Radiology

Article Type: Original Research—Thoracic Imaging

AI Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Etiology on Chest CT

Harrison X. Bai[3,5]*,M.D.; Robin Wang[2]*, B.A.; Zeng Xiong[1], M.D.; Ben Hsieh[3], M.S.; Ken Chang[4], M.S.E.; Kasey Halsey[3,5], B.A.; Thi My Linh Tran[5], B.S.; Ji Whae Choi[5], B.S.; Dong-Cui Wang[1], M.D.; Lin-Bo Shi[6], M.D.; Ji Mei[7], M.D.; Xiao-Long Jiang[8], M.D.; Ian Pan[3,5], M.A.; Qiu-Hua Zeng[9], M.D.; Ping-Feng Hu[10], M.D.; Yi-Hui Li[11], M.D.; Fei-Xian Fu[12], M.D.; Raymond Y. Huang[13], M.D., Ph.D.; Ronnie Sebro[14], M.D.; Qi-Zhi Yu[15], M.D.; Michael K. Atalay[3], M.D.; Wei-Hua Liao[1], M.D.

1. Department of Radiology, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China
2. Perelman School of Medicine at University of Pennsylvania, Philadelphia, Pennsylvania 19104
3. Department of Diagnostic Imaging, Rhode Island Hospital, Providence, Rhode Island, 02903, United States
4. Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
5. Warren Alpert Medical School at Brown University, Providence, Rhode Island, 02903, United States
6. Department of Radiology, Yongzhou Central Hospital, Yongzhou, Hunan, 425006, China
7. Department of Radiology, Changde Second People's Hospital, Changde, Hunan, 415001, China
8. Department of Radiology, Affiliated Nan Hua Hospital, University of South China, Hengyang, Hunan, 421002, China
9. Department of Radiology, Loudi Central Hospital, Loudi, Hunan, 417000, China
10. Department of Radiology, Chenzhou Second People's Hospital, Chenzhou, Hunan, 423000, China
11. Department of Radiology, Zhuzhou Central Hospital, Zhuzhou, Hunan, 412002, China
12. Department of Radiology, Yiyang City Center Hospital, Yiyang, Hunan, 413000, China
13. Department of Radiology, Brigham and Women's Hospital, Boston, Massachusetts, 02115, United States
14. Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania, 19104, United States
15. Department of Radiology, The First Hospital of Changsha, Changsha, Hunan, 410005, China

Correspondence:
Wei-Hua Liao
Department of Radiology, Xiangya Hospital, Central South University, Changsha, Hunan, China 410008
Email: liaoweihua2017@163.com

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

Summary: AI assistance improved radiologists’ performance in distinguishing COVID-19 from pneumonia of other etiology on chest CT.

Key Results:

- An AI model had higher test accuracy (96% vs 85%, $p<0.001$), sensitivity (95% vs 79%, $p<0.001$), and specificity (96% vs 88%, $p=0.002$) than radiologists.
- In an independent test set, our AI model achieved an accuracy of 87%, sensitivity of 89% and specificity of 86%.
- With AI assistance, the radiologists achieved a higher average accuracy (90% vs 85%, $p<0.001$), sensitivity (88% vs 79%, $p<0.001$) and specificity (91% vs 88%, $p=0.001$).

Abbreviations:

COVID-19: Coronavirus Disease 2019

AI: Artificial Intelligence

ROC: Receiver Operating Characteristic

PR: Precision-Recall

RT-PCR: Reverse Transcription Polymerase Chain Reaction

RPP: Respiratory Pathogen Panel

RIH: Rhode Island Hospital

HUP: Hospital of the University of Pennsylvania

GGO: Ground-Glass Opacities

Grad-CAM: Gradient-weighted Class Activation Mapping

Abstract

Background: COVID-19 and pneumonia of other etiology share similar CT characteristics, contributing to the challenges in differentiating them with high accuracy.

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

Methods: 521 patients with positive RT-PCR for COVID-19 and abnormal chest CT findings were retrospectively identified from ten hospitals from January 2020 to April 2020. 665 patients with non-COVID-19 pneumonia and definite evidence of pneumonia on 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 two-layer fully-connected neural network to pool slices together. Our final cohort of 1,186 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 on separate hospitals. Studies were blindly reviewed by six radiologists without and then with AI assistance.

Results: Our final model achieved a test accuracy of 96% (95% CI: 90-98%), sensitivity 95% (95% CI: 83-100%) and specificity of 96% (95% CI: 88-99%) with Receiver Operating Characteristic (ROC) AUC of 0.95 and Precision-Recall (PR) AUC of 0.90. On independent testing, our model achieved an accuracy of 87% (95% CI: 82-90%), sensitivity of 89% (95% CI: 81-94%) and specificity of 86% (95% CI: 80-90%) with ROC AUC of 0.90 and PR AUC of 0.87. Assisted by the models' probabilities, the radiologists achieved a higher average test accuracy (90% vs. 85%, Δ=5, p<0.001), sensitivity (88% vs. 79%, Δ=9, p<0.001) and specificity (91% vs. 88%, Δ=3, p=0.001).

Conclusion: AI assistance improved radiologists' performance in distinguishing COVID-19 from non-COVID-19 pneumonia on chest CT.

Introduction

It has been hypothesized that COVID-19 infection is difficult to contain because of its potential transmission from asymptomatic carriers (1, 2). Common symptoms include fever, cough, and dyspnea while the disease has potential to cause a host of severe and potentially fatal cardiorespiratory complications in vulnerable populations—particularly the elderly with comorbid conditions (3, 4). While distinguishing COVID-19 from normal lung or other lung pathologies such as cancer on chest CT may be straightforward, a major hurdle in controlling the current pandemic is making out subtle radiological differences between COVID-19 and pneumonia of other etiology. For example, manual radiologist interpretation of chest CT is a specific modality for recognizing COVID-19 by its characteristic patterns including peripheral ground-glass opacities (GGO), but unfortunately this measure often has low specificity in distinguishing COVID-19 from other pneumonia (5, 6). The exception to this is in screening populations with high disease prevalence, such as in Wuhan at the beginning of the outbreak and in Italy presently. In these cases, the sensitivity of chest CT for COVID-19 is high while specificity is low due to an abundance of false positives (7, 8).

Current literature has revealed that it is possible for artificial intelligence (AI) to identify 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, no gold standard for the “other pneumonia” diagnosis, etc. (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 on chest CT.

The purpose of this study was to establish and evaluate an AI system in differentiating COVID-19 and other pneumonia on chest CT and assess radiologist performance without and with AI assistance.

