Radiology

Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT

*Harrison X. Bai, MD** • *Robin Wang, BA** • *Zeng Xiong, MD* • *Ben Hsieh, MS* • *Ken Chang, PhD* • *Kasey Halsey, BA* • *Thi My Linh Tran, BS* • *Ji Whae Choi, BA* • *Dong-Cui Wang, MD* • *Lin-Bo Shi, MD* • *Ji Mei, MD* • *Xiao-Long Jiang, MD* • *Ian Pan, MA* • *Qiu-Hua Zeng, MD* • *Ping-Feng Hu, MD* • *Yi-Hui Li, MD* • *Fei-Xian Fu, MD* • *Raymond Y. Huang, MD, PhD* • *Ronnie Sebro, MD* • *Qi-Zhi Yu, MD* • *Michael K. Atalay, MD, PhD* • *Wei-Hua Liao, MD, PhD*

From the Department of Radiology, Xiangya Hospital, Central South University, Changsha 410008, China (H.X.B., Z.X., D.C.W., W.H.L.); Department of Diagnostic Imaging, Rhode Island Hospital, Providence, RI (H.X.B., B.H., K.H., I.P., M.K.A.); Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pa (R.W.); Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology. Massachusetts General Hospital, Boston, Mass (K.C.); Warren Alpert Medical School at Brown University, Providence, RI (H.X.B., K.H., T.M.L.T., J.W.C., I.P.); Department of Radiology, Yongzhou Central Hospital, Yongzhou, China (L.B.S.); Department of Radiology, Changde Second People's Hospital, Changde, China (J.M.); Department of Radiology, Affiliated Nan Hua Hospital, University of South China, Hengyang, China (X.L.J.); Department of Radiology, Loudi Central Hospital, Loudi, China (Q.H.Z.); Department of Radiology, Chenzhou Second People's Hospital, Chenzhou, China (P.F.H.); Department of Radiology, Zhuzhou Central Hospital, Zhuzhou, China (Y.H.L.); Department of Radiology, Yiyang City Center Hospital, Yiyang, China (F.X.F.); Department of Radiology, Brigham and Women's Hospital, Boston, Mass (R.Y.H.); Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, Pa (R.S.); and Department of Radiology, The First Hospital of Changsha, Changsha, China (Q.Z.Y.). Received April 8, 2020; revision requested April 15; revision received April 19; accepted April 23. **Address correspondence to** W.H.L. (e-mail: *owenliao@csu.edu.cn*).

H.X.B. supported by the Brown COVID-19 Research Seed Award (GR300196), the Amazon Web Services Diagnostic Development Initiative, Research Scholar Grant by RSNA Research and Education Foundation, and the National Cancer Institute of the National Institutes of Health under Award Number R03CA249554. W.H.L. supported by National Natural Science Foundation of China (81671676, 91959117). K.C. supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (5T32EB1680) and the National Cancer Institute of the National Institutes of Health (F30CA239407).

Conflicts of interest are listed at the end of this article.

*H.X.B. and R.W. contributed equally to this work.

Radiology 2020; 296:E156–E165 • https://doi.org/10.1148/radiol.2020201491 • Content codes: CH IN

Background: Coronavirus disease 2019 (COVID-19) and pneumonia of other diseases share similar CT characteristics, which contributes to the challenges in differentiating them with high accuracy.

Purpose: To establish and evaluate an artificial intelligence (AI) system for differentiating COVID-19 and other pneumonia at chest CT and assessing radiologist performance without and with AI assistance.

Materials and Methods: A total of 521 patients with positive reverse transcription polymerase chain reaction results for COVID-19 and abnormal chest CT findings were retrospectively identified from 10 hospitals from January 2020 to April 2020. A total of 665 patients with non–COVID-19 pneumonia and definite evidence of pneumonia at chest CT were retrospectively selected from three hospitals between 2017 and 2019. To classify COVID-19 versus other pneumonia for each patient, abnormal CT slices were input into the EfficientNet B4 deep neural network architecture after lung segmentation, followed by a two-layer fully connected neural network to pool slices together. The final cohort of 1186 patients (132 583 CT slices) was divided into training, validation, and test sets in a 7:2:1 and equal ratio. Independent testing was performed by evaluating model performance in separate hospitals. Studies were blindly reviewed by six radiologists without and then with AI assistance.

Results: The final model achieved a test accuracy of 96% (95% confidence interval [CI]: 90%, 98%), a sensitivity of 95% (95% CI: 83%, 100%), and a specificity of 96% (95% CI: 88%, 99%) with area under the receiver operating characteristic curve of 0.95 and area under the precision-recall curve of 0.90. On independent testing, this model achieved an accuracy of 87% (95% CI: 82%, 90%), a sensitivity of 89% (95% CI: 81%, 94%), and a specificity of 86% (95% CI: 80%, 90%) with area under the receiver operating characteristic curve of 0.90 and area under the precision-recall curve of 0.87. Assisted by the probabilities of the model, the radiologists achieved a higher average test accuracy (90% vs 85%, $\Delta = 5$, $P < .001$), sensitivity (88% vs 79%, $\Delta = 9$, $P < .001$), and specificity (91% vs 88%, $\Delta = 3$, $P = .001$).

Conclusion: Artificial intelligence assistance improved radiologists' performance in distinguishing coronavirus disease 2019 pneumonia from non–coronavirus disease 2019 pneumonia at chest CT.

Online supplemental material is available for this article.

It has been hypothesized that coronavirus disease 2019 (COVID-19) infection is difficult to contain because of its potential transmission from asymptomatic carriers (1,2). Common symptoms include fever, cough, and dyspnea, although the disease has the potential to cause a host of severe and potentially fatal cardiorespiratory complications in vulnerable populations—particularly the elderly with comorbid conditions (3,4). Although distinguishing COVID-19 from normal lung or other lung diseases, such as cancer at chest CT, may be straightforward, a major hurdle in controlling the current pandemic is making out subtle radiologic differences between

Abbreviations

AI = artificial intelligence, CI = confidence interval, COVID-19 = coronavirus disease 2019, RIH = Rhode Island Hospital, RT-PCR = reverse transcription polymerase chain reaction

Summary

Artificial intelligence assistance improved radiologists' performance in distinguishing coronavirus disease 2019 from pneumonia of other origin at chest CT.

Key Results

- An artificial intelligence (AI) model had higher test accuracy (96% vs 85%, $P < .001$), sensitivity (95% vs 79%, $P < .001$), and specificity (96% vs 88%, $P = .002$) than radiologists.
- In an independent test set, the AI model achieved accuracy of 87%, sensitivity of 89%, and specificity of 86%.
- With AI assistance, the radiologists achieved higher average accuracy (90% vs 85%, $P < .001$), sensitivity (88% vs 79%, $P < .001$), and specificity (91% vs 88%, $P = .001$).

COVID-19 and pneumonia of other origins. For example, manual radiologist interpretation of chest CT is a specific modality for recognizing COVID-19 by its characteristic patterns that include peripheral ground-glass opacities, but unfortunately this measure often has low specificity in distinguishing COVID-19 from other pneumonia (5,6). The exception to this is by screening populations with high disease prevalence, such as in Wuhan, China, at the beginning of the outbreak and in Italy presently. In these cases, the sensitivity of chest CT for COVID-19 is high, whereas specificity is low because of an abundance of false-positive results (7,8).

