

► Additional supplemental

material is published online

only. To view, please visit the

journal online (http://dx.doi.

org/10.1136/thorax-2021-

For numbered affiliations see

218581).

end of article.

Dr Lian Yang;

Correspondence to

Dr Yang Jin, Department of

Respiratory and Critical Care

Medicine. NHC Key Laboratory

of Pulmonary Diseases, Wuhan

Union Hospital, Wuhan, Hubei,

China; whuhjy@126.com

vanglian@hust.edu.cn

Professor Qingjia Bao;

## Original research

# Differentiation of malignant from benign pleural effusions based on artificial intelligence

Sufei Wang,<sup>1</sup> Xueyun Tan,<sup>1</sup> Piqiang Li,<sup>2</sup> Qianqian Fan,<sup>3</sup> Hui Xia,<sup>1</sup> Shan Tian,<sup>4</sup> Feng Pan,<sup>3</sup> Na Zhan,<sup>5</sup> Rong Yu,<sup>6</sup> Liang Zhang,<sup>7</sup> Yanran Duan,<sup>8</sup> Juanjuan Xu,<sup>1</sup> Yanling Ma,<sup>1</sup> Wenjuan Chen,<sup>1</sup> Yan Li,<sup>9</sup> Zilin Zhao,<sup>1</sup> Chaoyang Liu,<sup>2</sup> Qingjia Bao,<sup>2</sup> Lian Yang <sup>(1)</sup>, <sup>3</sup> Yang Jin<sup>1</sup>

## ABSTRACT

Introduction This study aimed to construct artificial intelligence models based on thoracic CT images to perform segmentation and classification of benign pleural effusion (BPE) and malignant pleural effusion (MPE).

Methods A total of 918 patients with pleural effusion were initially included, with 607 randomly selected cases used as the training cohort and the other 311 as the internal testing cohort; another independent external testing cohort with 362 cases was used. We developed a pleural effusion segmentation model (M1) by combining 3D spatially weighted U-Net with 2D classical U-Net. Then, a classification model (M2) was built to identify BPE and MPE using a CT volume and its 3D pleural effusion mask as inputs.

**Results** The average Dice similarity coefficient, Jaccard coefficient, precision, sensitivity, Hausdorff distance 95% (HD95) and average surface distance indicators in M1 were 87.6±5.0%, 82.2±6.2%, 99.0±1.0%, 83.0±6.6%,  $6.9\pm3.8$  and  $1.6\pm1.1$ , respectively, which were better than those of the 3D U-Net and 3D spatially weighted U-Net. Regarding M2, the area under the receiver operating characteristic curve, sensitivity and specificity obtained with volume concat masks as input were 0.842 (95% CI 0.801 to 0.878), 89.4% (95% CI 84.4% to 93.2%) and 65.1% (95% CI 57.3% to 72.3%) in the external testing cohort. These performance metrics were significantly improved compared with those for the other input patterns.

**Conclusions** We applied a deep learning model to the segmentation of pleural effusions, and the model showed encouraging performance in the differential diagnosis of BPE and MPE.



## INTRODUCTION

Effusions, including pleural effusions, ascites, pericardial effusions and abscesses, are commonly observed in many diseases, such as infections and various cancers. The most common effusions are malignant pleural effusions (MPEs) caused by lung cancer, breast cancer, lymphoma and so on, and benign pleural effusions (BPEs) caused by Mycobacterium tuberculosis infection, heart failure, parapneumonic infections and so on.<sup>1-3</sup> The most common conditions leading to ascites are liver disease, cirrhosis and cancer.<sup>4</sup> Because pleural effusions are representative effusions, we chose MPE and BPE as our study objects. The gold standard

## WHAT IS ALREADY KNOWN ON THIS TOPIC

 $\Rightarrow$  The limitations of the gold standard in the diagnosis of benign pleural effusion (BPE) and malignant pleural effusion (MPE) suggest opportunities for more convenient, highly sensitive and non-invasive methods to improve diagnostic performance. Although many previous studies have explored other examinations to help diagnose pleural effusion. no available studies have focused on the differential diagnosis of pleural effusion based on thoracic CT image analysis using deep learning algorithms.

## WHAT THIS STUDY ADDS

 $\Rightarrow$  The artificial intelligence (AI) model proposed in this study showed encouraging performance in the segmentation of pleural effusion areas and differential diagnosis of BPE and MPE based on thoracic CT images.

## HOW THIS STUDY MIGHT AFFECT RESEARCH, **PRACTICE OR POLICY**

 $\Rightarrow$  In the present study, we proposed an AI model to implement the segmentation of pleural effusion regions. It would be worthwhile to apply these segmentation and classification deep learning models to other sorts of effusions.

in the diagnosis of MPE and BPE depends on pleural effusion pathogenic/cytological examinations and thoracentesis with pleural biopsy.<sup>56</sup> However, the low positivity rates for pathogenic diagnosis, the invasive-ness and high costs of pleural biopsy, and the risk of complications represent the limitations of these gold-standard techniques, although their high specificity is their most important advantage.<sup>78</sup> These limitations their most important advantage.78 These limitations suggest opportunities for more convenient, highly sensitive and non-invasive methods to improve the diagnostic performance of BPE and MPE.

Thoracic CT is an appropriate method for the further assessment of pleural effusion.9 Because the features extracted from images by radiologists are limited, artificial intelligence (AI) deep learning algorithms are helpful tools for automatically analysing complex medical images thanks to their strong feature-learning ability.<sup>10</sup> However, no

▶ http://dx.doi.org/10.1136/ thorax-2022-219608

Check for updates

© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY. Published by BMJ.

To cite: Wang S, Tan X, Li P, et al. Thorax 2023;78:376-382.



baogingjiahaoba@gmail.com SW, XT, PL and QF contributed equally. Received 8 December 2021

Accepted 30 August 2022 Published Online First 30 September 2022

available studies have focused on the differential diagnosis of pleural effusion based on thoracic CT image analysis using deep learning algorithms.