Material and Methods

Patient Cohorts

The institutional review board of all nine hospitals in Hunan Providence, China and Rhode Island Hospital (RIH) in Providence, RI and Hospital of the University of Pennsylvania (HUP) in Pennsylvania, PA from the U.S. approved this retrospective study, and written informed consent was waived. A total of 521 patients with confirmed positive COVID-19 by RT-PCR and chest CT imaging were retrospectively identified from RIH and 9 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 Co., Ltd or Shanghai BioGerm Medical Biotechnology Co., Ltd, both of which have approved their use by the China Food and Drug Administration for the Chinese cohorts and the COVID-19 RT-PCR test (Laboratory Corporation of America) for US cohorts. For patients with multiple RT-PCR assays, positive result on the last performed test was adopted as a confirmation of diagnosis.

The radiology search engine MONTAGE (Nuance Communications, Burlington, MA) at RIH and HUP 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 (BH) followed by verification by a radiologist (HXB) board-certified in general diagnostic radiology and interventional radiology with one year of practice experience to identify cases with final CT impression being "consistent with" or "highly suspicious for" pneumonia. Then, the images were further reviewed by a radiologist (HXB) 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 Providence, China from 2017 to 2019, followed by verification by a radiologist (DW). The identified CT scans were directly downloaded from the hospital Picture Archiving and Communications Systems (PACS), and non-chest CTs were excluded.

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

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

Lung Segmentation

To exclude non-pulmonary regions of the CT, the lungs were first segmented on the basis of Hounsfield Unit (HU) with -320 HU used as the thresholding value. Manual editing of the lung segmentation was performed by radiologists using the manual active contour segmentation method with 3D Slicer software (v4.6) when auto-segmentation was insufficient.

Image Preprocessing

The whole dataset was preprocessed by setting the CT window width and level to the lung window (WL: -400; WW: 1500). The slices with lesions (COVID-19 or pneumonia) were manually labeled by radiologists (HXB and DW) in consensus and used as gold standard for training the deep neural network for identifying slices with abnormal lung findings. Images were padded, if necessary, to equal height and width and rescaled to 224x224 pixels. Lung windowing was applied to the Hounsfield units to generate an 8-bit image for each individual 2D axial slice in a CT scan. Images were preprocessed by first normalizing pixel values from the range [0, 255] to [0, 1], then standardizing using the normalized ImageNet mean and standard deviation.

Development of the deep learning model

A classification model was trained to distinguish between slices with and without pneumonia-like findings (both COVID and non-COVID). The EfficientNet architecture (15), which consists of mobile inverted bottleneck MBConv blocks (16) was employed for the classification task. It possessed a smaller number of model parameters and improved the accuracy and efficiency over the existing convolutional networks. Pretrained on ImageNet, an EfficientNet-B3 convolutional neural network with a single fully connected 2-class classification layer was used. Dropout with probability 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 x 10^-6 to a max of 1.0 x 10^-4. Validation was carried out on a separate validation set every 2 epochs. The area under the curve (AUC) was used to track model performance, and the checkpoint with the highest validation AUC 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, a smaller network was used to maintain computational efficiency.

Pneumonia Classification

The EfficientNet B4 architecture was employed for the pneumonia classification task. Each slice was stacked to three channels as the input of EfficientNet to use the pre-trained 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, 32) combined with 0.5 dropout using rectified linear unit (ReLU) 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 vs. non-COVID pneumonia slices. Then, the slices were pooled using a 2 layer fully-connected neural network to make prediction at the patient level. Stochastic gradient descent optimizer with 0.0001 learning rate was utilized. Batch size was a set to 64. Figures 2-3 illustrate our deep learning workflow. A heatmap 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 (Grad-CAM) (17).

Radiologist Interpretation

Six radiologists with 10, 10, 20, 20 (XZ), 20, and 10 years of chest CT experience reviewed the test set consisting of 119 chest CT images and scored each case as COVID-19 or pneumonia of other etiology. All identifying information was removed from the CT studies, which were shuffled and uploaded to 3D 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 one day. The radiologists were not given feedback on their performance after the first session.

Statistical Analysis

Demographic and clinicopathologic characteristics were compared between COVID-19 and non-COVID-19 pneumonia groups by means of chi-square test for categorical variables and 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 Receiver Operating Characteristic curve (ROC AUC), and area under Precision-Recall curve (PR AUC) were calculated for classification model. 95% confidence intervals on accuracy, sensitivity, and specificity were determined using the adjusted Wald method (18). Model performance was compared to average radiologist performance. Radiologist performance without AI assistance was compared with that with AI assistance. The p-values were calculated using the permutation method. All analyses were performed with the use of the R statistical computing language (R version 3.4.2, The R Foundation for Statistical Computing, Vienna, Austria; http://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 (version 1.12.3) on the backend. The models were trained on a computer with two NVIDIA V100 GPUs. To allow other researchers to develop their models, the code is publicly available on Github at http://github.com/robinwang08/COVID19.

Results

Patient Characteristics

Our final cohort consisted of 1,186 patients, of which 521 were patients with COVID-19 and 665 were patients with non-COVID pneumonia. The average age of patients with COVID-19 was lower than that of patients with non-COVID-19 pneumonia (48 vs. 62 years, p<0.001). Patients with COVID-19 were less likely to have an elevated white blood cell count than patients with non-COVID-19 illness (2% vs. 51%, p<0.001) or reduced lymphocyte (36% vs. 55%, p<0.001). The clinical characteristics of the COVID-19 and non-COVID-19 pneumonia patients including co-morbidities are shown in Table 1. A diagram illustrating the breakdown of the viral pathogen species for the RIH cohort is shown in Supplementary Figure E1. The average number of days between chest CT and COVID-19 diagnosis was 2.8±4.0. A diagram illustrating the breakdown of days between symptom onset and CT scan is shown in Supplementary Figure E2. There was no significant difference in median CT slice thickness COVID-19 and non-COVID-19 cases (COVID-19: 1.25 mm, non-COVID-19: 1.00 mm, p=0.869). Supplementary Table E3 shows CT slice thickness among training, validation and test sets for the different splits.

Slice Identification

The classifier to distinguish between slices with and without pneumonia-like findings (both COVID and non-COVID) achieved a final test AUC of 0.83. A naïve 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 1,186 patients (132,583 slices) were divided into training, validation, and test sets in a 7:2:1 ratio (i.e., 830, 237 and 119 patients, respectively). The number of patients and slices in training, validation and testing sets is shown in Supplementary Table E4. Our final model achieved a test accuracy of 96% (95% CI: 90-98%), sensitivity of 95% (95% CI: 83-100%) and specificity of 96% (95% CI: 88-99%) with ROC AUC of 0.95 and PR AUC of 0.90. Compared to average radiologist, our model had higher test accuracy (96% vs. 85%, p<0.001), sensitivity (95% vs. 79%, p<0.001), and specificity (96% vs. 88%, p=0.002) (Table 2). The ROC curve comparing model with radiologist performance is shown in Figure 4. A model trained on random equal split of training, validation and test sets (i.e., 396, 395 and 395 patients, respectively) achieved a test accuracy of 91% (95% CI: 87-93%), sensitivity of 94% (95% CI: 90-97%) and specificity of 87% (95% CI: 82-91%) with ROC AUC of 0.95 and PR AUC of 0.92. The number of patients and slices in training, validation and testing sets is shown in Supplementary Table E4. Model performance on training, validation and test sets for the different splits is shown in Supplementary Table E5.

