

## RESEARCH ARTICLE

# An original deep learning model using limited data for COVID-19 discrimination: A multicenter study

Fangyi Xu<sup>1</sup> | Kaihua Lou<sup>1</sup> | Chao Chen<sup>1</sup> | Qingqing Chen<sup>1</sup> | Dawei Wang<sup>2</sup> |  
 Jiangfen Wu<sup>2</sup> | Wenchao Zhu<sup>1</sup> | Weixiong Tan<sup>2</sup> | Yong Zhou<sup>3,4</sup> | Yongjiu Liu<sup>5</sup> |  
 Bing Wang<sup>5</sup> | Xiaoguo Zhang<sup>6</sup> | Zhongfa Zhang<sup>6</sup> | Jianjun Zhang<sup>7</sup> |  
 Mingxia Sun<sup>7</sup> | Guohua Zhang<sup>8</sup> | Guojiao Dai<sup>8</sup> | Hongjie Hu<sup>1</sup>

<sup>1</sup>Department of Radiology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China

<sup>2</sup>Institute of Advanced Research, Infervision Medical Technology Co. Ltd., Beijing, China

<sup>3</sup>Department of Pulmonary and Critical Care Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China

<sup>4</sup>China National Respiratory Regional Medical Center (East China), Hangzhou, China

<sup>5</sup>Department of Radiology, Jingmen No.1 People's Hospital, Jingmen, China

<sup>6</sup>Department of Respiratory Medicine, Jinan Infectious Disease Hospital, Shandong University, Jinan, China

<sup>7</sup>Department of Radiology, Zhejiang Hospital, Hangzhou, China

<sup>8</sup>Department of Radiology, Taizhou No.1 People's Hospital, Taizhou, China

## Correspondence

Hongjie Hu, Department of Radiology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, No.3, Qingchun East Road, Shangcheng District, Hangzhou, Zhejiang, 310016, China.  
 Email: hongjiehu@zju.edu.cn

Fangyi Xu and Kaihua Lou contributed equally to this work.

## Funding information

Zhejiang Provincial Natural Science Foundation of China, Grant/Award Number: 2020XGZX051; Zhejiang University special scientific research fund for COVID-19 prevention and control, Grant/Award Number: LQ20F030018; Beijing Bethune Charitable Foundation Run Wu You Sheng, Grant/Award Number: BJ-RW2020006J; National Natural Science Foundation of China, Grant/Award Number: 82071988; Key Research and Development Program of Zhejiang Province, Grant/Award Number: 2019C03064; Program Co-sponsored by Province and Ministry, Grant/Award Number: WKJ-ZJ-1926

## Abstract

**Objectives:** Artificial intelligence (AI) has been proved to be a highly efficient tool for COVID-19 diagnosis, but the large data size and heavy label force required for algorithm development and the poor generalizability of AI algorithms, to some extent, limit the application of AI technology in clinical practice. The aim of this study is to develop an AI algorithm with high robustness using limited chest CT data for COVID-19 discrimination.

**Methods:** A three dimensional algorithm that combined multi-instance learning with the LSTM architecture (3DMTM) was developed for differentiating COVID-19 from community acquired pneumonia (CAP) while logistic regression (LR), k-nearest neighbor (KNN), support vector machine (SVM), and a three dimensional convolutional neural network set for comparison. Totally, 515 patients with or without COVID-19 between December 2019 and March 2020 from five different hospitals were recruited and divided into relatively large (150 COVID-19 and 183 CAP cases) and relatively small datasets (17 COVID-19 and 35 CAP cases) for either training or validation and another independent dataset (37 COVID-19 and 93 CAP cases) for external test. Area under the receiver operating characteristic curve (AUC), sensitivity, specificity, precision, accuracy, F1 score, and G-mean were utilized for performance evaluation.

**Results:** In the external test cohort, the relatively large data-based 3DMTM-LD achieved an AUC of 0.956 (95% confidence interval, 95% CI, 0.929~0.982) with 86.2% and 98.0% for its sensitivity and specificity. 3DMTM-SD got an AUC of 0.937 (95% CI, 0.909~0.965), while the AUC of 3DCM-SD decreased dramatically to 0.714 (95% CI, 0.649~0.780) with training data reduction. KNN-MMSD, LR-MMSD, SVM-MMSD, and 3DCM-MMSD benefited significantly from the inclusion of clinical information while models trained with relatively large dataset got slight performance improvement in COVID-19 discrimination.

3DMTM, trained with either CT or multi-modal data, presented comparably excellent performance in COVID-19 discrimination.

**Conclusions:** The 3DMTM algorithm presented excellent robustness for COVID-19 discrimination with limited CT data. 3DMTM based on CT data performed comparably in COVID-19 discrimination with that trained with multi-modal information. Clinical information could improve the performance of KNN, LR, SVM, and 3DCM in COVID-19 discrimination, especially in the scenario with limited data for training.

#### KEYWORDS

artificial intelligence, coronavirus disease 2019, deep learning, spiral computed, tomography

## 1 | INTRODUCTION

The novel coronavirus disease 2019 (COVID-19) has spread as a pandemic all over the world since its first outbreak in the late of 2019, with great threats and economic implications to human life.<sup>1</sup> As of February 2021, there have been more than 110 million confirmed cases worldwide with almost 2.5 million deaths included according to the latest report from the World Health Organization.<sup>2</sup> Presently, the reverse transcriptase polymerase chain reaction (RT-PCR) is widely used for the diagnosis of patients with COVID-19.<sup>3</sup> Nevertheless, RT-PCR might not be sensitive enough for COVID-19 screening, especially for early detection of the suspicious patients.<sup>4-7</sup> As a fast imaging technology, computed tomography (CT) could show the pulmonary structure and certain abnormalities of patients rapidly without any invasive operations, which had been proved to be able to provide complement information for early detection in suspicious COVID-19 patients and severity assessment in confirmed cases.<sup>5,8-10</sup> However, demand for chest CT examinations in COVID-19 screening among highly suspected cohorts increased the interpretation burden of radiologists dramatically and led to certain consumption of limited medical resource in emergent scenarios. Furthermore, COVID-19 could present heterogeneous imaging findings and may share some similar radiological features with pneumonia caused by other infection, making it challenging to discriminate between COVID-19 and other types of pneumonia.<sup>5</sup>