U-Net, a convolutional neural network, has become an increasingly important basis for many deep learning models in medical image analysis.<sup>11</sup> It can achieve remarkable generalisation performance when trained with a limited number of images, which makes it especially suitable for our research.<sup>12</sup> In previous studies, U-Net was used for the segmentation of solid organs and lesion regions, such as pancreas segmentation,<sup>13</sup> 3D cardiac segmentation<sup>14</sup> and automatic ground-glass nodule detection.<sup>15</sup> In our study, we applied U-Net to the segmentation of effusion regions, and specifically pleural effusions. We combined the 3D spatially weighted U-Net with the 2D classical U-Net for pleural effusion segmentation in thoracic CT images to obtain fine masks. The high precision of pleural effusion segmentation identifies predictive features which can be subsequently used to train deep learning models for lesion classification. We thus proposed a deep learning algorithm, based on the global and partial analvsis of thoracic CT image features to diagnose BPE and MPE, which can potentially play a critical role in improving patients' clinical prognosis.

## MATERIALS AND METHODS

## Patients and study design

In this consecutive study, 918 pleural effusion cases retrospectively collected from Wuhan Union Hospital between January 2016 and December 2021 were enrolled, with 311 cases randomly selected as the internal testing cohort and the other 607 as the training cohort. Another independent cohort including 362 patients with pleural effusion collected from Renmin Hospital of Wuhan University between January 2020 and May 2022 was used as the external testing cohort. Patients who met the following inclusion criteria were enrolled: (1) diagnosed with pleural effusion by CT scan of the chest and (2) underwent pleural effusion pathogenic/cytological examinations and diagnostic thoracentesis with or without pleural biopsy. The exclusion criteria were (1) pleural effusion whose cause could not be determined, (2) under the age of 18 and (3) unavailable clinical information. The diagnostic criteria for MPE and BPE adopted in this study are based on our previous studies,<sup>16 17</sup> and the criteria are further described in online supplemental methods.

Two professional physicians, HX and WC, collected clinical information, including demographic characteristics, radiological features and laboratory testing results of the enrolled patients from electronic medical records. The volume of pleural effusion was classified as mild (<500 mL), moderate (500–1000 mL) or severe (>1000 mL). The parameters of the CT scanner are presented in online supplemental methods.

## Architecture of the pleural effusion segmentation model (M1)

The pleural effusion segmentation model (M1) is a cascaded two-step deep-learning model (figure 1A). The initial coarse segmentation results are obtained from the spatial attention information based on the 3D spatially weighted U-Net. We used a 3D spatial attention mechanism to capture large-scale contextual information, thus enhancing the representative ability of the



**Figure 1** Classification architecture for the differential diagnosis of BPE and MPE based on AI-processed CT images. In order to diminish the effects of unrelated image regions, we first extract the results of the pleural effusion segmentation models (the coarse segmentation model based on the 3D spatially weighted U-Net and the fine segmentation model based on the 2D classical U-Net). Differential diagnosis is then achieved by the pleural effusion classification model based on the residual blocks and SE blocks. (A) Schematic flowchart of the proposed algorithm for pleural effusion segmentation. (B) Schematic flowchart of the proposed algorithm for predicting the probability of pleural effusion being MPE. (C) Architecture of the 3D attention layer in the segmentation model. (D) Architecture of the SE block in the classification model. BPE, benign pleural effusion; MPE, malignant pleural effusion; SE, squeeze-and-excitation.

Protected

by copyright, including

ð

uses related to text and data mining,

Al training, and similar technologies

model (figure 1C). However, owing to the huge amount of information per sample for 3D U-Net, the region of interest needs to be cropped into small image patches to be used as inputs. Using this method, natural contour information may be lost to some degree. Therefore, the model's learning of natural contour information was enhanced through the 2D classical U-Net. The concatenation of 3D spatially weighted U-Net and 2D classical U-Net helps obtain a fine segmentation of pleural effusion. Details about the algorithms of the coarse and fine segmentation models are shown in online supplemental methods.

## Architecture of the pleural effusion classification model (M2)

We developed a 3D deep convolutional neural network to identify patients with MPE from thoracic CT volumes. As shown in figure 1B, this classification model (M2) uses CT volume and its 3D pleural effusion masks as inputs (details in online supplemental methods). The 3D pleural effusion fine masks are obtained by assembling 2D fine masks generated by the fine segmentation model. This component takes advantage of both stacked bottleneck blocks and squeeze-and-excitation (SE) blocks. The bottleneck block is introduced to extract deeper features from the CT volumes and to solve the problem of degradation in the network training process. The SE block is introduced to improve the representational power of the network by enabling it to perform dynamic channel-wise feature recalibration (figure 1D; online supplemental methods). Inputting the 3D fine masks of pleural effusion according to the 3D thoracic CT volume helps reduce the effects of background information and improves the classification of BPE and MPE. Details about the training process of the pleural effusion segmentation and classification model are shown in online supplemental methods.

## Quantitative assessment indicators

For the pleural effusion segmentation model, the Dice similarity coefficient (DSC) and Jaccard coefficient were used to evaluate the spatial overlap between the model-generated contour (M) and the ground truth contour (G). In our study, G means the sets defined by these boundaries of pleural effusion area drawn by a professional radiologist (15 years of experience) on CT images; M means the sets defined by these boundaries of pleural effusion area generated by the AI model. Precision and sensitivity measure the detection capability for identifying the correct regions. The Hausdorff distance 95% (HD95) and average surface distance (ASD) measure the boundary similarity between the model-generated contour and the ground truth contour. Details about the above indicators are described in online supplemental methods.

The area under the receiver operator characteristic (ROC) curve (AUC), sensitivity and specificity were used to evaluate the predictive performance of M2 as a pleural effusion classification model.

## Statistical analysis

Implementation details are described thoroughly in online supplemental methods. Statistical analyses were performed using SPSS Statistics (V.22). Comparisons were performed using the Mann-Whitney U test for continuous variables and  $\chi^2$  or Fisher's exact test for categorical variables, as appropriate. ROC curves were generated to evaluate the classification performance. Statistical significance was defined as a two-sided p value <0.05.

