 (0.52, 0.80)	0.76 (0.67, 0.83)	0.91 (0.77, 0.97)	0.84		
Resident 1	0.85 (0.77, 0.92)	0.82 (0.69, 0.91)	0.89 (0.76, 0.96)	0.89 (0.78, 0.95)	0.82 (0.71, 0.89)	0.85	0.0648	0.0115
Resident 2	0.86 (0.78, 0.93)	0.78 (0.64, 0.88)	0.96 (0.85, 0.99)	0.95 (0.83, 0.99)	0.80 (0.70, 0.87)	0.86	0.0211	0.0005
Resident 3	0.84 (0.76, 0.91)	0.76 (0.62, 0.87)	0.93 (0.82, 0.99)	0.93 (0.81, 0.97)	0.78 (0.69, 0.86)	0.84	0.0117	0.0016
Right lower lobe								
Model	0.77 (0.67, 0.85)	0.98 (0.91, 1.00)	0.46 (0.30, 0.63)	0.73 (0.67, 0.78)	0.95 (0.71, 0.99)	0.84		
Resident 1	0.86 (0.78, 0.93)	0.86 (0.74, 0.94)	0.87 (0.73, 0.96)	0.91 (0.81, 0.96)	0.81 (0.69, 0.89)	0.88	0.0151	0.0001
Resident 2	0.86 (0.78, 0.93)	0.84 (0.72, 0.93)	0.90 (0.76, 0.97)	0.92 (0.82, 0.97)	0.80 (0.68, 0.88)	0.88	0.0081	< 0.0001
Resident 3	0.86 (0.78, 0.93)	0.79 (0.66, 0.89)	0.97 (0.87, 1.00)	0.98 (0.87, 1.00)	0.76 (0.66, 0.84)	0.87	0.0012	< 0.0001
Left upper lobe								
Model	0.84 (0.76, 0.91)	0.96 (0.87, 1.00)	0.70 (0.54, 0.83)	0.80 (0.71, 0.86)	0.94 (0.79, 0.98)	0.87		
Resident 1	0.89 (0.80, 0.94)	0.89 (0.77, 0.96)	0.88 (0.75, 0.96)	0.90 (0.80, 0.96)	0.86 (0.75, 0.93)	0.90	0.1413	0.0340
Resident 2	0.83 (0.74, 0.90)	0.77 (0.64, 0.88)	0.91 (0.78, 0.97)	0.91 (0.80, 0.96)	0.76 (0.66, 0.84)	0.84	0.0041	0.0148
Resident 3	0.88 (0.79, 0.93)	0.79 (0.69, 0.87)	0.98 (0.86, 1.00)	0.97 (0.86, 1.00)	0.79 (0.69, 0.87)	0.88	0.0077	0.0005
Left lower lobe								
Model	0.85 (0.77, 0.92)	0.97 (0.89, 1.00)	0.61 (0.42, 0.78)	0.84 (0.77, 0.89)	0.90 (0.70, 0.97)	0.90		
Resident 1	0.89 (0.80, 0.94)	0.86 (0.75, 0.93)	0.94 (0.79, 0.99)	0.97 (0.88, 0.99)	0.76 (0.64, 0.86)	0.91	0.0274	0.0023
Resident 2	0.89 (0.80, 0.94)	0.86 (0.75, 0.93)	0.94 (0.79, 0.99)	0.97 (0.88, 0.99)	0.76 (0.64, 0.86)	0.91	0.0274	0.0023
Resident 3	0.89 (0.80, 0.94)	0.85 (0.74, 0.92)	0.97 (0.83, 1.00)	0.98 (0.89, 1.00)	0.75 (0.63, 0.84)	0.91	0.0154	0.0006

Dates in parentheses are 95% confidence intervals. *p* value for sensitivity represents the *p* value of sensitivity in comparison between algorithm and residents, *p* value for specificity represents the *p* value of specificity in comparison between algorithm and residents. *PPV* positive predictive value, *NPV* negative predictive value

Volume information extracted by deep learning model

The algorithm in this study specifically extracted the detailed volume and density of each abnormality, distance of lesion from pleura from chest CT images. The median volume of all the detected lesions on per-patient basis is 40.10 cm^3 (interquartile range 7.67, 116.16). The median volume of single lesion is 0.64 cm^3 (interquartile range 0.11, 3.06). The median CT value of the lesion is − 555 HU with the interquartile range of − 6980 HU and − 401 HU. The median distance of lesion from pleura is 2.90 mm with the interquartile range of 0.93 and 10.83.

