Deep learning for differentiating novel coronavirus pneumonia and influenza pneumonia

Min Zhou[1,2#], Dexiang Yang[3#], Yong Chen[4#], Yanping Xu[1,2#], Jin-Fu Xu[5], Zhijun Jie[6], Weiwu Yao[7], Xiaoyan Jin[8], Zilai Pan[9], Jingwen Tan[4], Lan Wang[4], Yihan Xia[4], Longkuan Zou[10], Xin Xu[10], Jingqi Wei[10], Mingxin Guan[10], Fuhua Yan[4], Jianxing Feng[10], Huan Zhang[4], Jieming Qu[1,2]

[1]Department of Respiratory and Critical Care Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [2]Institute of Respiratory Diseases, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [3]Department of Respiratory Medicine, Tongling People's Hospital, Tongling, China; [4]Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [5]Department of Respiratory and Critical Care Medicine, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China; [6]Department of Respiratory and Critical Care Medicine, Shanghai Fifth People's Hospital, Fudan University, Shanghai, China; [7]Department of Radiology, Shanghai Tongren Hospital Affiliated to Jiao Tong University School of medicine, Shanghai, China; [8]Department of Pulmonary and Critical Care Medicine, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [9]Department of Radiology, Ruijin North Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [10]Haohua Technology Co., Ltd., Shanghai, China

Contributions: (I) Conception and design: J Feng, H Zhang, J Qu; (II) Administrative support: F Yan, H Zhang, J Qu; (III) Provision of study materials or patients: F Yan, H Zhang, J Qu; (IV) Collection and assembly of data: M Zhou, D Yang, JF Xu, Z Jie, X Jin; (V) Data analysis and interpretation: Y Chen, Y Xu, W Yao, Z Pan, J Tan, L Wang, Y Xia, L Zou, X Xu, J Wei, M Guan; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Jieming Qu. Department of Respiratory and Critical Care Medicine, Ruijin hospital, Shanghai Jiao Tong University School of Medicine, No. 197, Rui Jin 2nd Road, Shanghai 200025, China. Email: jmqu0906@163.com; Huan Zhang. Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No. 197, Rui Jin 2nd Road, Shanghai 200025, China. Email: huanzhangy@163.com; Jianxing Feng. Haohua Technology Co., Ltd., Weihai International Group Building, No. 511 Weihai Road, Shanghai 200041, China. Email: fengjianxing@harmon.health; Fuhua Yan. Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No. 197, Rui Jin 2nd Road, Shanghai 200025, China. Email: yanfuhua@yahoo.com.

Background: Chest computed tomography (CT) has been found to have high sensitivity in diagnosing novel coronavirus pneumonia (NCP) at the early stage, giving it an advantage over nucleic acid detection during the current pandemic. In this study, we aimed to develop and validate an integrated deep learning framework on chest CT images for the automatic detection of NCP, focusing particularly on differentiating NCP from influenza pneumonia (IP).

Methods: A total of 148 confirmed NCP patients [80 male; median age, 51.5 years; interquartile range (IQR), 42.5–63.0 years] treated in 4 NCP designated hospitals between January 11, 2020 and February 23, 2020 were retrospectively enrolled as a training cohort, along with 194 confirmed IP patients (112 males; median age, 65.0 years; IQR, 55.0–78.0 years) treated in 5 hospitals from May 2015 to February 2020. An external validation set comprising 57 NCP patients and 50 IP patients from 8 hospitals was also enrolled. Two deep learning schemes (the Trinary scheme and the Plain scheme) were developed and compared using receiver operating characteristic (ROC) curves.

Results: Of the NCP lesions, 96.6% were >1 cm and 76.8% were of a density <−500 Hu, indicating them to have less consolidation than IP lesions, which had nodules ranging from 5–10 mm. The Trinary scheme accurately distinguished NCP from IP lesions, with an area under the curve (AUC) of 0.93. For patient-level classification in the external validation set, the Trinary scheme outperformed the Plain scheme (AUC: 0.87 *vs.* 0.71) and achieved human specialist-level performance.

Conclusions: Our study has potentially provided an accurate tool on chest CT for early diagnosis of NCP with high transferability and showed high efficiency in differentiating between NCP and IP; these findings could help to reduce misdiagnosis and contain the pandemic transmission.

Keywords: Deep learning; novel coronavirus pneumonia (NCP); influenza pneumonia (IP); chest computed tomography (CT)

Submitted Jul 15, 2020. Accepted for publication Oct 30, 2020.
doi: 10.21037/atm-20-5328
View this article at: http://dx.doi.org/10.21037/atm-20-5328

Introduction

In December 2019, a cluster of idiopathic pneumonia cases emerged in Wuhan, China. These cases were eventually identified as novel coronavirus pneumonia (NCP), namely coronavirus disease 2019 (COVID-19). In late January 2020, the outbreak was declared a global health emergency by the World Health Organization. The virus responsible for NCP, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (1), can be transmitted from person to person, and in severe cases, can progress rapidly, leading to sepsis or multiple organ failure (2). The human health crisis that has resulted from the spread of SARS-CoV-2 is unprecedented.

Emerging seasonal influenza viruses are also a critical cause of contagious respiratory disease across the globe that could potentially lead to hospitalization and mortality (3,4). As well as sharing non-specific onset symptoms, such as sudden fever, cough, sore throat, and headache, patients with viral pneumonia (VP) caused by SARS-CoV-2 and influenza infection have similar laboratory findings; consequently, the viruses have been widely considered to be clinically indistinguishable (5,6). Therefore, early diagnosis and differentiation between NCP and influenza pneumonia (IP) are of the utmost importance.

Real-time reverse transcriptase-polymerase chain reaction (RT-PCR) was initially used to confirm the clinical diagnosis of NCP; however, it was reported to have a high false negative rate, owing to a low viral load at the early stage of infection or possibly genetic mutations, and further testing using specimens from multiple sites was usually required for confirmation (7-9). Meanwhile, some early-onset NCP patients who had presented with abnormal findings on chest computed tomography (CT) were found to return negative results on the initial nucleic acid test. Therefore, early differentiation of IP patients from NCP patients based on chest CT would provide another diagnostic approach that could potentially reduce mortality rates and the risk of cross infection while patients await laboratory confirmation.

Imaging is routinely performed to help detect and differentiate disease in patients suspected of having VP, with common radiological features being unilateral or patchy bilateral areas of consolidation, nodular opacities, bronchial wall thickening, and lobar consolidation (10). However, the radiologic manifestations of VP are nonspecific, and it is challenging for clinicians to differentiate NCP from IP due to their similar imaging features. Despite its high sensitivity (97%) in screening for NCP, chest CT has a poor performance in differential diagnosis, with a specificity of 25% (7).

The problem of differentiating between NCP and IP could potentially be alleviated by deep learning, a technique that was formerly used to automatically detect pneumonia based on chest X-ray images and to discriminate usual interstitial pneumonia from nonspecific interstitial pneumonia based on chest CT images (11-13). Deep learning is considered to be worthwhile in terms of differentiation, as it can achieve expert-level performance in medical image analysis with minimal demands on time and labor, thus optimizing the allocation of medical resources.

In this study, we developed and validated an integrated deep learning framework on chest CT images for the automatic detection of NCP, particularly focusing on differentiating NCP from IP at the early stage to ensure prompt implementation of isolation. An intrinsic difference was found in deep learning classification models trained by different devices from multiple centers (13,14). To solve the problem, we proposed a deep learning model (the Trinary scheme) to discriminate IP lesions from NCP lesions by learning image features that reflect differences among devices or hospitals. We present the following article in accordance with the STARD reporting checklist (available at http://dx.doi.org/10.21037/atm-20-5328).

Methods

Patients

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was reviewed and approved by the Ruijin Hospital Ethics Committee (2017-

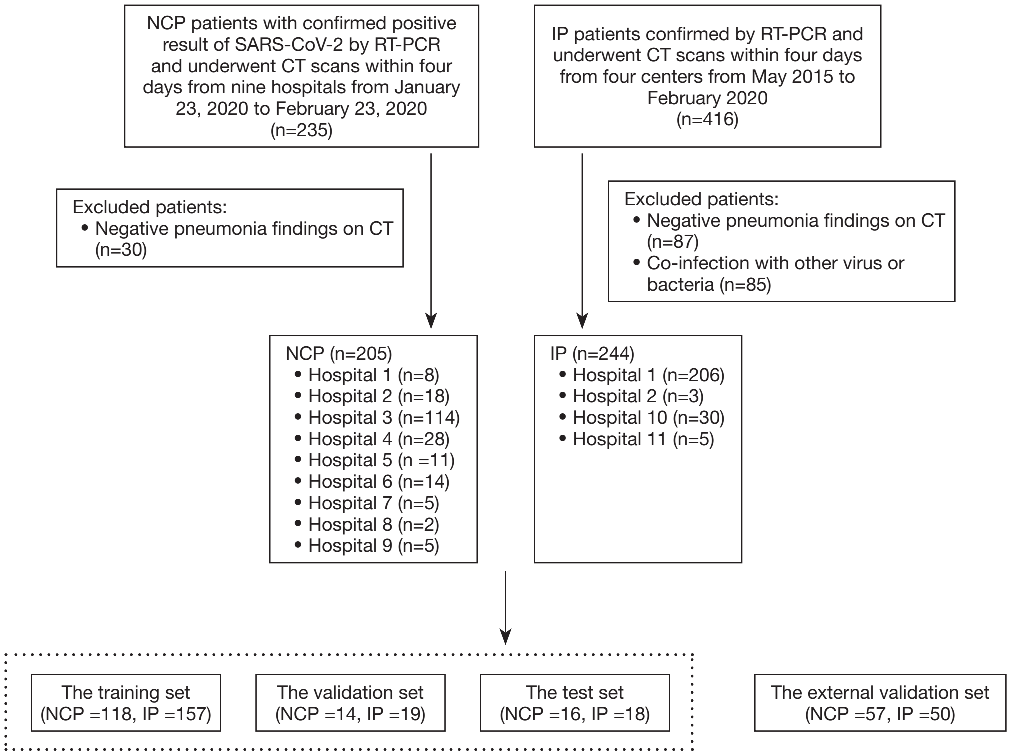
NCP patients with confirmed positive result of SARS-CoV-2 by RT-PCR and underwent CT scans within four days from nine hospitals from January 23, 2020 to February 23, 2020 (n=235)
IP patients confirmed by RT-PCR and underwent CT scans within four days from four centers from May 2015 to February 2020 (n=416)
Excluded patients:
• Negative pneumonia findings on CT (n=30)
Excluded patients:
• Negative pneumonia findings on CT (n=87)
• Co-infection with other virus or bacteria (n=85)
NCP (n=205)
• Hospital 1 (n=8)
• Hospital 2 (n=18)
• Hospital 3 (n=114)
• Hospital 4 (n=28)
• Hospital 5 (n =11)
• Hospital 6 (n=14)
• Hospital 7 (n=5)
• Hospital 8 (n=2)
• Hospital 9 (n=5)
IP (n=244)
• Hospital 1 (n=206)
• Hospital 2 (n=3)
• Hospital 10 (n=30)
• Hospital 11 (n=5)
The training set (NCP =118, IP =157)
The validation set (NCP =14, IP =19)
The test set (NCP =16, IP =18)
The external validation set (NCP =57, IP =50)