## RESULTS

## **Baseline characteristics of patients**

The baseline characteristics of the enrolled patients are presented in table 1. We compared the distribution of sex, age, volume of pleural effusion (mild/moderate/severe) and unilateral/bilateral pleural effusion between the two groups (BPE vs MPE). There was no significant difference between the two groups in any of the three cohorts in terms of age and volume of pleural effusion. However, the distribution of gender in patients with MPE and BPE was significantly distinct (p<0.001) in all three cohorts. Lung cancer was the leading cause of MPE (24.2% in the training cohort, 24.7% in the internal testing cohort, 40.3% in the external testing cohort, yield in the internal testing cohort, 19.3% in the internal testing cohort, 24.3% in the external testing cohort; table 2).

## 3D spatially weighted attention mechanism in coarse segmentation for pleural effusion area discovery

To highlight the importance of the attention mechanism inserted in 3D U-Net, figure 2 depicts the pleural effusion areas delineated by 3D U-Net and 3D spatially weighted U-Net, respectively, and displays heatmaps to indicate the importance of each part of the pleural effusion areas. The cut-off value used to acquire the

Table 1         Patient characteristics in the training, internal testing and external testing cohorts									
	Training cohort			Internal testing cohort			External testing cohort		
Characteristics	MPE (n=341)	BPE (n=266)	P value	MPE (n=179)	BPE (n=132)	P value	MPE (n=204)	BPE (n=158)	P value
Gender			<0.001			<0.001			<0.001
Female	177 (51.9%)	81 (30.5%)		101 (56.4%)	37 (28.0%)		108 (52.9%)	39 (24.7%)	
Male	164 (48.1%)	185 (69.5%)		78 (43.6%)	95 (72.0%)		96 (47.1%)	119 (75.3%)	
Age, median (IQR), years	60.0 (50.0–68.0)	61.0 (48.0–71.0)	0.694	59.0 (52.0–68.0)	57.0 (46.0–68.0)	0.115	64.0 (56.3–71.0)	65.0 (53.8–75.0)	0.755
Unilateral pleural effusion	179 (52.5%)	127 (47.7%)	0.246	107 (59.8%)	63 (47.7%)	0.035	152 (74.5%)	71 (44.9%)	< 0.001
Bilateral pleural effusion	162 (47.5%)	139 (52.3%)		72 (40.2%)	69 (52.3%)		52 (25.5%)	87 (55.1%)	
Volume of pleural effusion			0.750			0.627			0.763
Mild	236 (69.2%)	186 (69.9%)		124 (69.3%)	97 (73.5%)		112 (54.9%)	81 (51.3%)	
Moderate	60 (17.6%)	50 (18.8%)		33 (18.4%)	19 (14.4%)		40 (19.6%)	35 (22.1%)	
Severe	45 (13.2%)	30 (11.3%)		22 (12.3%)	16 (12.1%)		52 (25.5%)	42 (26.6%)	

Data shown are the number and percentage of patients, with the exception of age (median and IQR). Statistical comparisons were performed for each clinical variable between the two groups (MPE vs BPE). P values were computed using the Mann-Whitney U test for age as a continuous variable and the  $\chi^2$  test or Fisher's exact test for categorical variables, as appropriate.

BPE, benign pleural effusion; MPE, malignant pleural effusion.

<u>s</u>

a

nolog

Ies

Table 2	Causes of pleural	effusion in the	e training, internal te	esting
and exterr	nal testing cohorts			

	-				
Causes	Training cohort (n=607)	Internal testing cohort (n=311)	External testing cohort (n=362)		
Malignant effusions	341 (56.1%)	179 (57.6%)	204 (56.3%)		
Lung cancer	147 (24.2%)	77 (24.7%)	146 (40.3%)		
Breast cancer	91 (15.0%)	46 (14.8%)	20 (5.5%)		
Lymphoma	64 (10.5%)	36 (11.6%)	2 (0.6%)		
Gastrointestinal cancer	28 (4.6%)	13 (4.2%)	24 (6.6%)		
Ovarian	4 (0.6%)	4 (1.3%)	7 (1.9%)		
Others	7 (1.2%)	3 (1.0%)	5 (1.4%)		
Benign effusions	266 (43.9%)	132 (42.4)	158 (43.7%)		
Parapneumonics	111 (18.3%)	60 (19.3%)	88 (24.3%)		
Heart failure	59 (9.7%)	25 (8.0%)	11 (3.1%)		
Tuberculous pleuritis	83 (13.7%)	43 (13.8%)	36 (10.0%)		
Pulmonary embolism	3 (0.5%)	0 (0.0%)	3 (0.8%)		
Pericardial diseases	1 (0.2%)	1 (0.3%)	3 (0.8%)		
Others	9 (1.5%)	3 (1.0%)	17 (4.7%)		
Data shown are the number and percentage of patients.					

high-response area was 0.5. It can be clearly observed that the highresponse areas were mainly concentrated at the pleural effusion boundary when using 3D spatially weighted U-Net, while they were gathered in the pleural effusion inner part when using 3D U-Net, showing that the attention mechanism significantly improved the accuracy of the pleural effusion area segmentation. Compared with



Figure 2 Pleural effusion area discovery. The cut-off value used to acquire the high-response area in heatmaps was 0.5. The second and third rows show the heatmap of 3D U-Net and the 3D spatially weighted U-Net, respectively. The fourth and fifth rows show the pleural effusion area discovered by 3D U-Net and the 3D spatially weighted U-Net, respectively.



Figure 3 Comparison of the segmentation results of a patient using the ground truth contour, 3D spatially weighted U-Net (one-step method) and M1 (two-step method: 3D spatially weighted U-Net and 2D classical U-Net) in 2D and 3D views. The first column shows the images in three different CT slices (A–C) and the 3D (D) view of a patient's thoracic CT scan. The second column (a1, b1, c1 and d1) shows the corresponding experts' ground truth contour (pleural effusion outline highlighted in red). The third column (a2, b2, c2 and d2) shows the pleural effusion segmentation results (pleural effusion outline highlighted in yellow) using 3D spatially weighted U-Net. The fourth column (a3, b3, c3 and d3) shows the pleural effusion segmentation results (pleural effusion outline highlighted in blue) using M1.