Independent testing was performed by leaving out cohorts from one US hospital (HUP) 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%), sensitivity of 89% (95% CI: 81-94%) and specificity of 86% (95% CI: 80-90%) with ROC AUC of 0.90 and PR AUC of 0.87.

Grad-CAM on representative CT slices from test set demonstrates that the model focused on the area of abnormality (Figure 5). Figure 6 shows five cases (Figure 6a-e) in the test set where the deep learning model was correct but most of the radiologists were incorrect (at least 4/6) as well as one case (Figure 6f) where both AI and most of the radiologists were incorrect. The three COVID-19 cases (Figure 6a-c) demonstrate atypical findings (e.g., focal abnormality) that could have been mistaken for non-COVID-19 pneumonia, while the three non-COVID-19 pneumonia cases (Figure 6d-f) demonstrate GGOs that mimic COVID-19 cases.

Radiologist Performance

For blind review on the test set without AI prediction, six radiologists had an average accuracy of 85% (95% CI: 77-90%), sensitivity of 79% (95% CI: 64-89%) and specificity of 88% (95% CI: 78-94%). Assisted by the model's probabilities, the radiologists achieved a higher average accuracy (90% vs. 85%, Δ=5, p<0.001), sensitivity (88% vs. 79%, Δ=9, p<0.001) and specificity (91% vs. 88%, Δ=3, p=0.001). Table 3 summarizes the comparison of radiologist performances without and with AI assistance.

Discussion

COVID-19 can be difficult to distinguish from other types of pneumonia on chest CT . It has been revealed that the gold standard diagnostic test, real time reverse-transcriptase polymerase chain reaction (RT-PCR), frequently produces false-negatives 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 build a more complete patient assessment. AI has shown efficacy in differentiating COVID-19 from pneumonia of other etiology on 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 to a radiologist only approach, AI augmentation significantly improved radiologists' performance in distinguishing COVID-19 from pneumonia of other etiology yielding higher measures of accuracy, sensitivity, and specificity.

The diagnostic accuracy produced by manual interpretation of COVID-19 chest CT is good but needs to be improved to make resource allocation and disease management during the current pandemic less strenuous on healthcare systems and economies worldwide. Current clinical algorithms for the management of COVID-19 patients are contingent on the amount of resources available and require definitive imaging results (20). While distinguishing COVID-19 from normal lung or other lung pathologies such as cancer on chest CT 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 patient 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 etiology on chest CT. The results that we present suggest that integrating AI into radiologists' routine workflow 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 due to differences in patient population and imaging 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 has 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 etiology 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 RPP testing or had the test results available. For those without RPP 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 etiology (e.g., 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. Lastly, there was significant difference in baseline characteristics between COVID-19 and non-COVID-19 pneumonia patients which could have introduced bias. For example, patients in the non-COVID-19 pneumonia cohort were predominately from the United States, and 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 multi-national, multi-institutional cohort, our model training could have benefited from a larger cohort size.

AI assistance improved radiologist performance in distinguishing COVID-19 from pneumonia of other etiology on chest CT. Future study will investigate integration of these algorithms into routine clinical workflow to assist radiologists in accurately diagnosing COVID-19.

References

1. Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, et al. Presumed Asymptomatic Carrier Transmission of COVID-19. JAMA. 2020. Epub 2020/02/23. doi: 10.1001/jama.2020.2565. PubMed PMID: 32083643; PubMed Central PMCID: PMCPMC7042844.

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. The Lancet Infectious Diseases. 2020.

3. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497-506. Epub 2020/01/28. doi: 10.1016/S0140-6736(20)30183-5. PubMed PMID: 31986264.

4. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. Jama. 2020.

5. Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran TML, et al. Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT. Radiology. 2020:200823-.

6. Choi H, Qi X, Yoon SH, Park SJ, Lee KH, Kim JY, et al. Extension of Coronavirus Disease 2019 (COVID-19) on Chest CT and Implications for Chest Radiograph Interpretation. Radiology: Cardiothoracic Imaging. 2020;2(2):e200107.

7. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology. 2020:200642. Epub 2020/02/27. doi: 10.1148/radiol.2020200642. PubMed PMID: 32101510.

8. Caruso D, Zerunian M, Polici M, Pucciarelli F, Polidori T, Rucci C, et al. Chest CT Features of COVID-19 in Rome, Italy. Radiology. 2020:201237.

9. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT. Radiology.0(0):200905. doi: 10.1148/radiol.2020200905. PubMed PMID: 32191588.

10. Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). medRxiv. 2020.

11. Xu X, Jiang X, Ma C, Du P, Li X, Lv S, et al. Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia. arXiv preprint arXiv:200209334. 2020.

12. Chen J, Wu L, Zhang J, Zhang L, Gong D, Zhao Y, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv. 2020.

13. Shi W, Peng X, Liu T, Cheng Z, Lu H, Yang S, et al. Deep Learning-Based Quantitative Computed Tomography Model in Predicting the Severity of COVID-19: A Retrospective Study in 196 Patients. 2020.

14. GenMark Diagnostics I. ePlex Pipeline: GenMarkDx; [cited 2020 March 9]. Available from: https://genmarkdx.com/solutions/panels/eplex-panels/eplex-pipeline/.

15. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv e-prints [Internet]. 2019 May 01, 2019:[arXiv:1905.11946 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190511946T.

16. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv e-prints [Internet]. 2018 January 01, 2018:[arXiv:1801.04381 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2018arXiv180104381S.

17. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, editors. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE

international conference on computer vision; 2017.

18. Agresti A, Coull BA. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician. 1998;52(2):119-26. doi: 10.1080/00031305.1998.10480550.

19. Li Y, Yao L, Li J, Chen L, Song Y, Cai Z, et al. Stability Issues of RT-PCR Testing of SARS-CoV-2 for Hospitalized Patients Clinically Diagnosed with COVID-19. Journal of Medical Virology. 2020.