As shown in Table 4, the algorithm exhibited a much faster diagnosis speed at a mean rate of 20.3 s ± 5.8 per case, while the residents executed the task with reading speed of 101.1 s ± 53.3, 68.3 s ± 18.5, and 112.4 s ± 44.7, respectively (all *p* values < 0.017).

Assistance of deep learning algorithm generated results

Figure 4 displays performance of residents with the aiding of AI system. The algorithm had AUCs of 0.86 (0.74, 0.98) and 0.87 (0.75, 0.98) in the identification of lesions on per-patient and per-lobe basis, which were slightly inferior to the residents (triangle markers). However, the assistance of AI system improved the diagnostic performance of three residents (circle markers). As shown in Table 5, the sensitivity was slightly improved with the assistance of AI system (0.94 vs. 0.98, 0.93 vs. 0.97, 0.89 vs. 0.97) without sacrifice of specificity on per-patient basis. For per-lobe basis, the diagnostic performance of three residents with the combination of AI system was also superior to their initial performance. Notably, the AI system can assist radiologists make quicker diagnosis with much faster diagnosis speeds (101.1 vs. 44.9 s, 68.3 vs. 39.2 s, 112.4 vs 48.8 s, all *p* values < 0.0001) (Table S1).

Discussion

In our study, we utilized and validated a deep learning approach for precise chest CT image feature identification and quantitative assessment in 96 consecutive patients diagnosed with COVID-19. In the survey of chest CT images, the algorithm specifically analyzed the volume of abnormalities and distance between lesion and pleura. Also, the algorithm presented a much faster rate in CT image reading than residents. In the detection of infected patients with COVID-19 pneumonia, the algorithm showed robust performance with sensitivity

Fig. 3 Confusion matrix comparing CT severity grading performance between deep learning model and radiological residents

of 1.00 (0.96, 1.00), which is significantly higher than residents. Based on per-patient or per-lung lobe level, it was demonstrated that algorithm was comparable with that of radiologists, with F1 scores of 0.97 vs. 0.97, 0.95, and 0.94, and 0.86 vs. 0.89, 0.89, and 0.89. This study highlights the usefulness of this deep learning model in actual clinical practice.

Utilization of chest CT scanning for suspected patients at admission has been recommended by Chinese health professionals for prompt diagnosis [21]. AI technology powers many aspects in medical research, especially the image processing [22, 23]. In the past years, there have been several deep learning–based automatic algorithms for detection of abnormalities in chest radiography and CT images, including lung cancer screening, malignant pulmonary nodule detection, and pulmonary tuberculosis classification [24–27]. These researches demonstrated the property of deep learning model in facilitating the screening and evaluation of pulmonary diseases. In this study, we applied a deep learning model which is comparable with radiologists in detecting abnormities on CT images from patients confirmed of COVID-19. The automatic detection and analysis make the diagnosis of COVID-19 pneumonia much faster than traditional reading process and reduces the burden of clinicals in repeated exposure in the new coronavirus. To some extent, the application of deep learning algorithm in medical imaging accelerates the diagnosis and reduces the human-to-human transmission in hospital.

Noteworthy, radiologists across the world have provided new insights by accessing the lung CT as additional diagnosis or screening tool of COVID-19 pneumonia. Basically, bilateral GGOs, consolidative pulmonary opacities, as well as the prominent subpleural distribution are regarded as classical features in chest CT images of patients diagnosed with COVID-19 pneumonia, which are similar to those reported with SARS-CoV and MERS-CoV [9–19]. In parallel with these findings, our study also demonstrated higher incidence of GGOs and consolidative opacities in the CT images from COVID pneumonia patients. Specially, as shown by Bernheim in a relatively larger retrospective study, lung abnormalities of COVID-19 pneumonia detected by CT was related with virus time course, and mostly, the lesion features progressed from GGO to crazy paving pattern [28]. The "Diagnosis and Treatment Program of 2019 New Coronavirus Pneumonia" (trial sixth version) released by Chinese Health Commission highlighted that the change of lesion volume larger than 50% in 24 to 48 h was suggested as severe disease in management [21]. The deep learning model we used here can automatically calculate the volume of lesion and precisely locate the lesions which may be of great importance in monitoring, evaluating disease severity, and guiding the treatment by collecting and analyzing data from baseline and follow-up CT images.

