Development and Evaluation of an AI System for COVID-19 Diagnosis

Cheng Jin[†1], Weixiang Chen[†1], Yukun Cao[†2,3], Zhanwei Xu[1], Xin Zhang[2,3], Lei Deng[1], Chuansheng Zheng[2,3], Jie Zhou[1], Heshui Shi[*2,3], and Jianjiang Feng[*1]

[1]Department of Automation, Tsinghua University
[2]Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology
[3]Hubei Province Key Laboratory of Molecular Imaging, China

March 21, 2020

Abstract

Early detection of COVID-19 based on chest CT will enable timely treatment of patients and help control the spread of the disease. With rapid spreading of COVID-19 in many countries, however, CT volumes of suspicious patients are increasing at a speed much faster than the availability of human experts. Here, we propose an artificial intelligence (AI) system for fast COVID-19 diagnosis with an accuracy comparable to experienced radiologists. A large dataset was constructed by collecting 970 CT volumes of 496 patients with confirmed COVID-19 and 260 negative cases from three hospitals in Wuhan, China, and 1,125 negative cases from two publicly available chest CT datasets. Trained using only 312 cases, our diagnosis system, which is based on deep convolutional neural network, is able to achieve an accuracy of 94.98%, an area under the receiver operating characteristic curve (AUC) of 97.91%, a sensitivity of 94.06%, and a specificity of 95.47% on an independent external verification dataset of 1,255 cases. In a reader study involving five radiologists, only one radiologist is slightly more accurate than the AI system. The AI system is two orders of magnitude faster than radiologists and the code is available at $https://github.com/ChenWWWeixiang/diagnosis_covid19$.

† Cheng Jin, Weixiang Chen and Yukun Cao — contributed equally to this work.
* Co-corresponding authors: Jianjiang Feng (jfeng@tsinghua.edu.cn) and Heshui Shi (heshuishi@hust.edu.cn)

1 Introduction

The new coronavirus disease, now known as COVID-19[1], was first detected in Wuhan, China, in December 2019. Outbreaks have now occurred in more than 166 countries. It has infected more than 210,000 people worldwide, killed more than 8,000 and is still spreading rapidly. The World Health Organization (WHO) officially announced that this is a global pandemic.

It is important to diagnosis COVID-19 as quick and accurate as possible for controlling the spread of the disease and treating patients. Even though reverse transcription-polymerase chain reaction (RT-PCR) is still ground truth of COVID-19 diagnosis, the sensitivity of RT-PCR is not high enough for low viral load present in test specimens or laboratory error[2], and kits of RT-PCR are in short of supply in many areas[3]. By contrast, chest CT can show early lesions in the lung and, if diagnosed by experienced radiologists, can achieve high sensitivity[4]. In addition, CT scans are widely available and economic. At present, the diagnosis of chest CT depends on the radiologists, which has some problems. Firstly, chest CT contains hundreds of slices, which takes a long time to diagnose. Secondly, COVID-19, as a new lung disease, has similar manifestations with various types of pneumonia4. Radiologists need to accumulate a lot of CT diagnostic experience to achieve a high diagnostic performance. In some outbreak areas, many suspected patients are in urgent need of diagnosis and proper treatment, and many CT scans are performed every day. There is an urgent shortage of radiologists with high diagnosis performance for COVID-19.

Artificial intelligence (AI) may be the unique preparation to take up this challenge. Powered by large labeled datasets[5] and modern GPUs, AI, especially deep learning technique[6], has achieved excellent performance in several computer vision tasks such as image classification[7] and object detection[8]. Recent research shows that AI algorithms can even achieve or exceed the performance of human experts in certain medical image diagnosis tasks[9-13]. The AI diagnosis algorithm also has the advantages of high efficiency, high repeatability and easy large-scale deployment. The current outbreak of COVID-19 is worldwide, and the shortage of specialist radiologists threatens the availability and adequacy of screening services for COVID-19 in affected areas. By deploying AI diagnosis algorithms, suspicious patients everywhere, especially in developing countries, will have equal access to right diagnosis, timely treatment and isolation.

As a very recent disease, we have not yet found AI studies for COVID-19 diagnosis in peer-reviewed publications, but a few reports about COVID-19 diagnosis algorithms based on chest CT in preprint form[14, 15]. Wang et al.[14] describe a COVID-19 diagnosis system with specificity of 67% and sensitivity of 74% on 216 slices extracted from CT volumes of patients (the whole dataset consists of 44 positive and 55 negative cases, but split strategy of dataset is unclear). Chen et al.[15] describe a COVID-19 diagnosis system with a performance comparable to that of an expert radiologist, however the system is validated based on a quite small dataset with only 19 confirmed COVID-19 patients and only one radiologist is compared. Clearly, the development and

rigorous testing of COVID-19 diagnosis algorithms remains an open topic.

In this study, we used clinically representative large-scale datasets from three centers in Wuhan and two publicly available chest CT datasets to develop and evaluate an AI system for the diagnosis of COVID-19. We compared the diagnostic performance of this system with that of five radiologists in a reader study of 200 cases, and the results showed the performance of this system is comparable to that of experienced radiologists in designated hospitals for COVID-19 in major epidemic areas. In addition, based on prediction score on every slice, we located the lesion areas in COVID-19 patients and statically analyzed lesion position in different subsets. We traced the specific phenotypic basis of the diagnosis output from the system on the original CT images through an interpretation network and apply radiomics analysis of the attentional region to understand the imaging characteristics of COVID-19. Figure 1 a shows the overall flow of the study.

2 Results

2.1 Datasets for System Development and Evaluation

We developed and evaluated a deep learning based COVID-19 diagnosis system, using multi-center data, which includes 756 subjects (496 COVID-19 positives and 260 negatives) collected in Wuhan Union Hospital, Western Campus of Wuhan Union Hospital, and Jianghan Mobile Cabin Hospital in Wuhan, the major epidemic area in China (detailed information is in Table 1). Two international public databases, LIDC-IDRI (1,012 subjects available)[16] and ILD-HUG (113 subjects available)[17] are also included as negative cases to help develop and evaluate our AI system. Negatives from the three centers in Wuhan are healthy people and ones from public databases are healthy or suffer from other lung diseases.

All CT data was divided into four independent parts with no overlapping subjects, a 312 subjects training cohort, a 104 subjects internal validation cohort, an external test cohort consist of 1,255 subjects, and a reader study cohort with 200 subjects (detailly described in Methods). Due to some positive subjects have multi-stage CTs and different stage CTs of one subject might be similar, the subset division was performed on subjects to make sure all multi-stage CTs of the same subject are in a same subset.