3D U-Net segmentation, the results of 3D spatially weighted U-Net segmentation better fit the ground truth.

## Comparison among 3D U-Net, 3D spatially weighted U-Net and segmentation deep learning model (M1)

For visual demonstration, representative pleural effusion area segmentation results of M1 (two-step method: 3D spatially weighted U-Net and 2D classical U-Net) are compared with the results of 3D spatially weighted U-Net (one-step method). Figure 3A1-C1 show an example of the radiologist's ground truth contours at three different CT slices, with the outline of contours shown by red lines. figure 3A2-C2 show the contours (yellow lines) obtained using only 3D spatially weighted U-Net, while figure 3A3-C3 show the contours (blue lines) obtained using M1. For 3D illustration, figure 3D1–D3 show the 3D views of the pleural effusion area discriminated by the radiologist, onestep method and two-step method, respectively. Compared with the one-step method, the segmentation results of the two-step method better fit the ground truth.

For quantitative assessment of the segmentation results, six indicators were used to evaluate the similarity, difference and segmentation performance of 3D U-Net, 3D spatially weighted U-Net and M1. The average DSC, Jaccard coefficient, precision, sensitivity, HD95 and ASD indicators for M1 were 87.6%, 82.2%, 99.0%, 83.0%, 6.9 and 1.6, respectively, which were better than those of 3D U-Net and 3D spatially weighted U-Net (table 3).

difference and positioning performance						
Methods	DSC (%)	Jaccard (%)	Precision (%)	Sensitivity (%)	HD95	ASD
3D U-Net	30.2±4.9	26.7±6.3	76.2±4.1	29.4±7.8	197.6±14.7	4.3±0.4
3D spatially weighted U-Net	67.1±2.4	69.8±2.7	81.0±3.3	83.8±5.3	72.7±11.2	4.7±0.8
M1	87.6±5.0	82.2±6.2	99.0±1.0	83.0±6.6	6.9±3.8	1.6±1.1

Comparison of the proposed segmentation model (M1) with the 3D U-Net and 3D spatially weighted U-Net in terms of similarity, Table 3

Data shown are mean±SD.

ASD, average surface distance; DSC, Dice similarity coefficient; HD95, Hausdorff distance 95%; M1, two-step method (3D spatially weighted U-Net and 2D classical U-Net).

## Diagnostic validation of the classification deep learning model (M2)

The proposed M2 model with volume concat mask as input consistently achieved the highest accuracy across the internal and external testing cohorts (figure 4A,B). In addition, the classification score indicated a notable distinction between BPE and MPE with different input patterns in the internal and external testing cohorts (all p<0.001). The input with volume concat mask bore the most significant distinction between BPE and MPE in both the internal and external testing cohorts, as revealed by the violin plots (figure 4C,D).

AUC, sensitivity and specificity were used as the main indicators for evaluating the diagnostic performance of M2. The three indicators for input with volume concat mask were 0.883 (95% CI 0.841 to 0.916), 78.4% (95% CI 71.6% to 84.2%) and 86.2% (95% CI 79.0% to 91.6%) in the internal testing cohort, and 0.842 (95%) CI 0.801 to 0.878), 89.4% (95% CI 84.4% to 93.2%) and 65.1% (95% CI 57.3% to 72.3%) in the external testing cohort, which were significantly improved compared with those for input with only volume and input with volume multiply mask (table 4). The similar AUC values of the internal and external testing cohorts suggested an encouraging level of generalisability of M2 for diagnosing BPE and MPE in new patients. The input with the volume concat mask significantly improved the classification performance of M2, while, notably, the decrease in the speed of the network compared with the other two input patterns was negligible.



Figure 4 Predictive performance of the classification deep learning model shown by ROC curves and violin plots. (A, B) ROC curves for different input patterns in the internal (A) and external (B) testing cohort, (C, D) Deep learning score difference between BPE and MPE with different input patterns in the internal (C) and external (D) testing cohorts. AUC, area under the ROC curve; BPE, benign pleural effusion; MPE, malignant pleural effusion; ROC, receiver operating characteristic.

Comparison of the activation heatmaps generated by M2 between two randomly selected patients with MPE (one with lung cancer, one with breast cancer) and two randomly selected patients with BPE (one with tuberculous pleuritis, one with heart failure) is shown in figure 5. The activation heatmaps indicated the importance of different parts of the pleural effusion regions and suggested that different areas drew the attention of M2 to various degrees. The important areas found by M2, which were considered closely associated with the nature of pleural effusion (BPE or MPE), varied in different patients. The difference in features between highimportance pleural effusion areas and other pleural effusion areas requires further research.

Comparison of the heatmaps between typical MPE and BPE

## DISCUSSION

In this study, we proposed a new architecture for the differential diagnosis of BPE and MPE based on pleural effusion segmentation of thoracic CT images. This deep learning architecture was trained using 607 CT images, and its performance was validated in an internal testing cohort (311 pleural effusion cases) and an external testing cohort (362 pleural effusion cases) from Wuhan Union Hospital and Renmin Hospital of Wuhan University. The encouraging diagnostic performance of the deep learning model was shown in both the internal (AUC 0.883, 95% CI 0.841 to cohorts. In addition, we combined this AI model with some clinical data, including gender, age, unilateral/bilateral pleural effusion and yolume of pleural effusion, to predict RPF and MDF. showed that combining clinical indicators could improve the AUC in all three cohort (training cohort: 0.903 vs 0.896, internal testing cohort: 0.895 vs 0.882, external testing cohort: 0.868 vs 0.842) (online supplemental figure S1). This deep learning model discovered suspect pleural effusion areas and produced fine segmentations in the first step, then identified BPE and MPE by holistically and partially analysing thoracic CT image features, revealing that the features of thoracic CT images were closely related to the nature of pleural effusions. Our study provides an alternative, easy-to-use method to achieve non-invasive and efficient diagnosis of BPE and MPE from original CT images without human assistance.