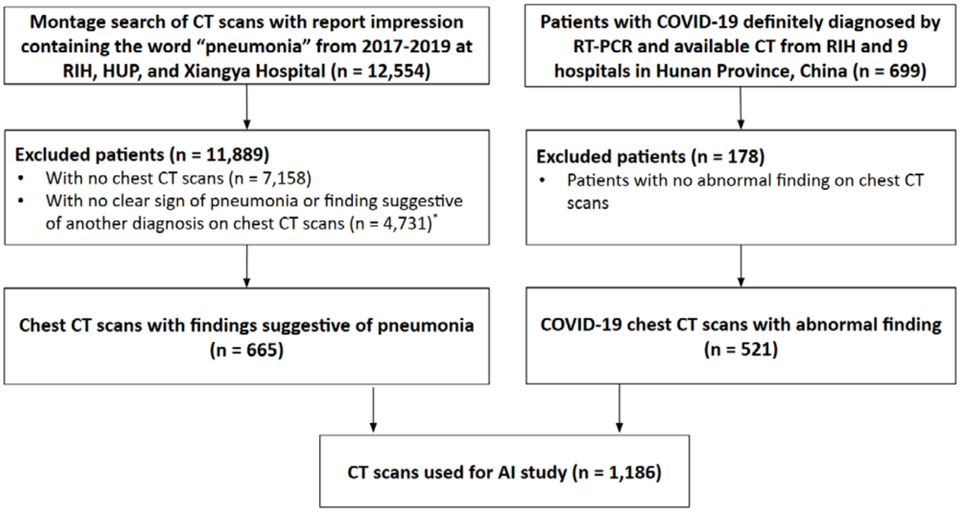
20. Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, et al. The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society. Radiology.0(0):201365. doi: 10.1148/radiol.2020201365. PubMed PMID: 32255413.

21. Zu ZY, Jiang MD, Xu PP, Chen W, Ni QQ, Lu GM, et al. Coronavirus Disease 2019 (COVID-19): A Perspective from China. Radiology.0(0):200490. doi: 10.1148/radiol.2020200490. PubMed PMID: 32083985.

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 medicine. 2018;15(11).

23. AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. Medical physics. 2018;45(3):1150-8.

24. Du Q. Clinical classification. Chinese Clinical Guidance for COVID-19 Pneumonia Diagnosis and Treatment (7th edition) [Internet]. 2020 April 6, 2020 April 8, 2020]. Available from: http://kjfy.meetingchina.org/msite/news/show/cn/3337.html.



Montage search of CT scans with report impression containing the word "pneumonia" from 2017-2019 at RIH, HUP, and Xiangya Hospital (n = 12,554)

Excluded patients (n = 11,889)
- With no chest CT scans (n = 7,158)
- With no clear sign of pneumonia or finding suggestive of another diagnosis on chest CT scans (n = 4,731)*

Chest CT scans with findings suggestive of pneumonia (n = 665)

Patients with COVID-19 definitely diagnosed by RT-PCR and available CT from RIH and 9 hospitals in Hunan Province, China (n = 699)

Excluded patients (n = 178)
- Patients with no abnormal finding on chest CT scans

COVID-19 chest CT scans with abnormal finding (n = 521)

CT scans used for AI study (n = 1,186)

*Excluded based on review of original radiology reports and images

Figure 1. Diagram illustrating patient inclusion and exclusion. Abbreviations: RIH, Rhode Island Hospital; HUP, Hospital of the University of Pennsylvania; AI, artificial intelligence; RT-PCR, reverse transcriptase polymerase chain reaction.

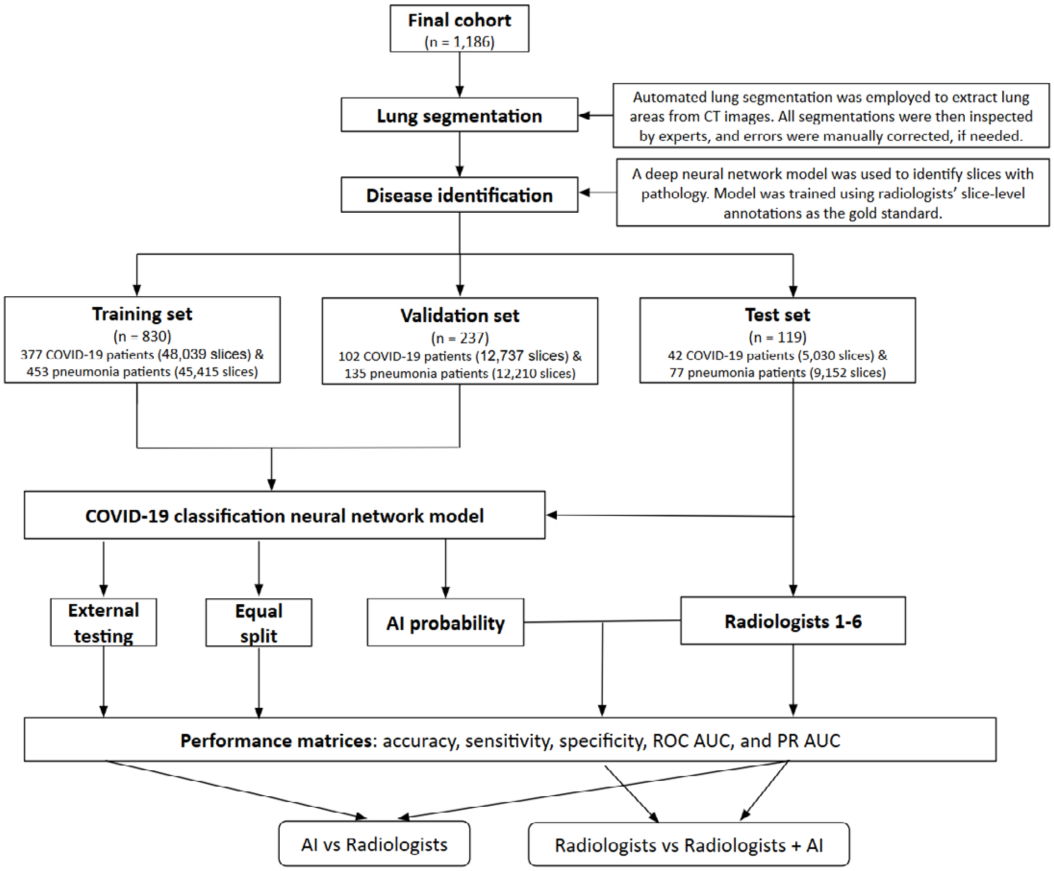

Final cohort
(n = 1,186)
Lung segmentation
Automated lung segmentation was employed to extract lung areas from CT images. All segmentations were then inspected by experts, and errors were manually corrected, if needed.
Disease identification
A deep neural network model was used to identify slices with pathology. Model was trained using radiologists' slice-level annotations as the gold standard.
Training set
(n = 830)
377 COVID-19 patients (48,039 slices) &
453 pneumonia patients (45,415 slices)
Validation set
(n = 237)
102 COVID-19 patients (12,737 slices) &
135 pneumonia patients (12,210 slices)
Test set
(n = 119)
42 COVID-19 patients (5,030 slices) &
77 pneumonia patients (9,152 slices)
COVID-19 classification neural network model
External testing
Equal split
AI probability
Radiologists 1-6
Performance matrices: accuracy, sensitivity, specificity, ROC AUC, and PR AUC
AI vs Radiologists
Radiologists vs Radiologists + AI

Figure 2. Flow diagram illustrating our AI model for distinguishing COVID-19 from non-COVID-19 pneumonia. Abbreviations: ROC AUC: Receiver Operator Characteristics Area Under the Curve; PR AUC: Precision Recall area under curve.