Recently, artificial intelligence (AI) is developing rapidly and has been extensively applied to clinical settings to do medical tasks, for example, the pulmonary nodule detection, the cerebral hemorrhage prediction, the malignancy identification of mass in human anatomic organs and the treatment management and prognosis prediction of tumor.<sup>11-15</sup> Regarding the COVID-19 diagnosis, AI has been proved to be a highly efficient and accurate tool.<sup>16</sup> Several studies have demonstrated the promise of machine learning and deep learning in COVID-19 relevant investigations.<sup>17-21</sup> A deep learning algorithm was developed with 19291 CT

scans from 14435 pneumonia patients with or without COVID-19 and achieved an accuracy of 94% for lesion detection in validation cohorts.<sup>19</sup> In another study, 1381 patients were used to build an automated radiomics CT signature for COVID-19 detection, which had an area under the receiver operating characteristic curve (AUC) of 0.882 (95% CI, 0.851~0.913) in the test cohort consisting of 641 patients.<sup>22</sup>

However, previous AI studies on COVID-19 usually required either enough label force or a large number of targeted cases for algorithm development, which was physically and emotionally exhausting. Considering the certain radiological similarity between COVID-19 and community-acquired pneumonia (CAP), specific clinical features like laboratory test results might provide critical supplemental information for COVID-19 diagnosis,<sup>1</sup> but the diversity of laboratory tests and the validity of responding results increased the difficulty of data collection, which to some extent limited its use in the field of COVID-19-related AI studies.

Therefore, the purpose of this study was to construct a diagnostic algorithm with high robustness using limited multi-modal data for the discrimination between COVID-19 and CAP.

## 2 | MATERIALS AND METHODS

The institutional review board of the five hospitals approved this multicenter retrospective study and waived the informed consent since patient information was anonymized to ensure privacy.

### 2.1 | Patient

A sum of 644 patients were enrolled between December 2019 and March 2020 from five different hospitals. The corresponding clinical information and CT data were collected and reviewed. Patients with positive RT-PCR results for severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) were included in

COVID-19 dataset. Patients with positive CT findings but diagnosed as other CAP by negative RT-PCR results since the COVID-19 outbreak were included for CAP dataset.

The exclusion criteria for COVID-19 and CAP datasets were as followed: (1) lack of corresponding laboratory test results; (2) the time interval between RT-PCR test and chest CT scans >14 days; (3) CT images with poor quality. Process of patient enrollment was showed in Figure 1, and detailed information of patient distribution and clinical types was summarized in Table 1 and Appendix S-1.

## 2.2 | Image acquisition

CT scans used in this study were acquired with multi-detector CT (Siemens SOMATOM Definition Flash, Siemens FORCE CT, Siemens Sensation16, Siemens Definition AS 40, Siemens Definition AS 20, GE Light-Speed VCT, GE MEDICAL SYSTEMS OPTIMA CT540, GE MEDICAL SYSTEMS OPTIMA CT660, uCT510). The scanning parameters were as follows: tube voltage, 80–120KV; current, automatic exposure control; reconstruction slice thickness, 1.25 mm; and interslice gap, 1.25 mm. All CT scans were saved in the picture archiving and communication system.

## 2.3 | Study design

A novel weakly supervised algorithm that combined multi-instance learning with the long and short-term memory (LSTM) architecture (MIL-LSTM) was designed for the discrimination between COVID-19 and CAP. The lesion layers in 3D CT scans, instead of one randomly selected slice from averaged groups or all slices in CT scans, were selected as the input instances for this novel 3D-MIL-LSTM (3DMTM) algorithm using a lesion instance generator based on a pneumonia segmentation model (constructed by Infervision Medical Technology Co., Ltd.),<sup>23</sup> so as to reduce the annotation label force and to enhance model performance by extracting more spatial information of lesions. Meanwhile, another three dimensional convolutional neural network (3D CNN) and three classic machine learning algorithms including logistic regression (LR), k-nearest neighbor (KNN), and support vector machine (SVM) were also developed using 3D CT data to validate the feasibility of newly proposed algorithm.

To verify the role of clinical information in identifying COVID-19, clinical and radiological features were also concatenated for training when exploring the effects of multi-modal information on the performance of algorithms in identifying COVID-19. Notably, the impact of

training data size on model performance was also studied by exchanging training and validation cohorts. Figure 1 showed the process of model development. Details of algorithm design were available in Figure 2.

## 2.4 | Data partition and modeling

Data from four of the five hospitals was used for model development and was divided into relatively large and small datasets as either training or validation cohorts through different combinations while the fifth hospital acted as the data supplier of the independent external test cohort. Detailed dataset combinations are showed in Table 2.

Three classic machine learning models, including KNN, SVM, and LR, were trained with selected clinical features or the combination of clinical and radiomics features. We first utilized the relatively large dataset as the training cohort and the relatively small dataset as the validation cohort. In subsequent, training and validation cohorts were switched in order to reveal the robustness of employed algorithms on different sized datasets. A sum of 12 machining learning models was obtained (Table 3).

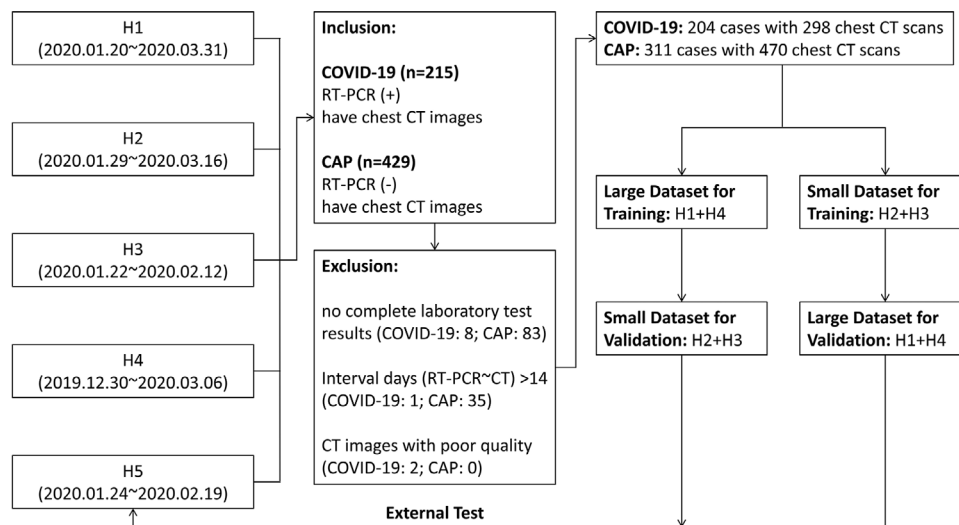
Of note, 3DCM and 3DMTM were both developed with the same procedure and dataset combinations (LD-CT and SD-CT, MMLD and MMSD). The corresponding developed models are listed in Table 3.