Previous studies have demonstrated that thoracic CT image features, such as fluid loculation, pleural lesions, pleural nodules and extrapleural fat, can help discriminate MPEs from BPEs.<sup>18</sup> Pleural nodules and nodular pleural thickening were reported to be associated to MPE, while circumferential pleural thickening was more common in tuberculous pleural effusion (TPE).<sup>18 20</sup> Zhang et al revealed that spectral CT imaging features combined with patient age and disease history could differentiate BPEs from MPEs with a sensitivity of 100% and a specificity of 71.4%, as well as an AUC of 0.933.<sup>21</sup>

Pleural effusions can be divided into transudates and exudates. Although no CT feature can accurately distinguish transudates from exudates, Abramowitz et al indicated that fluid loculation and

Protected

ş

copyright

including

tor uses

related to text

and

data

D

<u>0</u>

a

nolog

les

5				
Input	Testing cohort	AUC (95% CI)	Sensitivity % (95% CI)	Specificity % (95% CI)
Only volume	Internal	0.830 (0.783 to 0.871)	67.1 (59.6 to 73.9)	85.4 (78.1 to 91.0)
	External	0.766 (0.719 to 0.808)	58.0 (50.9 to 64.8)	82.5 (75.9 to 88.0)
Volume multiply mask	Internal	0.825 (0.778 to 0.866)	83.5 (77.2 to 88.7)	73.1 (64.6 to 80.5)
	External	0.766 (0.720 to 0.808)	91.8 (87.2 to 95.1)	50.6 (42.7 to 58.4)
Volume concat mask	Internal	0.883 (0.841 to 0.916)	78.4 (71.6 to 84.2)	86.2 (79.0 to 91.6)
	External	0.842 (0.801 to 0.878)	89.4 (84.4 to 93.2)	65.1 (57.3 to 72.3)
Data are presented as % (95% CI).				

Comparison about classification performance with different inputting patterns of the classification model (M2) in the internal and external Table 4 testing cohorts

AUC, area under the receiver operating characteristic curve.

pleural thickening were more common in exudates than in transudates.<sup>22</sup> Discrimination of a pleural effusion as transudate or exudate is important for further evaluation and treatment. Some causes of BPE, such as heart failure and cirrhosis, generate transudates. However, some causes of BPE, such as infections and pulmonary embolism, generate exudates, as does MPE.<sup>3 23</sup> The immunological microenvironment and inflammatory responses in MPE are two important factors that lead to the production of different components. In addition to neoplastic cells, cytokines and chemokines produced by immune cells, signalling molecules generated by tumour-associated macrophages, and fibroblasts are the main components of the surviving environment of tumour cells in pleural effusions.<sup>24</sup> In early-stage TPE, lymphocyte predominance characterises a large proportion of the fluid; in the meantime, a higher mycobacterial burden appears in effusions that have loculations.<sup>2</sup> Li et al identified different peptide profiles between BPE and MPE through proteomic analysis and established a model to discriminate between BPE and MPE.<sup>26</sup> The different pleural effusion components



Figure 5 Comparison of the activation heatmaps generated by the pleural effusion classification model between four randomly selected patients with lung cancer, breast cancer, tuberculous pleuritis and heart failure, respectively.

for BPE and MPE may be a crucial cause of different thoracic CT image features, making it feasible to classify BPE and MPE using a deep learning model based on thoracic CT image features.

In the present study, we proposed an AI model to implement the segmentation of pleural effusion regions. The deep learning model for segmentation proposed in our study successfully integrated 3D spatially weighted U-Net and 2D classical U-Net. Our results showed that the cascaded segmentation architecture combining 3D spatially weighted U-Net with 2D classical U-Net (M1) was superior to the other two segmentation methods (only 3D U-Net and only 3D spatially weighted U-Net). Applying the spatial attention mechanism to 3D U-Net not only focuses the deep learning model on the regions of interest for input thoracic CT images, avoiding the interference of background information, but also extracts both shallowlevel and deep-level attention information, which can improve the feature extraction ability of the model.<sup>15 27</sup> However, in order to reduce the cubically growing number of network parameters caused by 3D convolution, using patches (crop region of interest into small image patches) as input may lose some natural contour information. 3D spatially weighted U-Net cascaded by a 2D classical U-Net can be conducive to supplementing natural contour information and i mining, excluding most error information about the pleural effusion region. In addition, in the deep learning model for pleural effusion classification, we input the holistic thoracic CT image and the fine segmentation region of pleural effusion generated by M1 at the same time. On the one hand, this approach stresses the features within the pleural effusion areas; while on the other hand, it does not neglect the related information within the areas outside the pleural effusion.

It has been reported that the primary tumour cannot be found in approximately 10% of MPEs.<sup>24</sup> Therefore, it is of vital importance to identify MPEs of unknown origin in a timely and non-invasive manner. It would be worthwhile to apply these segmentation and classification deep learning models to other sorts of effusions. Since the causes of ascites and abscess vary depending on the type of technolog tumour and pathogen infection, a single AI model able to determine which type of cancer or bacteria is the reason for effusion production would represent remarkable progress. Further research and efforts are required to achieve this goal.

Although the proposed deep learning model of pleural effusion segmentation and classification showed encouraging performance, our study has several limitations. First, the data source only derived from two hospitals which may have limited the generalisability and robustness of the deep learning model. Second, high model interpretability of deep learning networks is considered valuable,<sup>28</sup> but the association between the imaging representations and the nature feature of pleural effusions cannot be fully understood in our study because of the end-to-end learning strategy. Third, despite the advantages of the proposed model, which uses exclusively thoracic CT images, in terms of convenience and time-saving, the predictive

and

data

, and

Protected by copy

Bul

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

## **Pleural disease**

performance may be improved by combining this model with other clinical models; however, this point was not clarified in this study. Future large-scale external validations from multiple centres are necessary to provide convincing evidence of the generalisability of the deep learning model proposed in this study.