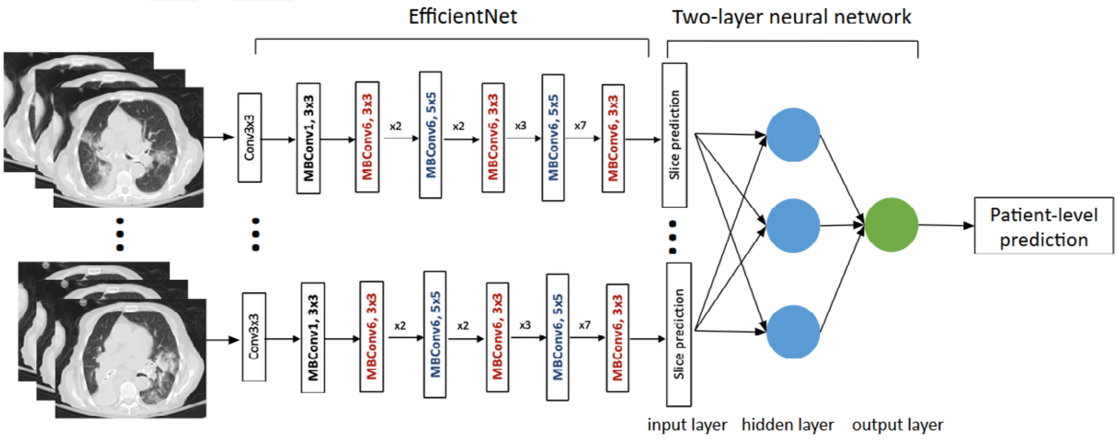


Figure 3: COVID-19 Classification Neural Network Model.

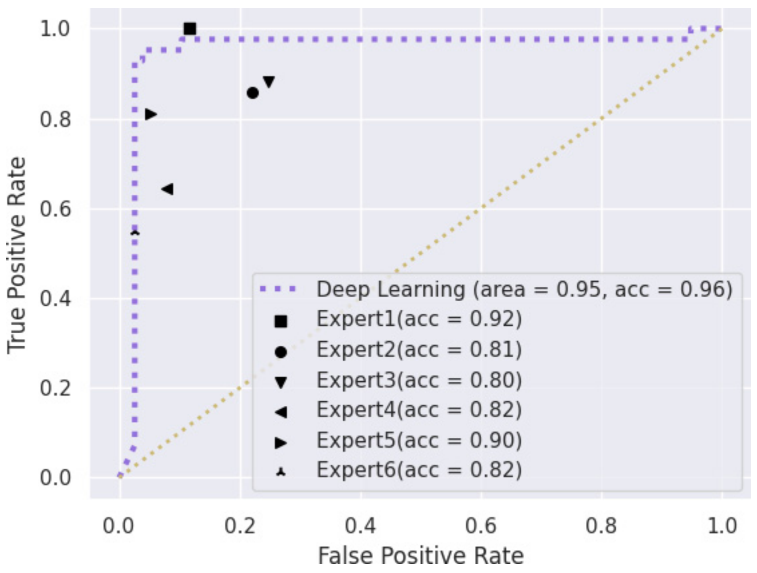


Figure 4. ROC curve of deep neural network on the test set compared to radiologist performance. ROC = receiver operating curve.

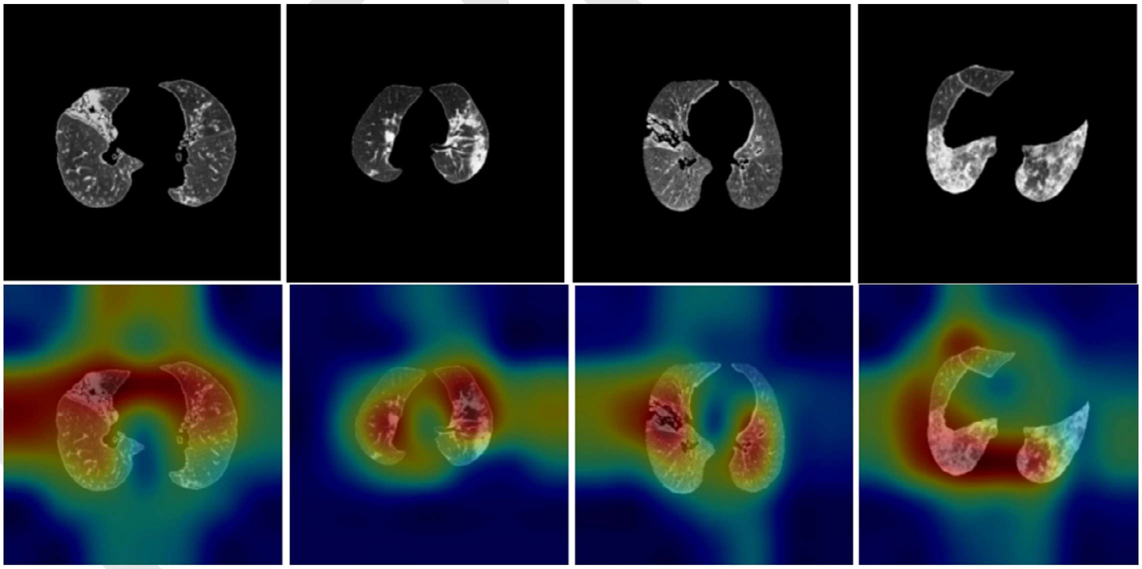


Figure 5. Representative slices corresponding to Grad-CAM images on the test set.