Figure 1 Illustration of the patient recruitment process. CT, computed tomography; IP, influenza pneumonia; NCP, novel coronavirus pneumonia; RT-PCT, real-time reverse transcriptase-polymerase chain reaction.

186). The requirement for written informed consent was waived due to the retrospective nature of this study.

From January 11, 2020 to February 23, 2020, NCP patients with a confirmed positive result of SARS-CoV-2 via laboratory testing of respiratory secretions by nucleic acid in double swab tests through RT-PCR from 4 hospitals were consecutively enrolled in this study. All included participants had undergone CT scans within 4 days of admission to hospital. Patients with negative CT findings for pneumonia were excluded. Finally, 148 NCP patients were recruited. Among them, 15 patients from one hospital underwent 2 CT scans within 4 days due to rapid disease progression, while the rest of the patients underwent CT scans only once. Chest abnormalities for NCP patients included bilateral pulmonary parenchymal ground-glass opacity and consolidation, sometimes presenting as rounded morphology and peripheral lung distribution (15). Each CT scan was considered as a case. Therefore, there were 163 cases (148 patients) analyzed.

From May 2015 to February 2020, patients confirmed with influenza infection by RT-PCR from 4 hospitals were enrolled as the IP group. Patients who presented negative pneumonia findings on CT and/or co-infection with other virus or bacteria were excluded. Finally, 194 IP patients were included in this study. All IP patients underwent CT scanning once.

The clinical indices collected for all participants included sex, age, and symptoms (including cough and fever). All the patients who participated in the training and tuning of the deep learning algorithms for classification were randomly divided into a training set, a validation set, and a testing set.

To further validate the generalization of the deep learning framework, 57 NCP patients from 8 hospitals and 50 IP patients from 1 hospital were enrolled as an external validation set.

The patient recruitment process is shown in *Figure 1*. The average time intervals between CT scans and positive RT-PCR test results for all NCP patients and IP patients were 10.8±3.2 and 3.3±0.8 days, respectively. *Figure 2* shows a flowchart of this study.

Computation process overview

First, lesion level computation was performed, followed

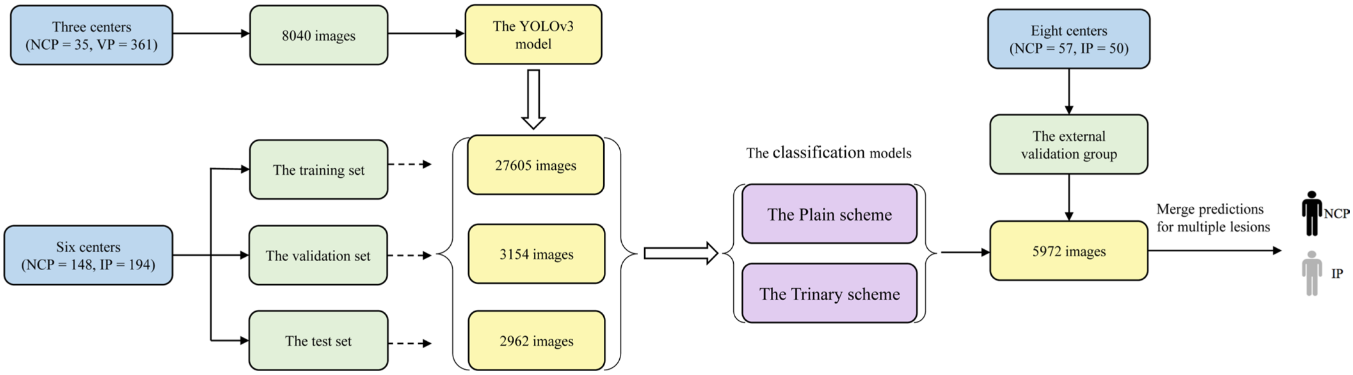

Figure 2 Flowchart illustrating the deep learning process for the differential diagnosis of NCP and IP from multiple centers. IP, influenza pneumonia; NCP, novel coronavirus pneumonia.

by patient-level computation based on the lesion level results. The lesion level computation comprised 4 parts: (I) lesion detection using YOLOv3; (II) lesion classification using a modified VGGNet model; (III) a Trinary training scheme was introduced, which had the aim of lessening the influence of CT machines on the model and improving lesion classification; (IV) a widely used transfer learning method that was adopted to achieve better model performance by fine-tuning a pre-trained classification model on public non-medical images. Each part is described in detail in the following sections.

Lesion detection

For lesion detection, 35 NCP patients (50 cases) randomly selected from 3 hospitals and 361 patients (361 cases) with VP were enrolled. The 361 VP patients were diagnosed according to the 2016 Clinical Practice guidelines by the Chinese Thoracic Society (16) and the 2007 Infectious Diseases Society of America (IDSA)/American Thoracic Society (ATS) guidelines (17). The CT parameters are presented in *Table 1*.

Annotation

In our study, the lesion regions of each CT image were independently annotated by 2 radiologists individually with more than 10 years of experience in pulmonary-thoracic disease. Both radiologists were aware of the clinical history of infection. All lesion annotations by the 2 radiologists were collected. If a lesion was annotated by only 1 of the radiologists, a consensus was reached between the pair.

Preprocessing

Before the lesions were detected, the CT images were preprocessed through the following steps to remove irrelevant parts. First, an in-house tool was used to segment the lung region. Second, the convex hull of the lung region was calculated independently for each slice of the CT image. Then, the smallest bounding box that could cover all the convex hulls was calculated. Next, the CT image was cropped using the calculated bounding box. Finally, only the slices that contained ≥1 lesions were preserved to train the lesion detection model.

To reduce the influence of CT image spacing, all CT images with a spacing of <5 mm were transformed into 5 mm by taking the average of continuous slices. Specifically, to transform CT images with a spacing of 1.25 mm, the mean of every 4 continuous slices was taken as 1 slice. The spacing distribution was as follows. First, all preprocessed images were resized to 256×256 to fit the data into the YOLOv3 model, and then further randomly placed on a 416×416 template. During training, all the input images were randomly flipped.

Dataset partition

Ten patients with NCP (15 CT series) and 40 patients (40 CT series) with VP were randomly selected as the test set. The remaining patients (24 patients with NCP and 321 patients with VP) were used for training and validation. After the removal of image slices without lesions, there were a total of 8,040 image slices in the training and validation sets, and 10% of them were randomly selected as the validation set. Due to randomness, only 1 NCP

Table 1 Detailed information of CT imaging protocol

Hospital	Manufacturers model name	Tube current (mAs)	Tube volume (kVp)	Pitch (mm)	Matrix	Slice thickness (mm)
Hospital 1	Aquilion ONE	100	100	0.828	512×512	5
	Discovery CT750 HD	300	120	0.984	512×512	5
	iCT 256	103	120	0.758	512×512	5
	IQon-Spectral CT	108	120	0.765	512×512	5
	LightSpeed16	280	120	1.375	512×512	5
	LightSpeed VCT	145	120	1.375	512×512	5
	SOMATOM Definition Flash	360	120	1.2	512×512	5
	uCT 528	106	120	1.325	512×512	1.5
	uCT 760	194	120	1.0875	512×512	5
	uCT S-160	100	120	1.1	512×512	5
Hospital 2	Perspective	150	110	0.6	512×512	1
	uCT 760	158	120	1.0875	512×512	1.25
Hospital 3	Discovery CT	185	120	1.375	512×512	0.625
	SOMATOM Definition AS+	138	120	1.2	512×512	1
	uCT 760	185	120	1.0875	512×512	0.625
	uCT 530	187	120	1.175	512×512	1
Hospital 4	Brilliance16	188	120	1.1	512×512	2
	Brilliance 64	308	120	1.2	512×512	1
Hospital 5	SOMATOM Definition AS+	151	120	1.2	512×512	1.5
	Brilliance16	188	120	1.1	512×512	2
Hospital 6	LightSpeed VCT	150	120	1.375	512×512	5
	SOMATOM Definition Flash	329	120	1.2	512×512	1
Hospital 7	LightSpeed16	250	120	1.375	512×512	1.25
	LightSpeed Ultra	250	120	0.875	512×512	2.5
	Revolution Frontier	219	120	0.984375	512×512	1.25
Hospital 8	Aquilion ONE	172	120	0.813	512×512	5
	SOMATOM Perspective	128	120	0.95	512×512	1
Hospital 9	uCT 510	174	120	1.0625	512×512	1.5
	uCT 528	42	120	1.15	512×512	1.2
Hospital 10	LightSpeed16	250	120	1.375	512×512	5
	LightSpeed VCT	349	120	1.375	512×512	5
	uCT 510	185	120	1.1875	512×512	1.5
Hospital 11	SOMATOM Definition AS+	279	120	1.2	512×512	0.6

A total of 33 CT scanners from 11 hospitals were used in this study. Tube current, tube volume, pitch, matrix and slice thickness were described. CT, computed tomography.

patient (1 CT series) was included in the validation set. Therefore, there were 23 NCP patients (35 CT series) in the training set. For NCP and VP, there were 793 and 2,021 images, respectively, in the test set, and 3,277 and 2,695 images, respectively, in the external validation set after preprocessing.