2.2 Construction of the AI System for COVID-19 Diagnosis

We propose a deep-learning based AI system for COVID-19 diagnosis, which directly takes CT data as input, performs lung segmentation, COVID-19 diagnosis and abnormal slices locating. In addition, we hope that the diagnosis results of AI system can be quantitatively explained in the original image to alleviate the drawback of deep neural networks as a black box. The system con-

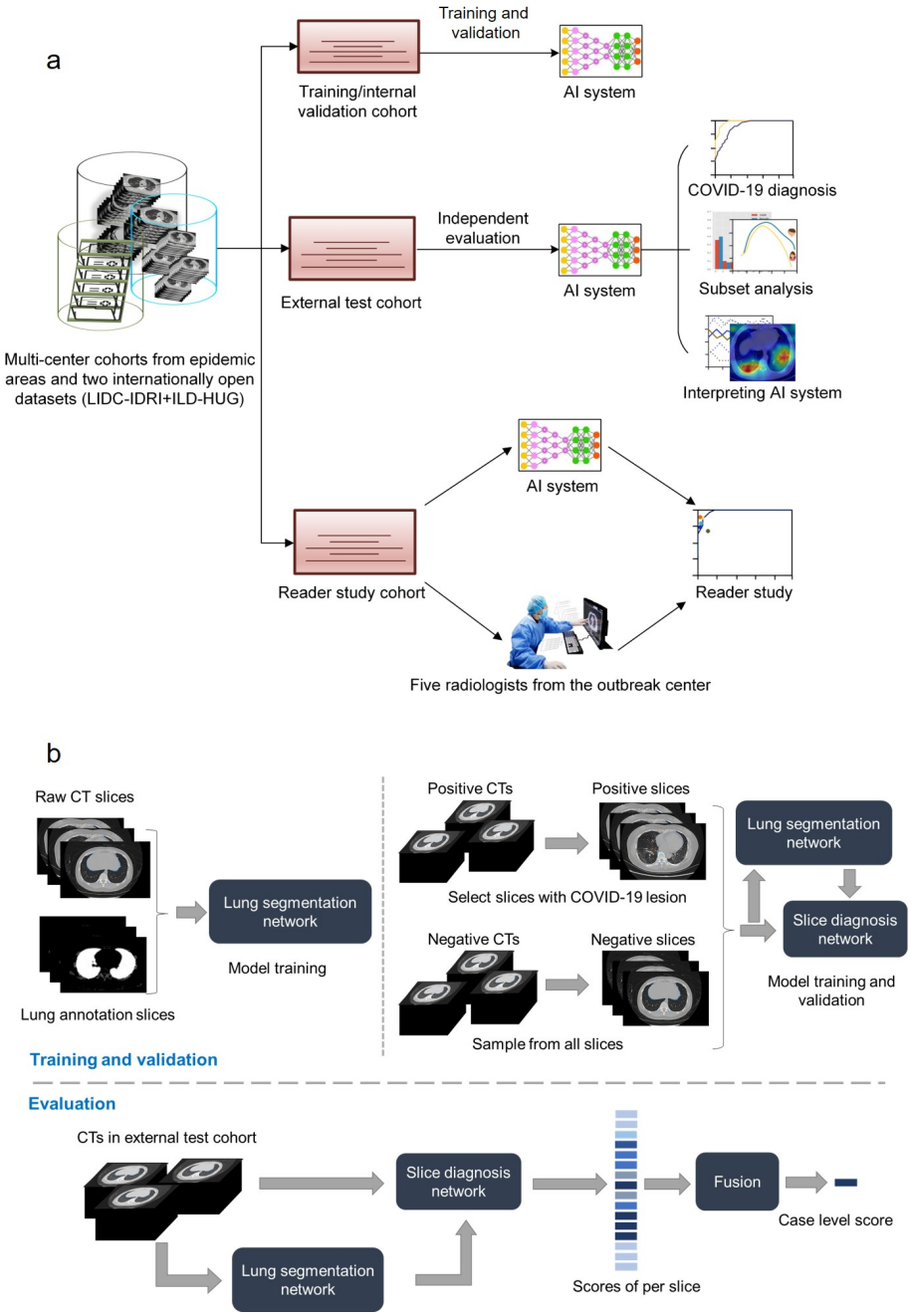


Figure 1: Workflows of the whole study and the proposed AI system. a. Workflow of the whole study. b. Construction and usage of the AI system.

Variables		Training cohort	Internal validation cohort	External test cohort	Reader study cohort
Gender	Female	140	47	198	95
	Male	82	27	101	66
Ages (years)	<20	1	0	1	0
	20-39	135	47	145	84
	40-59	58	14	88	49
	≥60	28	13	65	78
Diagnosis result	Positive	147	49	183	117
	Negative	75	25	116	44

Table 1: Characteristics of patients from Wuhan Union hospital, Western Campus of Wuhan Union Hospital, Jianghan Mobile Cabin Hospital.

sists of five key components (Figure 1 a), (1) a 2D convolutional neural network (CNN) model for segmenting the lung, (2) a COVID-19 diagnosis model, (3) an abnormal slices locating block to locate abnormal slices in positive cases, (4) a network visualization module for interpreting the attentional region of deep networks, and (5) an image phenotype analysis module for explaining the features of the attentional region.

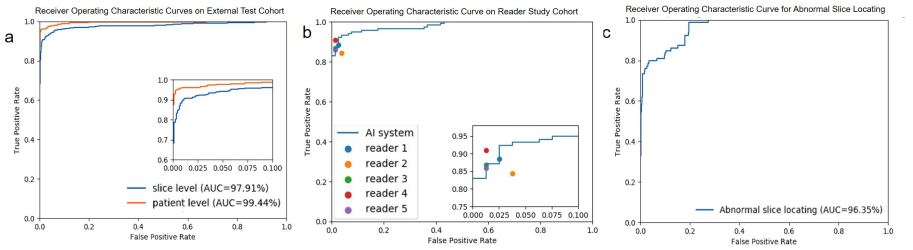
The workflow of deep-learning based diagnosis model is shown in Figure 1 b. CT cases were firstly divided to different cohorts and extracted to slices since our model takes 2D slices as input. Then after slice level training, our model can accurately predict whether the input slices come from COVID-19 subjects. With a top-k average block, our model finally fused slice results into case level diagnosis. The model was implemented in 2D not only because 2D network was easily to train with more training samples, but also because slice-level scores can be used for abnormal slice locating. We fine-tuned our diagnosis model on a training dataset consisting of normal and abnormal slices from COVID-19 positive cases and obtained the abnormal slice locating model. Other parts of our system are described in Methods.

2.3 Performances of AI System

The trained AI system was evaluated on the external test cohort. We used the receiver operating characteristic (ROC) curves (in Figure 2 a) to evaluate the diagnostic accuracy of the proposed AI system. The PR curves of evaluating the prediction accuracy were also illustrated in Figure 7 a.

On the external test cohort, the ROC curve showed AUC of 0.9791, sensitivity of 0.9406, and specificity of 0.9547. In the open data set LIDC-IDRI and ILD-HUG, the false positive rates of AI system were 3.12% and 11.85%, and the system showed good generalization ability (Figure 7 c).