Current literature has revealed that it is possible for artificial intelligence (AI) to distinguish COVID-19 from other pneumonia with good accuracy (9). However, published studies have limitations, such as small sample size, lack of external validation, no comparison with radiologist performance, and no standard of reference for the "other pneumonia" diagnosis (10–13).

To capture and properly manage all cases of COVID-19, it is essential to develop testing methods that accurately recognize the disease as distinct from other causes of pneumonia at chest CT.

The purpose of this study was to establish and evaluate an AI system that differentiates COVID-19 and other pneumonia at chest CT and assesses radiologist performance without and with AI assistance.

Materials and Methods

Patient Cohorts

The institutional review board of all nine hospitals in Hunan Providence, China, and the Rhode Island Hospital (RIH) in Providence and the Hospital of the University of Pennsylvania in Philadelphia in the United States approved this retrospective study, and the written informed consent requirement was waived. A total of 521 patients with confirmed positive COVID-19 reverse transcription polymerase chain reaction (RT-PCR) results and chest CT images were retrospectively identified from RIH and the nine hospitals in Hunan Providence, China, from January 6 to April 1, 2020. The RT-PCR results were extracted from the patients' electronic medical records in the hospital information system. The RT-PCR assays were performed by using TaqMan One-Step RT-PCR Kits from Shanghai Huirui Biotechnology (Shanghai, China) or Shanghai BioGerm Medical Biotechnology (Shanghai, China), both of which have been approved for use by the China Food and Drug Administration for the Chinese cohorts and the COVID-19 RT-PCR test (LabCorp, Burlington, NC) for U.S. cohorts. For patients with multiple RT-PCR assays, a positive result on the last performed test was adopted as a confirmation of diagnosis.

The radiology search engine MONTAGE (Nuance Communications, Burlington, Mass) at RIH and Hospital of the University of Pennsylvania was used to identify cases that contain the word *pneumonia* in the impression section of the radiology CT reports from January 1, 2017, to December 30, 2019. The impression sections of these CT reports were initially reviewed by a research assistant (B.H.) followed by verification by a board-certified radiologist (H.X.B., with 5 years of experience) in general diagnostic radiology and interventional radiology with 1 year of practice experience to identify cases with the final CT impression being consistent with or highly suspicious for pneumonia. Then, the images were further reviewed by a radiologist (H.X.B.) to ensure agreement with the original report. A Chinese radiology search engine was used to identify a similar non–COVID-19 pneumonia cohort from Xiangya Hospital in Hunan Provence, China, from 2017 to 2019, followed by verification by a radiologist (D.W., with 5 years of experience). The identified CT scans were directly downloaded from the hospital picture archiving and communications systems, and nonchest CT images were excluded.

Data on the respiratory pathogen were collected from respiratory pathogen panel results for the RIH cohort, as described in a previous study (5). The tests of ePlex Respiratory Pathogen panel (GenMark Diagnostics, Carlsbad, Calif) were performed in the Microbiology Laboratory of Rhode Island Hospital Pathology Department according to the manufacturer's protocol (14).

The final cohort consisted of 665 patients with non–COVID-19 pneumonia. A diagram illustrating patient inclusion and exclusion is shown in Figure 1. The number of cases included from each hospital is shown in Table E1 (online). The chest CT protocols from all 11 hospitals are shown in Table E2 (online). A total of 214 patients from the COVID-19 group and 202 patients from the non–COVID-19 pneumonia group (RIH cohort) overlapped with a previous study (5).

Lung Segmentation

To exclude nonpulmonary regions of the CT, the lungs were first segmented on the basis of attenuation, with −320 HU used as the threshold value. Manual editing of the lung segmentation was performed by radiologists using the manual active contour segmentation method with three-dimensional Slicer software (version 4.6; Brigham and Women's Hospital, Boston, Mass) when autosegmentation was insufficient.

Image Preprocessing

The whole data set was preprocessed by setting the CT window width and level to the lung window (1500 HU and −400 HU,

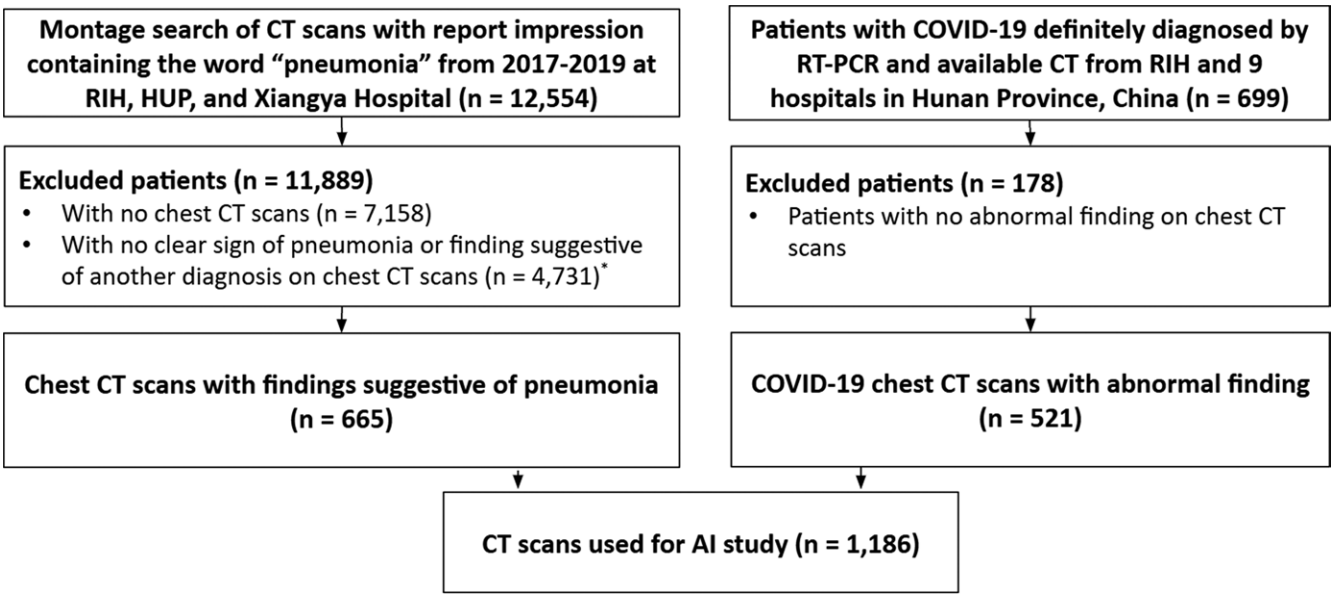

Figure 1: Diagram shows patient inclusion and exclusion criteria. AI = artificial intelligence, COVID-19 = coronavirus disease 2019, HUP = Hospital of the University of Pennsylvania, RIH = Rhode Island Hospital, RT-PCR = reverse transcriptase polymerase chain reaction.

respectively). The slices with lesions (COVID-19 or pneumonia) were manually labeled by radiologists (H.XB. D.W.) in consensus and were used as the reference standard for training the deep neural network to identify slices with abnormal lung findings. Images were padded, if necessary, to equal height and width and were rescaled to 224 × 224 pixels. Lung windowing was applied to the Hounsfield units to generate an 8-bit image for each individual two-dimensional axial slice in a CT scan. Images were preprocessed by first normalizing pixel values from the range (0–255) to 0–1 and then by standardizing use of the normalized ImageNet mean and standard deviation.