Lesion detection model and training details

YOLOv3, a highly efficient and widely used object detection network in computer vision, was employed to perform lesion detection on the preprocessed images (18). In our dataset, lesions were found in only 34% and 36% of the slices for NCP and VP, respectively. The YOLOv3 model can identify a limited number of candidate regions by ruling out most of the irrelevant healthy regions. Removing slices and regions without lesions can reduce the potential of the model to overfit, especially when the training data are limited, as they were in our case. The YOLOv3 architecture pre-trained to 33% mAP on the COCO dataset was used in this study, and we retrained it on our dataset. The parameters were fine-tuned across all layers with a learning rate of 0.0001, where the learning rate was adjusted by step policy. A decay factor of 0.0005 and a momentum factor of 0.9 were used in our model. Due to the multiple-scale method of feature pyramid networks (FPN), YOLOv3 output feature maps are of 3 different scales. The output of YOLOv3 may contain multiple bounding boxes, which may be false positive results or overlapping results. Therefore, the score threshold was set at 0.2 to reduce the number of bounding boxes of the output.

Lesion level classification

Preprocessing

In our work, a classification model for classifying NCP and IP was built. We randomly selected 80% of patients as the training set, 10% as the validation set, and 10% as the test set. Specifically, 118 patients with NCP (132 CT series) and 157 patients with IP (157 CT series) for training, 14 patients with NCP (15 CT series) and 19 patients with IP (19 CT series) for validation, and 16 patients with NCP (18 CT series) and 18 patients with IP (18 CT series) were incorporated for testing. After preprocessing, there were 27,605 and 3,154 images in the training and validation sets, respectively. There were 2,962 images (4,683 annotated lesions) containing ≥1 lesions in the test set. All lesions in an NCP patient were considered as NCP lesions, and all lesions in an IP patient were considered to be IP lesions. For an annotated lesion, a random 128×128 patch was cropped from the CT image around the lesion so that the lesion was always in the cropped patch. If the lesion was larger than 128×128, the CT image was magnified before cropping.

Classification model

Due to the limited number of training samples, VGGNet was chosen as the classification model (19). The neural network structure VGGNet is relatively old and simple, and can be improved using AlexNet. The network consisted of 5 convolution groups and 1 max pooling layer, followed by 3 fully connected layers and 1 softmax layer. The entire network used the same convolution kernel size (3×3) and maximum pooling size (2×2).

To better address our problem, some modifications were made to the original VGGNet. Specifically, first, the 3 fully connected layers and one softmax layer were removed, followed by the max pooling layer. Then, 1 convolution layer with rectified linear unit (ReLU), 1 fully connected layer with ReLU, and 1 softmax layer were added. ReLU is the most commonly used activation function in deep learning methods and can be written as:

$$f(x) = \max(0, x) \tag{1}$$

where x is the input to a neuron.

These added modules are denoted as “head” (*Figure 3*).

Trinary classification scheme

To distinguish NCP and IP, the classification model was trained using NCP and IP image patches. This training process is referred to as the Plain scheme herein. We found that the Plain scheme may have been influenced by device-specific features and did not generalize well. Therefore, a novel training scheme, referred to as the Trinary scheme, was introduced as follows.

In the Trinary scheme, we started by collecting random regions from the CT images and then generated random bounding boxes. These random regions were assumed to share the same device-specific features with lesions from the same device, and the randomly generated bounding box regions were not overlapping with lesions with a high probability. The only difference between the random regions and the lesions was whether a lesion was present. Then, the binary classification problem was revised to a trinary classification problem, and the 3 classes corresponded to NCP lesions, IP, lesions, and random

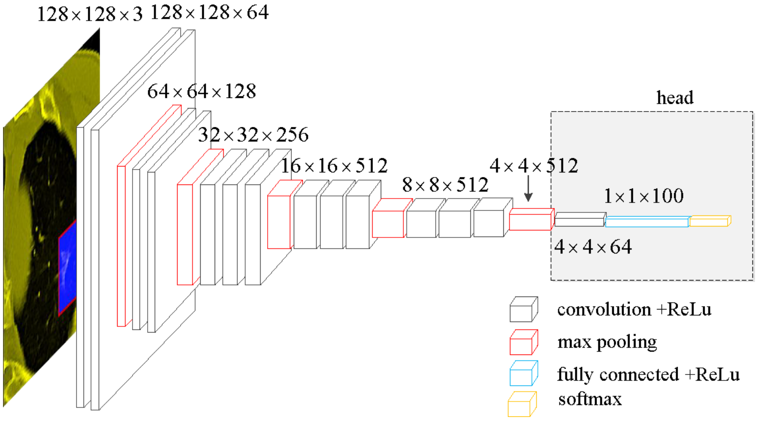

Figure 3 Network structure of the modified VGGNet.

regions.

Cross-entropy, which characterizes the distance between 2 probability distributions and is commonly used in multi-classification problems, was used as the classification loss function (20):

$$CE(p,y)=\sum_{i=0}^{2}-\log(pi)*\delta(y-i) \qquad [2]$$

Where p_i is the output probability for class i, y is the benchmark label from {0, 1, 2}, and $\delta(\cdot)$ is the Dirac's delta function. For hard instance mining, we used focal loss (21):

$$FL(p.y)=-(1-p_i)^2\log(p_i) \qquad [3]$$

Where $(1-p_i)^y$ is a modulating factor, y is a focusing parameter, and we set y to 2. Focal loss suppressed the weight of samples that are easy to classify during training calculation loss, so that the model could focus on learning to distinguish those samples that could not be distinguished easily. The final loss function is:

$$CE(p,y)=\sum_{i=0}^{2}-(1-p_i)^{\gamma}log(p_i)*\delta(y-i) \qquad [4]$$

Transfer learning

Transfer learning was used based on previous reports. The training of the modified VGGNet model was divided into two steps (22,23). Firstly, the preserved original VGGNet (5 convolution groups and 1 max pooling layer) was frozen (without back propagation or updated parameters), and only the "head" module was trained. An Adam optimizer with a learning rate of 0.0001 was used. Secondly, after 70 iterations, the entire network was fine-tuned. Because the training sets for NCP or IP were imbalanced, the NCP lesion images were up-sampled to balance the training data. For the Trinary scheme, 1 random region was inserted for every 3 NCP or IP lesions.

Patient-level classification

Taking the Plain model as an example, the classification model outputs two probabilities corresponding to NCP and IP for each lesion. To obtain patient-level prediction from lesion level predictions, the lesion level probabilities of a patient were averaged. For Trinary scheme, the probabilities for NCP and IP were further normalized such that they were summed to be one.

The rationale behind taking the average was as follows. If each lesion classification was considered to be the result of a weak classifier for the patient (24), taking the average provides an efficient ensemble method of obtaining a strong classifier. Many popular machine learning methods have utilized such an ensemble method. For instance, the well-known random forest classification model would take the average of the probabilities of multiple decision trees to achieve a better performance than a single decision tree (25). In our experiments, the patient-level classification was also observed to have higher accuracy than the lesion-level classification.

Comparison with experts on the external validation set

To compare the performances of the deep learning framework and radiologists in the external validation set, a panel of 10 radiologists from two groups (group 1, 5 specialist-level radiologists with more than 15 years of experience; group 2, 5 resident radiologists with 3–5 years of experience) in thoracic imaging were recruited.

The radiologists were completely blinded to the clinical information and histological findings of the participants (they were aware that the patients had either NCP or IP but were completely blinded to the location and distribution). Each radiologist had access to digital imaging and communications in medicine (DICOM) series. All cases were anonymized, randomly assigned case numbers from 1 to 107, and randomly divided into two groups with 54 cases and 53 cases respectively. The 10 radiologists interpreted 1 group of images each time. Each reading session was separated by 1 day. To mimic the routine diagnostic process, each session was required to be finished within 1 day. The evaluation was repeated after 1 week with disorganized groups different from the first time to avoid possible interobserver variation. The radiologists were instructed to independently provide a classification decision of NCP or IP each time. The final decision was not made until a consensus was reached between the two assessments. The performance of both individual radiologists and the radiologists as a group were compared with that of the Plain scheme and Trinary scheme. The consistency between the radiologist group and the schemes was also recorded.

To determine which scheme was more akin to the judgement of human experts, we also performed the lesion re-classification process by radiologists, for there was no standard for reference. First, any lesion with inconsistent classification between the two schemes was selected for further evaluation. To eliminate potential interference factors (e.g., other lesions), only an individual lesion, instead of the whole image, was selected. Second, the lesions were re-assessed by two senior specialists, who had the highest area under the curves (AUCs) and were blinded to the results of the algorithms. The classification of NCP or IP was recorded for all selected lesions, and the final decision for lesion triage was made by consensus of the specialists. Lastly, the consistency was calculated between human experts and the algorithms.

Statistical analysis

Clinical indices were analyzed based on the variable type. Continuous variables were measured by means or medians and were compared using two-sided independent *t* tests or Wilcoxon rank-sum tests. Categorical variables were measured as proportions and were compared using chi-square tests or Fisher's exact tests. The classification metrics used included AUC, sensitivity, specificity, accuracy, precision, and F1 score. 95% confidence interval (CI) was used to handle indeterminacy. The intraclass correlation coefficient (ICC) was used to evaluate interobserver agreement among radiologists. The mean values of the classification metrics were computed across the specialist group as well as the resident group. A statistically significant difference was indicated when results with two-sided P<0.05. Data analysis was performed using Python (version 3.7.4, http://www.r-project.org) and MedCalc (version 1.13.1, MedCalc Software Ltd., Ostend, Belgium). The deep learning models were implemented based on Tensorflow (version 1.13.1, https://www.tensorflow.org) framework.