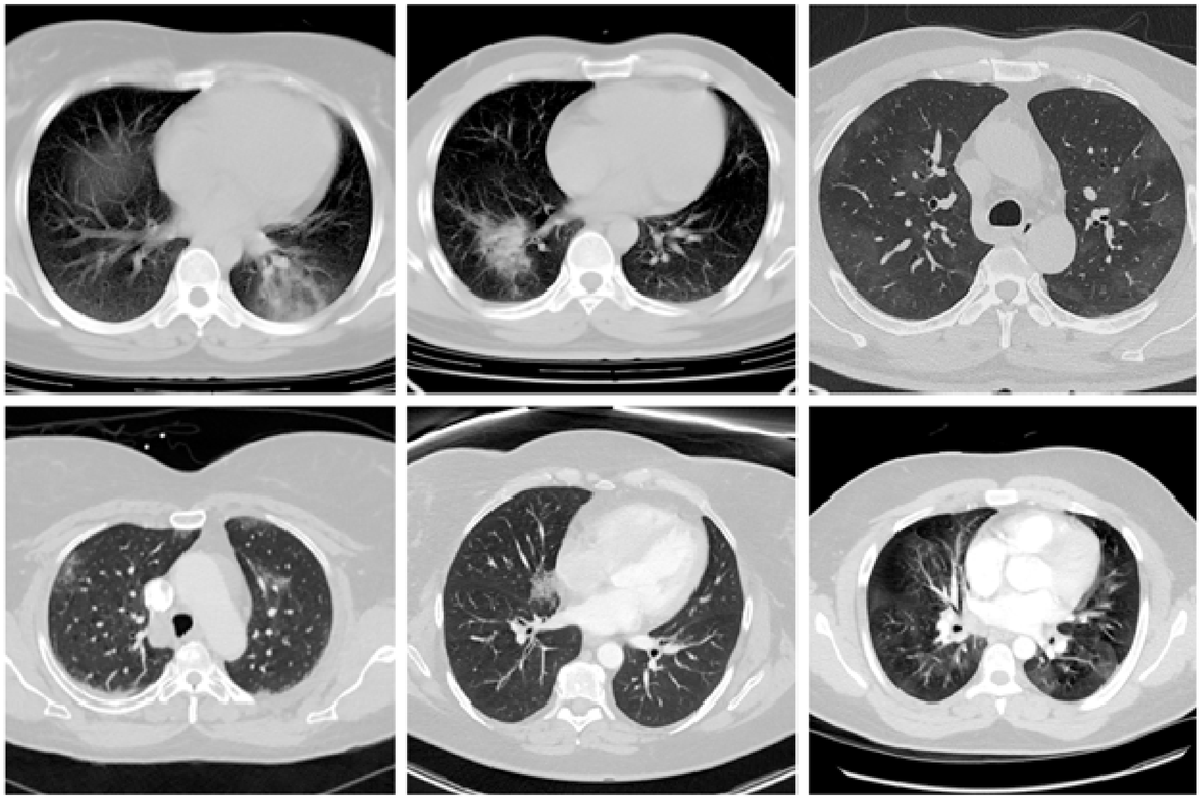


Figure 6. Representative cases that the majority of radiologists misclassified.

A-C (top row, left to right): COVID-19 pneumonia. Our model correctly classified all three cases.

A. 4/6 radiologists (radiologists 3-6) said it was non-COVID-19. With AI assistance, 2/6 radiologists (radiologists 5 and 6) continued to say it was non-COVID-19.

B. 4/6 radiologists (radiologists 3-6) said it was non-COVID-19. With AI assistance, 3/6 radiologists (radiologists 3-5) continued to say it was non-COVID-19.

C. 4/6 radiologists (radiologists 2 and 4-6) said it was non-COVID-19. With AI assistance, 1/6 radiologist (radiologist 2) continued to say it was non-COVID-19.

D-F (bottom row, left to right): Non-COVID-19 pneumonia. Our model correctly classified D and E.

D. 5/6 radiologists (radiologists 1-5) said it was COVID-19. With AI assistance, all 5/6 radiologists continued to say it was COVID-19.

E. 4/6 radiologists (radiologists 1, 2, 4, and 6) said it was COVID-19. With AI assistance, 3/6 radiologists (radiologists 1, 2, and 4) continued to say it was COVID-19.

F. 4/6 radiologists (radiologists 1-3 and 6) said it was COVID-19. With AI assistance, 5/6 radiologists (radiologists 1-4 and 6) said it was COVID-19.

Table 1. Clinical Characteristics of COVID-19 and non-COVID-19 pneumonia patient cohorts

	COVID-19	Non-COVID-19	
	(n=521)	(n=665)	p-value
Age (year)			<0.001
Mean age	46 ± 16	62 ± 19	
<20	11 (2)	12 (2)	
20-39	151 (29)	76 (11)	
40-59	222 (43)	166 (25)	
>=60	136 (26)	411 (62)	
Sex			0.03
Male	268 (51)	385 (58)	
Female	251 (48)	280 (42)	
Presence of Fever			<0.001
Fever	303 (58)	361 (54)	
No Fever	147 (28)	280 (42)	
White Blood Cell Count			<0.001
Elevated	12 (2)	337 (51)	
Normal	441 (85)	325 (49)	
Lymphocyte Count			<0.001
Normal	303 (58)	293 (44)	
Decreased	186 (36)	365 (55)	
Comorbidities			
Cardiovascular Disease	18 (3)	230 (35)	<0.001
Hypertension	65 (12)	258 (39)	<0.001
COPD	22 (4)	157 (24)	<0.001
Diabetes	29 (6)	116 (17)	<0.001
Chronic Liver Disease	11 (2)	17 (3)	0.62
Chronic Kidney Disease	6 (1)	70 (11)	<0.001
Malignant Tumor	2 (0)	84 (13)	<0.001
HIV	0 (0)	15 (2)	<0.001
Time from Onset to Presentation (days)*	11 ± 12		
<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: Data are patients with percentages in parentheses. Age and Time from Onset to Presentation are mean ± standard deviation.
COVID-19: Coronavirus Disease 2019
COPD: Chronic Obstructive Pulmonary Disease
HIV: Human Immunodeficiency Virus
Severity is defined by the Chinese Clinical Guidance for COVID-19 Pneumonia Diagnosis and Treatment (7th edition), published by the China National Health Commission on March 4, 2020 (24).
Formal diagnosis was done using reverse transcription polymerase chain reaction (RT-PCR) for COVID-19 cases.