Development of the Deep Learning Model

A classification model was trained to distinguish between slices with and those without pneumonia-like findings (both COVID-19 and non–COVID-19). The EfficientNet architecture (15), which consists of mobile inverted bottleneck MBConv blocks (16), was used for the classification task. It possessed a smaller number of model parameters and improved the accuracy and efficiency over those of the existing convolutional networks. Pretrained on ImageNet, an EfficientNet-B3 convolutional neural network with a single fully connected two-class classification layer was used. Dropout with probability of 0.5 was applied to the fully connected layer. Data augmentation was performed dynamically during training and included flips, scaling, rotations, random brightness and contrast manipulations, random noise, and blurring. Training was performed for 20 epochs, where each epoch was defined as 16 000 slices. The AdamW optimizer was used with default parameters. A one-cycle policy was used for the learning rate schedule, with an initial learning rate of 4.0×10^{-6} to a maximum of 1.0×10^{-4}. Validation was performed on a separate validation set every two epochs. The area under the curve was used to track model performance, and the checkpoint with the highest validation area under the curve was selected as the final model. The choice of compound scaling metrics was made empirically based on validation set performance. Specifically, a larger network was used when it resulted in high performance on the validation set. If increasing the network size did not result in higher performance on the validation set, then a smaller network was used to maintain computational efficiency.

Pneumonia Classification

The EfficientNet B4 architecture was used for the pneumonia classification task. Each slice was stacked to three channels as the input of EfficientNet to use the pretrained weights on ImageNet. EfficientNets with dense top fully connected layers were used. The configuration of dense top layers was as follows: four fully connected layers of 256, 128, 64, and 32 neurons, respectively, combined with 0.5 dropout using rectified linear unit activations with batch normalization layers replacing the top fully connected layers of EfficientNet. A fully connected layer with 16 neurons with batch normalization and a classification layer (one neuron) with sigmoid activation were at the end of EfficientNet to make predictions of COVID-19 versus non–COVID-19 pneumonia slices. Then, the slices were pooled using a two-layer fully connected neural network to make predictions at the patient level. Stochastic gradient descent optimizer with a 0.0001 learning rate was used. Batch size was set to 64. Figures 2 and 3 show our deep learning workflow. A heat map for important image regions that lead our model to classify a case as COVID-19 or non–COVID-19 was generated using gradient-weighted class activation mapping (17).

Radiologist Interpretation

Six radiologists with 10 years (F.F.X), 10 years (S.L.), 20 years (Y.X.), 20 years (X.Z.), 20 years (H.Z.), and 10 years (S.Y.) of chest CT experience reviewed the test set consisting of 119

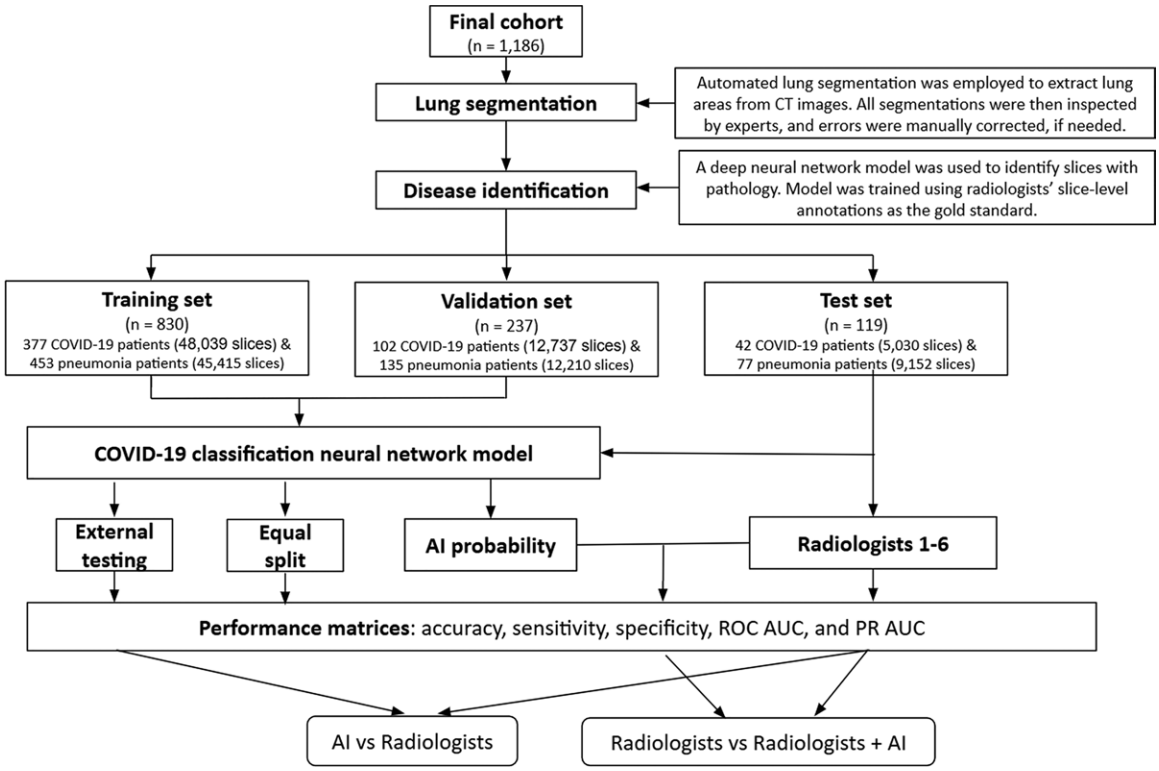

Figure 2: Flowchart shows the artificial intelligence (AI) model used to distinguish coronavirus disease 2019 (COVID-19) from non–COVID-19 pneumonia. PR AUC = precision recall area under curve, ROC AUC = receiver operator characteristics area under the curve.