Results

Patient information

Among the 148 patients with NCP, 53.7% were men, which showed no significant difference with the IP patients (57.7%, P=0.43). With a median age of 51.5 years (IQR, 42.5–63.0), the NCP patients were significantly younger than the IP patients, (P<0.01). Fewer NCP patients than IP patients exhibited coughing (50% *vs.* 73.2%, P<0.01). The information for the external validation set is presented in *Table 2*.

Comparison of imaging features between IP and NCP

A joint analysis was further performed of the imaging features of annotated lesions from 148 NCP patients and 194 IP patients. Because a three-dimensional lesion may appear in multiple layers of the CT images, to perform comparison, the annotated lesions that had overlapped bounding boxes were merged. There were 1,669 and 1,568 NCP and IP merged lesions, respectively. To simplify the description, 'merged lesion' is simplified as 'lesion' in this section.

Of the NCP lesions, 96.3% were >1 cm, were significantly more NCP lesions than IP lesions >1 cm (P<0.01). However, NCP was more likely to form moderate-sized lesions, and only 28.8% of these lesions were >3 cm. There was a similar distribution of abnormal features in the lower lobes between NCP (42%) and IP (42.8%). Lesions with an intensity <−500 Hu accounted for 80.4% of NCP lesions. In comparison, 40.1% of IP patients had a focus density >−500 Hu, indicating that NCP had less consolidation than IP. Significantly more extensive distribution was demonstrated by NCP than IP, with 86.5% of patients

Table 2 Clinical information for patients in the training, validation, test, and external validation sets

	The training, validation and test set			The external validation set		
	NCP (n=148)	IP (n=194)	P value	NCP (n=57)	IP (n=50)	P value
Gender			0.43			0.22
Male	80 (53.7%)	112 (57.7%)		31 (54.4%)	33 (66%)	
Female	68 (46.3%)	82 (42.3%)		26 (45.6%)	17 (34%)	
Age (y)	51.5 (42.5, 63.0)	65.0 (55.0, 78.0)	<0.01	49.0 (35.0, 63.3)	66.5 (60.0, 79.0)	<0.01
Cough	74 (50%)	142 (73.2%)	<0.01	28 (49.1%)	31 (62%)	0.18
Fever	108 (73.0%)	143 (73.7%)	0.88	41 (71.9%)	37 (74%)	0.81

Data are n (%) or median (IQR). The P values indicate differences between NCP and IP patients, with P<0.05 considered statistically significant. Categorical variables including sex, cough, and fever were compared using chi-square tests. Age differences between the NCP and IP patients were measured using Wilcoxon rank-sum tests. NCP, novel coronavirus pneumonia; IP, influenza pneumonia; IQR, interquartile range.

presenting bilateral pulmonary invasion and 35.0% showing involvement of 5 lobes. Of 1568 IP lesions, 4.4% were nodules (Hu >0), of which 60.9% had CT values >40 Hu, and 10 (0.6%) IP nodules were >5 mm. In contrast, only 0.4% (6/1,669) of NCP lesions presented as nodules (*Table 3*). Examples of differences in CT imaging features between an NCP patient and an IP patient are shown in *Figures 4* and *5*.

Lesions could be effectively detected by YOLOv3

To evaluate the YOLOv3 model, intersection over union (IOU) was used as the match measure for measuring the detection accuracy. IOU is an evaluation metric which computes the intersection over the union of two bounding boxes. For a lesion annotated by specialists, if its bounding box was overlapped with a predicted bounding box with an IOU ≥0.2, this lesion was considered as having been detected. Examples of detected lesions are presented in *Figure 6*. For each predicted candidate lesion, the model gave a confidence score. Generally, a high confidence score led to a low false positive rate as well as a low recall rate. However, when the confidence score was extremely low, the recall rate also dropped due to highly overlapping predictions (*Table 4*). The results showed that the detection performance was not sensitive to the confidence score as long as the cutoff for confidence score was in a reasonable range. For example, if the confidence score cutoff was set to 0.1, the false positive rate was 0.216, the recall rate was 0.704, and the F1 was 0.742. If the confidence score was in the range of (0.03–0.25), the F1 score was always >0.7. In the external validation set, the model detected 2,984 and 3,811 NCP and IP lesions, respectively. These detected lesions were used in the following analysis.

Lesions could be accurately classified by the modified VGGNet model

The annotated lesions were used to train and evaluate the modified VGGNet classification model. Both Plain and Trinary training schemes were applied. The Trinary scheme (AUC: 0.93) performed better than the Plain scheme (AUC: 0.85) (*Figure 7*). The F1 scores for the Plain and Trinary schemes were 0.65 and 0.79, respectively (*Table 5*).

Trinary scheme was more consistent with specialists on lesion classification

To better understand the difference between the Plain and Trinary schemes, the predicted NCP probability from both schemes was compared using lesions identified in the external validation set. Generally, we found that the Trinary scheme is less influenced by the source of the data. More specifically, for lesions from hospitals in the training set, we found that the Plain scheme may give a very high NCP probability, even if for lesions were not very typical of NCP (*Figure 8A*). On the contrary, for lesions from hospitals not in the training set, the Plain scheme may give a very low NCP probability, even if the lesions were extremely typical of NCP (*Figure 8B*). In contrast, the Trinary scheme seemed to give more intuitively reasonable predictions. To quantify the difference, detected lesions on which the absolute difference of predicted NCP probability between

Table 3 Comparison of NCP and IP in radiological manifestations

	NCP	IP	P value
Lesions	1,669	1,568	
Diameter (cm)			
>1	1,608 (96.3%)	1,382 (88.1%)	<0.01
>2	868 (53.1%)	1,011 (64.5%)	<0.01
>3	481 (28.8%)	775 (49.4%)	<0.01
Involved lobe			
Right upper lobe	462 (27.7%)	469 (29.9%)	0.17
Right middle lobe	225 (13.5%)	178 (11.4%)	0.07
Right lower lobe	409 (24.5%)	359 (22.9%)	0.28
Left upper lobe	281 (16.8%)	354 (22.6%)	<0.01
Left lower lobe	292 (17.5%)	312 (19.9%)	0.08
Intensity (HU)			
<–500	1342 (80.4%)	940 (59.9%)	<0.01
–500–0	321 (19.2%)	559 (35.7%)	<0.01
>0	6 (0.4%)	69(4.4%)	<0.01
>20	6 (0.4%)	54(3.4%)	<0.01
>40	6 (0.4%)	39 (2.5%)	<0.01
>0, D <3[a]	3 (0.2%)	57 (3.6%)	<0.01
>0, 3≤ D <5[a]	1 (0.1%)	2 (0.1%)	0.61
>0, 5≤D <10[a]	2 (0.1%)	6 (0.4%)	0.17
>0, D ≥10[a]	0 (0.0%)	4 (0.3%)	0.06
Cases	163	194	
Distribution 1			<0.01
Unilateral	22 (13.5%)	41 (21.1%)	
Bilateral	141(86.5%)	153 (78.9%)	
Distribution 2			0.03
Single	4 (2.5%)	15 (7.7%)	
Multiple	159 (97.5%)	179 (92.3%)	
No. of involved lobe			
1	5 (3.1%)	27 (8.2%)	0.04
2	22 (13.5%)	38 (19.6%)	0.13
3	32 (19.6%)	47 (24.2%)	0.30
4	47 (28.8%)	42 (21.6%)	0.12
5	57 (35.0%)	40 (20.6%)	<0.01

The P values indicate differences between NCP and IP patients; P<0.05 was considered statistically significant. Variables including the diameter, involved lobes, intensity, intensity (HU) (<–500), intensity (HU) (–500–0), and number of involved lobes [3,4,5] were measured using chi-square tests. The remaining variables in this table were measured using Fisher's exact test. [a], D in the Intensity (HU) presents diameter of lesions (mm). IP, influenza pneumonia; NCP, novel coronavirus pneumonia.

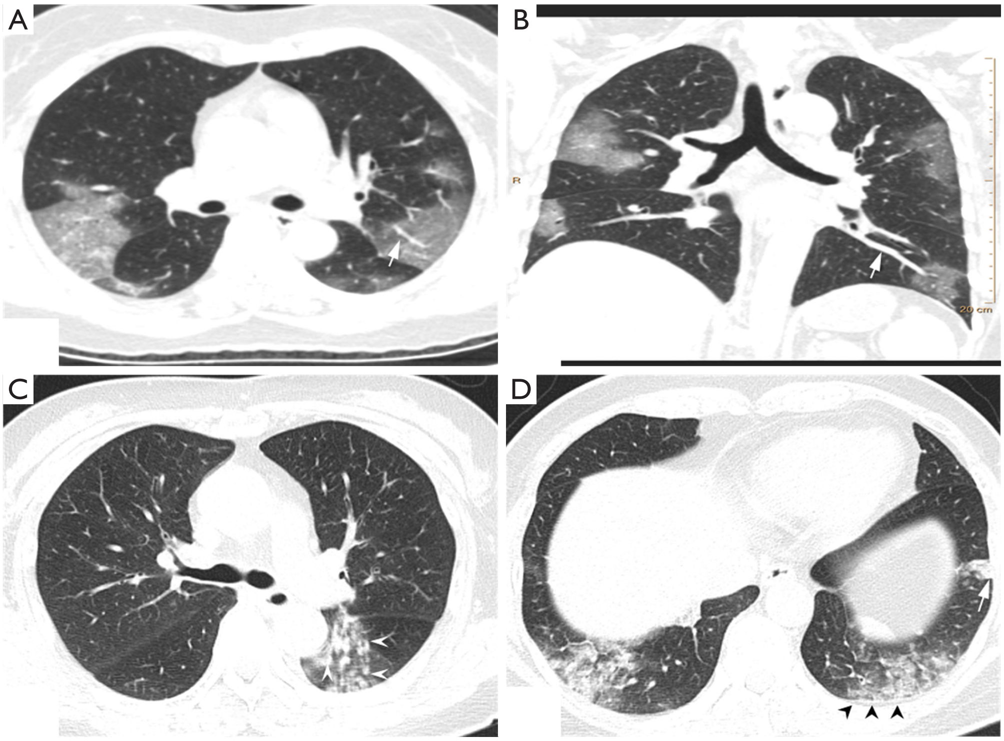

Figure 4 CT image features of an NCP patient and an IP patient. (A,B) Chest CT of an NCP patient: a 45-year-old female with history of visiting Wuhan for 2 days. Presenting with fever and cough for 4 days, she was confirmed as having NCP. The CT scan (A) shows bilateral GGO scatted in 4 lobes, with an obvious peripheral distribution and bilateral lobular or subsegmental GGO involving mainly the subpleural lung regions. Vascular dilation with GGO surrounding was more evident, and a pulmonary venous branch passed through the lesion with luminal dilation (white arrow). The features are more obvious in maximum-intensity projection imaging (B). (C,D) The CT images of an IP patient: CT scan shows multiple, bilateral, and randomly distributed small ill-defined nodules (white arrow head) with small branch opacities indicating the bronchiolitis. Peripheral subpleural consolidation (white arrow) in the left lower lobe with the interlobular septum and pleura thickening (black head). CT, computed tomography; GGO, ground-glass opacity; IP, influenza pneumonia; NCP, novel coronavirus pneumonia.