## 2.5 | Visualization of lesion features learned by 3DCM and 3DMTM models

To understand how 3DCM and 3DMTM models identified COVID-19, we visualized the most informative regions for these models on CT images using gradient-weighted class activation mapping (Grad-CAM).<sup>24</sup> As an output, attention heat maps were generated to indicate the suspicious area in CT images that contributed most to identify COVID-19.

## 2.6 | Statistical analysis

Area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, precision, accuracy, F1 score, and G-Mean were utilized to evaluate the diagnostic performance of these proposed models. Categorical variables were expressed in terms of frequency and statistically analyzed by chi-square test or Fisher exact probability test. Continuous variables would be analyzed by two-sample *t*-test if they distributed normally with homogeneous variance; if not, Wilcoxon signed rank test would be adopted. Continuous variables are represented by the median (interquartile range, IQR).



**FIGURE 1** Flow diagram of patient enrollment. A sum of 644 patients with or without COVID-19 were collected from five hospitals in this study. Based on inclusion and exclusion criteria, 204 COVID-19 patients (298 CT scans) and 311 CAP patients (470 CT scans) were finally recruited for model development. Patients from four hospitals (H1~4) were used for model development while patients from the fifth hospital (H5) as independent external test data. During model development, large and small datasets were exchanged once from training to validation sets for robustness assessment. CAP, community-acquired pneumonia; COVID-19, coronavirus disease 2019; CT, computed tomography; H1~5, hospital 1~5; RT-PCR, reverse transcriptase polymerase chain reaction

**TABLE 1** Summary of demographic information in recruited patients

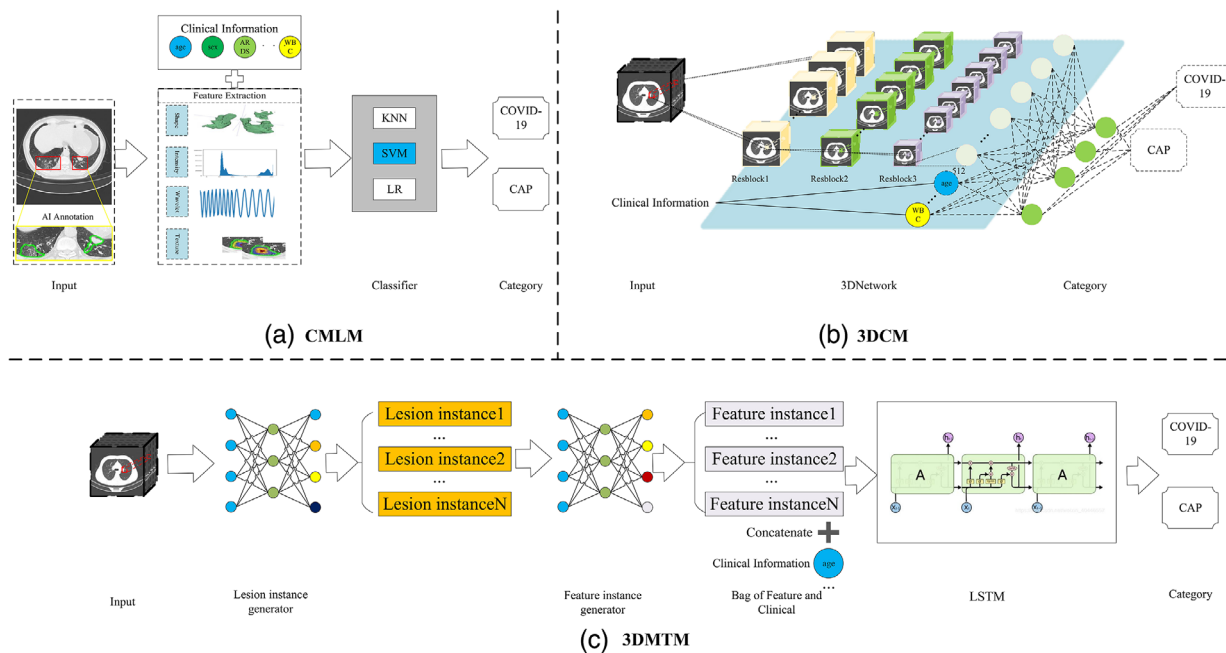
Patients (CT scans)	Total 515 (768)	COVID-19 204 (298)	CAP 311 (470)	p-value
Gender				
Male	270 (52.43%)	97 (47.55%)	173 (55.63%)	0.073
Female	245 (47.57%)	107 (52.45%)	138 (44.37%)	
Age	38.00 (23.00)	44.50 (22.00)	38.50 (26.00)	0.004
Smoking status				
Smoking or ever smoker	52 (10.48%)	12 (5.88%)	40 (13.70%)	<0.001
None	444 (89.52%)	192 (94.12%)	252 (86.30%)	
Fever				
Yes	304 (79.0%)	133 (79.6%)	171 (78.4%)	0.775
None	81 (21.0%)	34 (20.4%)	47 (21.3%)	
Laboratory test				
Leucocytes ( $\times 10^9/L$ ; 3.5–9.5)	6.10 (3.80) <sup>§</sup>	4.60 (1.90)	7.70 (3.40)	<0.001
Neutrophils ( $\times 10^9/L$ ; 1.8–6.3)	3.26 (3.44) <sup>§</sup>	3.00 (1.90)	2.02 (3.14)	0.029
Lymphocytes ( $\times 10^9/L$ ; 1.1–3.2)	1.76 (3.31) <sup>§</sup>	1.22 (0.69)	7.90 (16.60)	<0.001
Lymphocytes percentage (%; 20–50)	14.80 (21.49) <sup>§</sup>	27.00 (16.83)	8.52 (11.06)	<0.001
Eosinophils ( $\times 10^9/L$ ; 0.02–0.52)	0.03 (0.07) <sup>†</sup>	0.01 (0.06)	0.06 (0.10)	<0.001
ALT (U/L; 7–40)	21.00 (19.00) <sup>†</sup>	21.00 (17.00)	20.00 (18.00)	0.189
AST (U/L; 13–35)	22.00 (9.75) <sup>†</sup>	20.00 (10.00)	22.00 (11.00)	0.029
LDH (U/L; 120–250)	195.00 (63.50) <sup>†</sup>	189.50 (72.00)	200.50 (57.00)	0.241
CK-MB (IU/L; 0–24)	11.00 (7.00) <sup>†</sup>	9.00 (4.00)	14.00 (8.80)	<0.001
CRP (mg/L; 0.2–4.0)	11.60 (25.15) <sup>§</sup>	8.75 (11.95)	21.90 (43.85)	<0.001

Note: Data are showed as *n* (%) or median (interquartile range, IQR).