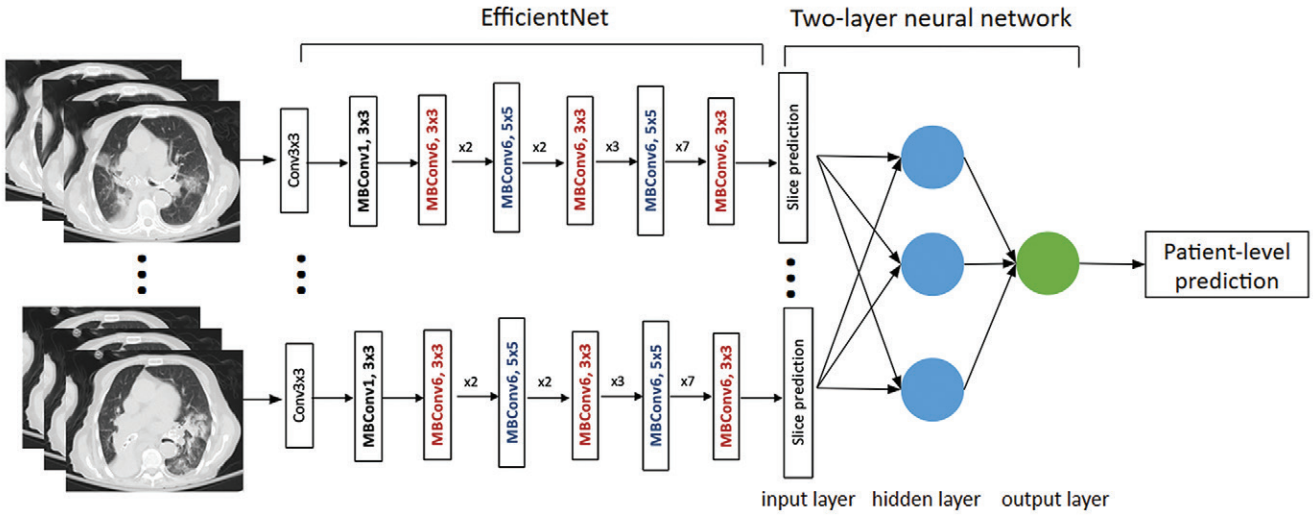

Figure 3: Image shows coronavirus disease 2019 classification neural network model.

chest CT images and scored each case as COVID-19 or pneumonia of other causes. All identifying information was removed from the CT studies, which were shuffled and uploaded to the three-dimensional slicer for interpretation. All radiologists were given information on patient age when reviewing images. All radiologists then reviewed the test set again, with prediction from the AI. The studies were shuffled between the two evaluation sessions. The two sessions were separated by at least 1 day. The radiologists were not given feedback on their performance after the first session.

Statistical Analysis

Demographic and clinical-pathologic characteristics were compared between COVID-19 and non–COVID-19 pneumonia groups by means of the χ^2 test for categorical variables and the Student *t* test for continuous variables. CT slice thickness was compared between the COVID-19 and non–COVID-19 pneumonia groups using the Mann-Whitney *U* test and among training, validation, and test sets using the Kruskal-Wallis H test. Accuracy, sensitivity, specificity, area under the receiver operating characteristic curve, and area under the precision

Table 1: Clinical Characteristics of COVID-19 and Non–COVID-19 Pneumonia Patient Cohorts

Characteristic	COVID-19 (n = 521)	Non–COVID-19 (n = 665)	P Value
Age (y)			<.001
Mean*	46 ± 16 (4–84)	62 ± 19 (0–99)	…
<20	11 (2)	12 (2)	…
20–39	151 (29)	76 (11)	…
40–59	222 (43)	166 (25)	…
≥60	136 (26)	411 (62)	…
Sex			.03
Male	268 (51)	385 (58)	…
Female	251 (48)	280 (42)	…
Presence of fever			<.001
Fever	303 (58)	361 (54)	…
No fever	147 (28)	280 (42)	…
White blood cell count			<.001
Elevated	12 (2)	337 (51)	…
Normal	441 (85)	325 (49)	…
Lymphocyte count			<.001
Normal	303 (58)	293 (44)	…
Decreased	186 (36)	365 (55)	…
Comorbidities			
Cardiovascular disease	18 (3)	230 (35)	<.001
Hypertension	65 (12)	258 (39)	<.001
Chronic obstructive pulmonary disease	22 (4)	157 (24)	<.001
Diabetes	29 (6)	116 (17)	<.001
Chronic liver disease	11 (2)	17 (3)	.62
Chronic kidney disease	6 (1)	70 (11)	<.001
Malignant tumor	2 (0)	84 (13)	<.001
Human immunodeficiency virus	0 (0)	15 (2)	<.001
Time from onset to presentation (d)*	11 ± 12 (0–54)	…	…
<10	313	…	…
10–19	37	…	…
20–29	62	…	…
≥30	65	…	…
Epidemiologic contact			
Wuhan	169 (32)	…	…
COVID-19	115 (22)	…	…
Severity			
Mild	34 (7)	…	…
Medium	405 (78)	…	…
Severe	53 (10)	…	…
Critical	24 (5)	…	…

Note.—Unless otherwise indicated, data are numbers of patients, with percentages in parentheses. Severity is defined by the Chinese Clinical Guidance for COVID-19 Pneumonia Diagnosis and Treatment (seventh edition), published by the China National Health Commission on March 4, 2020. Formal diagnosis was done using reverse transcription polymerase chain reaction for coronavirus disease 2019 (COVID-19) cases. Source.—Reference 24.

* Data are mean ± standard deviation, and data in parentheses are the range.

recall curve were calculated for the classification model. The 95% confidence intervals (CIs) for accuracy, sensitivity, and specificity were determined with the adjusted Wald method (18). Model performance was compared with average radiologist performance. Radiologist performance without AI assistance was compared with radiologist performance with AI assistance. The P values were calculated with the permutation method. All analyses were performed with the use of R statistical computing language (R, version 3.4.2; *https://www.r-project.org*).

Code Availability

The implementation of the deep learning models was based on the Keras package (version 2.2.5) with the TensorFlow library

Table 2: AI Model and Six Radiologists with AI Assistance Test Set Results

Radiologist No. and Statistic	Radiologist Performance (%)	AI Performance (%)	AI Performance Minus Radiologist Performance (%)	*P* Value
1				
Accuracy	92 (86, 96)	96 (90, 98)	3 (−2, 8)	.34
Sensitivity	100 (90, 100)	95 (83, 100)	−5 (−15, 0)	.50
Specificity	88 (79, 94)	96 (88, 99)	8 (1, 15)	.07
2				
Accuracy	81 (72, 87)	96 (90, 98)	15 (8, 23)	<.001
Sensitivity	86 (71, 94)	95 (83, 100)	10 (0, 21)	.22
Specificity	78 (67, 86)	96 (88, 99)	18 (9, 28)	.001
3				
Accuracy	80 (71, 86)	96 (90, 98)	16 (9, 23)	<.001
Sensitivity	88 (74, 95)	95 (83, 100)	7 (−3, 18)	.36
Specificity	75 (64, 84)	96 (88, 99)	21 (12, 30)	<.001
4				
Accuracy	82 (74, 88)	96 (90, 98)	13 (6, 21)	.002
Sensitivity	64 (49, 77)	95 (83, 100)	31 (15, 47)	.001
Specificity	92 (83, 97)	96 (88, 99)	4 (−4, 12)	.51
5				
Accuracy	90 (83, 94)	96 (90, 98)	6 (0, 13)	.12
Sensitivity	81 (66, 90)	95 (83, 100)	14 (2, 27)	.07
Specificity	95 (87, 98)	96 (88, 99)	1 (−5, 8)	>.99
6				
Accuracy	82 (74, 88)	96 (90, 98)	13 (6, 21)	.001
Sensitivity	55 (40, 69)	95 (83, 100)	40 (23, 58)	<.001
Specificity	97 (90, 100)	96 (88, 99)	−1 (−6, 3)	>.99
Radiologist Average				
Accuracy	85 (77, 90)	96 (90, 98)	11 (7, 16)	<.001
Sensitivity	79 (64, 89)	95 (83, 100)	16 (8, 24)	.001
Specificity	88 (78, 94)	96 (88, 99)	8 (3, 14)	.002

Note.—Data in parentheses are 95% confidence intervals. Table shows the results of an artificial intelligence (AI) model and six radiologists without AI assistance on the test set (n = 119) in differentiating between coronavirus disease 2019 (COVID-19) pneumonia and non–COVD-19 pneumonia.