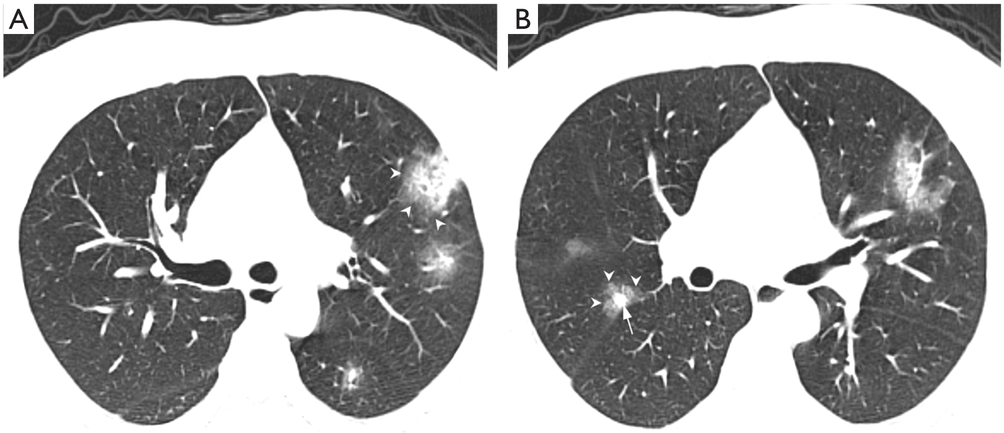

Figure 5 CT images of an NCP patient. A 37-year-old male with NCP of unknown exposure history, presenting with fever and cough. WBC 6.96×10^9/L, N 82.70%, L 13.10%, C-reactive-protein 13.9 mg/dL. Axial non-contrast CT scan shows multiple (5 lobes involved), bilateral, and randomly distributed pulmonary nodules (arrow) surrounded by a halo of GGO (arrow head). CT, computed tomography; GGO, ground-glass opacity; IP, influenza pneumonia; NCP, novel coronavirus pneumonia.

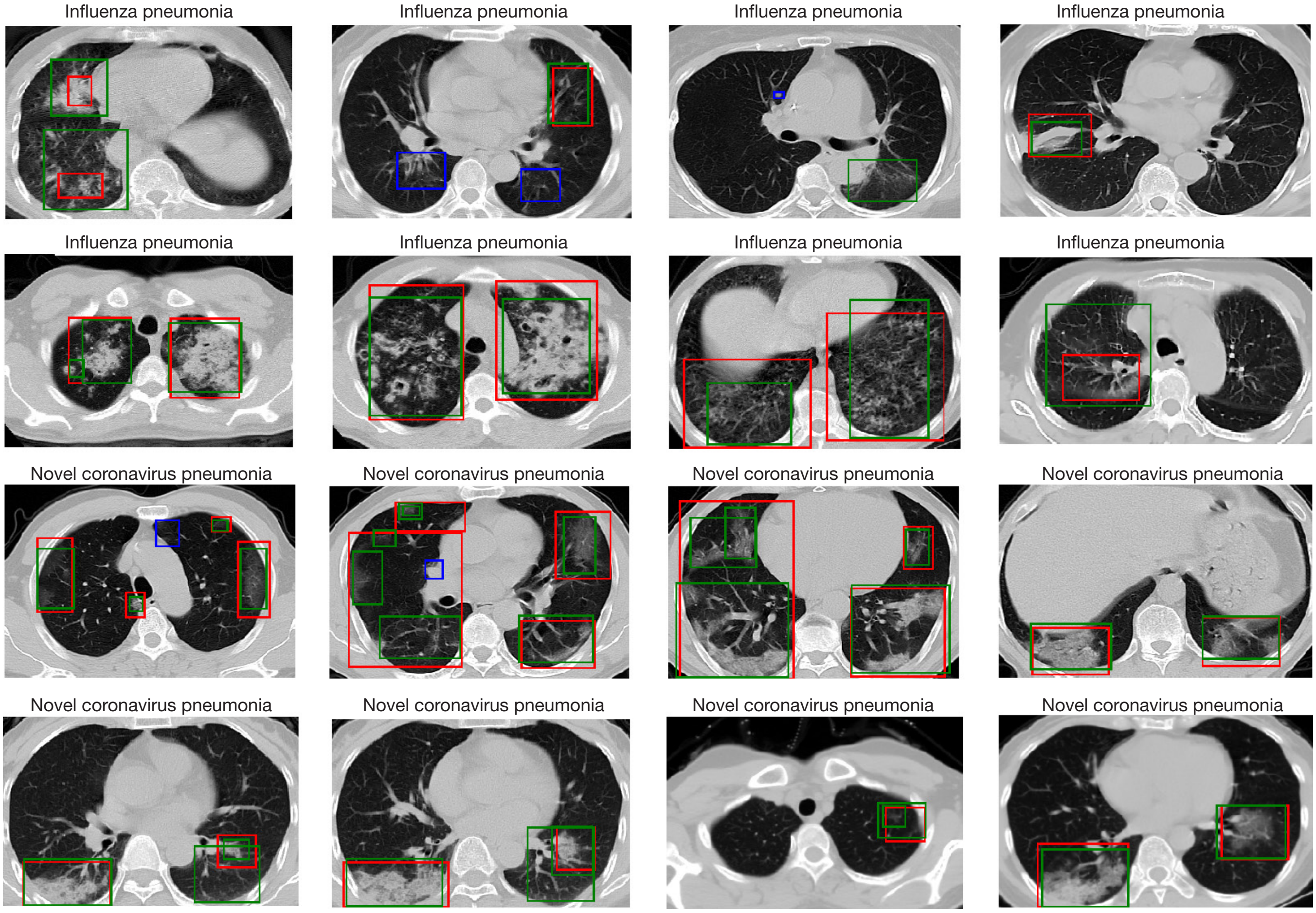
Influenza pneumonia
Influenza pneumonia
Influenza pneumonia
Influenza pneumonia
Influenza pneumonia
Influenza pneumonia
Influenza pneumonia
Influenza pneumonia
Novel coronavirus pneumonia
Novel coronavirus pneumonia
Novel coronavirus pneumonia
Novel coronavirus pneumonia
Novel coronavirus pneumonia
Novel coronavirus pneumonia
Novel coronavirus pneumonia
Novel coronavirus pneumonia

Figure 6 Examples of detected lesions performed by YOLOv3 model for IP and NCP patients. The green boxes represent the lesions predicted by the detection model. The red boxes represent the ground truth of lesions. While the blue boxes represent the lesions that the model failed to detect. IP, influenza pneumonia; NCP, novel coronavirus pneumonia.

Table 4 Performance of the YOLOv3 model for lesion detection in the test set under different confidence scores

Confidence	Predict	Benchmark	FPR	Recall rate	F1
0.01	3,033	1,337	0.44	0.72	0.63
0.02	2,267	1,337	0.36	0.73	0.68
0.03	1,931	1,337	0.32	0.74	0.71
0.04	1,749	1,337	0.29	0.75	0.73
0.05	1,617	1,337	0.27	0.74	0.74
0.1	1,301	1,337	0.22	0.70	0.74
0.15	1,135	1,337	0.17	0.67	0.74
0.2	1,038	1,337	0.15	0.64	0.73
0.25	946	1,337	0.13	0.61	0.72

The Confidence indicates the confidence score of each predicted candidate lesion. The Predict indicates the predicted number of lesions. The Benchmark indicates the annotated number of true positive lesions. FPR, false positive rate.

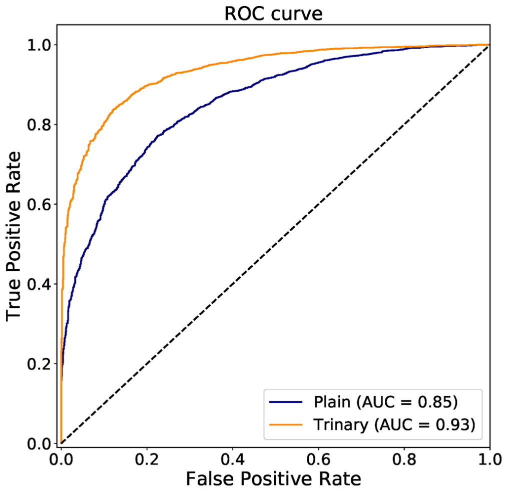

Figure 7 ROC curve of the Plain and Trinary schemes of lesion-level classification on the test set. AUC, area under the curve; ROC, receiver operating characteristic.