Abbreviations: ALT, alanine transaminase; AST, aspartate aminotransferase; CAP, community-acquired pneumonia; COVID-19, coronavirus disease 2019; CK-MB, creatine kinase isoenzyme-MB; CRP, C-reactive protein; CT, computed tomography; LDH, lactate dehydrogenase.

<sup>†</sup>Patients with available data were less than 360.

<sup>§</sup>Patients with available data were more than 360.



**FIGURE 2** Illustration of the machine learning (ML) Model and deep learning (DL) Models. (a) Classic machine learning models (CMLM) were trained with radiomics features or the combination of clinical and radiomics features to differentiate COVID-19 and community acquired pneumonia (CAP), including k-nearest neighbor (KNN), support vector machine (SVM), and logistic regression (LR). (b) An improved three dimensional convolutional neural network (3D CNN) model (3DCM), which constituted three convolutional blocks of Resnet and three fully connected layers, was employed to distinguish between COVID-19 and CAP using CT images with or without the addition of clinical information. (c) A novel algorithm based on multi-instance learning and long and short-term memory (LSTM) (3DMTM) was proposed for COVID-19 identification. Lesion instance generator enabled efficient selection of instance (slices) with lesions; feature instance generator based on Resnet-18 extracted features from input instances. Clinical information could be concatenated after feature extraction. Long and short-term memory (LSTM) helped obtain the spatial information by combining features from different layers

**TABLE 2** Data partition for model development and validation

	H1		H2		H3		H4		H5	
	CT	CI	CT	CI	CT	CI	CT	CI	CT	CI
Small dataset (SD-CT)			※		※					
Small dataset (SD-CI)				※		※				
Multi-modal small dataset (MMSD)			※	※	※	※				
Large dataset (LD-CT)	※						※			
Large dataset (LD-CI)		※						※		
Multi-modal large dataset (MMLD)	※	※					※	※		
External test set (ETS-CT)									※	
External test set (ETS-CI)										※
Multi-modal external test set (MMETS)									※	※

Abbreviations: CI, clinical information; CT, computed tomography.

A two-sided 95% confidence interval for AUC was constructed following the approach of Hanley and McNeil (1982).<sup>25</sup> Model performance was compared using DeLong test.<sup>26,27</sup> All statistical analyses were performed with the R statistical package (The R Foundation for Statistical Computing, Vienna, Austria). Figures in our study were made with GraphPad Prism 5 (GraphPad Software Inc., San Diego, CA, USA) and R statistical package.  $p < 0.05$  was considered statistically significant.

## 3 | RESULTS

### 3.1 | Patient overview

Totally, 204 patients (298 chest CT scans) with COVID-19 and 311 patients (470 chest CT scans) with CAP were finally recruited for further analysis and model development. Details of the recruited patients are summarized in Table 1 and Appendix S-1. Briefly, most of COVID-19 patients (184 cases, 90.20%) were clinically

**TABLE 3** Data partition for modeling and the corresponding model names

Type	Algorithms	Data	Training set	Test set	Model-name	
CMLM	SVM	(+)Clinical info	MMLD	MMSD	SVM-MMLD	
			MMSD	MMLD	SVM-MMSD	
		(-)Clinical info	LD-CT	SD-CT	SVM-LD	
			SD-CT	LD-CT	SVM-SD	
		KNN	(+)Clinical info	MMLD	MMSD	KNN-MMLD
				MMSD	MMLD	KNN-MMSD
	(-)Clinical info		LD-CT	SD-CT	KNN-LD	
			SD-CT	LD-CT	KNN-SD	
	LR	(+)Clinical info	MMLD	MMSD	LR-MMLD	
			MMSD	MMLD	LR-MMSD	
		(-)Clinical info	LD-CT	SD-CT	LR-LD	
			SD-CT	LD-CT	LR-SD	
3DCM	3D CNN	(+)Clinical info	MMLD	MMSD	3DCM-MMLD	
			MMSD	MMLD	3DCM-MMSD	
		(-)Clinical info	LD-CT	SD-CT	3DCM-LD	
			SD-CT	LD-CT	3DCM-SD	
3DMTM	3D-MIL-LSTM	(+)Clinical info	MMLD	MMSD	3DMTM-MMLD	
			MMSD	MMLD	3DMTM-MMSD	
		(-)Clinical info	LD-CT	SD-CT	3DMTM-LD	
			SD-CT	LD-CT	3DMTM-SD	

Abbreviations: CI, clinical information; CMLM, classic machine learning models; CNN, convolutional neural network; CT, computed tomography; KNN, k-nearest neighbor; LD, large dataset; LR, logistic regression; MMLD, multi-modal large dataset; MMSD, multi-modal small dataset; SD, small dataset; SVM, support vector machine; 3DCM, 3D CNN model; 3DMTM, 3D-MIL-LSTM algorithm.

diagnosed as moderate type while seven (3.43%), six (2.94%), seven (3.43%) were diagnosed as mild, severe, and critical types, respectively. Seven clinical features including gender, age, smoking status, fever, leucocytes, neutrophils, and lymphocytes were selected and normalized with the z-score normalization for further algorithm training.

### 3.2 | Performance evaluation for proposed models trained on relatively large datasets

In addition to the newly proposed 3DMTM algorithm, we also utilized 3D CNN algorithm and classical machine learning models (KNN, SVM, and LR) to identify COVID-19. A relatively large dataset (150 COVID-19 cases with 251 CT scans and 183 CAP cases with 334 CT scans from H1 and H4) and a relatively small dataset (17 COVID-19 cases with 17 CT scans and 35 CAP cases with 35 CT scans from H2 and H3) were firstly utilized as the training and validation datasets, respectively.