(version 1.12.3) on the back end. The models were trained on a computer with two NVIDIA V100 graphics processing units. To allow other researchers to develop their models, the code is publicly available on Github at *https://github.com/robinwang08/COVID19*.

Results

Patient Characteristics

Our final cohort consisted of 1186 patients, of whom 521 were patients with COVID-19 and 665 were patients with non–COVID-19 pneumonia. The average age of patients with COVID-19 was lower than that of patients with non–COVID-19 pneumonia (48 years vs 62 years, $P < .001$). Patients with COVID-19 were less likely to have an elevated white blood cell count (2% vs 51%, $P < .001$) or a reduced lymphocyte count (36% vs 55%, $P < .001$) than were patients with non–COVID-19 illness. The clinical characteristics of the patients with COVID-19 pneumonia and those

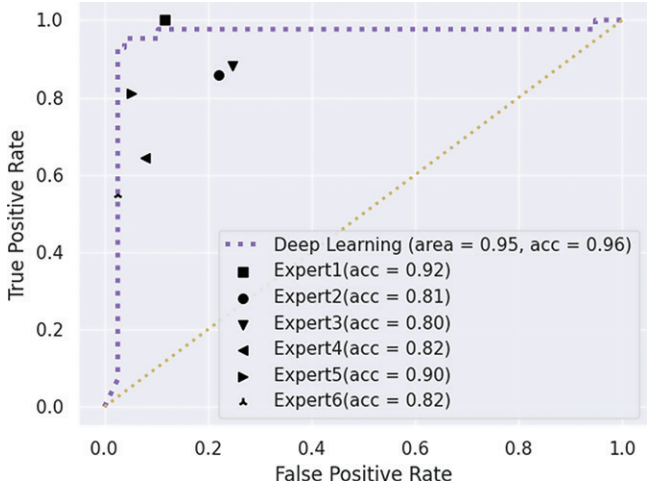

Figure 4: Image shows receiver operating characteristic (blue) of deep neural network on the test set compared with radiologist performance. acc = accuracy.

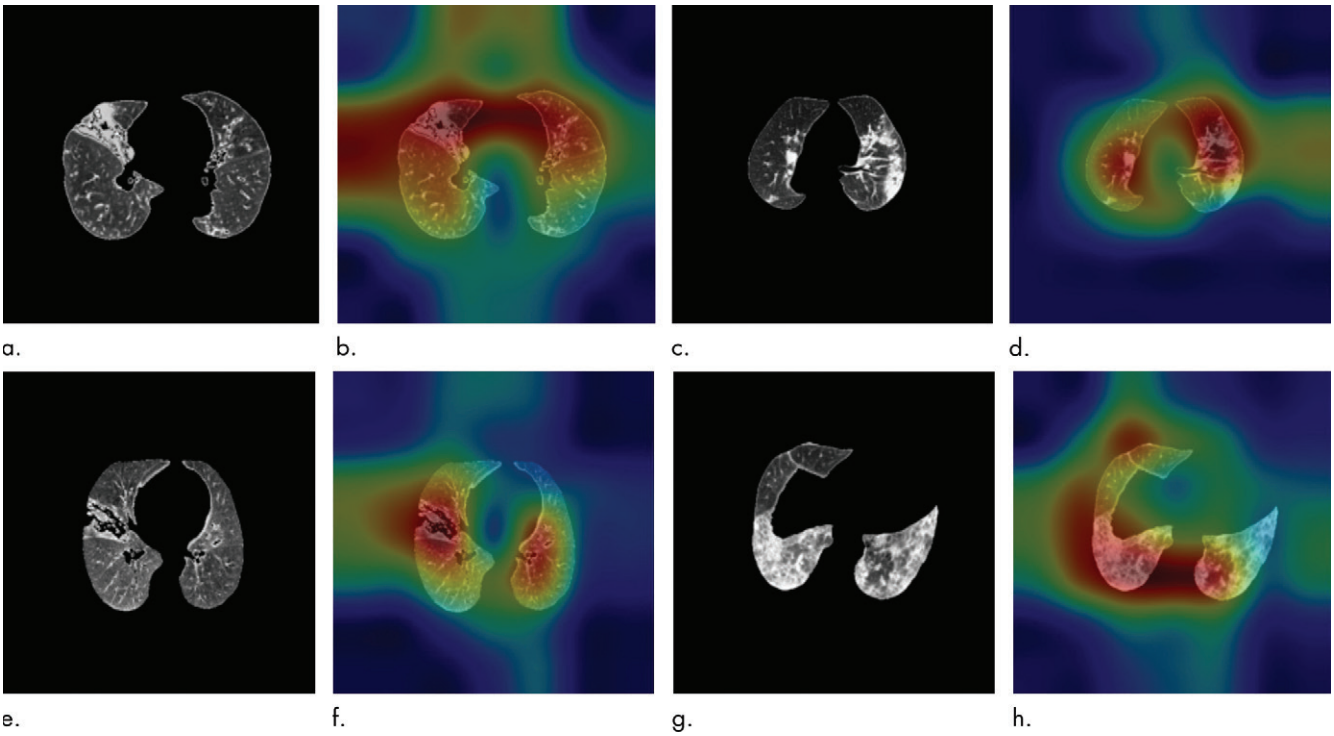

Figure 5: **(a–h)** Images show representative slices corresponding to gradient-weighted class activation mapping images on the test set.

with non–COVID-19 pneumonia, including comorbidities, are shown in Table 1. A diagram illustrating the breakdown of the viral pathogen species for the RIH cohort is shown in Figure E1 (online). The average time between chest CT and COVID-19 diagnosis was 2.8 days ± 4.0 (standard deviation). A diagram showing the breakdown of days between symptom onset and CT scanning is shown in Figure E2 (online). There was no significant difference in median CT slice thickness for COVID-19 and non–COVID-19 cases (COVID-19, 1.25 mm; non–COVID-19, 1.00 mm; P = .869). Table E3 (online) shows CT slice thickness among training, validation, and test sets for the different splits.

Slice Identification

The classifier to distinguish between slices with and those without pneumonia-like findings (both COVID-19 and non–COVID-19) achieved a final test area under the curve of 0.83. A naive classification threshold of 0.5 was used to binarize predictions. Additional metrics included average mean precision (0.675), F1 score (0.675), and positive predictive value (0.795).