Table 5 Performance of the Plain and Trinary schemes for lesion-level classification in the test set

	Accuracy	Precision	Sensitivity	Specificity	F1 score
Plain scheme	79.9%	68.0%	62.2%	87.5%	0.65
Trinary scheme	86.8%	76.6%	80.6%	89.5%	0.79

the Trinary and Plain schemes was >0.5 were filtered out. These lesions were subsequently annotated by 2 specialists. A lesion was considered as NCP or IP only when it was annotated as such by both specialists. Of 812 filtered lesions, 540 and 102 of them were annotated as NCP and IP, respectively. Out of 540 NCP lesions, the Trinary and Plain schemes correctly identified 366 and 174, respectively. Out of 102 IP lesions, 61 and 41 lesions were correctly

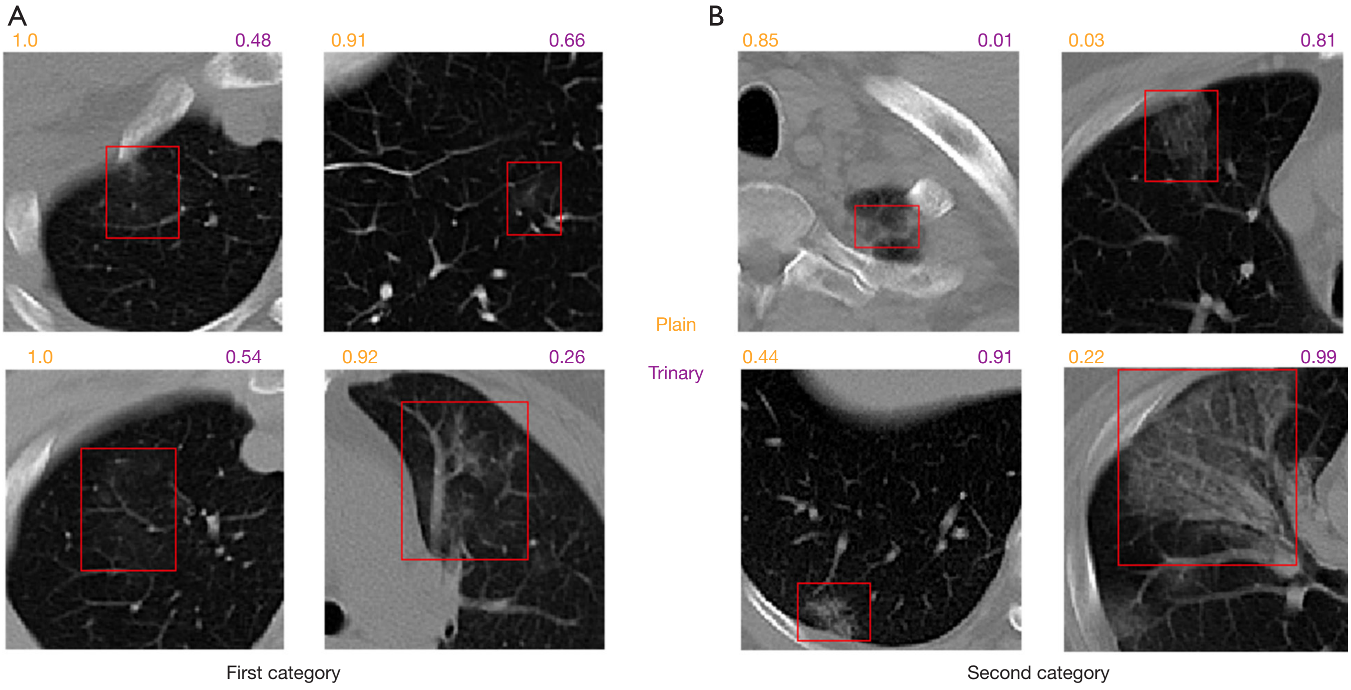

Figure 8 Graphical examples for lesion classification on NCP patients from the external validation set. The numbers above each CT image patch represent the predicted probabilities of being NCP by the Plain scheme (orange) and the Trinary scheme (purple), respectively. (A) presents 4 detected lesions in the first category (20 NCP patients from hospital 1–3). On the 4 non-typical NCP lesions, the Plain scheme gave much higher probability than the Trinary scheme did. (B) presents 4 detected lesions on the second category (37 NCP patients from hospital 4–8). The top left lesion is a false positive. CT, computed tomography; NCP, novel coronavirus pneumonia.

Table 6 Performance of both deep learning schemes and human experts for patient-level classification in the external validation set

	S1	S2	S3	S4	S5	R1	R2	R3	R4	R5	Plain	Trinary
IP recall	45	45	42	41	50	34	38	43	38	39	39	43
NCP recall	43	48	41	40	31	31	33	37	32	30	33	44
Precision (%)	89.6	90.6	83.7	81.6	100.0	66.0	73.3	84.1	72.7	73.2	61.9	76.7
Sensitivity (%)	75.4	84.2	71.9	70.2	54.4	54.4	57.9	64.9	56.1	52.6	78	86
Specificity (%)	90	90	84	82	100	68	76	86	76	78	57.9	77.2
F1 score	0.819	0.873	0.774	0.755	0.705	0.596	0.647	0.733	0.634	0.612	0.690	0.811
Average F1 score	0.790 (specialists)					0.640 (residents)					–	–

Generally, the specialist group (S1–S5) outperformed the resident group (R1–R5) while the Trinary scheme outperformed the Plain scheme for patient-level classification in the external validation set. The performance of the Trinary scheme was close to the specialist group.

Table 7 Correlation of radiologists in the specialist group (S1–S5) and the resident group (R1–R5) for patient-level classification in the external validation set

	S2	S3	S4	S5	R1	R2	R3	R4	R5
S1	0.76	0.76	0.64	0.62	0.41	0.56	0.66	0.47	0.29
S2		0.67	0.55	0.65	0.33	0.48	0.58	0.35	0.33
S3			0.70	0.65	0.51	0.59	0.68	0.41	0.32
S4				0.53	0.43	0.43	0.49	0.53	0.28
R1					0.31	0.37	0.51	0.35	0.22
R2						0.58	0.56	0.45	0.35
R3							0.71	0.52	0.22
R4								0.58	0.20
R5									0.24

The highest correlation between radiologists appeared in the comparison of specialist 1 and 2 (0.759). Generally, members in the specialist group presented a higher correlation with each other than members in the specialist group or inter-group members.

identified by the Trinary and Plain schemes, respectively. The differences between the two schemes were statistically significant with P=0.0076 and P<2.2e-16 for IP and NCP lesions, respectively. Therefore, the Trinary scheme was more consistent with specialists than the Plain scheme on lesion-level classification.

Performance of patient-level classification by human experts

For the external validation set, an average of 8.1 h was required for 10 radiologists to differentiate NCP and IP in all 107 patients. For the specialist group, the average time was 6.8 h (5.3–7.4 h). For the resident group, the average time was 8.9 h (7.5–10.2 h). For the specialist group, micro-average sensitivity and specificity of 71.2% (95% CI, 57.8–84.7%) and 89.2% (95% CI, 80.5–97.9%) were reached, respectively. For the resident group, the mean sensitivity and specificity were 57.2% (95% CI, 51.3–63.1%) and 76.8% (95% CI, 68.8–84.8%), respectively. The mean AUC of the specialist group was 0.802, which was significantly better than that of the resident group (AUC 0.67, P<0.0001). The mean F1 scores for the specialist group and the resident group were 0.79 and 0.64, respectively (*Table 6*). Both the specialist group and the resident group were found to have good consistency, with ICCs of 0.899 and 0.798, respectively. The correlation for 10 radiologists is presented in *Table 7*.

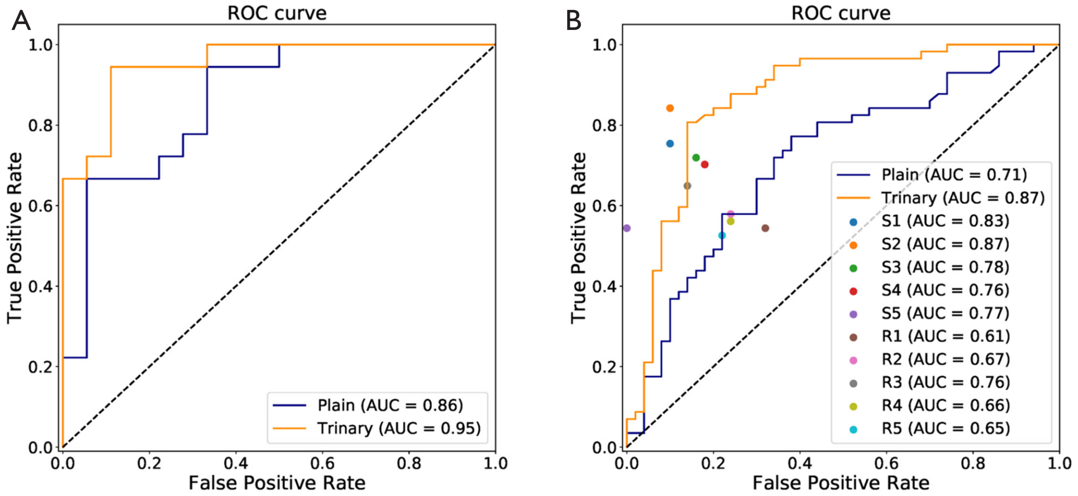

Figure 9 ROCs of patient level classification. (A) The performance of the Trinary and Plain scheme for differential diagnosis of NCP and IP patients in the test set. (B) AUCs of deep learning schemes and human experts for differential diagnosis of the external validation set. The Trinary scheme (AUC 0.87) performed better than the Plain scheme (AUC 0.71) and achieved a specialist-level (S1–S5) performance, while the Plain scheme provided similar capability to the resident group (R1–R5). AUC, area under the curve; ROC, receiver operating characteristic; IP, influenza pneumonia; NCP, novel coronavirus pneumonia.