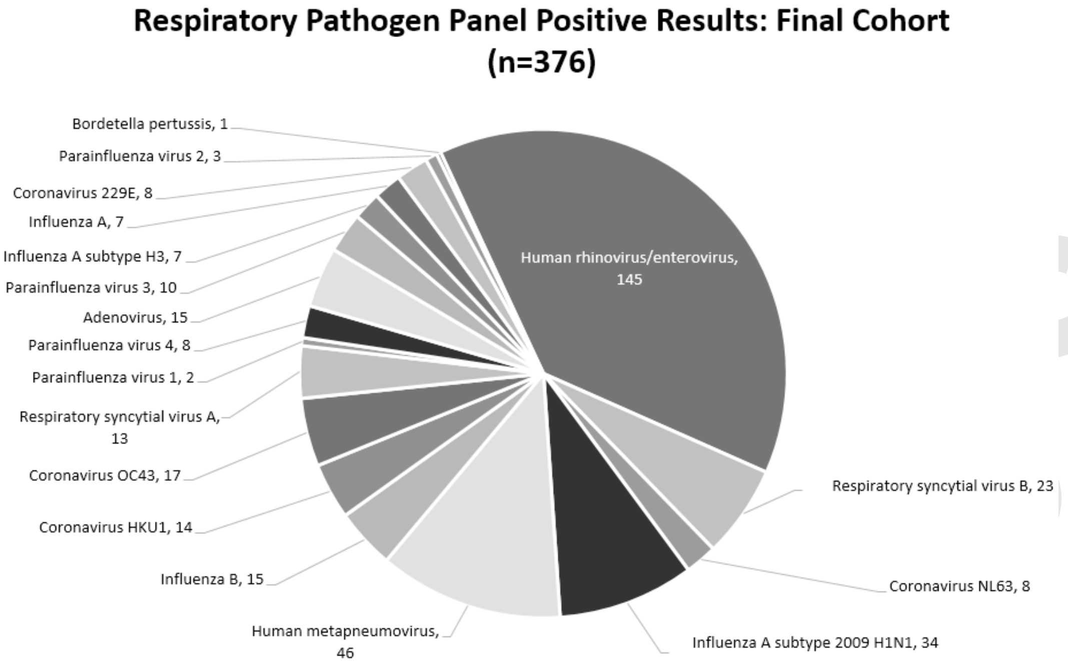
Table 2. The results of an artificial intelligence (AI) model and six radiologists without AI assistance on the test set (n=119) in differentiating between COVID-19 pneumonia and non-COVID-19 pneumonia

		Radiologist performance (%)	AI performance (%)	AI performance minus radiologist performance (%)	p-value
Radiologist #1	Accuracy	92 (86-96)	96 (90-98)	3 (-2-8)	0.34
	Sensitivity	100 (90-100)	95 (83-100)	-5 (-15-0)	0.50
	Specificity	88 (79-94)	96 (88-99)	8 (1-15)	0.07
Radiologist #2	Accuracy	81 (72-87)	96 (90-98)	15 (8-23)	<0.001
	Sensitivity	86 (71-94)	95 (83-100)	10 (0-21)	0.22
	Specificity	78 (67-86)	96 (88-99)	18 (9-28)	0.001
Radiologist #3	Accuracy	80 (71-86)	96 (90-98)	16 (9-23)	<0.001
	Sensitivity	88 (74-95)	95 (83-100)	7 (-3-18)	0.36
	Specificity	75 (64-84)	96 (88-99)	21 (12-30)	<0.001
Radiologist #4	Accuracy	82 (74-88)	96 (90-98)	13 (6-21)	0.002
	Sensitivity	64 (49-77)	95 (83-100)	31 (15-47)	0.001
	Specificity	92 (83-97)	96 (88-99)	4 (-4-12)	0.51
Radiologist #5	Accuracy	90 (83-94)	96 (90-98)	6 (0-13)	0.12
	Sensitivity	81 (66-90)	95 (83-100)	14 (2-27)	0.07
	Specificity	95 (87-98)	96 (88-99)	1 (-5-8)	1.00
Radiologist #6	Accuracy	82 (74-88)	96 (90-98)	13 (6-21)	0.001
	Sensitivity	55 (40-69)	95 (83-100)	40 (23-58)	<0.001
	Specificity	97 (90-100)	96 (88-99)	-1 (-6-3)	1.00
Radiologist Average	Accuracy	85 (77-90)	96 (90-98)	11 (7-16)	<0.001
	Sensitivity	79 (64-89)	95 (83-100)	16 (8-24)	0.001
	Specificity	88 (78-94)	96 (88-99)	8 (3-14)	0.002

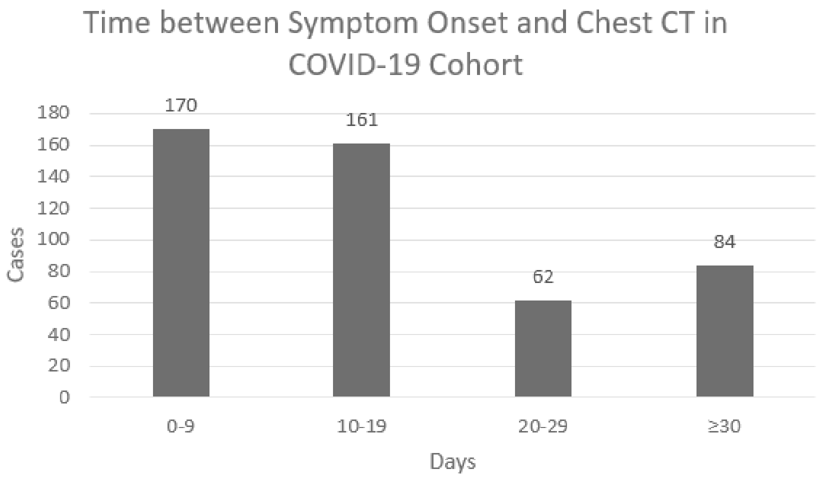
Table 3. Comparison of six radiologists without and with assistance of an artificial intelligence (AI) model in differentiating between COVID-19 pneumonia and non-COVID-19 pneumonia

		Without AI assistance (%)	With AI assistance (%)	Radiologist with AI assistance minus radiologist without AI assistance (%)	p-value
Radiologist #1	Accuracy	92 (86-96)	92 (86-96)	0 (-3-3)	1.00
	Sensitivity	100 (90-100)	98 (86-100)	-2 (-8-0)	1.00
	Specificity	88 (79-94)	90 (80-95)	1 (-3-6)	1.00
Radiologist #2	Accuracy	81 (72-87)	89 (82-94)	8 (4-13)	0.002
	Sensitivity	86 (71-94)	86 (71-94)	0 (0-0)	1.00
	Specificity	78 (67-86)	91 (82-96)	13 (6-21)	0.002
Radiologist #3	Accuracy	80 (71-86)	83 (75-89)	3 (1-7)	0.13
	Sensitivity	88 (74-95)	93 (80-98)	5 (0-12)	0.49
	Specificity	75 (64-84)	78 (67-86)	3 (0-7)	0.50
Radiologist #4	Accuracy	82 (74-88)	90 (83-94)	8 (3-13)	0.01
	Sensitivity	64 (49-77)	83 (69-92)	19 (8-32)	0.01
	Specificity	92 (83-97)	94 (85-98)	1 (-3-6)	1.00
Radiologist #5	Accuracy	90 (83-94)	93 (87-97)	3 (1-7)	0.12
	Sensitivity	81 (66-90)	88 (74-95)	7 (0-16)	0.25
	Specificity	95 (87-98)	96 (88-99)	1 (0-4)	1.00
Radiologist #6	Accuracy	82 (74-88)	92 (86-96)	10 (5-16)	<0.001
	Sensitivity	55 (40-69)	81 (66-90)	26 (13-40)	0.001
	Specificity	97 (90-100)	99 (92-100)	1 (0-4)	1.00
Radiologist Average	Accuracy	85 (77-90)	90 (83-94)	5 (4-7)	<0.001
	Sensitivity	79 (64-89)	88 (74-95)	9 (5-13)	<0.001
	Specificity	88 (78-94)	91 (82-96)	3 (2-5)	0.001

Supplement



Supplementary Figure E1. Distribution of viral pathogens based on RPP test in the final included RIH cohort. Note: Three patients each have two pathogens.