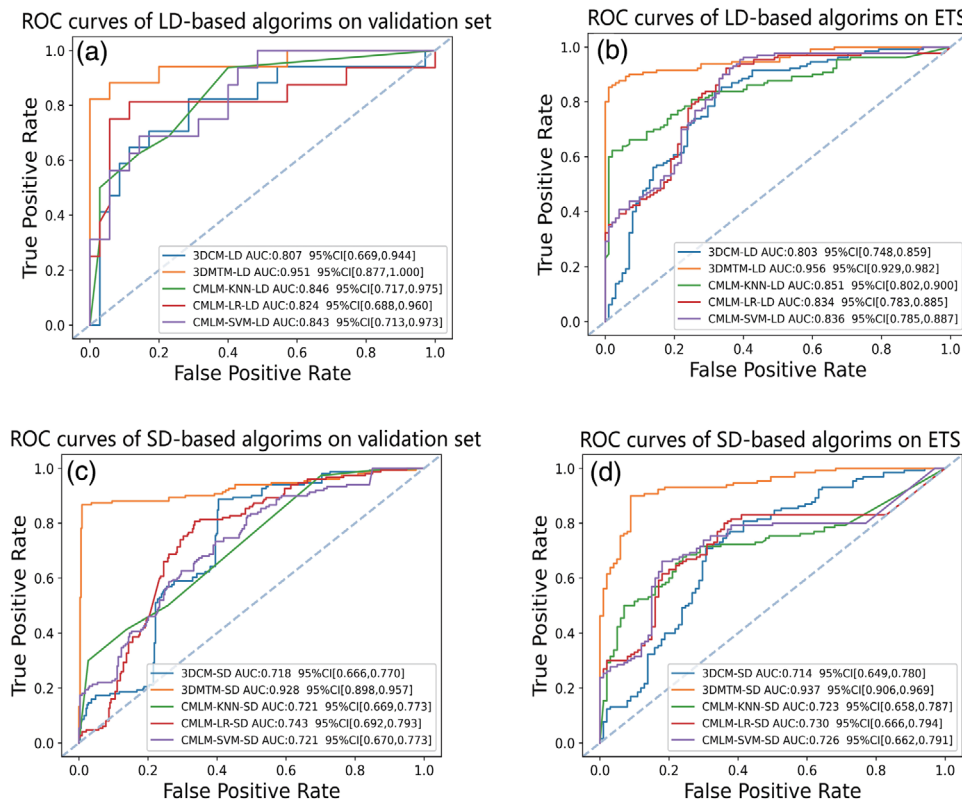
As shown in Figures 3A and 4, KNN-LD, SVM-LD, and LR-LD trained only with radiomics features achieved AUCs of 0.846 (95% CI, 0.717~0.975), 0.843 (95% CI, 0.713~0.973), and 0.824 (95% CI, 0.688~0.960) with

F1-scores of 0.667, 0.688, and 0.788 in the validation cohort, respectively. Meanwhile, AUCs of 3DCM-LD and 3DMTM-LD trained with CT scans reached 0.807 (95% CI, 0.669~0.944) and 0.951 (95% CI, 0.877~1.000) with 0.683 and 0.882 for their F1-scores.

In addition, model performance was further evaluated on another independent external test dataset (37 COVID-19 and 93 CAP cases with totally 231 CT scans from H5, the fifth participated hospital). The highest AUC of 0.956 (95% CI, 0.929~0.982) was achieved by 3DMTM-LD, followed by KNN-LD of 0.851 (95% CI, 0.802~0.900), SVM-LD of 0.836 (95% CI, 0.785~0.887), and LR-LD of 0.834 (95% CI, 0.783~0.885), while 3DCM-LD had the worst performance as well on the external test set (AUC, 0.803, 95% CI, 0.748~0.859) (Figures 3B and 4).

### 3.3 | Performance evaluation for proposed models trained on relatively small datasets

To explore the feasibility of the proposed algorithms in different scenarios, the relatively small and large datasets were switched once as training set to simulate the data-insufficient scenario and to explore the impact



**FIGURE 3** Receiver operating characteristic (ROC) analysis of different models in this study for the discrimination between COVID-19 and community acquired pneumonia (CAP). (a) Performance of large dataset trained models on validation set (SD); (b) performance of large dataset trained models on external test set; (c) performance of small dataset trained models on validation set (LD); (d) performance of small dataset trained models on external test set. KNN, k-nearest neighbor; LD, large dataset; LR, logistic regression; MIL-LSTM, multi-instance learning with the long and short-term memory; SD, small dataset; SVM, support vector machine; 3D CNN, 3 dimensional convolutional neural network; 3DMTM, 3D-MIL-LSTM model

of data size on model performance. Although performance decrease was noted in all small data-based models in the validation cohort, 3DMTM-SD still presented excellent ability in differentiating COVID-19 from CAP (AUC, 0.928, 95%CI, 0.898~0.957) with an increased F1-score of 0.919 (Figures 3C and 4).

In the independent external test cohort, 3DMTM-SD outperformed other small data-based algorithms with a comparable AUC of 0.937 (95% CI, 0.906~0.965) and a F1-score of 0.910 to 3DMTM-LD (Figures 3D and 4 and Appendix S-5). 3DCM-SD showed significantly inferior diagnostic performance with the reduction of training data (Figure 3D and Appendix S-5).

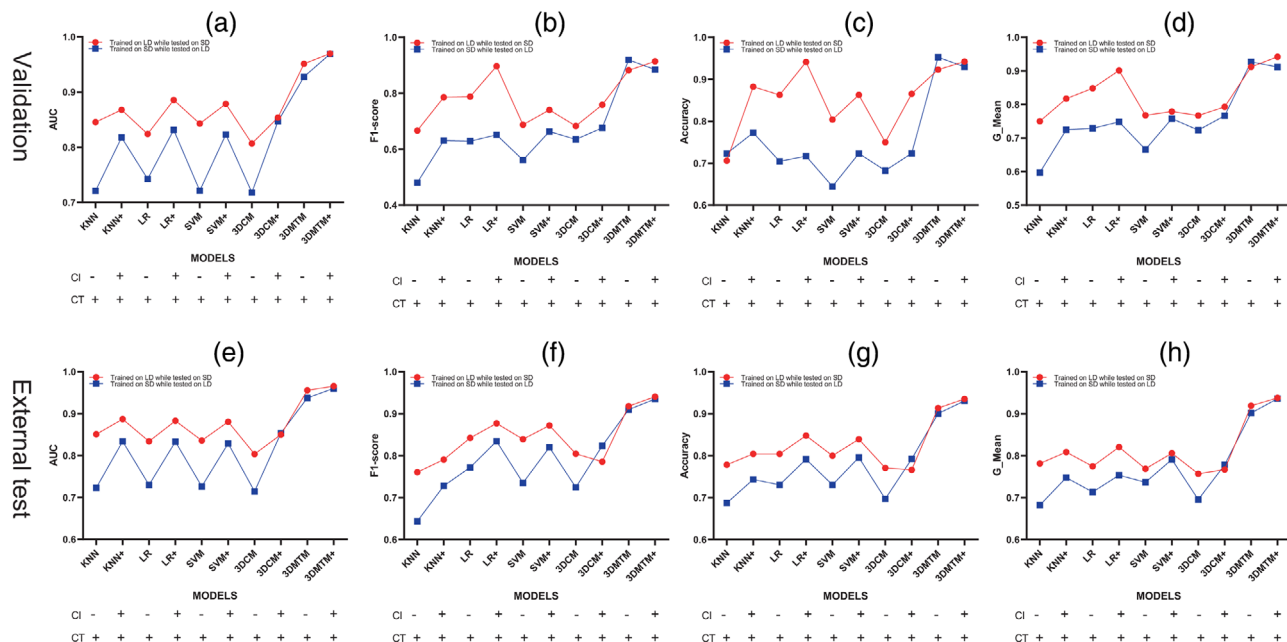
### 3.4 | Enhanced performance of proposed models by training with multi-modal information