Table 8 Performance of the Plain and Trinary schemes for patient-level classification in the test set

	Accuracy	Precision	Sensitivity	Specificity	F1 score
Plain scheme	72.2%	72.2%	72.2%	72.2%	0.72
Trinary scheme	91.7%	94.1%	88.9%	94.4%	0.91

Trinary scheme outperformed the specialist group on patient-level classification

For the Plain and Trinary schemes, it took an average of 10 seconds to detect and classify all detected lesions for a single patient. The receiver operating characteristic (ROC) curves for the test set and the external validation set of both training schemes are shown in *Figure 9*. For the test set, the Plain and Trinary schemes achieved AUCs of 0.86 and 0.95, respectively (*Figure 9A*). The AUCs were similar those for the lesion level. For the Plain and Trinary schemes, the sensitivities were 72.2% and 88.9%, respectively, and the specificities were 72.2% and 94.4%, respectively (*Table 8*).

For the external validation set, the sensitivity and specificity of the Plain scheme were 78% and 57.9%, respectively, compared with 86% and 77.2%, respectively, for the Trinary scheme. Compared with the mean sensitivity (71.2%) and specificity (89.2%) of 5 specialists, the deep learning model achieved higher sensitivity but lower specificity. The F1 scores for the Trinary scheme and Plain scheme reached 0.690 and 0.811, respectively, and the Trinary outperformed 9 human radiologists. Generally, the Trinary scheme achieved specialist-level performance while the Plain scheme achieved resident-level performance (*Figure 9B*). Our Trinary scheme correctly classified 13 (22.8%) participants with NCP that were misdiagnosed by at least 3 specialists. Among them, 2 patients displayed CT findings less frequently reported in other NCP cases, such as a small mixed ground-glass opacity in the center, or solitary consolidation. These patients were misdiagnosed by 5 specialists (*Figure 10*). The results indicated that both of the schemes achieved human expert level (*Table 6*). The Trinary scheme achieved an F1 score of 0.811, which was higher than that of the Plain scheme (0.690), and also achieved a similar level of performance to the human experts (specialist group, average 0.790; resident group, average 0.640) (*Table 6*).

The correlations between the predictions of the two

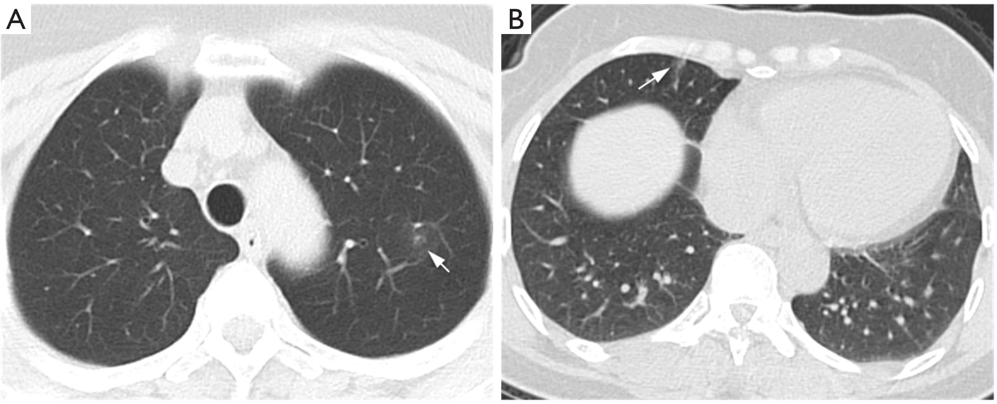

Figure 10 CT images of two NCP patients misdiagnosed by specialists but correctly classified by the Trinary scheme. Axial CT plain scan in patients suffering from NCP who were misdiagnosed as IP by the specialist panel. (A) A 50-year-old female with a small piece of pure GGO in left upper lobe (arrow); she was twice found negative by nucleic acid tests but tested positive the third time (predicted probability: 0.573); (B) A 49-year-old female also with a small GGO (arrow) in right middle lobe (predicted probability 0.56). Their common characteristics were either small lesions located in one segment or a lack of specificity of typical findings in NCP. The patients above were correctly classified as NCP by the Trinary scheme (predicted probability represented the probabilities of being NCP by the Trinary scheme). CT, computed tomography; GGO, ground-glass opacity; NCP, novel coronavirus pneumonia.

Table 9 Correlation between the two schemes and the specialist group (S1–S5) and the resident group (R1–R5) in the external validation set for patient-level classification

	Plain scheme	Trinary scheme
S1	0.39	0.61
S2	0.35	0.63
S3	0.26	0.48
S4	0.11	0.33
S5	0.09	0.42
R1	0.10	0.21
R2	0.33	0.44
R3	0.03	0.50
R4	0.19	0.31
R5	0.12	0.33

The Trinary scheme showed stronger correlation with both the human groups than the Plain scheme, especially with the specialist group.

schemes and radiologists were further examined in the external validation set (*Table 9*). The Trinary scheme outperformed the Plain scheme, as higher correlations were observed in all radiologists. Generally, both the Plain and the Trinary scheme were better correlated with 5 experienced specialists, especially with specialist 1 and specialist 2 in both categories. Specialists 1 and 2 performed best in differential diagnosis for the external validation set; therefore, the Trinary scheme was better correlated with specialists in the external validation set.

Discussion

In this study, we established an integrated artificial intelligence (AI) framework for differential diagnosis between NCP and IP, which showed a good performance in the automatic detection of disease lesions, classification of lesions, and differential diagnosis of patients even at the early stage. This AI effort was driven by the phenomenon of misdiagnosis of influenza in COVID-19 deaths in multiple countries during the escalating pandemic. The similarities in clinical symptoms between these two types of VP, along with the shortage and high false negative rate of nucleic acid detection kits, make the differential diagnosis difficult (26-28) and has prompted the search for new diagnostic methods. Meanwhile, the high sensitivity of chest CT gives it an advantage in diagnosing NCP at an early stage, which encouraged the Hubei Provincial Government to adopt characteristic chest CT findings as an important criterion for diagnosis of NCP at the peak of outbreak. In the present climate, our AI framework indicated improvements in accuracy and speed in identifying specific lesions on chest CT images, and thus provides another predominant diagnostic tool to assist clinicians and radiologists.

Recent studies have reported the value of deep learning in differentiating NCP from various other types of pneumonia (29-31). In this study, we focused on discriminating NCP and pneumonia infected with influenza virus, which present extremely similar symptoms and demonstrations. Both viruses can be quickly transmitted through person-to-person contact, droplets, and fomites. The serial interval for SARS-CoV-2 is estimated to be 6–9 days, compared with 3 days for influenza (32,33). Further, pre-symptomatic viral transmission is possible in both types of viruses, which increases the difficulty of precautionary measures. While there are several clinical trials and more than 20 vaccines in development for SARS-CoV-2, there is currently no licensed vaccine or antiviral medication for NCP. On the contrary, medicaments and vaccines are available for influenza, but some patients who have been vaccinated can still be infected by influenza, and early use of antivirals is essential (34). The high transmissibility, lethality, and difficulty in treating both viruses make the early clinical identification particularly important.

Unfortunately, the diagnosis of VP still relies heavily on clinician suspicion, which is based on host risk factors, presentation, and exposure, and thus makes different subtypes of viruses rather indistinguishable. Patients who are infected with influenza or SARS-CoV-2 could show a similar range of symptoms, such as sudden onset fever, cough, sore throat, fatigue, and myalgia (1,35). Laboratory findings also show similar results, including normal or lower levels of leucocytes, decreased lymphocyte count, increased C-reactive protein, lower albumin, and higher D-dimer and urea levels (36). Despite some efforts to compare hematological indices between patients with NCP and IP, it is not possible to differentiate based on them due to individual biases. Moreover, patients with severe pneumonia may present with abnormal laboratory findings of damage to multiple organs, including to the heart, liver, or kidney. In a prospective study, clinicians failed to clinically diagnose influenza in approximately 2/3 of influenza-confirmed patients (37), not to mention the difficulty in differentiating it from other types of respiratory viruses on clinical grounds.

Meanwhile, despite the fact that the advent of PCR testing has enormously facilitated the identification of respiratory viruses, the test results usually take days to return, and involve an unavoidable false negative rate (10). In view of that, closer attention to patients' radiological features is needed from clinicians and radiologists in the search for evidence of influenza or SARS-CoV-2 infection.

Our integrated AI framework provides the possibility for early and differential diagnosis of NCP. Previous research has shown that at the early stage, NCP is difficult to distinguish from IP. Our integrated framework performed well in the automatic detection of disease lesions, the classification of lesions, and patient differential diagnosis, even at the early stage.

Deep learning also suffers from the "black box" problem, which essentially means the results are difficult to explain (38-40). One option is to explain black boxes using various techniques (41), and another is to develop explainable models (42). Although the deep learning pipelines proposed in this work are based on black box models, taking the average of individual lesions to get patient-level prediction obtained some extent of interpretability. The proposed Trinary scheme behaved more similarly to specialists than the Plain deep learning classification model did, providing extra interpretability to the pipeline. The Trinary classification scheme is designed to improve the network from extracting device-specific features during learning. The idea behind the scheme is that if the network only extracted device-specific features, it would lead to a high cost on the random region inputs. Therefore, the scheme forced the trained model to instead extract lesion-specific features. Although it is possible that device-specific features could also be extracted, we demonstrated that such a scheme could clearly improve the classification results and make prediction by the model more consistent with the judgement of experienced specialists. Because the proposed Trinary scheme is task independent, we believe that it can be widely applied to many other medical image classification problems and accelerate the application of deep learning systems to wider clinical usage.

In our study, we found that chest CT images of early-stage NCP manifest a more obvious distribution of ground-glass opacity (GGO) in the lungs, with fewer nodules and consolidation than IP. This result was also confirmed by a recent study comparing CT characteristics between IP and NCP, with NCP showing patchy areas or GGO combined consolidation opacities, with peripheral distribution and balanced lobe predomination, while in IP the lesions were predominantly located in the inferior lobe (7). These findings support the point that chest CT could play an early warning role in the diagnosis and differentiation of NCP, and also provide a basis for further research. At the same time, these subtle differences are quite difficult to uncover through artificial reading and can be easily missed, which makes it necessary to establish an AI model to assist clinical work based on machine learning. We are obliged to admit the fact that the exposure to ionizing radiation is an intrinsic defect of CT. However, compared with other radiological modalities, the role of CT in the detection of NCP is irreplaceable, not only in diagnosis but also in assessing the disease severity, as it provides critical guidance for patients' treatment. In fact, accumulated evidence has demonstrated the risk related to routine usage of CT in medical care to be low (43,44). Besides, the carcinogenic risk from CT screening for lung cancer, even if non-negligible, should be considered acceptable in light of the substantial reduction in screening-associated mortality (45).