The AI system shows good performances and it can be used with different diagnosis thresholds according to different policies or prior probabilities. The



d

Metrics	External Test Cohort	Reader Study Cohort	Abnormal Slice Locating
AUC (95% CI)	0.9791 (0.9753-0.9841)	0.9805 (0.9746-0.9864)	0.9635 (0.9569-0.9722)
ACC (95% CI)	0.9498 (0.9446-0.9554)	0.9337 (0.9231-0.9487)	0.8901 (0.8796-0.9011)
SPEC (95% CI)	0.9547 (0.9495-0.9631)	0.9139 (0.8871-0.9483)	0.9008 (0.8916-0.9116)
SENS (95% CI)	0.9406 (0.9325-0.9496)	0.9470 (0.9341-0.9604)	0.8219 (0.7895-0.8615)
PPV (95% CI)	0.9153 (0.9070-0.9286)	0.9425 (0.9247-0.9674)	0.5646 (0.5269-0.6087)
NPV (95% CI)	0.9686 (0.9640-0.9739)	0.9204 (0.9016-0.9394)	0.9700 (0.9626-0.9782)
F1 (95% CI)	0.9278 (0.9203-0.9352)	0.9447 (0.9326-0.9588)	0.6691 (0.6358-0.7037)
Youden (95% CI)	0.8953 (0.8842-0.9056)	0.8608 (0.8312-0.8980)	0.7226 (0.6886-0.7690)

e

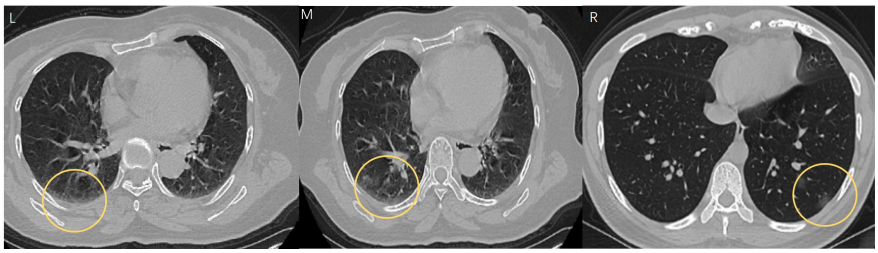


Figure 2: Performance evaluations of AI system for COVID-19 diagnosis in External Test Cohorts. a. ROC curves of COVID-19 diagnosis on external test cohort. b. Performances of the AI system and five readers in COVID-19 diagnosis on reader study cohort. c. ROC curve for abnormal slice locating. This result was test on 12 COVID-19 positives cases from internal validation cohort which have manual lesion segmentation. d. Metrics of proposed AI system for different cohorts and tasks. e. Discrepancies between the AI system and human readers. L) M) Two COVID-19 cases identified by the AI system but missed by all five readers. R) A COVID-19 case identified by a reader but missed by the AI system. (The yellow circles denote possible lesion area)

sensitivity of our system is about 84.76% when specificity is 99.5%, and specificity is 80.02% when sensitivity is 97%. Besides, because patients in our external test cohort have multi-stage CT volumes, some of the stages of positive subjects might be in the recovery state whose CT may have no abnormalities but are still regarded as positive in experiments. Figure 2 a shows the results after roughly filtering out these cases by only keeping the maximum predicted value of multi-stage CTs in the same patient, in which the specificity is about 96.74% at sensitivity of 97%. The decision curve analysis (DCA) for the AI system are presented in Figure 7 b, which indicated that the AI system adds benefit than the "diagnose all" or "diagnose none" strategies when the threshold is within a wide range 1.82-97.6% in COVID-19.

Abnormal slice locating results are showed in Figure 2 c, d. The slice locating block took in COVID-19 positive cases and predicted where abnormal slices are with AUC of 96.35%, specificity of 90.08% and sensitivity of 82.19%.

2.4 Comparison of AI System to Radiologists

We conducted a reader study with five board-certified radiologists (Average of 8 years clinical experience, range 5-11 years, Table 2 a). These radiologists interpreted 200 CT volumes in reader study cohort. All readers were asked to read independently without other information about patients.

The AI system performs slightly better than the average of five radiologists. The ROC curve had AUC of 0. 9805, sensitivity of 0. 9470, and specificity of 0. 9139 on the cohort of reader study (Figure 2 b, d). In 46% (6/13) of cases, when the AI system missed, the radiologist also missed (Table 2 b), indicating that the diagnosis of these missed cases is challenging. Among the five readers, one reader performed better than the AI system, one reader performed worse, and the rest three have similar performance as the AI system at different operating points. Performance of the AI system in COVID-19 diagnosis compared to five readers is shown in Figure 2 b and Table 2 c.

The left, middle of Figure 2 e shows two COVID-19 cases that all five radiologists missed but were correctly identified by the AI system. It is hard to notice the evidence of COVID-19 in these two cases, but AI system can still identify them. It shows that the AI system may capture the information of subtle and continuous changes in slices and make a comprehensive judgment. The right of Figure 2 e shows an example that was detected by all five radiologists but missed by the AI system. These cases show that the AI system and human readers are potentially complementary.

2.5 Subset Analysis

For an in-depth understanding of the AI system and characteristics of different populations with COVID-19, we evaluated the AI system on subsets of the external test cohort divided by gender, age and number of CT scans. Figure 3 a shows the ROC curves of these three subsets. To understand the cause for different diagnosis performances, we analyzed the abnormal slice locating results

a

Reader Experience			
	Reads per year	Years of experience	Reads on COVID-19
Reader 1	3000-4000	5	300-400
Reader 2	4000-5000	10	300-400
Reader 3	4000-5000	11	400-500
Reader 4	3000-4000	8	400-500
Reader 5	3000-4000	7	400-500

b

Reader study		Any of five readers	
		Correct	Wrong
AI system	Correct	172	15
	Wrong	7	6

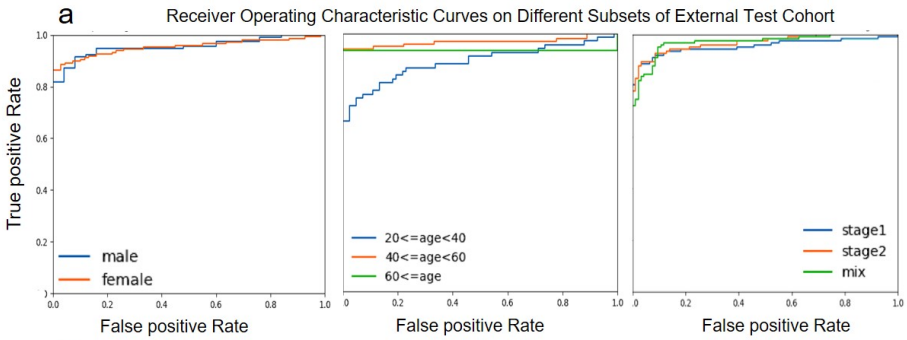
c

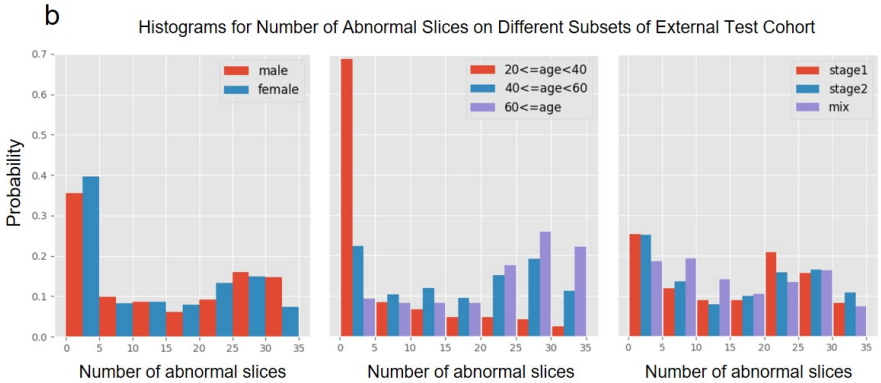
Reader study		First reader		Second reader		Third reader		Fourth reader		Fifth reader	
		Correct	Wrong	Correct	Wrong	Correct	Wrong	Correct	Wrong	Correct	Wrong
AI system	Correct	178	9	173	14	176	11	179	8	175	12
	Wrong	8	5	8	5	8	5	9	4	7	6