In conclusion, our research proposed an original deep learning model: a combination of 3D spatially weighted U-Net and 2D classical U-Net were used for the segmentation of pleural effusion. Subsequently, a deep learning model was established for the differential diagnosis of BPE and MPE based on thoracic CT images with masks. The non-invasiveness and high efficiency of the segmentation and classification models suggest their potential clinical utility. Our work shows the potential of AI to assist radiologists in identifying malignant disease and thereby improving patient care.

### Author affiliations

<sup>1</sup>Department of Respiratory and Critical Care Medicine, NHC Key Laboratory of Pulmonary Diseases, Wuhan Union Hospital, Wuhan, Hubei, China

<sup>2</sup>State Key Laboratory of Magnetic Resonance and Atomic and Molecular Physics, Chinese Academy of Sciences Wuhan Institute of Physics and Mathematics, Wuhan, Hubei, China

<sup>3</sup>Department of Radiology, Wuhan Union Hospital, Wuhan, Hubei, China

<sup>4</sup>Department of Infectious Diseases, Wuhan Union Hospital, Wuhan, Hubei, China <sup>5</sup>Department of Pathology, Renmin Hospital of Wuhan University, Wuhan, Hubei, China

<sup>6</sup>Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, Hubei, China

<sup>7</sup>Department of Radiology, Renmin Hospital of Wuhan University, Wuhan, Hubei, China

<sup>8</sup>School of Public Health, Tongji Medical College of Huazhong University of Science and Technology, Wuhan, Hubei, China

<sup>9</sup>Department of Pathology, Wuhan Union Hospital, Wuhan, Hubei, China

**Correction notice** This article has been corrected since it was first published. The open access licence has been updated to CC BY.

**Acknowledgements** We are grateful to Dr Jiazheng Wang (Philips Healthcare), Dr Peng Sun (Philips Healthcare) and Mr Yuyang Chen (Putnam Science Academy, USA) for many useful discussions through the formation of this work. We would like to thank Editage (www.editage.cn) for English language editing.

**Contributors** SW: conceptualisation, project administration, writing—original draft. XT: investigation, methodology, writing—original draft. PL: formal analysis, methodology, software, writing—original draft. QF: methodology, data curation. HX: data curation, investigation. ST: methodology, writing—review and editing. FP: methodology, writing—review and editing. NZ: data curation, methodology. RY: data curation, methodology. LZ: data curation, methodology. YD: data curation, supervision. WC: data curation, methodology. YL: investigation, supervision. ZZ: methodology. CL: methodology. SU: methodology. SU: methodology, supervision, RY: data curation, methodology. CL: methodology. SU: methodology. YD: investigation, supervision, writing—review and editing. LY: data curation, methodology, supervision, YJ: conceptualisation, funding acquisition, supervision, writing—review and editing. YJ is the quarantor of this publication.

**Funding** This work was supported by the National Natural Science Foundation of China (No. 81770096; No. 82102496; No. 82172034).

## Competing interests None declared.

Patient consent for publication Not applicable.

**Ethics approval** The institutional ethics committees of Wuhan Union Hospital (No. S1041) and Renmin Hospital of Wuhan University (No.WDRY2019-K014) reviewed and approved this study protocol and waived the need for informed consent due to the study design.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines,

terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

## ORCID iD

Lian Yang http://orcid.org/0000-0003-4664-2474

## REFERENCES

- Bibby AC, Dorn P, Psallidas I, et al. ERS/EACTS statement on the management of malignant pleural effusions. Eur Respir J 2018;52:1800349.
- 2 Shaw JA, Diacon AH, Koegelenberg CFN. Tuberculous pleural effusion. *Respirology* 2019;24:962–71.
- 3 Beaudoin S, Gonzalez AV. Evaluation of the patient with pleural effusion. CMAJ 2018;190:E291–5.
- 4 Gupta A, Sedhom R, Beg MS. Ascites, or fluid in the belly, in patients with cancer. *JAMA Oncol* 2020;6:308.
- 5 Ren Z, Hu Y, Xu L. Identifying tuberculous pleural effusion using artificial intelligence machine learning algorithms. *Respir Res* 2019;20:220.
- 6 Zhang M, Yan L, Lippi G, *et al*. Pleural biomarkers in diagnostics of malignant pleural effusion: a narrative review. *Transl Lung Cancer Res* 2021;10:1557–70.
- 7 Hallifax RJ, Corcoran JP, Ahmed A, et al. Physician-based ultrasound-guided biopsy for diagnosing pleural disease. Chest 2014;146:1001–6.
- 8 Wang X-J, Yang Y, Wang Z, *et al*. Efficacy and safety of diagnostic thoracoscopy in undiagnosed pleural effusions. *Respiration* 2015;90:251–5.
- 9 Hallifax RJ, Talwar A, Wrightson JM, et al. State-of-the-art: radiological investigation of pleural disease. *Respir Med* 2017;124:88–99.
- 10 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
- 11 Falk T, Mai D, Bensch R, et al. U-Net: deep learning for cell counting, detection, and morphometry. Nat Methods 2019;16:67–70.
- 12 Ö Çiçek, Abdulkadir A, Lienkamp SS. 3D U-Net: learning dense volumetric segmentation from sparse annotation. Cham: Springer International Publishing, 2016: 424–32.
- 13 Liu Y, Lei Y, Fu Y, et al. CT-based multi-organ segmentation using a 3D self-attention U-net network for pancreatic radiotherapy. *Med Phys* 2020;47:4316–24.
- 14 Ding X, Peng Y, Shen C, *et al*. CAB U-Net: an end-to-end category attention boosting algorithm for segmentation. *Comput Med Imaging Graph* 2020;84:101764.
- 15 Ni Y, Yang Y, Zheng D, et al. The invasiveness classification of ground-glass nodules using 3D attention network and HRCT. J Digit Imaging 2020;33:1144–54.
- 16 Wang S, Tian S, Li Y, et al. Development and validation of a novel scoring system developed from a nomogram to identify malignant pleural effusion. *EBioMedicine* 2020;58:102924.
- 17 Luo P, Mao K, Xu J, et al. Metabolic characteristics of large and small extracellular vesicles from pleural effusion reveal biomarker candidates for the diagnosis of tuberculosis and malignancy. J Extracell Vesicles 2020;9:1790158.
- 18 Arenas-Jiménez J, Alonso-Charterina S, Sánchez-Payá J, et al. Evaluation of CT findings for diagnosis of pleural effusions. *Eur Radiol* 2000;10:681–90.
- 19 Porcel JM, Pardina M, Bielsa S, *et al*. Derivation and validation of a CT scan scoring system for discriminating malignant from benign pleural effusions. *Chest* 2015;147:513–9.
- 20 Kim JS, Shim SS, Kim Y, et al. Chest CT findings of pleural tuberculosis: differential diagnosis of pleural tuberculosis and malignant pleural dissemination. Acta Radiol 2014;55:1063–8.
- 21 Zhang X, Duan H, Yu Y, *et al*. Differential diagnosis between benign and malignant pleural effusion with dual-energy spectral CT. *PLoS One* 2018;13:e0193714.
- 22 Abramowitz Y, Simanovsky N, Goldstein MS, et al. Pleural effusion: characterization with CT attenuation values and CT appearance. AJR Am J Roentgenol 2009;192:618–23.
- 23 Jany B, Welte T. Pleural effusion in adults—etiology, diagnosis, and treatment. *Dtsch Arztebl Int* 2019;116:377–86.
- 24 Giarnieri E, Bellipanni G, Macaluso M, *et al*. Review: Cell dynamics in malignant pleural effusions. *J Cell Physiol* 2015;230:272–7.
- 25 Shaw JA, Irusen EM, Diacon AH, et al. Pleural tuberculosis: a concise clinical review. Clin Respir J 2018;12:1779–86.
- 26 Li H, Tang Z, Zhu H, et al. Proteomic study of benign and malignant pleural effusion. J Cancer Res Clin Oncol 2016;142:1191–200.
- 27 Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: learning to leverage salient regions in medical images. *Med Image Anal* 2019;53:197–207.
- 28 Q-s Z, S-c Z. Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering 2018;19:27–39.