Further analysis of our model results can help us to obtain a better understanding of the pathophysiological origin of NCP. It is presumed that SARS-CoV-2 is able to bind to the ACE-2 receptor in humans, a key component of the renin-angiotensin system (46,47). The ACE-2 receptor is mainly expressed on the surface of alveolar type-2 epithelial cells which, when infected, undergo apoptosis, leading to diffuse alveolar damage and interstitial fluid absorption disorder. This may explain why we found that 80.4% of lesions in NCP were <–500 Hu, 86.5% of patients had bilateral lung damage, and 35.0% of patients had all 5 lung lobes affected. Unlike SARS-CoV-2, influenza viruses primarily damage the tracheal epithelial cells, leading to necrotizing bronchitis and diffuse alveolar damage of the upper respiratory tract (10). This suggests that the trachea and main bronchus should be affected first, resulting in bronchitis and neutrophilic bronchopneumonia at the early stage of infection.

The clinical applicability of our developed AI model was improved by the inclusion of data from multiple centers and machines. In the current study, the training and test data of NCP were from 4 machines in 3 hospitals, and the AI model performed well. The data of our independent verification group were from different machines in 8 hospitals, and both schemes performed well, which suggests that our model has good clinical applicability.

Most of the existing deep learning studies on pneumonia were performed using chest X-ray due to its easy accessibility (11,14,48). However, CT is the major clinical diagnostic method for NCP. A natural concern for the current study was therefore the potentially scarce sample size from CT, which would have supplied insufficient training data for the deep learning models. However, each CT scan can generate multiple images for analysis, and the amount of data can be further increased by manipulations such as image rotations to avoid overfitting. Overall, the performance of our AI system in a multi-center on the independent validation group from multiple centers and machines demonstrates the reliability of our deep learning results.

Conclusions

Currently, SARS-CoV-2 is rampantly spreading around the world, and efficient and accurate diagnosis of NCP is crucial for prevention and control. Our deep learning model potentially supplies an accurate early diagnostic tool for NCP, especially when nucleic acid test kits are in short supply, which commonly happens during outbreaks. This tool could help to reduce the rate of misdiagnosis and diagnosis time, ensure prompt patient isolation and early treatment, improve prognosis, and considerably reduce transmission. The high efficiency of our model in differentiating NCP from IP could be very beneficial to reducing the rate of misdiagnosis and optimizing the allocation of medical resources, particularly in areas with a high prevalence of both NCP and IP. The Trinary scheme not only improved the performance of the model in discriminating NCP from IP, but it also behaved more similarly to specialists than the Plain scheme. Because the proposed Trinary scheme is designed for general use, it could potentially be applied for classifying a wide range of medical images.

Acknowledgments

Funding: This work was supported by National Key R&D Program of China [2017YFC1309701 and 2017YFC1309700 to MZ], Shanghai Key Discipline for Respiratory Diseases [2017ZZ02014 to JQ] and Shanghai Science and Technology Committee [18411953000 to HZ] and Zhejiang University special scientific research fund for COVID-19 prevention and control.

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at http://dx.doi.org/10.21037/atm-20-5328

Data Sharing Statement: Available at http://dx.doi.org/10.21037/atm-20-5328

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-20-5328). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was reviewed and approved by the Ruijin Hospital Ethics Committee (2017-186). The requirement for written informed consent was waived due to the retrospective nature of this study.

References

1. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 2020; 382:1199-207.
2. 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;323:1061-9.
3. Ly S, Arashiro T, Ieng V, et al. Establishing seasonal and alert influenza thresholds in Cambodia using the WHO method: implications for effective utilization of influenza surveillance in the tropics and subtropics. Western Pac Surveill Response J 2017;8:22-32.
4. Livingston E, Bucher K, Rekito A. Coronavirus Disease 2019 and Influenza. JAMA 2020;323:1122.
5. Chen L, Han XD, Li YL, et al. Severity and outcomes of influenza-related pneumonia in type A and B strains in China, 2013-2019. Infect Dis Poverty 2020;9:42.
6. Kilbourne ED. Influenza pandemics of the 20th century. Emerg Infect Dis 2006;12:9-14.
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 2020;296:E32-40.
8. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020;395:565-74.
9. Wang W, Xu Y, Gao R, et al. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. JAMA 2020; 323:1843-4.
10. Franquet T. Imaging of pulmonary viral pneumonia. Radiology 2011;260:18-39.
11. Hwang EJ, Park S, Jin KN, et al. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. JAMA Netw Open 2019;2:e191095.
12. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Med 2018;15:e1002686.
13. Walsh SLF, Calandriello L, Silva M, et al. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir Med 2018;6:837-45.
14. Zech JR, Badgeley MA, Liu M, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Med 2018;15:e1002683.
15. Chung M, Bernheim A, Mei X, et al. CT Imaging Features of 2019 Novel Coronavirus (2019-nCoV). Radiology 2020;295:202-7.
16. Cao B, Huang Y, She DY, et al. Diagnosis and treatment of community-acquired pneumonia in adults: 2016 clinical practice guidelines by the Chinese Thoracic Society, Chinese Medical Association. Clin Respir J 2018;12:1320-60.
17. Mandell LA, Wunderink RG, Anzueto A, et al. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. Clin Infect Dis 2007;44 Suppl 2:S27-72.
18. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement (serial online) 2018 April (Cited 2020 Mar 19). Available online: https://arxiv.org/abs/1804.02767
19. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition (serial online) 2014 Sept (Cited 2020 Mar 19). Available online: http://arxiv.org/abs/1409.1556
20. Boer P, Kroese D, Mannor S, et al. A Tutorial on the Cross-Entropy Method. Ann Oper Res 2005;134:19-67.
21. Lin TY, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell

2020;42:318-27.
22. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019;1:e271-97.
23. Toba S, Mitani Y, Yodoya N, et al. Prediction of pulmonary to systemic flow ratio in patients with congenital heart disease using deep learning-based analysis of chest radiographs. JAMA Cardiol 2020;5:449-57.
24. Sagi O, Rokach L. Ensemble learning: A survey. Wiley Interdiscip Rev Data Min Knowl Discov 2018;8:e1249.
25. Breiman L. Random Forests. Mach Learn 2001;45;5-32.
26. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020;395:507-13.
27. Ruuskanen O, Lahti E, Jennings LC, et al. Viral pneumonia. Lancet 2011;377:1264-75.
28. Guan WJ, Ni ZY, Liang WH, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med 2020;382:1708-20.
29. Wang S, Zha Y, Li W, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 2020;56:2000775.
30. Ni Q, Sun ZY, Qi L, et al. A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images. Eur Radiol 2020;30:6517-27.
31. Li L, Qin L, Xu Z, et al. Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT. Radiology 2020;296:200905.
32. Sanche S, Lin YT, Xu C, et al. High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg Infect Dis 2020;26:1470-7.
33. Biggerstaff M, Cauchemez S, Reed C, et al. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: A systematic review of the literature. BMC Infect Dis 2014;14:480.
34. Orzeck EA, Shi N, Blumentals WA. Oseltamivir and the risk of influenza-related complications and hospitalizations in patients with diabetes. Clin Ther 2007;29:2246-55.
35. Stein J, Louie J, Flanders S. Performance characteristics of clinical diagnosis, a clinical decision rule, and a rapid influenza test in the detection of influenza infection in a community sample of adults. Ann Emerg Med 2005;46:412-9.
36. Yun H, Sun Z, Wu J, et al. Laboratory data analysis of novel coronavirus (COVID-19) screening in 2510 patients. Clin Chim Acta 2020;507:94-7.
37. Dugas AF, Valsamakis A, Atreyam MR, et al. Clinical diagnosis of influenza in the ED. Am J Emerg Med 2015;33:770-5.
38. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 2019;29:102-27.
39. Sahiner B, Pezeshk A, Hadjiiski LM, et al. Deep learning in medical imaging and radiation therapy. Med Phys 2019;46:e1-36.
40. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44-56.
41. Wang F, Casalino LP, Khullar D. Deep Learning in Medicine-Promise, Progress, and Challenges. JAMA Intern Med 2019;179:293-4.
42. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1;206-15.
43. Kuo W, Ciet P, Tiddens HA, et al. Monitoring cystic fibrosis lung disease by computed tomography. Radiation risk in perspective. Am J Respir Crit Care Med 2014;189:1328-36.
44. Marant-Micallef C, Shield KD, Vignat J, et al. The risk of cancer attributable to diagnostic medical radiation: Estimation for France in 2015. Int J Cancer 2019;144:2954-63.
45. Rampinelli C, De Marco P, Origgi D, et al. Exposure to low dose computed tomography for lung cancer screening and risk of cancer: secondary analysis of trial data and risk-benefit analysis. BMJ 2017;356:j347.
46. Wan Y, Shang J, Graham R, et al. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. J Virol 2020;94:e00127-20.
47. Zhao Y, Zhao Z, Wang Y, et al. Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov. bioRxiv 2020. doi: 10.1101/2020.01.26.919985
48. Kermany, DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 2018;172:1122-31.e9.

Cite this article as: Zhou M, Yang D, Chen Y, Xu Y, Xu JF, Jie Z, Yao W, Jin X, Pan Z, Tan J, Wang L, Xia Y, Zou L, Xu X, Wei J, Guan M, Yan F, Feng J, Zhang H, Qu J. Deep learning for differentiating novel coronavirus pneumonia and influenza pneumonia. Ann Transl Med 2021;9(2):111. doi: 10.21037/atm-20-5328