Table 2: Reader study statistics and results. a. Experience levels of the five radiologists involved in the reader study. b. Comparison of diagnostic error between AI systems and human readers. Notes: wrongs of any of five readers means any one or more readers were wrong, while corrects means no any of five readers were wrong. c. Comparison of diagnostic performances between AI system and every human reader.

in different subsets (Figure 3 b, c). We found that the different performances of different subsets were highly correlated to the number of abnormal slices, which means smaller lesion with fewer abnormal slices are more difficult to diagnose (Figure 3 c,Figure 4 b). Together with the position of abnormal slices and the voxel numbers of lungs (Figure 4 a), we concluded that reason for worse performance between 20 and 40 years old might be that younger people may have smaller lesions and less abnormal slices, while the worse performance on women might come from the smaller lungs and lesions.

Part of the patients in the database have multi-stage CTs. We compared the diagnostic performance of stage I and stage II and fusion of stage I, II in the external test cohort (Figure 3 a, b, d). The experiment suggested that the performance of the AI system is independent of the progress of the disease because of no significant differences between performances of different stages. The statistical results also showed that fusion of stage I and II could slightly improve the performance of diagnosis, in which the fusion method we adopted is to simply average the scores of two stages. We did not test more complex fusion methods which may overestimate the performance since each negative case has





c

		# Subjects	Sensitivity	# Abnormal Slices
Gender	**Male**	81	0.9656	24.35
	Female	102	0.9183	23.21
Age	**20-39**	50	0.8222	17.15
	40-59	68	0.9793	24.30
	>=60	64	0.9891	27.94
All		183	0.9391	23.72

d

Stage	AUC	Accuracy	Specificity	Sensitivity
Stage I	0.9548	0.9124	0.9043	0.9186
Stage II	0.9682	0.9170	0.9042	0.9268
Fusion of Stage I & II	0.9693	0.9308	0.9042	0.9512

Figure 3: Statistics on different subsets of positive subjects in the external test cohort. a. ROC curves of the diagnosis system on different subsets. b. Normalized histograms of the number of abnormal slices in different subsets. c. Subject numbers, sensitivities (with prediction threshold of 0.5), and average numbers of abnormal slices of different subsets. d. Diagnosis performances for each of the two stages and their fusion.

only one CT.

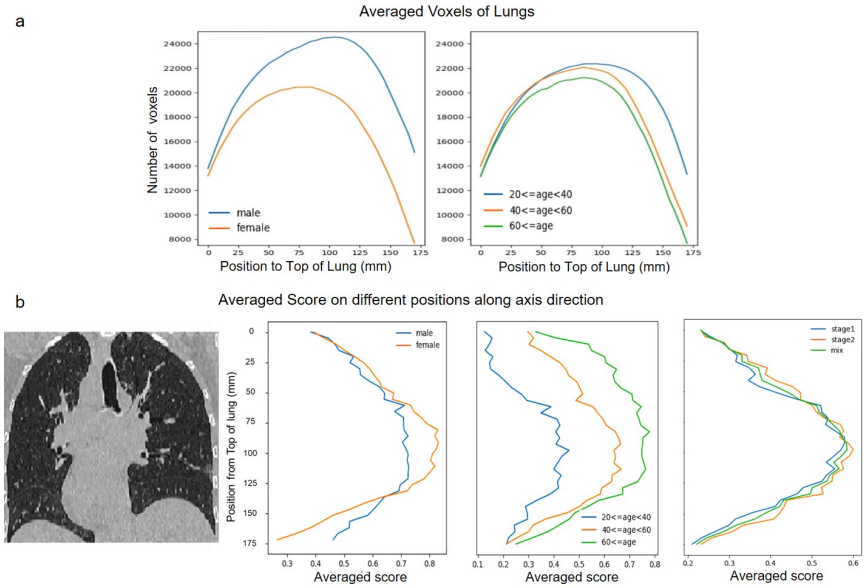


Figure 4: Statics on subset analysis. a. Averaged number of voxels of lung along axis direction. b. Averaged score in different position along axis direction for different subsets.

2.6 Interpreting the AI System

After proper training of the deep network, Guided gradient-weighted Class Activation Mapping (Guided Grad-CAM)[18] was exploited to explain the "black box" system and extract attentional areas which is connected to the back end of the diagnostic model. Figure 5 shows some representative cases for the visualization of Guided Grad-CAM to determine the attentional regions. The original CT slices are in the first column. The second column is the result of pseudo-color display of the feature map. The third column is the gradient map in the region of attention. We found that the spatial distribution of the attentional region, morphology and the texture within it are consistent with the characteristics of COVID-19 as reported in previous manual diagnosis studies[4,19].

In order to further verify our conjecture, we performed radiomics[20] feature extraction on these attentional regions, and obtained a total of 752-dimensional imaging features. The Least Absolute Shrinkage and Selection Algorithm (LASSO) were used to find the most discriminative features (Figure 6). A total of 15 features were finally chosen whose absolute values of coefficients are higher than 10-3. The cluster-heatmap and heat map of correlation coefficient matrix before and after LASSO show that LASSO filtered out redundant and unimportant

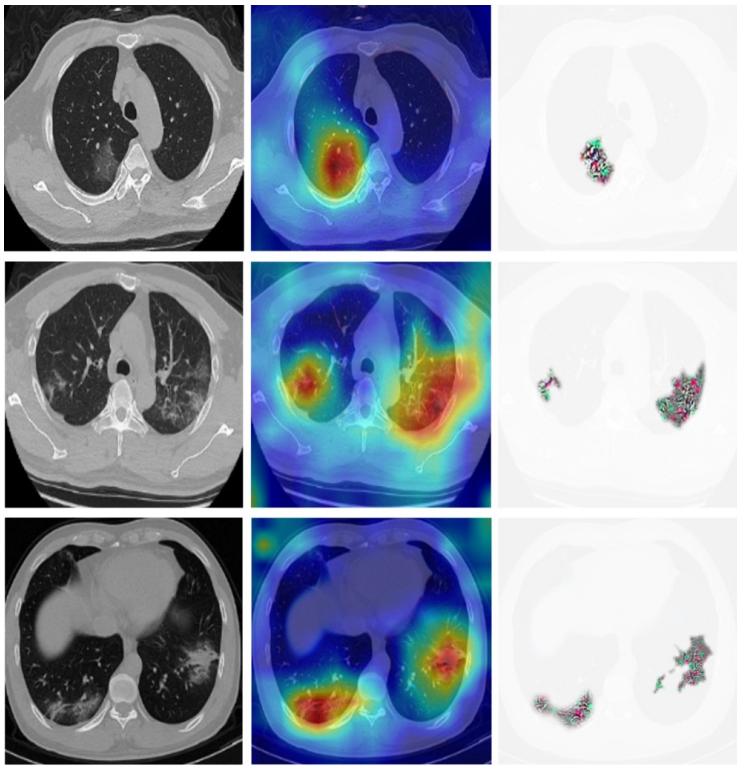


Figure 5: Visualization of Guided Grad-CAM. Notes: Three representative cases for the visualization of AI diagnosis. From left to right: Original CT image; Coarse-resolution attentional regions overlaid on CT image; High-resolution attentional regions with fine granularity.

features and the remaining features have better identification ability and lower correlation (Figure 9, Figure 10). The selected 15 features were used to explain the imaging characteristics in CT (Feature Analysis in Method). The extracted features show more separable statistical distribution between lesion and normal regions (Figure 6 b). We extracted three additional features for the attentional regions, distance feature, 2-D margin fractal dimension, and 3-D grayscale mesh fractal dimension (Figure 8). According to previous literature[21] on the pathogenesis and morphology of COVID-19, we believe that there may be a statistical rule in the pathogenesis (see Feature Analysis section in Methods).