Another advantage of our study is that we evaluated the performance of deep learning models in abnormality detection from chest CT images of COVID-19 pneumonia patients. It is confirmed that the algorithm we used was non-inferior to experienced radiologists in lesion detection and identification. Currently, there is a study by Xu which retrospectively analyzed the performance of inception migration-learning model in distinguishing COVID-19 with other pathogen infection [15]. In the external test, their algorithm model showed a total accuracy of 73% with sensitivity of 74% and specificity of 67%. Unlike it, our algorithm was specifically developed for detailed structure information extraction and precise lesion detection. For all the 96 patients with chest CT images involved, this algorithm exhibited high sensitivity in pneumonia diagnosis both the per-patient and per-lung lobe basis. High sensitivity of algorithm would be especially important in prompt screening of COVID-19-infected patients. When

Table 4 Running time comparison (unit in second)

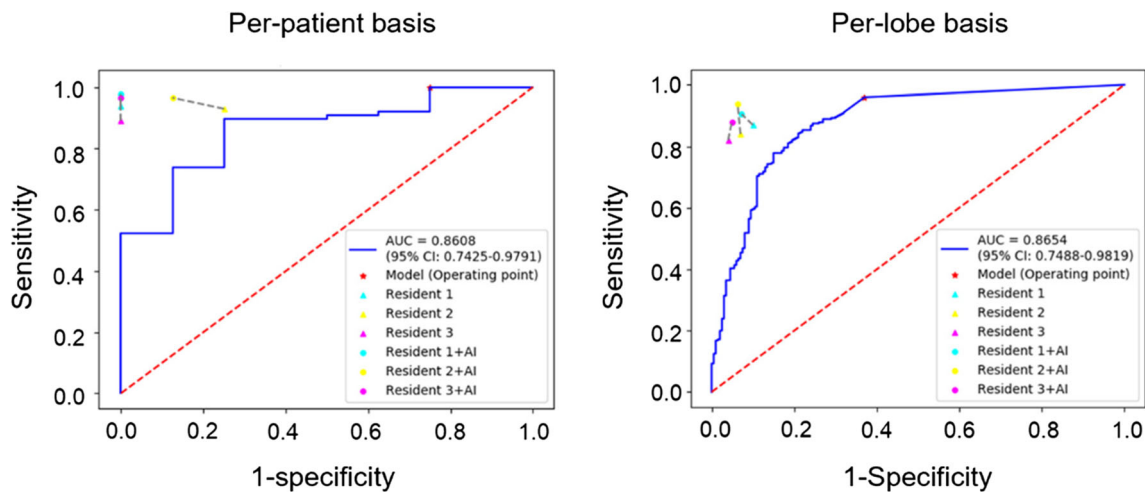| Variables	Mean	STD	Max	Min	*p* value
Deep leaning model	20.3	5.8	38.9	10.5	
Senior Radiologist	82.7	17.5	108	52	< 0.0001
Resident 1	101.1	53.3	218	23	< 0.0001
Resident 2	68.3	18.5	100	32	< 0.0001
Resident 3	112.4	44.7	186	34	< 0.0001

STD standard deviation

Fig. 4 Receiver operating characteristic (ROC) diagram for AI system versus radiologists. The blue curve was created by taking different thresholds over the predicted probability, showing the macro-average AUC of AI system. The asterisk showed the performance of model in a balanced setting. The filled markers showed residents' performance. Dashed line connected performance of radiologists with and without the assistance of AI system

compared with radiology residents' report, we found the specificity of algorithm is inferior to clinicians, which is attributed to metallic or respiratory marked artifact ($n = 3$) and fibrosis ($n = 3$) easily recognized by human experts. Objectively, the deep learning model we utilized here improved the sensitivity with the sacrifice of specificity in lesion detection. Despite the trade-off between sensitivity and specificity, considering the global outbreak and fast spread, prompt diagnosis and quarantine should be the most imperative action; sensitivity, instead of specificity, should play a more important role in identifying patients infected with the new coronavirus. We believe the application of deep learning model will

accelerate the speeds of patient screening and effectively stop the human-to-human transmission.

Due to the development of computer science, AI techniques have been widely applied in biological and medical researches in recent years. So far, there have been some successful cases based on AI which have made great contributions to epidemic alert and infected patient screening [13]. Li et al recently reported a COVID-19 detection neural network (COVNet) which successfully distinguished COVID-19 pneumonia from community-acquired pneumonia [29]. To the best of our knowledge, our study first applied deep learning model to comprehensively analyze lesion features from chest CT images of COVID-19 patients. Notably, the