Pneumonia Classification

The CT images of the 1186 patients (132 583 slices) were divided into training, validation, and test sets in a 7:2:1 ratio (ie, 830, 237, and 119 patients, respectively). The number of patients and slices in training, validation, and testing sets is shown in Table E4 (online). Our final model achieved a test accuracy of 96% (95% CI: 90%, 98%), a sensitivity of 95% (95% CI: 83%, 100%), and a specificity of 96% (95% CI: 88%, 99%) with an area under receiver operating characteristic curve of 0.95 and an area under the precision recall curve of 0.90. Compared with results from an average radiologist, our model had higher test accuracy (96% vs 85%, $P < .001$), sensitivity (95% vs 79%, $P < .001$), and specificity (96% vs 88%, $P = .002$) (Table 2). The receiver operating characteristic curve comparing model with radiologist performance is shown in Figure 4. A model trained on random equal split of training, validation, and test sets (ie, 396, 395, and 395 patients, respectively) achieved a test accuracy of 91% (95% CI: 87%, 93%), a sensitivity of 94% (95% CI: 90%, 97%), and a specificity of 87% (95% CI: 82%, 91%), with an area under the receiver operating characteristic curve of 0.95 and an area under the precision recall curve of 0.92. The number of patients and slices in training, validation, and testing sets is shown in Table E4 (online). Model performance on training, validation, and test sets for the different splits is shown in Table E5 (online).

Independent testing was performed by leaving out cohorts from one U.S. hospital (Hospital of the University of Pennsylvania) and three Chinese hospitals (Yongzhou Central Hospital, Zhuzhou Central Hospital, and Yiyang No. 4 Hospital). Our model achieved a test accuracy of 87% (95% CI: 82%, 90%), a sensitivity of 89% (95% CI: 81%, 94%), and a specificity of 86% (95% CI: 80%, 90%), with an area under receiver operating characteristic curve of 0.90 and an area under the precision recall curve of 0.87.

Gradient-weighted class activation mapping on representative CT slices from the test set shows that the model focused on the area of abnormality (Fig 5). Figure 6 shows five cases (Fig 6a–6e) in the test set where the deep learning model was correct but most of the radiologists were incorrect (at least four of six), as well as one case (Fig 6f) where both AI and most of the radiologists were incorrect. The three COVID-19 cases (Fig

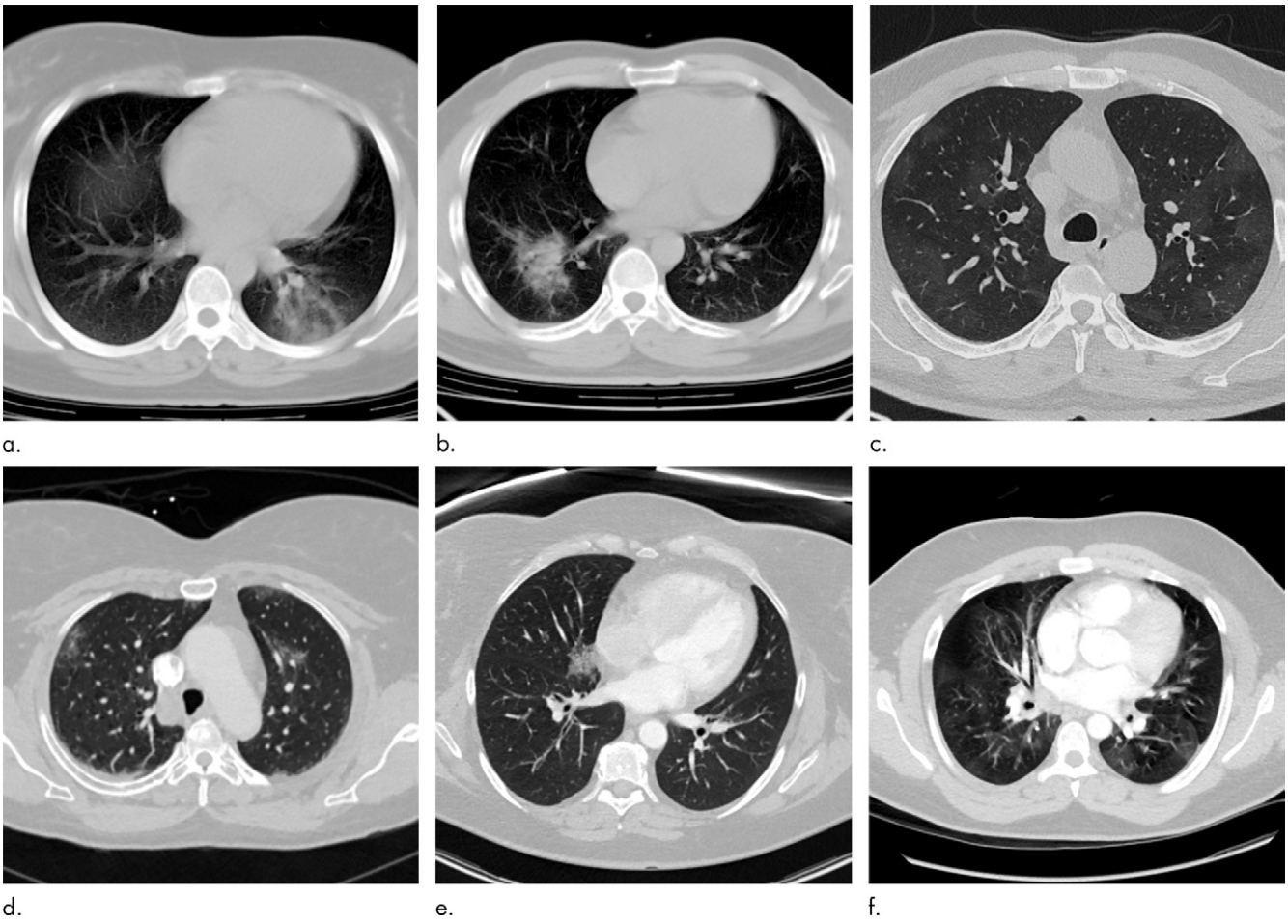

Figure 6: Images show representative cases that the majority of radiologists misclassified. **(a–c)** Coronavirus disease 2019 (COVID-19) pneumonia. Our model correctly classified all three cases. **(a)** Four of six radiologists (radiologists 3–6) said it was non–COVID-19. With artificial intelligence (AI) assistance, two of six radiologists (radiologists 5 and 6) continued to say it was non–COVID-19. **(b)** Four of six radiologists (radiologists 3–6) said it was non–COVID-19. With AI assistance, three of six radiologists (radiologists 3–5) continued to say it was non–COVID-19. **(c)** Four of six radiologists (radiologists 2 and 4–6) said it was non–COVID-19. With AI assistance, one of six radiologists (radiologist 2) continued to say it was non–COVID-19. **(d–f)** Non–COVID-19 pneumonia. Our model correctly classified **d** and **e**. **(d)** Five of six radiologists (radiologists 1–5) said it was COVID-19. With AI assistance, the same five radiologists continued to say it was COVID-19. **(e)** Four of six radiologists (radiologists 1, 2, 4, and 6) said it was COVID-19. With AI assistance, three of six radiologists (radiologists 1, 2, and 4) continued to say it was COVID-19. **(f)** Four of six radiologists (radiologists 1–3 and 6) said it was COVID-19. With AI assistance, five of six radiologists (radiologists 1–4 and 6) said it was COVID-19.