3 Discussion

In this study, we developed an AI system for diagnosis of COVID-19. The system showed good sensitivity (94.06%), specificity (95.47%) and AUC (97.91%) in external test cohort. Furthermore, in the reader study, the diagnostic accura-

c

Feature names	P-value
original_glrlm_RunLengthNonUniformity	<0.001
original_glszm_LargeAreaEmphasis	<0.001
log-sigma-1-0-mm-3D_glszm_LargeAreaLowGrayLevelEmphasis	0.096694578
log-sigma-1-0-mm-3D_gldm_LargeDependenceHighGrayLevelEmphasis	<0.001
log-sigma-3-0-mm-3D_gldm_LargeDependenceHighGrayLevelEmphasis	<0.001
wavelet-LH_firstorder_Variance	0.001969931
wavelet-HL_glcm_ClusterProminence	<0.001
wavelet-HL_ngtdm_Complexity	<0.001
wavelet-HH_glrlm_RunLengthNonUniformity	0.584389978
wavelet-HH_gldm_DependenceNonUniformity	0.093045225
wavelet-LL_firstorder_Maximum	<0.001
wavelet-LL_glrlm_GrayLevelNonUniformity	<0.001
wavelet-LL_glrlm_LongRunHighGrayLevelEmphasis	<0.001
wavelet-LL_glrlm_ShortRunHighGrayLevelEmphasis	<0.001
wavelet-LL_ngtdm_Complexity	<0.001
score	<0.001

Figure 6: Coefficients and Distribution of LASSO selected features. a. Coefficients of the 15 selected features. b. Feature distribution for 2 classes. The solid lines mean averaged value and dotted lines mean value with one standard deviation. Value has been normalized to Normal distribution for easier visualization and comparison. c. T-test results of all 15 features and the weighted sum score in distinguishing abnormal and normal lungs. Notes: The score was computed by summing up 15 features weighted with their coefficients.

cy of the AI system was comparable to that of experienced radiologists from the outbreak center, who achieved higher sensitivity (94.70%), specificity (91.39%) and AUC (98.05%). Among the five professional readers in the radiology department, only one was able to produce a higher diagnostic accuracy than the AI system.

This automatic, high-precision, non-invasive diagnostic system was developed to provide clinicians with easy-to-use tools. Given the chest CT of a suspected patient as input, the AI system can automatically output the diagnosis result. In the reader study, the average reading time of radiologists was 6.5 min, while that of AI system was 2.73 s, which can significantly improve the productivity of radiologists. Meanwhile, we found that 71% (15/21) of errors made by radiologists could be corrected by AI system. It means that AI system can be used as an effective secondary reader to provide reference suggestions when the radiologist is not sure about the case or when multiple radiologists are inconsistent. In general, AI can be adapted to different requirements. According to the highly sensitive settings, it can screen out suspicious patients for confirmation by doctors; In accordance with the highly specific settings, it can warn possible diagnosis errors made by the doctor; or an optimal threshold value is chosen according to the prior probability of infectious diseases and the local prevention and control strategy.

To further understand the performance of the AI system, we evaluated it on subsets divided by gender, age and number of CT scans. In the subsets divided by gender, the diagnostic performance of men was higher than that of women. We noticed an obvious difference in the size of lungs and lesions between men and women. This is consistent with the conclusion of Xiong et al.[22] that women have higher antiviral immunity than men, leading to a longer incubation period of COVID-19. In the subsets divided by age, the disease developed rapidly in the elderly patients and the CT findings were more obvious than young patients. This suggests that different thresholds should be adopted according to gender and age. The results on the subsets divided by the number of CT scans showed that diagnosis results by the AI system have little correlation with the stages of CT scans. These observations may provide decision-making references for the diagnosis of COVID-19 in different populations.

Further, we provided a visual explanation of the system's decision by performing a radiomics analysis to obtain diagnostically relevant phenotypic characteristics of the attentional regions that are fully traceable on the original CT image. This is important for an in-depth study of pulmonary imaging findings in patients with COVID-19. For the AI system, by visualizing the diagnostic results of 200 subjects from the reader study cohort and comparing them with human reader in the reader study, together with the subsequent radiomics analysis, we were able to perform detailed imaging phenotype analysis on the diagnosis of COVID-19, and subsequently make pathophysiological and anatomical speculations on the viral infection process (see Feature Analysis in Methods).

There are still some drawbacks and future works of this research. First, collecting more data on other types of viral pneumonias or lung lesions can help improve its specificity further. Second, based on many chest CTs with

detailed labelled lesions, a semantic segmentation algorithm can be trained to locate the outline of the lesion more accurately than Guided Grad-GAM, and distinguish the detailed category of the lesion. Overall, the proposed AI system has been comprehensively validated on large dataset with diagnosis performance comparable to human experts in diagnosing COVID-19. Unlike classical black-box deep learning approaches, by visualizing AI system and applying radiomics analysis, it can decode effective representation of COVID-19 on CT imaging, and potentially lead to the discovery of new biomarkers. Radiologists could perform an individualized diagnosis of COVID-19 with the AI system, adding new driving force for fighting the global spread of outbreak.

4 Methods

4.1 Development and Validation Datasets

Under insitutional review board (IRB) approval , data used in our experiments come from three centers in Wuhan, which are Wuhan Union hospital, Western Campus of Wuhan Union Hospital, and Jianghan Mobile Cabin Hospital, and two public databases, LIDC-IDRI of the American National Cancer Institute (NCI), and ILD-HUG data of University Hospitals of Geneva.

The datasets from the three centers in Wuhan contain both positive (confirmed COVID-19) and negative cases. Database collected from Jianghan Mobile Cabin Hospital in Wuhan includes chest CT volumes of patients with confirmed COVID-19 from February 5th, 2020 to February 29th, 2020. Chest CT volumes from Wuhan Union hospital, Western Campus of Wuhan Union Hospital are collected from January 11th, 2020 to February 29th, 2020. Data from three centers contain 756 subjects, of which 496 were positives and 260 negatives. Some positive subjects have multi-stage CTs, so that datasets from three centers consist of 710 CT volumes of positive cases and 260 CT volumes of negative cases. Stage I means the first chest CT of a patient, and the intervals to stages II are 4-10 days.