Noticing the value of radiological information in identifying COVID-19, we further studied if multi-modal data would improve the model diagnostic performance

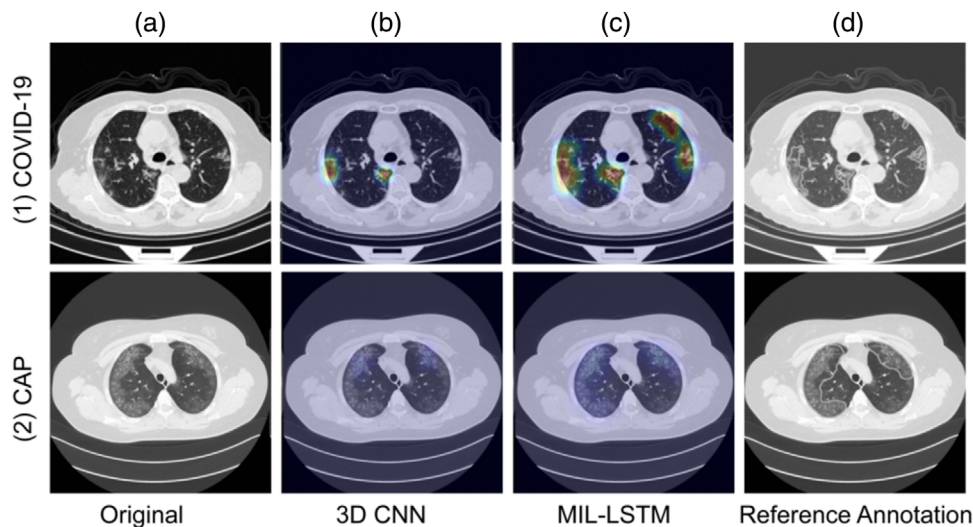
in discriminating between COVID-19 and CAP by combing CT imaging features with selected clinical features. It turned out that all models in our study benefited from the additional clinical features in the validation cohort, no matter which dataset (the relatively small or large datasets) they were trained on (Figure 4). In the external test cohort, the performance of KNN-SD, LR-SD, SVM-SD, and 3DCM-SD got improved dramatically while and 3DMTM benefited slightly from the inclusion of clinical information (Appendix S-5).

### 3.5 | Grad-CAM visualization of 3DCM and 3DMTM-enabled identification of COVID-19

Attention heat maps were generated in our study to interpret the diagnostic process of 3DCM and 3DMTM, which could provide visual information like lesion location and the probability of targeted lesion to be COVID-19. As can be seen in Figure 5, inflammation lesions focused by



**FIGURE 4** Performance evaluation of proposed models in the discrimination between COVID-19 and CAP on validation and external test sets. Area under the receiver operating characteristic curve (AUC), F1-score, accuracy and G-mean were utilized to evaluate the model performance. Of note, two sets of models were developed with switched training and validation sets and both tested on another external test set. (a–d) models performance on validation sets (plotted on diagram according to the metrics values); (e–h) models performance on external test set (plotted on diagram according to the metrics values). CAP, community-acquired pneumonia; CI, clinical information; COVID-19, coronavirus disease 2019; CT, computed tomography; KNN, k-nearest neighbor; LD, large dataset; LR, logistic regression; SD, small dataset; SVM, support vector machine; 3D CNN, 3 dimensional convolutional neural network; 3DMTM, 3D-MIL-LSTM model



**FIGURE 5** Representative of visualized COVID-19 and CAP cases using gradient-weighted class activation mapping (Grad-CAM) on the small dataset trained model in the independent external test set. (a) (1~2), original axial computed tomography (CT) images of COVID-19 and CAP cases; (b) (1~2) and (c) (1~2), attention heat maps generated using Grad-CAM for three dimensional convolutional neural network (3D CNN) and MIL-LSTM in the discrimination between COVID-19 and CAP; (d) (1~2), reference annotation by senior radiologists. 3DMTM detected more inflammation lesions than 3DCM and shared a good consistency with the gold standardization annotated by senior radiologists. CAP, community-acquired pneumonia; COVID-19, coronavirus disease 2019; Grad-CAM, gradient-weighted class activation mapping; MIL-LSTM, multi-instance learning with the long and short-term memory; 3D CNN, 3 dimensional convolutional neural network; 3DMTM, 3D-MIL-LSTM algorithm; 3DCM, 3D CNN model



3DMTM were much larger than that noted by 3DCM and shared a decent consistency with gold standard lesions annotated by senior radiologists.

## 4 | DISCUSSION

In this study, a novel weakly supervised 3DMTM algorithm was developed for the discrimination between COVID-19 and CAP. Compared to the previous studies, this study owned four innovations. First, the original 3DMTM algorithm was developed with limited multi-modal and multicenter data; second, no manual annotation was required for algorithm training; third, we systematically evaluated the performance of 3DMTM, classic machine learning algorithms, and 3D-CNN in identifying COVID-19 from CAP; last, the impact of sample size on the performance of those algorithms was investigated, and an independent external dataset was used to verify the model robustness in this study.

Many scholars have demonstrated the promising value of machine learning or deep learning technology in diagnosis, prognosis prediction, and medical management of COVID-19 since its outbreak.<sup>20,28–31</sup> Li et al. used a dataset consisting of 3322 patients with 4356 chest CT exams to develop a deep learning model, which could fully automatically detect COVID-19 with an AUC of 0.96 in the test set.<sup>4</sup> In another study, which included 1020 chest CT images from 108 COVID-19 patients and 86 non-COVID-19 pneumonia patients, 10 well-known CNNs were trained and showed good performance to differentiate COVID-19 and non-COVID-19 pneumonia with AUCs of 0.894–0.994.<sup>32</sup> Xu et al. established an early screening system to differentiate COVID-19 from influenza-A viral pneumonia (IAPV) and normal patients with 618 CT samples (219 COVID-19, 224 IAPV and 175 normal cases), of which the overall accuracy was up to 86.7% from the perspective of CT cases as a whole.<sup>33</sup>

Compared with previous deep learning researches about COVID-19, which tended to recruit a large number of data or annotation for algorithm training, the novel deep learning algorithm in our study, 3DMTM-LD, was trained with less than 500 chest CT scans (150 COVID-19 cases with 251 CT scans and 183 CAP cases with 334 CT scans) and showed comparable excellent performance for differentiating COVID-19 from CAP in both validation (AUC = 0.951, accuracy = 92.3%) and external test (AUC = 0.956, accuracy = 91.3%) sets. In addition, no manual annotation was required during the model development. What's more, 3DMTM also demonstrated a decent feasibility when trained on the small dataset (17 COVID-19 cases with 17 CT scans and 35 CAP cases with 35 CT scans) and validated on the relatively large dataset, as evidenced by the unaffected diag-

nostic performance of 3DMTM-SD (AUC = 0.928, accuracy = 95.3%). In contrast, an obvious decrease was noted in the performance of 3DCM-SD to differentiate COVID19 from CAP when trained on the relatively small dataset.