## **Supplementary Methods**

# Diagnostic criteria for malignant pleural effusion (MPE) and benign pleural effusion (BPE)

According to the criteria adopted in our previous study[1, 2], if malignant cells were detected in the pleural effusion based on cytologic examination or pleural biopsy, the effusion was classified as malignant. BPE was identified by a known aetiology, such as tuberculous pleural effusion (TPE) or parapneumonic effusion, without any signs of cancer. TPE was diagnosed if acid-fast stains or Lowenstein–Jensen cultures of pleural effusion, sputum, bronchoalveolar lavage fluid, or pleural biopsy specimens were positive.

## Computed tomography (CT) imaging protocols

For the establishment of the training and internal testing dataset, non-enhanced chest CT data were obtained from Wuhan Union Hospital. All these chest CT examinations were performed on two commercial multi-detector CT scanners: Philips Ingenuity Core128 (Philips Medical Systems) (n=410) and SOMATOM Definition AS (Siemens Healthineers) (n=508) with the routine mediastinal window reconstruction: axial images with a matrix size of 512 × 512, slice thickness of 5 mm; and mediastinal kernels of iDose5 in Philips Ingenuity Core128 and b30f in SOMATOM Definition AS. Before the scanning, patients were instructed on breath-holding in order to minimize motion artifacts. Afterwards, CT images were acquired during a single breath-hold. For external validation, additional chest CT data were obtained from a third-party hospital (Renmin Hospital of Wuhan University). These routine non-contrast chest CT scans were performed on four commercial multi-detector CT scanners: GE Optima CT680 CT (n=155), 64-slice LightSpeed VCT (n=95), BrightSpeed Elite CT (n=89) and Revolution CT (n=23; GE Medical Systems). The standard reconstruction of axial mediastinal-window images was implemented following similar commercial scanning protocols from the CT manufacturers[3].

## Model training

The specific training process is as follows:

First, we train the coarse segmentation component on the dataset to generate coarse pleural effusion masks. Then, we concatenate the CT images with the corresponding coarse pleural effusion masks generated by the coarse segmentation component as the input to the fine segmentation component. We train the fine segmentation component to generate the fine pleural effusion masks. Finally, we fuse the CT images and the corresponding fine pleural effusion masks refined by the fine segmentation component and feed them to the classification component (M2) to train the classification component to deliver more accurate diagnosis results.

## Mathematical description of the pleural effusion classification model

## SE block

SE block is derived from SENet, which mainly learns the correlation between channels and filters out the attention for the channels, and slightly increases the computational effort, but the result of it is better. By processing the feature map from the convolution, a one-dimensional vector with the same number of channels is obtained as the evaluation score of each channel, and then the score is applied to the corresponding channel separately to get its result. It contains three main operations as follows:

Squeeze: Compressing the features along the spatial dimension, turning each twodimensional feature channel into a real number, which somehow has a global perceptual field, and the output dimension matches the number of input feature channels. Excitation: Based on the correlation between the feature channels, a weight is generated for each feature channel, which represents the importance of the feature channels. Reweight: The weights from Excitation are considered as the importance of each feature channel, and then weighted to the previous features by multiplying them channel by channel to complete the rescaling of the original features in the channel dimension.

## Details of the pleural effusion segmentation model

## Coarse segmentation model

The coarse segmentation model plays a major role in the generation of coarse pleural effusion segmentation masks. The architecture of this component is based on a 3D spatially weighted U-Net. The 3D U-Net is commonly used for medical image segmentation tasks because of its multiscale feature fusion capability. At the same time, we utilized a 3D spatial attention mechanism to capture large-scale contextual information, thus enhancing the representative ability of the model.