Two public databases contain only negative cases since these data were collected before COVID-19 outbreak. The Lung Image Database Consortium Image collection (LIDC-IDRI) is a collaboration between seven academic centers and eight medical imaging companies initiated by the national cancer institute (NCI) in the United States. This database contains 1,012 subjects[16]. ILD-HUG is an international public dataset on interstitial lung diseases (ILDs) built at the University Hospitals of Geneva (HUG). The dataset contains 113 chest CTs (reported to have 128 but only 113 available) with three-dimensional annotated regions of pathological lung lesions[17]. Most of the cases taken from these two public datasets are abnormal cases and some of which are similar in CTs as COVID-19.

All CT data was divided into four independent parts with no overlapping subjects:

- Training cohort: 312 subjects were assigned to the training cohort, includ-

ing 147 positive and 75 negative cases from three centers in Wuhan, and 75 cases from LIDC-IDRI and 15 cases from ILD-HUG). This cohort was used to train parameters of model.

- Internal validation cohort: 104 subjects were assigned to the internal validation cohort, including 49 positive and 25 negative cases from three centers in Wuhan, and 25 cases from LIDC-IDRI and 5 cases from ILD-HUG. This cohort was used to validate the performances and turn super-parameters of model.

- External test cohort: There were 183 positive and 116 negative cases from three centers in Wuhan, 873 cases from LIDC-IDRI and 83 cases from ILD-HUG. To test performances for different genders and ages, we used this cohort excluded data of public databases. To test the performances of different stage, we used 123 patients with 2 stages and all the negative cases collected in Wuhan in this cohort. This cohort was used to evaluate and analyze performances of AI system.

- Reader study cohort: 200 subjects (117 positive and 44 negative cases from three centers in Wuhan, and 39 cases from LIDC-IDRI database) were assigned to the reader study cohort. This cohort was used to compare diagnosis results with human radiologists.

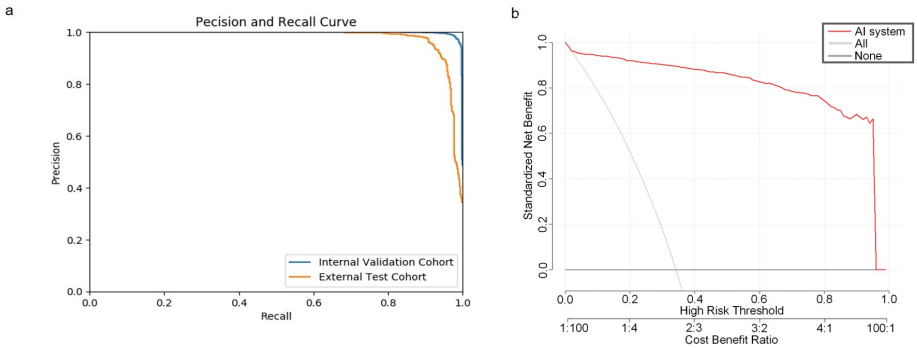
4.2 Overview of Our AI System

The proposed AI system takes as input a whole CT volume and outputs COVID-19 diagnosis along with abnormal slices (if diagnosed as positive). The whole system consists of five parts: lung segmentation block, COVID-19 classification network, abnormal slice locating block for COVID-19 positives, AI system interpreting block using Guided Grad-CAM, and feature analysis block. The first four blocks are deep-learning based blocks and the last one is traditional feature extraction block.

4.3 Development and Training of Deep Learning Blocks

The lung segmentation block is implemented based on Deeplab v1[23], which is a 2D semantic segmentation network. All CTs are in 3D, so we trained and tested the segmentation model slice by slice. The training slices were extracted from chest CTs in the training cohort and annotations of lung segmentation were obtained manually. The segmentation results were used as masks to determined lung areas, and they were concatenated to the raw CT slices as a different channel before feeding into the next block. We used this input-with-mask method to improve diagnosis results which has better performance according to experiments.

Our COVID-19 diagnosis block is a 2D classification deep network whose backbone is ResNet152[24], a deep network with 152 convolutional, pooling or fully-connected layers. The parameters of ResNet152 are pretrained on a



c

	LIDC-IDRI			ILD-HUG		
	TN	FP	Error rate	TN	FP	Error rate
AI system	846	27	3.12%	82	11	11.85%

Figure 7: Extend Performances of proposed AI system. a. PR curves was employed to assess the AI system performance of COVID-19 diagnosis. b. Decision curve analyses (DCA) of AI system in independent external test set for COVID-19 diagnosis. c. Errors count in external public databases in external test cohort.

huge dataset ImageNet7 for better and faster convergence. We tested a 3D classification network but this 2D scheme showed much better performance. The input of classification model is lung-masked slices, which means the input slices including training, internal validation and external test cohort, have been segment by segmentation model to get lung masks. The outputs of classification model are two scores respectively representing confidence levels of being normal and COVID-19 affected. Loss function of this block is cross entropy. The block was trained using 2D slices with batch size 32 for 100 epochs which costed about 5 hours under learning rate 10-5. Slices for training this block were extracted from training cohort, and the extraction process is detailed explained in supplementary methods. Since the predictions are based on 2D slices, an extra step is done to get a volume-level prediction. Because one volume is COVID-19 positive when any one of its slices is COVID-19 positive, we averaged the top 3 highest scores of all slices of a volume as the volume score. As a result, though training and validation were done on slice level, the block can take the whole CT volumes (with the whole lung segmentation volumes) and output a single prediction on volume level.

To measure the performance of classification model, AUC and some other metrics are computed on both internal validation and external test cohort. The metrics of internal validation cohort is computed on slice-level because the training is on slice-level and turning of super-parameters will be easier if validation is also done on slice-level. While metrics for external test cohort and reader study cohort are on case level, consistent with clinical application.

Block to locate abnormal slices is in the same structure of diagnosis block but trained especially on COVID-19 positive cases whose lesions have been marked manually. We used 28 cases with slice-level annotations in training cohort to train the block and the rest 12 cases in internal validation cohort with annotations to test performances of locating.

We used Guided Grad-CAM to obtain attentional regions as our system interpreting block. Guided Grad-CAM has the advantage that it not only generates a heat map to locate the relevant area, but also produces a coarse localization map highlighting the important regions in the image for predicting the result. Generally, the features used for classification judgment can be found, such as edges and spots in specific areas. Guided Grad-CAM is important because the areas it focused on are a secondary output of our system together with diagnosis result, giving more detailed diagnosis suggestions. Also, the attentional regions were used in latter feature extraction and analysis to get more detailed information about lesion areas. We extracted region of attention by binarizing output of Grad-CAM and then some morphological operations were done on binarization map.

All the deep learning blocks were implemented using PyTorch[25].

4.4 Features Extraction

Features were extracted in the attentional region determined by Guided Grad-CAM. We also extracted the same feature in normal lung in controlled cases for comparison. Due to no valid lesions attentional region for controlled cases is computed by Guided Grad-CAM, we used the shape of attentional region of COVID-19 cases and randomly choose positions within lung area as the attentional regions of controlled cases. We did not use shape features because the shape of attentional regions between COVID-19 and controlled cases are the same.