Table 5 Performance of residents with assistance of deep learning model

Variables	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score	p value sensitivity	p value specificity
Patient level								
Resident 1	0.95 (0.88, 0.98)	0.94 (0.87, 0.98)	1.00 (0.63, 1.00)	1.00 (0.95, 1.00)	0.62 (0.41, 0.79)	0.97		
Resident 1+ AI	0.98 (0.92, 1.00)	0.98 (0.92, 1.00)	1.00 (0.63, 1.00)	1.00 (0.95, 1.00)	0.80 (0.48, 0.95)	0.99	0.440	1.00
Resident 2	0.92 (0.84, 0.96)	0.93 (0.86, 0.97)	0.75 (0.35, 0.97)	0.98 (0.93, 0.99)	0.50 (0.30, 0.70)	0.95		
Resident 2+ AI	0.96 (0.89, 0.99)	0.97 (0.90, 0.99)	0.88 (0.51, 1.00)	0.99 (0.93, 1.00)	0.70 (0.39, 0.90)	0.98	0.494	0.519
Resident 3	0.90 (0.82, 0.95)	0.89 (0.80, 0.94)	1.00 (0.63, 1.00)	1.00 (0.94, 1.00)	0.44 (0.31, 0.60)	0.94		
Resident 3+ AI	0.97 (0.91, 0.99)	0.97 (0.90, 0.99)	1.00 (0.63, 1.00)	1.00 (0.95, 1.00)	0.73 (0.43, 0.91)	0.98	0.044	1.00
Lung lobe level								
Resident 1	0.88 (0.85, 0.91)	0.87 (0.82, 0.91)	0.90 (0.84, 0.93)	0.92 (0.89, 0.95)	0.83 (0.78, 0.87)	0.89		
Resident 1+ AI	0.91 (0.89, 0.94)	0.91 (0.86, 0.94)	0.93 (0.88, 0.96)	0.94 (0.91, 0.97)	0.88 (0.83, 0.92)	0.92	0.203	0.337
Resident 2	0.88 (0.84, 0.90)	0.84 (0.79, 0.88)	0.93 (0.88, 0.96)	0.94 (0.91, 0.96)	0.80 (0.76, 0.84)	0.89		
Resident 2+ AI	0.94 (0.91, 0.96)	0.94 (0.90, 0.96)	0.94 (0.89, 0.96)	0.95 (0.92, 0.97)	0.91 (0.87, 0.95)	0.94	< 0.001	0.791
Resident 3	0.88 (0.85, 0.91)	0.82 (0.77, 0.86)	0.96 (0.92, 0.98)	0.97 (0.94, 0.98)	0.79 (0.75, 0.83)	0.89		
Resident 3+ AI	0.91 (0.88, 0.93)	0.88 (0.84, 0.91)	0.95 (0.91, 0.97)	0.96 (0.93, 0.98)	0.86 (0.80, 0.90)	0.92	0.053	0.668

involvement of AI markedly accelerates the reading process without the sacrifice of sensitivity. And the assistance of AI system improves the diagnostic performance of radiologists. We believe the application of AI system will effectively accelerate the diagnosis of pneumonia and provide the precise location of pneumonia lesions. COVID-19 will not be the last epidemic to challenge public health experts. The growth of AI-driven techniques to identify epidemiologic risks early will be key to our improvement of prediction, prevention, and detection of future global health risks.

There are several limitations of this study. First, since this is a retrospective study, the performance of deep learning model on an actual clinical situation is not validated. Real-time application of this model in clinical practice is needed. Second, we used experienced radiologists' reading reports as reference standard. Although it is a routine practice, there might still be some variabilities. Third, we involved a total of 96 patients from three hospitals across China, whereas 87.5% are from one single institution, so the reproducibility of the performance of our algorithm remains unclear. Fourth, because of the small sample size and outbreak of epidemic, our study suffered the imbalanced database problem. Appropriate statistical evaluation was not applied because commonly used probabilistic metric or ranking metric is not applicable in this deep learning algorithm. Also, the testing results from a small dataset might not generalize well to all the unseen cases, we expect larger database from multi-centers across the world to test our deep learning model in COVID-19 pneumonia detection. Finally, this deep learning model showed worse specificity than radiologists in lesion detection, which will lead to more false positive cases. However, these results are easily recognized by human experts.

In conclusion, we utilized a deep learning model in specific feature extraction and quantitative lesion detection from chest CT images of patients diagnosed with COVID-19 pneumonia. The precise lesion identification such as volume may provide valuable information for clinical classification and treatment selection. Moreover, the algorithm we used in this study presented superior diagnostic performance in quantitatively detecting abnormalities on per-patient and per-lung lobe basis compared with radiologists, making rapid referral suggestions that deep learning algorithm should be a standard care in real-time application.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Long Jiang Zhang.

Conflict of interest All authors have no conflicts of interest to disclose.

Statistics and biometry Zhen Zhou kindly provided statistical advice for this manuscript.
     One of the authors has significant statistical expertise.
     No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap Study subjects or cohorts have not been previously reported.