6a–6c) demonstrate atypical findings (eg, focal abnormality) that could have been mistaken for non–COVID-19 pneumonia, whereas the three non–COVID-19 pneumonia cases (Fig 6d–6f) demonstrate ground-glass opacities that mimic COVID-19 cases.

Radiologist Performance

For blind review on the test set without AI prediction, six radiologists had average accuracy of 85% (95% CI: 77%, 90%), average sensitivity of 79% (95% CI: 64%,89%), and average specificity of 88% (95% CI: 78%, 94%).

Assisted by the probabilities of the model, the radiologists achieved a higher average accuracy (90% vs 85%, $\Delta = 5$, $P < .001$), sensitivity (88% vs 79%, $\Delta = 9$, $P < .001$), and specificity (91% vs 88%, $\Delta = 3$, $P = .001$). Table 3 summarizes the comparison of radiologist performances without and with AI assistance.

Discussion

Coronavirus disease 2019 (COVID-19) can be difficult to distinguish from other types of pneumonia at chest CT. It has been revealed that the standard diagnostic test, real-time reverse transcriptase polymerase chain reaction (RT-PCR), frequently produces false-negative findings or fluctuating results that make it difficult to diagnose and contain active COVID-19 infections with confidence (19). Therefore, chest CT is often relied on as a supplementary diagnostic measure that helps physicians to build a more complete patient assessment. Artificial intelligence (AI) has shown efficacy in differentiating COVID-19 from pneumonia of other origin at chest CT, yet the practical application of AI augmentation to radiologists' COVID-19 diagnostic workflow has not been explored in the literature (9). Our study revealed that when compared with a radiologist-only approach, AI augmentation significantly improved radiologists' performance in distinguishing COVID-19 from

Table 3: Comparison of Results from Six Radiologists with and without AI assistance

Radiologist No. and Statistic	Without AI Assistance (%)	With AI Assistance (%)	Radiologist with AI Assistance Minus Radiologist Without AI Assistance (%)	*P* Value
1				
Accuracy	92 (86–96)	92 (86–96)	0 (−3 to 3)	>.99
Sensitivity	100 (90–100)	98 (86–100)	−2 (−8 to 0)	>.99
Specificity	88 (79–94)	90 (80–95)	1 (−3 to 6)	>.99
2				
Accuracy	81 (72–87)	89 (82–94)	8 (4–13)	.002
Sensitivity	86 (71–94)	86 (71–94)	0 (0–0)	>.99
Specificity	78 (67–86)	91 (82–96)	13 (6–21)	.002
3				
Accuracy	80 (71–86)	83 (75–89)	3 (1–7)	.13
Sensitivity	88 (74–95)	93 (80–98)	5 (0–12)	.49
Specificity	75 (64–84)	78 (67–86)	3 (0–7)	.50
4				
Accuracy	82 (74–88)	90 (83–94)	8 (3–13)	.01
Sensitivity	64 (49–77)	83 (69–92)	19 (8–32)	.01
Specificity	92 (83–97)	94 (85–98)	1 (−3 to 6)	>.99
5				
Accuracy	90 (83–94)	93 (87–97)	3 (1–7)	.12
Sensitivity	81 (66–90)	88 (74–95)	7 (0–16)	.25
Specificity	95 (87–98)	96 (88–99)	1 (0–4)	>.99
6				
Accuracy	82 (74–88)	92 (86–96)	10 (5–16)	<.001
Sensitivity	55 (40–69)	81 (66–90)	26 (13–40)	.001
Specificity	97 (90–100)	99 (92–100)	1 (0–4)	>.99
Radiologist average				
Accuracy	85 (77–90)	90 (83–94)	5 (4–7)	<.001
Sensitivity	79 (64–89)	88 (74–95)	9 (5–13)	<.001
Specificity	88 (78–94)	91 (82–96)	3 (2–5)	.001

Note.—Data in parentheses are 95% confidence intervals. Comparison of six radiologists without and with assistance of an artificial intelligence (AI) model in differentiating between coronavirus disease 2019 (COVID-19) pneumonia and non–COVID-19 pneumonia

pneumonia of other origin, yielding higher measures of accuracy, sensitivity, and specificity.

The diagnostic accuracy produced by manual interpretation of COVID-19 chest CT scans is good but needs to be improved to make resource allocation and disease management during the current pandemic less strenuous on health care systems and economies worldwide. Current clinical algorithms for the management of patients with COVID-19 are contingent on the amount of resources available and require definitive imaging results (20). Although distinguishing COVID-19 from healthy lung or from other lung diseases, such as cancer, on chest CT scans may be straightforward, differentiation between COVID-19 and other pneumonia can be particularly troublesome for physicians because of the radiographic similarities (21). Inaccurate imaging interpretation makes it harder for disease management strategies to work efficiently.

Our study is relevant and novel for demonstrating the effect of AI augmentation on radiologist performance in distinguishing COVID-19 from pneumonia of other cause on chest CT scans. Our results suggest that integrating AI into radiologists' routine workflows has potential to improve diagnostic outcomes related to COVID-19. In addition, external validation was used in our study while other recent AI studies either lacked external validation completely or had poor outcomes associated with external validation. The slight decrease in performance on external validation is secondary to some lack of generalization, which is expected across institutions because of differences in patient population and image acquisition (22,23). This research makes progress on the practical use of AI in COVID-19 diagnosis, and a future study will explore the prospective use of AI in real time to assist physician diagnosis.

Our study had several limitations. First, there could be bias as a product of the radiologists in this study evaluating the same cases twice, first without and then with AI assistance. However, this limitation cannot be overcome without a prospective design. Second, our COVID-19 cohort was heterogenous in the distribution of time between symptom onset and CT. Although this reflected a spectrum of chest CT presentations that likely represented the real-world scenario, the most difficult distinction between COVID-19 and pneumonia of other origin remains during the early disease stage. The limited sample size of early stage COVID-19 CT prevented us from performing a subgroup

analysis focusing on this cohort. Third, the composition of other pneumonia cases is heterogeneous, and not all the patients in the non–COVID-19 pneumonia cohort underwent respiratory pathogen panel testing or had the test results available. For those without respiratory pathogen panel testing, the cases were selected by searching the impression section of the original report and further review of the images by a second radiologist, which could have introduced selection bias. Furthermore, there is a possibility of pneumonia of other origin (eg, viral pneumonia by influenza) superimposed on COVID-19. Although we did our best to standardize CT images, a possibility remains for AI or radiologists to notice subtle differences between scans from different countries, institutions, or CT instruments. Fourth, there was significant difference in baseline characteristics between patients with COVID-19 and those with non–COVID-19 pneumonia that could have introduced bias. For example, patients in the non–COVID-19 pneumonia cohort were predominately from the United States and were significantly older with more comorbid conditions than those in our COVID-19 cohort, which by contrast mainly contained patients from China. Any of these factors could have complicated the appearance of non–COVID-19 chest CTs and influenced performance measures in our study. Lastly, although we had a multinational multi-institutional cohort, our model training could have benefited from a larger cohort size.