We extracted radiomics features which are widely used in lesion diagnosis these years. These features are composed of different image transforms and feature matrix calculations. We adopted three image transforms: original image, transformed image by Laplacian of Gaussian (LoG) operator, and transformed image by wavelet. For each image after the operation of a transform, six series of features are extracted, including first order features, Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), Neighboring Gray Tone Difference Matrix (NGTDM), Gray Level Dependence Matrix (GLDM). Radiomics analysis was performed using python version 3.6 and the "pyradiomics" package[26]. We designed three other features which are distance feature and fractal features of 2D contour and 3D gray level mesh of attentional region. The distance feature was defined as the distance between the center of gravity of the region of interest (obtained by the classification network after Grad-GAM) and the edge of the lung (obtained by the edge of the lung automatically segmentation results). Besides, 2D contour fractal dimension and 3D grayscale mesh fractal dimension of the attentional region was extracted. The fractal dimension describes the de-

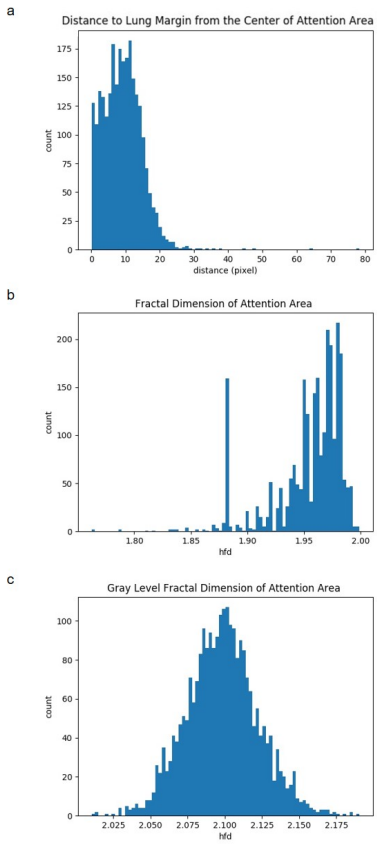


Figure 8: Distribution of Features. a. Distances from the center of attention area to lungs margin. b. Margin Fractal Dimension of Attention Area. c. Gray Level Mesh Fractal Dimension of Attention Area.

gree of curvature of the curve and surface. These three extra features were only extracted from the CT images of COVID-19 patients and were not analyzed and compared on the controlled cases. LASSO logistic regression model, heat map of cluster and correlation coefficient matrix were used to extract, select and verify the radiological features of the attentional region in the original CT images, which can interpret AI system. LASSO analysis was performed using python version 3.6 and the "scikit-learn" package.

4.5 Features Analysis

First, we located the distribution of the attentional region traced by the AI system, which mainly consisted of the subpleural distribution, the fragmentary distribution of patchy based on the secondary lobules, and the diffuse distribution of the fusing above two. The distances feature shows that the centers

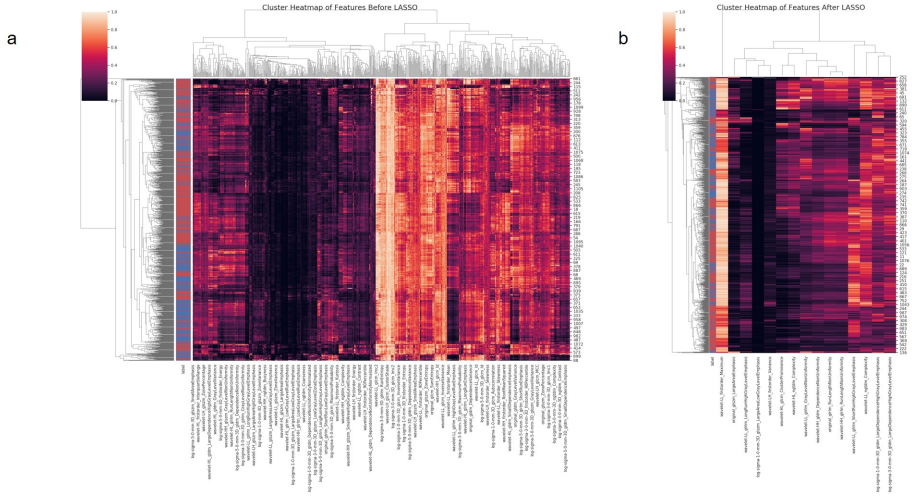


Figure 9: Cluster Heatmap of features. a. 752 radiomics features before LASSO. b. Cluster heatmap of 15 radiomics features after LASSO. Note: In label column, red means COVID-19 positives while blue means negatives.

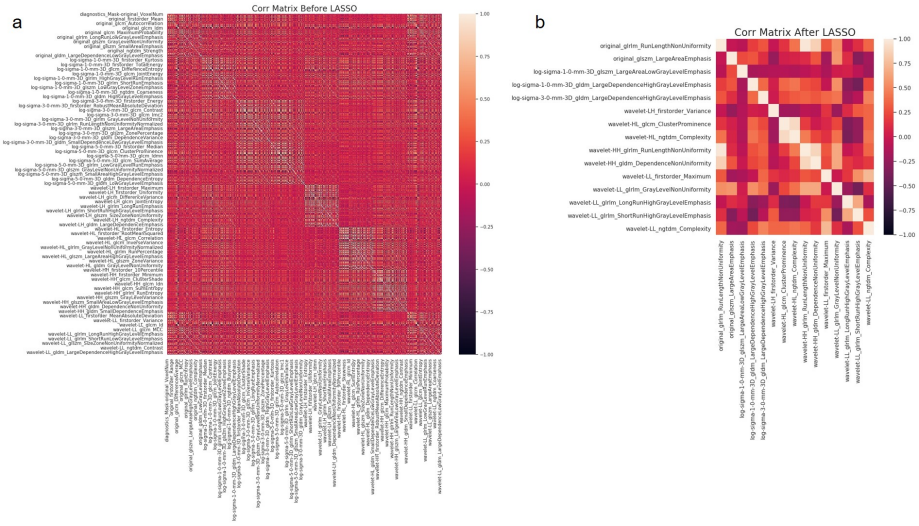


Figure 10: Correlation heat map. a. Before LASSO. b. After LASSO.

of attentional region are generally 0-20 pixels (2.5 mm/pixel) from the pleura (Figure 8 a), which is consistent with anatomical findings on COVID-19. The SARS-Cov2 is small (60-140 nm in diameter), and when inhaled through the airways, it mainly invades the deep bronchioles, causes inflammation of the bronchioles and their surroundings to damage alveolar[27,28]. Also, from the pathophysiology and etiology[29-31], the SARS-Cov2 mainly invades the bronchioles and lung parenchyma. Lesions tend to occur in the lower and peripheral areas of the lung with abundant capillaries and lymphatics. These areas have well-established immune system and well-developed pulmonary lobules, leading to a strong inflammatory response[32,33]. Secondly, fractal dimensions of most regional contour are between 1.95-2.00, shows a tortuous outline and conforms to the diffuse description, which corresponds to the fact that most of the cases in the dataset are at early or developmental stage. A few distributions are between 1.87-1.89, the overall contour of attentional regions are smoother or sub-parts are flat and narrow. The appearance of partial flatness at the boundary of the attentional region is a sign of consolidation, indicating the lesion is in the mend. Finally, we obtained some findings on the texture (homogeneity and heterogeneity) of the attentional regions that features for normal and lesion regions show different distributions (Figure 6 b, c).