Supplementary Figure E2. Distribution of time from symptom onset to CT scan for COVID-19 cases. Note: this information is not available for 44 COVID-19 cases

Supplementary Table E1. A list of participating hospitals with contributed number of cases

Hospital	Location	COVID-19 CT Scans	Non-COVID-19 CT Scans
Xiangya Hospital	Hunan, China	31	124
Yongzhou Central Hospital	Hunan, China	36	0
Changde Second People's Hospital	Hunan, China	61	0
Affiliated Nan Hua Hospital	Hunan, China	27	0
Loudi Central Hospital	Hunan, China	27	0
Chengzhou Second People's Hospital	Hunan, China	34	0
Zhuzhou Central Hospital	Hunan, China	61	0
Yiyang No.4 Hospital	Hunan, China	44	0
The First Hospital of Changsha	Hunan, China	188	0
Hospital of the University of Pennsylvania	Pennsylvania, United States	0	168
Rhode Island Hospital	Rhode Island, United States	12	373

Supplementary Table E2. Description of Chest CT protocols and parameters of 11 hospitals

Hospital	CT System	Tube Volume (kVp)	Tube Current (mAs)	Pitch (mm)	Matrix	Slice Thickness (mm)	Field of View (mm x mm)	Reconstructed Slice Thickness (mm)	Increment
Xiangya Hospital	SIEMENS SOMATOM Definition	120	90-120	0.8	512 x 512	1	300 x 300	1	-1
	SIEMENS SOMATOM go.Now	130	30-100	1.5	512 x 512	1	350 x 350	1	1
Yongzhou Central Hospital	SIEMENS Emotion 16	130	35-100	0.8	512 x 512	1.5	350 x 350	1.5	-1
Changde Second People's Hospital	GE BrightSpeed	120	100-160	1.3750	512 x 512	1.25	350 x 350	1.25	-0.6
Affiliated Nan Hua Hospital	GE LightSpeed Ultra	120	75-150	1.35	512 x 512	1.25	350 x 350	1.25	-1.25
Loudi Central Hospital	Philips Access CT	100	60-150	0.8125	512 x 512	1.5	300 x 300	1.5	1.5
Chenzhou Second People's Hospital	Hitachi ECLOS	120	50-100	1	512 x 512	2.5/5/10	350 x 350	2.5/5/10	-1.25/-5/-10
Zhuzhou Central Hospital	GE Optica CT680	120	90-320	1.375	512 x 512	0.625/1.25	350 x 350	0.625/1.25	
Yiyang No.4 Hospital	GE Optica CT520	120	50-200	1.375	512 x 512	10	360 x 360	10	-10
The First Hospital of Changsha	GE BRIVO CT325	120	100	1.5	512 x 512	5	380 x 380	5	-5
Hospital of the University of Pennsylvania	SIEMENS SOMATOM Definition AS Siemens Healthineers Global	120	varies	1	512 x 512	1.25	380 x 380	1	1.25
Rhode Island Hospital	SIEMENS SOMATOM Definition AS20	120	150	0.6	512 x 512	1.5/5	380 x 380	1.5/5	1.25
	SIEMENS SOMATOM Definition AS+	120	150	0.6	512 x 512	0.75/5	380 x 380	0.75/5	0.625
	GE LightSpeed VCT/Resolution	100	120-450	0.984	512 x 512	0.6/5	380 x 380	0.6/5	0.625
	GE Lightspeed 16/Optima CT580	120	120-450	1.375	512 x 512	1.25/5	380 x 380	1.25/5	1.25

Supplementary Table E3. Comparison of CT slice thickness across data sets

	Training set		Validation set		Test set		p-value
	N	Median and range (mm)	N	Median and range (mm)	N	Median and range (mm)	
Internal testing	830	1.25 (0.60-10.00)	237	1.25 (0.60-10.00)	119	1.00 (0.60-10.00)	0.38
External testing	637	2.00 (0.60-10.00)	274	1.25 (0.60-10.00)	275	1.00 (0.63-10.00)	<0.001
Equal split	396	1.25 (0.60-10.00)	395	1.00 (0.60-10.00)	395	1.25 (0.60-10.00)	0.72

Supplementary Table E4. Number of patients and CT slices for the different training, validation and test splits

	Training set		Validation set		Test set	
	# of patients	# of slices	# of patients	# of slices	# of patients	# of slices
Internal testing	377 / 453	48,039 / 45,415	102 / 135	12,737 / 12,210	42 / 77	5,030 / 9,152
External testing	270 / 367	31,660 / 30,769	144 / 130	16,355 / 14,765	107 / 168	17,791 / 21,243
Equal split	174 / 222	22,201 / 22,513	167 / 228	20,406 / 23,415	180 / 215	23,199 / 20,849

Data are presented as COVID-19/non-COVID-19

External test set is composed of non-COVID-19 patients from HUP and COVID-19 patients from Yongzhou Central Hospital, Zhuzhou Central Hospital, and Yiyang No.4 Hospital.

Supplementary Table E5. Model performance across training, validation and test sets for different splits

		Accuracy (%)	Sensitivity (%)	Specificity (%)	ROC AUC	PR AUC
Internal testing	Training set (n=830)	99 (97-99)	98 (96-99)	99 (98-100)	1.00	0.99
	Validation set (n=237)	97 (94-99)	96 (90-99)	99 (94-100)	1.00	0.98
	Test set (n=119)	96 (90-98)	95 (83-100)	96 (89-99)	0.95	0.95
External testing	Training set (n=637)	99 (98-100)	98 (95-100)	100 (98-100)	1.00	0.99
	Validation set (n=274)	98 (96-99)	98 (94-100)	98 (94-100)	0.99	0.99
	Test set (n=275)	87 (82-90)	89 (81-94)	86 (80-90)	0.90	0.87
Equal split	Training set (n=396)	98 (97-99)	97 (93-99)	100 (97-100)	0.99	0.99
	Validation set (n=395)	93 (90-95)	94 (89-97)	93 (88-95)	0.97	0.93
	Test set (n=395)	91 (87-93)	94 (90-97)	87 (82-91)	0.95	0.92