The decent robustness of 3DMTM algorithm in differentiating COVID-19 from CAP benefited from its key components consisting of MIL-LSTM architecture. The automatic segmentation algorithm in lesion instance generator enabled efficient selection of instances (slices) with lesions from whole CT scans to improve the signal noise ratio (SNR). MIL, in which labels are associated with bags rather than the instances in the bag, greatly reduces label requirement while CNN is a fully supervised deep learning model that asks for fully labeled samples for training.<sup>34–37</sup> LSTM is one special type of recurrent neural networks (RNNs), and it has better control in long-term memory to reduce the signal loss during the process of conventional RNN architectures and to provide spatial information among layers.<sup>18,38,39</sup>

Thus, the combination of those two algorithms allowed 3DMTM to extract more spatial information with high SNR from targeted lesion without any manual annotation. Especially in the case of insufficient training data, 3DMTM could effectively extract useful information from limited data for training without any manual annotation.

Given that the novel SARS-COV-2 may coexist with human in our daily life for a long time, radiological manifestations may vary with the mutation of virus or the regional divergence in COVID-19 patients all over the world. The robustness of our algorithm with different data size may allow the timely diagnosis and treatment management for those patients with mutated SARS-COV-2 from different regions, which may also have potential value in medical management of rare diseases.

Epidemiological investigations verified the role of clinical information in the diagnosis and management of COVID-19 patients.<sup>1,40–43</sup> Li et al. discovered several new associations between clinical features by reviewing COVID-19 data from 151 published studies and developed an AI model to discriminate COVID-19 from influenza cases with a sensitivity of 92.5% and a specificity of 97.9%.<sup>44</sup> Zhang et al. developed an AI system for the differentiation of COVID-19 from common pneumonia and normal controls with 3777 patients and demonstrated that clinical data could improve the performance of the system in prognosis prediction significantly.<sup>45</sup>

No matter which dataset was used for training and validation, the inclusion of clinical information could improve the diagnostic performance of all models proposed in our study, which confirmed the importance of clinical data for COVID-19 diagnosis. Meanwhile, in the external test cohort, KNN-MMSD, LR-MMSD, SVM-MMSD, and 3DCM-MMSD benefited significantly from

clinical information while models trained with relatively large data just achieved slight enhancement in performance of COVID-19 discrimination, indicating the essential roles of multi-modal information when sample size was limited. Of note, the slightly enhancement of 3DMTM with multi-modal data might result from its ability to effectively extract key and extra spatial information from lesions on CT images, which equalized the impact of multi-modal data on model performance. Considering the difficulty of clinical data collation, the 3DMTM algorithm in this study might be useful in the early screening of COVID-19, especially in the case without comprehensive clinical information.

The black box mechanism of deep learning technology leads to the lack of the transparency of its operation process.<sup>4</sup> To improve the interpretability of deep learning algorithms in this study, attention heat maps were generated using Grad-CAM to indicate suspicious areas that contribute most to the identification of COVID-19.<sup>24</sup> The visualization of 3DCM and 3DMTM was realized in our study to show not only the judgment process of 3DCM and 3DMTM models, but also the more precise recognition of inflammation lesion in CT scans of 3DMTM. Without manual annotation, more lesion area was noted by 3DMTM rather than 3DCM, and a higher SNR was obtained by 3DMTM, which might explain its outstanding diagnostic performance in identifying COVID-19. This visual output provided relatively intuitive information about lesion location and reference proportion in the deep learning process, which might be especially useful for the detection of subtle pathological changes in asymptomatic patients with no obvious macroscopic imaging findings.

There were also several limitations in our study. First, pneumonia could be caused by different factors like bacteria, virus, fungus, and medicine, we only focused on the binary discrimination between COVID-19 and CAP instead of a detailed etiology classification due to the lack of etiological confirmation of CAP cases involved in this study. Second, the 3DMTM algorithm was just trained for COVID-19 diagnosis in our study. Subsequently, we would further expand our data collection for the severity classification, prognosis prediction of COVID-19, and the detailed etiological analysis of pneumonia. Third, 3DMTM was not compared with radiologists in COVID-19 diagnosis, and we would then make a systematic analysis on the potential value of 3DMTM in clinical practice.

In conclusion, the weakly supervised algorithm 3DMTM developed in this study showed excellent robustness in discrimination between COVID-19 and CAP with limited chest CT data. Clinical information could significantly improve the performance of KNN, LR, SVM, and 3DCM in COVID-19 discrimination in the scenario with limited data for training. 3DMTM based on CT data performed comparably in COVID-

19 discrimination with that trained with multi-modal information.

## Abbreviations

AUC	area under the receiver operating characteristic curve
CAP	community acquired pneumonia
CI	confidence interval
COVID-19	coronavirus disease 2019
CT	computed tomography
KNN	k-nearest neighbor
LR	logistic regression
MIL-LSTM	multi-instance learning with the long and short-term memory
RT-PCR	reverse transcriptase polymerase chain reaction
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SVM	support vector machine
3D CNN	three-dimensional convolutional neural network
3DMTM	three-dimensional MIL-LSTM algorithm

## ACKNOWLEDGMENTS

This research was supported by Zhejiang University special scientific research fund for COVID-19 prevention and control under grant number: 2020XGZX051, Zhejiang Provincial Natural Science Foundation of China under grant number: LQ20F030018, Beijing Bethune Charitable Foundation Run Wu You Sheng under grant number: BJ-RW2020006J, National Natural Science Foundation of China under grant number: 82071988, Key Research and Development Program of Zhejiang Province under grant number: 2019C03064, program co-sponsored by Province and Ministry under grant number: WKJ-ZJ-1926.

## CONFLICT OF INTEREST

The authors have no conflict to disclose.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available upon request to the corresponding author.