According to all the extracted features, we can describe in depth the relationship between the medical findings and typical patterns. I) Halo pattern and anti-halo pattern were easily formed in the attentional regions. The halo pattern was speculated to be that the lesions (mainly the central node of the lobular) infiltrated into the surrounding interstitium and developed the aggregation of inflammatory cells in the interstitium. Anti-halo pattern is of the center of the ground glass shadow, almost completely surrounding by the high-density consolidation. The appearance of this sign may be that the inflammatory repair is dominated by the edge, leading to the formation of a band shadow tending to consolidation at the edge, while the central repair is relatively slow. II) The attentional region presents pleural parallel signs. The formation mechanism was speculated as follows: when the SARS-Cov2 invaded the interstitium around the alveoli, the lymphatic return direction was subpleural and interlobular septa, and diffused into pleural side and bilateral interlobular septum[34]. Because of the limitation of the pleura at the distal end, the lymph can only cling to the pleura and spread along the reticular structure of the interlobular septal margin on both sides. In addition, the fusion of the subpleural lesions results in the long axis of the lesions parallel to the pleura. III) Vascular thickening was found in the attentional region, which was consistent with the rules of inflammation production, inflammatory stimulation, increased vascular permeability, telangiectasia, further pulmonary artery thickening[33,35]. IV) The fine mesh feature of large area develops in the attentional region. The SARS-Cov2 mainly invades the interstitium in the lobules, so it appears as confluent fine mesh (crazy paving). V) The density of the ground-glass opacity (GGO) in the attentional regions increases, transforming to consolidation, the consolidation edges are flat or contracted, and fiber strands appear. These are all signs of disease outcome. In addition, when multiple regions of attention are recognized in the lungs of

some cases, the conditions of some regions will be improved, and that of other regions is lagging. The phenomenon of fluctuation is caused by unsynchronized development of lesions.

4.6 Reader Study

For this study, data from three centers was used, and negative data from public database is added, after which the reader study cohort statistically consisted of 58.5% positive cases and 41.5% negative cases. All the five readers are professional radiologists in the radiology department of Wuhan Union Hospital, who have rich clinical diagnosis experience and is in the center of the epidemic area with the most patients in this outbreak in China. They have all read over four-hundred CTs of COVID-19 in the past three months. Five radiologists had an average of 8 years of clinical experience in the imaging diagnosis of pulmonary diseases, as detailed in Table 2 a.

Readers can choose any window of gray value and zoom in or out when reading CT volumes using Slicer 4.10.2 software while our system used fixed size recased images (224 x 224 x 35) with fixed gray value window (-1200, 700) for all volumes.

Acknowledgements

We would like to acknowledge the radiologists participating the reader study. This study was supported by Zhejiang University special scientific research fund for COVID-19 prevention and control.

Reference

1 WHO. Laboratory testing for coronavirus disease 2019 (COVID-19) in suspected human cases: interim guidance, 2 March 2020. (World Health Organization, 2020).

2 Bai, H. X. et al. Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT. Radiology, 200823 (2020).

3 Ai, T. et al. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology, 200642, doi:10.1148/radiol.2020200642 (2020).

4 Shi, H. et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. Lancet Infect Dis, doi: 10.1016/S1473-3099(20)30086-4 (2020).

5 Deng, J. et al. Imagenet: A large-scale hierarchical image database. in 2009 IEEE conference on computer vision and pattern recognition, 248-255 (2009).

6 LeCun, Y., Bengio, Y. and Hinton, G. J. n. Deep learning. Nature 521, 436-444 (2015).

7 Krizhevsky, A., Sutskever, I. and Hinton, G. E. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems, 1097-1105 (2012).

8 Ren, S., He, K., Girshick, R. and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. in Advances in neural information processing systems. 91-99 (2015).

9 Esteva, A. et al. Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. Nature 546, 686, doi:10.1038/nature22985 (2017).

10 Shen, D., Wu, G. and Suk, H. I. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng 19, 221-248, doi:10.1146/annurev-bioeng-071516-044442 (2017).

11 Esteva, A. et al. A guide to deep learning in healthcare. Nat Med 25, 24-29, doi:10.1038/s41591-018-0316-z (2019).

12 Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25, 44-56, doi:10.1038/s41591-018-0300-7 (2019).

13 Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 25, 954-961, doi:10.1038/s41591-019-0447-x (2019).

14 Wang, S. et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). medRxiv. (2020).

15 Chen, J. et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv. (2020).

16 Armato, S. G., 3rd et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38, 915-931, doi:10.1118/1.3528204 (2011).

17 Depeursinge, A. et al. Building a reference multimedia database for interstitial lung diseases. Comput Med Imaging Graph 36, 227-238, doi:10.1016/j.compmedimag.2011.07.003 (2012).

18 Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. in Proceedings of the IEEE international conference on computer vision. 618-626 (2017).

19 Kanne, J. P. Chest CT Findings in 2019 Novel Coronavirus (2019-nCoV) Infections from Wuhan, China: Key Points for the Radiologist. Radiology, 200241, doi:10.1148/radiol.2020200241 (2020).

20 van Griethuysen, J. J. M. et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 77, e104-e107, doi:10.1158/0008-5472.CAN-17-0339 (2017).

21 Bernheim, A. et al. Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of Infection. Radiology, 200463, doi:10.1148/radiol.2020200463 (2020).

22 Xiong, Q. et al. Women May Play a More Important Role in the Transmission of the Corona Virus Disease (COVID-19) than Men. Lancet (2020).

23 Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans Pattern Anal Mach Intell 40, 834-848, doi:10.1109/TPAMI.2017.2699184 (2018).

24 He, K., Zhang, X., Ren, S. and Sun, J. Deep residual learning for image recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778 (2016).

25 Adam, P. et al. An imperative style, high-performance deep learning library. in Proceedings of Neural Information Processing Systems. 8024-8025 (2019).

26 Van Griethuysen, J. J. et al. Computational radiomics system to decode the radiographic phenotype. Cancer research 77, e104-e107 (2017).

27 Cascella, M., Rajnik, M. and Cuomo, A. Features, Evaluation and Treatment Coronavirus (COVID-19) [Updated 2020 Mar 8], (2020).

28 Dail, D. H. and Hammar, S. P. Dail and Hammar's pulmonary pathology. (Springer Science and Business Media, 2013).

29 Nicholls, J. M. et al. Lung pathology of fatal severe acute respiratory syndrome. Lancet 361, 1773-1778 (2003).

30 Cui, J., Li, F. and Shi, Z.-L. J. N. r. M. Origin and evolution of pathogenic coronaviruses. Nature reviews Microbiology 17, 181-192 (2019).

31 Su, S. et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends in microbiology 24, 490-502 (2016).

32 Cotes, J. E., Chinn, D. J. and Miller, M. R. Lung function: physiology, measurement and application in medicine. (John Wiley and Sons, 2009).

33 Panagiotou, M., Church, A. C., Johnson, M. K. and Peacock, A. J. Pulmonary vascular and cardiac impairment in interstitial lung disease. Eur Respir Rev 26, doi:10.1183/16000617.0053-2016 (2017).

34 Breslin, J. W. et al. Lymphatic Vessel Network Structure and Physiology. Compr Physiol 9, 207-299, doi:10.1002/cphy.c180015 (2018).

35 Moldoveanu, B. et al. Inflammatory mechanisms in the lung. Journal of inflammation research 2, 1 (2009).