Artificial intelligence assistance improved radiologist performance in distinguishing coronavirus disease 2019 (COVID-19) from pneumonia of other cause on chest CT scans. A future study will investigate integration of these algorithms into routine clinical workflow to assist radiologists in accurately diagnosing COVID-19.

Author contributions: Guarantors of integrity of entire study, H.X.B., Z.X., T.M.L.T., J.M., X.L.J., F.X.F., Q.Z.Y., W.H.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, H.X.B., K.C., K.H., J.M., X.L.J., F.X.F., R.Y.H., Q.Z.Y., W.H.L.; clinical studies, H.X.B., Z.X., B.H., K.H., J.W.C., D.C.W., L.B.S., J.M., X.L.J., Q.H.Z., P.F.H., Y.H.L., F.X.F., R.Y.H., Q.Z.Y., M.K.A., W.H.L.; statistical analysis, H.X.B., R.W., B.H., K.C., T.M.L.T., J.M., X.L.J., I.P., F.X.F., R.Y.H., Q.Z.Y., W.H.L.; and manuscript editing, H.X.B., R.W., B.H., K.C., K.H., T.M.L.T., J.W.C., J.M., X.L.J., I.P., F.X.F., R.Y.H., R.S., Q.Z.Y., M.K.A., W.H.L.

Disclosures of Conflicts of Interest: **H.X.B.** disclosed no relevant relationships. **R.W.** disclosed no relevant relationships. **Z.X.** disclosed no relevant relationships. **B.H.** disclosed no relevant relationships. **K.C.** disclosed no relevant relationships. **K.H.** disclosed no relevant relationships. **T.M.L.T.** disclosed no relevant relationships. **J.W.C.** disclosed no relevant relationships. **D.C.W.** disclosed no relevant relationships. **L.B.S.** disclosed no relevant relationships. **J.M.** disclosed no relevant relationships. **X.L.J.** disclosed no relevant relationships. **I.P.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a consultant for MD.ai. Other relationships: disclosed no relevant relationships. **Q.H.Z.** disclosed no relevant relationships. **P.F.H.** disclosed no relevant relationships. **Y.H.L.** disclosed no relevant relationships. **F.X.F.** disclosed no relevant relationships. **R.Y.H.** disclosed no relevant relationships. **R.S.** disclosed no relevant relationships. **Q.Z.Y.** disclosed no relevant relationships. **M.K.A.** disclosed no relevant relationships. **W.H.L.** disclosed no relevant relationships.

References

1. Bai Y, Yao L, Wei T, et al. Presumed asymptomatic carrier transmission of COVID-19. JAMA 2020;323(14):1406–1407.
2. Qiu H, Wu J, Hong L, Luo Y, Song Q, Chen D. Clinical and epidemiological features of 36 children with coronavirus disease 2019 (COVID-19) in Zhejiang, China: an observational cohort study. Lancet Infect Dis 2020;20(6):P689–P696.
3. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395(10223):497–506 [Published correction appears in Lancet 2020;395(10223):496.].
4. Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA 2020;323(11):1061–1069.
5. Bai HX, Hsieh B, Xiong Z, et al. Performance of radiologists in differentiating COVID-19 from viral pneumonia at chest CT. Radiology 2020 Published online March 10, 2020. https://doi.org/10.1148/radiol.2020200823
6. Choi H, Qi X, Yoon SH, et al. Extension of coronavirus disease 2019 (COVID-19) on chest CT and implications for chest radiograph interpretation. Radiol Cardiothorac Imaging 2020;2(2):e200107.
7. Ai T, Yang Z, Hou H, et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology Published online February 26, 2020. https://doi.org/10.1148/radiol.2020200642
8. Caruso D, Zerunian M, Polici M, et al. Chest CT features of COVID-19 in Rome, Italy. Radiology Published online April 3, 2020. https://doi.org/10.1148/radiol.2020201237
9. Li L, Qin L, Xu Z, et al. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology Published online March 19, 2020. https://doi.org/10.1148/radiol.2020200905
10. Wang S, Kang B, Ma J, et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). medRxiv 2020.02.14.20023028 [preprint] https://www.medrxiv.org/content/10.1101/2020.02.14.20023028v5. Posted April 24, 2020.
11. Xu X, Jiang X, Ma C, et al. Deep learning system to screen coronavirus disease 2019 pneumonia. ArXiv 200209334 [preprint] https://arxiv.org/abs/2002.09334. Posted February 21, 2020.
12. Chen J, Wu L, Zhang J, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv 2020.02.25.20021568 [preprint] https://www.medrxiv.org/content/10.1101/2020.02.25.20021568v2. Posted March 1, 2020.
13. Shi W, Peng X, Liu T, et al. Deep learning-based quantitative computed tomography model in predicting the severity of COVID-19: a retrospective study in 196 patients. SSRN 3546089 [preprint] https://ssrn.com/abstract=3546089. Posted March 3, 2020.
14. GenMark Diagnostics I. ePlex Respiratory Pathogen Panel Assay Manual. GenMark Diagnostics I, editor. Carlsbad, CA: GenMark Diagnostics, Inc; November 2016.
15. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. arXiv 1905.11946 [preprint] https://arxiv.org/abs/1905.11946. Posted May 28, 2019.
16. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. arXiv1801.04381 [preprint] https://arxiv.org/abs/1801.04381. Posted January 13, 2018.
17. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, October 22–29, 2017. Piscataway, NJ: IEEE, 2017.
18. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 1998;52(2):119–126.
19. Li Y, Yao L, Li J, et al. Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19. J Med Virol 2020;92(7):903–908.
20. Rubin GD, Ryerson CJ, Haramati LB, et al. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society. Radiology 2020;296(1):172-180.
21. Zu ZY, Jiang MD, Xu PP, et al. Coronavirus disease 2019 (COVID-19): a perspective from China. Radiology Published online February 21, 2020. https://doi.org/10.1148/radiol.2020200490.
22. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018;15(11):e1002683.
23. AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing. Med Phys 2018;45(3):1150–1158.
24. Du Q. Clinical classification. Chinese clinical guidance for COVID-19 pneumonia diagnosis and treatment (7th edition). http://kjfy.meetingchina.org/msite/news/show/cn/3337.html. Published March 4, 2020. Updated March 16, 2020. Accessed April 8, 2020.

Radiology

Erratum

Originally published in:

Radiology 2020; 296:E156–E165

https://doi.org:/10.1148/radiol.2020201491

Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT

Harrison X. Bai, Robin Wang, Zeng Xiong, Ben Hsieh, Ken Chang, Kasey Halsey, Thi My Linh Tran, Ji Whae Choi, Dong-Cui Wang, Lin-Bo Shi, Ji Mei, Xiao-Long Jiang, Ian Pan, Qiu-Hua Zeng, Ping-Feng Hu, Yi-Hui Li, Fei-Xian Fu, Raymond Y. Huang, Ronnie Sebro, Qi-Zhi Yu, Michael K. Atalay, Wei-Hua Liao

Erratum in:

https://doi.org:/10.1148/radiol.2021219004

Harrison X. Bai and **Robin Wang** share equal contribution in this article and should have been listed as **Harrison X. Bai*, Robin Wang***, with the asterisks indicating equal contribution.