Methodology
• retrospective
• multicenter study

References

1. Kickbusch I, Leung G (2020) Response to the emerging novel coronavirus outbreak. BMJ 368:m406
2. Phelan AL, Katz R, Gostin LO (2020) The nNovel cCoronavirus oOriginating in Wuhan. China: cChallenges for gGlobal hHealth gGovernance. JAMA. https://doi.org/10.1001/jama.2020.1097
3. Zhu N, Zhang D, Wang W et al (2020) A nNovel cCoronavirus from pPatients with pPneumonia in China, 2019. N Engl J Med 382:727–733
4. Holshue ML, DeBolt C, Lindquist S et al (2020) First cCase of 2019 nNovel cCoronavirus in the United States. N Engl J Med https://doi.org/10.1056/NEJMoa2001191
5. World Health Organization (2020) Coronavirus disease 2019 (COVID-19): situation report-39. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200303-sitrep-43-ncov.pdf. Published (March 03, 2020) [Epub ahead of print]
6. Mahase E (2020) Coronavirus covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. BMJ 368:m641
7. Chen N, Zhou M, Dong X et al (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan. China: a descriptive study. Lancet. https://doi.org/10.1016/S0140-6736(20)30211-7
8. Phan T (2020) Novel coronavirus: fFrom discovery to clinical diagnostics. Infect Genet Evol 79:104211
9. Chung M, Bernheim A, Mei X et al (2020) CT iImaging fFeatures of 2019 nNovel cCoronavirus (2019-nCoV). Radiology. https://doi.org/10.1148/radiol.2020200230:200230
10. Fang Y, Zhang H, Xie J et al (2020) Sensitivity of cChest CT for COVID-19: cComparison to RT-PCR. Radiology. https://doi.org/10.1148/radiol.2020200432:200432
11. Long JB, Ehrenfeld JM (2020) The rRole of aAugmented iIntelligence (AI) in dDetecting and pPreventing the sSpread of nNovel cCoronavirus. J Med Syst 44:59
12. Health LD (2020) COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread. Lancet Digital Health. S2589-7500(20):30054–30056
13. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:44–56
14. Tomasev N, Glorot X, Rae JW et al (2019) A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 572:116–119

15. Wang S, Kang B, Ma J et el (2020a2020) A deep learning algorithm using CT images to screen for cCorona vVirus dDisease (COVID-19). Preprint available via https://www.medrxiv.org/content/10.1101/2020.02.14.20023028v2

16. Yu Q, Wang Y, Huang S et al (2020) Multicenter cohort study demonstrates more consolidation in upper lungs on initial CT increases the risk of adverse clinical outcome in COVID-19 patients. Theranostics 10(12):5641–5648

17. Song F, Shi N, Shan F et al (2020) Emerging cCoronavirus 2019-nCoV pPneumonia. Radiology. https://doi.org/10.1148/radiol.2020200274:200274

18. Li Z, Zhang S, Zhang J, Huang K, Wang Y, Yu Y (2019) MVP-net: mMulti-view FPN with position-aware attention for deep universal lesion detection. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)

19. Ronneberger O, Fischer P, Brox T (2015) U-net: cConvolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention.

20. Wang XQ, Zhang QY, Zhou Z et al (2020) Evaluating multi-class segmentation errors with anatomical prior. IEEE International Symposium on Biomedical Imaging.

21. General Office of National Health Committee (2020) Notice on the issuance of a program for the diagnosis and treatment of novel coronavirus (2019-nCoV) infected pneumonia (trial sixth edition) In Press [EB/OL]. General Office of National Health Committee. Available via http://yzs.satcm.gov.cn/zhengcewenjian/2020-02-19/13221.html]).

22. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444

23. Ardila D, Kiraly AP, Bharadwaj S et al (2019a2019) End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 25:954–-961

24. Nam JG, Park S, Hwang EJ et al (2019) Development and vValidation of dDeep lLearning-based aAutomatic dDetection aAlgorithm for mMalignant pPulmonary nNodules on cChest rRadiographs. Radiology 290:218–228

25. Ardila D, Kiraly AP, Bharadwaj S et al (2019b2019) End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 25:954–961

26. Rajpurkar P, Irvin J, Ball RL et al (2018) Deep learning for chest radiograph diagnosis: aA retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15:e1002686

27. Lakhani P, Sundaram B (2017) Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 284:574–582

28. Bernheim A, Mei X, Huang M et al (2020) Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. Radiology. https://doi.org/10.1148/radiol.2020200463:200463

29. Li L, Qin L, Xu Z et al (2020) Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology. https://doi.org/10.1148/radiol.2020200905:200905