The 3D attention layer is placed before and after the concatenation operation to fully exploit the spatial contextual information on the intra-plane level and leverage it for volumetric spatial weighting on the inter-plane level. This emphasizes the regions of interest in volumetric feature maps.

Under the 3D convolutional network architecture, we assume that the volumetric feature tensors Fi input to the 3D attention layer in the i-th layer is of size I×J×K×P, where I, J, and K are the length, width, and height of a feature tensor, respectively, and Fi contains P channels. To acquire the spatial statistical information representation on a chosen plane, we applied global average pooling (GAP) to each plane of the 3D space. The three resulting orthogonal vectors compress the statistical information in the entire slice along each plane. Moreover, we also adopt fully connected (FC) layers, and rectified linear unit (ReLU) and sigmoid activation functions, to introduce additional nonlinearity in the generation of the weight vectors. The structure is similar to a bottleneck architecture with two fully connected layers. The first fully connected layer is used for dimensionality reduction, with a reduction ratio of 1/4 to limit capacity and improve model generalization. With the weighting vectors for each plane, we weighted the three dimensions of the feature tensors Fi. Mathematically, if the feature tensor is weighted along the three planes, this is equivalent to a tensor product of the orthogonal weighting vectors.

## Fine segmentation model

The input of this component contains cropped CT image patches and coarse pleural effusion masks. As aforementioned, 3D spatially weighted U-Nets can capture large-scale contextual information. However, because of GPU memory limits, the input

cannot cover the complete CT images, and the coarse segmentation model loses contour information during the learning process. Thus, we make use of a 2D classical U-Net to enhance the learning of natural contour information so that the precision of fine segmentation can be improved.

## Details of the pleural effusion classification model

The classification model is similar to the 3D ResNet but with several modifications. In this model, a 3D convolutional layer, a 3D pooling layer and a 3D SE block is defined as a group. Skip connection is added to input and output of each group. Two stacks of groups with a batch normalization layer is defined as a SE-ResBlock. First, the input will pass through a 3D convolutional layer, a 3D BN layer and a 3D pooling layer, then through four stacks of SE-ResBlocks for feature extraction, and finally, through the GAP and FC layers to predict the probability of MPE.

## Details of the quantitative assessment indicators

For the pleural effusion segmentation model, the Dice similarity coefficient (DSC) (1) and Jaccard coefficient (2) were used to evaluate the spatial overlap between the model-generated contour (M) and the ground truth contour (G):

$$DSC(M,G) = \frac{2|M \cap G|}{|M| + |G|},$$
 (1)

$$Jaccard = \frac{|M \cap G|}{|M \cup G|},\tag{2}$$

Precision (3) and sensitivity (4) measure the detection capability for identifying the correct regions.

$$Precision = \frac{TP}{TP + FP},$$
(3)

$$Sensitivity = \frac{TP}{TP+FN},$$
(4)

where true positives (TP) are defined as the regions of the segmentation results consistent with the ground truth, and false positives (FP) as the regions of the segmentation results that are not consistent with the ground truth. False negatives (FN) are defined as the ground truth regions that are not included in the segmentation results. The Hausdorff distance 95% (HD95) (5) and average surface distance (ASD) (6) measure the boundary similarity between the model-generated contour and the ground truth contour:

$$HD95 = max_{k95\%}[d(G, M), d(M, G)]$$
  
=  $max_{k95\%}[\max_{g \in G} \min_{m \in M} d\{g, m\}, \max_{m \in M} \min_{g \in G} d\{m, g\}],$  (5)

$$ASD = \frac{1}{S(G) + S(M)} \left( \sum_{g \in S(G)} d(g, S(M)) + \sum_{m \in S(M)} d(m, S(G)) \right),$$
(6)

where g represents points in G and m represents points in M. d(g,m) represents the distance between point g and point m. S(M) and S(G) represent the surfaces of M and G, respectively. d(g, S(M)) and d(m, S(G)) represent the shortest distance from any point g to S(M) and from any point m to S(G), respectively.

## Implementation

The Adam algorithm with a batch size of 16,  $\beta 1 = 0.9$ ,  $\beta 2 = 0.999$ , and decay = 1e-6 was adopted to optimize the segmentation and classification models[4]. The initial learning rate for both the segmentation and classification models was set at 0.001. The weights of the networks were initialized using the default initialization mechanism of the Keras framework. All experiments were performed in the Tensorflow and Keras framework. The training strategies were optimized in the same computer system with 32 GB RAM and a GeForce GTX 2080 graphics processing unit (GPU).

## Compute performance measures on an hourly basis

First step: Convert dicom data to nii data for CT images, use python code to extract window width and window level, and convert thin layer (thickness of <5mm) to thick layer (thickness of 5mm) (if necessary).

Second step: Input all the thick layer CT image data into the trained 3D spatially weighted U-Net to generate the coarse segmentation results. This step takes about 50 minutes to perform coarse segmentation for 104 patients.

Third step: Input the nii data and the coarse segmentation results for CT images into the trained 2D classical U-Net to generate the fine segmentation results. This step takes

about 10 minutes to perform fine segmentation for 104 patients.

Fourth step: Input the CT volume and its 3D pleural effusion fine masks into the classification model to generate the classification scores. This step takes about 44 seconds to perform classification for 104 patients.

Excluding the intermediate data processing steps, it takes about an hour to process 104 patients through the segmentation and classification model.

## Reference

1. Wang S, Tian S, Li Y, et al. Development and validation of a novel scoring system developed from a nomogram to identify malignant pleural effusion. EBioMedicine 2020;58:102924.

2. Luo P, Mao K, Xu J, et al. Metabolic characteristics of large and small extracellular vesicles from pleural effusion reveal biomarker candidates for the diagnosis of tuberculosis and malignancy. J Extracell Vesicles 2020;9(1):1790158.

3. Bhalla AS, Das A, Naranje P, et al. Imaging protocols for CT chest: A recommendation. Indian J Radiol Imaging 2019;29(3):236-246.

4. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv preprint 2014;arXiv:14126980.