## REFERENCES

1. 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:P507-P513.
2. Organization WH. WHO coronavirus disease (COVID-19) dashboard[EB/OL]. (2021-02-25). <https://covid19.who.int/>
3. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill*.2020;25(3):2000045.

4. Li L, Qin L, Xu Z, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology*.2020;296(2):E65-E71.
5. Ai T, Yang Z, Hou H, et al. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 Cases. *Radiology*.2020;296(2):E32-E40.
6. Fang Y, Zhang H, Xie J, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology*.2020;296(2):E115-E117.
7. Chae KJ, Jin GY, Lee CS, et al. Positive conversion of COVID-19 after two consecutive negative RT-PCR results: a role of low-dose CT. *Eur J Radiol*.2020;129:109122.
8. Rubin GD, Ryerson CJ, Haramati LB, et al. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the fleischner society. *Chest*.2020;158(1):106-116.
9. Ding X, Xu J, Zhou J, et al. Chest CT findings of COVID-19 pneumonia by duration of symptoms. *Eur J Radiol*.2020;127:109009.
10. Leonardi A, Scipione R, Alfieri G, et al. Role of computed tomography in predicting critical disease in patients with covid-19 pneumonia: a retrospective study using a semiautomatic quantitative method. *Eur J Radiol*.2020;130:109202.
11. Nagarajan N, Yapp E, Le NQK, et al. Application of computational biology and artificial intelligence technologies in cancer precision drug discovery. *Biomed Res Int*.2019;2019:8427042.
12. Wang S, Yang DM, Rong R, et al. Artificial intelligence in lung cancer pathology image analysis. *Cancers (Basel)*.2019;11(11):1673.
13. Huang S, Yang J, Fong S, et al. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. *Cancer Lett*.2020;471:61-71.
14. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*.2019;290(1):218-228.
15. Xu W, Ding Z, Shan Y, et al. A nomogram model of radiomics and satellite sign number as imaging predictor for intracranial hematoma expansion. *Front Neurosci*.2020;14:491.
16. Shi F, Wang J, Shi J, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE Rev Biomed Eng*.2021;14:4-15.
17. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review. *Chaos Solitons Fractals*.2020;139:110059.
18. Sedik A, Iliyasa AM, Abd EB, et al. Deploying machine and deep learning models for efficient data-augmented detection of COVID-19 infections. *Viruses*.2020;12(7):769.
19. 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(12):6517-6527.
20. Javor D, Kaplan H, Kaplan A, et al. Deep learning analysis provides accurate COVID-19 diagnosis on chest computed tomography. *Eur J Radiol*.2020;133:109402.
21. Li Z, Zhong Z, Li Y, et al. From community-acquired pneumonia to COVID-19: a deep learning-based method for quantitative analysis of COVID-19 on thick-section CT scans. *Eur Radiol*.2020;30(12):6828-6837.
22. Guiot J, Vaidyanathan A, Deprez L, et al. Development and validation of an automated radiomic CT signature for detecting COVID-19. *Diagnostics (Basel)*.2020;11(1):41.
23. Zhang X, Wang D, Shao J, et al. A deep learning integrated radiomics model for identification of coronavirus disease 2019 using computed tomography. *Sci Rep*.2021;11(1):3938.
24. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vision*.2020;128(2):336-359.
25. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*.1982;143(1):29-36.
26. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett*.2014;21(11):1389-1393.
27. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*.1988;44(3):837-845.
28. Qin L, Yang Y, Cao Q, et al. A predictive model and scoring system combining clinical and CT characteristics for the diagnosis of COVID-19. *Eur Radiol*.2020;30(12):6797-6807.
29. Bansal A, Padappayil RP, Garg C, et al. Utility of artificial intelligence amidst the COVID 19 pandemic: a review. *J Med Syst*.2020;44(9):156.
30. Ko H, Chung H, Kang WS, et al. COVID-19 Pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation. *J Med Internet Res*.2020;22(6):e19569.
31. Tabatabaei S, Rahimi H, Moghaddas F, et al. Predictive value of CT in the short-term mortality of coronavirus disease 2019 (COVID-19) pneumonia in nonelderly patients: a case-control study. *Eur J Radiol*.2020;132:109298.
32. Ardakani AA, Kanafi AR, Acharya UR, et al. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks. *Comput Biol Med*.2020;121:103795.
33. Xu X, Jiang X, Ma C, et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering (Beijing)*.2020;6(10):1122-1129.
34. Ke-Lei HE, Ying-Huan S, Yang G, et al. A prototype learning based multi-instance convolutional neural network. *Chinese J Comput*.2017;40(6):1265-1274.
35. Xiao Y, Liang F, Liu B. A transfer learning-based multi-instance learning method with weak labels. *IEEE Trans Cybern*.2020;52(1):287-300.
36. Tennakoon R, Bortsova G, Orting S, et al. Classification of volumetric images using multi-instance learning and extreme value theorem. *IEEE Trans Med Imaging*.2020;39(4):854-865.
37. Huang SJ, Gao W, Zhou ZH. Fast multi-instance multi-label learning. *IEEE Trans Pattern Anal Mach Intell*.2019;41(11):2614-2627.
38. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*.1997;9(8):1735-1780.
39. Zhao F, Zhang T, Wu Y, et al. Antidecay LSTM for siamese tracking with adversarial learning. *IEEE Trans Neural Netw Learn Syst*.2020;32:4475-4489.
40. Chen X, Tang Y, Mo Y, et al. A diagnostic model for coronavirus disease 2019 (COVID-19) based on radiological semantic and clinical features: a multi-center study. *Eur Radiol*.2020;30(9):4893-4902.
41. Sarkar K, Khajanchi S, Nieto JJ. Modeling and forecasting the COVID-19 pandemic in India. *Chaos Solitons Fractals*.2020;139:110049.
42. Zhang L, Huang B, Xia H, et al. Retrospective analysis of clinical features in 134 coronavirus disease 2019 cases. *Epidemiol Infect*.2020;148:e199.
43. Zhang M, Zeng X, Huang C, et al. An AI-based radiomics nomogram for disease prognosis in patients with COVID-19 pneumonia using initial CT images and clinical indicators. *Int J Med Inform*.2021;154:104545.

44. Li WT, Ma J, Shende N, et al. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. *BMC Med Inform Decis Mak.*2020;20(1):247.
45. Zhang K, Liu X, Shen J, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell.*2020;182(5):1360.

**How to cite this article:** Xu F, Lou K, Chen C, et al. An original deep learning model using limited data for COVID-19 discrimination: A multicenter study. *Med Phys.* 2022;49:3874–3885. <https://doi.org/10.1002/mp.15549>