



# Detection and staging of chronic obstructive pulmonary disease using a computed tomography–based weakly supervised deep learning approach

Jiaxing Sun<sup>1,2</sup> · Ximing Liao<sup>1</sup> · Yusheng Yan<sup>3</sup> · Xin Zhang<sup>4</sup> · Jian Sun<sup>5</sup> · Weixiong Tan<sup>6</sup> · Baiyun Liu<sup>6</sup> · Jiangfen Wu<sup>6</sup> · Qian Guo<sup>1</sup> · Shaoyong Gao<sup>1</sup> · Zhang Li<sup>7</sup> · Kun Wang<sup>1</sup> · Qiang Li<sup>1</sup>

Received: 5 October 2021 / Revised: 25 January 2022 / Accepted: 7 February 2022 / Published online: 24 February 2022  
© The Author(s), under exclusive licence to European Society of Radiology 2022, corrected publication 2022

## Abstract

**Objectives** Chronic obstructive pulmonary disease (COPD) is underdiagnosed globally. The present study aimed to develop weakly supervised deep learning (DL) models that utilize computed tomography (CT) image data for the automated detection and staging of spirometry-defined COPD.

**Methods** A large, highly heterogeneous dataset was established, consisting of 1393 participants retrospectively recruited from outpatient, inpatient, and physical examination center settings of four large public hospitals in China. All participants underwent both inspiratory chest CT scans and pulmonary function tests. CT images, spirometry data, demographic information, and clinical information of each participant were collected. An attention-based multi-instance learning (MIL) model for COPD detection was trained using CT scans from 837 participants. External validation of the COPD detection was performed with 620 low-dose CT (LDCT) scans acquired from the National Lung Screening Trial (NLST) cohort. A multi-channel 3D residual network was further developed to categorize GOLD stages among confirmed COPD patients.

**Results** The attention-based MIL model used for COPD detection achieved an area under the receiver operating characteristic curve (AUC) of 0.934 (95% CI: 0.903, 0.961) on the internal test set and 0.866 (95% CI: 0.805, 0.928) on the LDCT subset acquired from the NLST. The multi-channel 3D residual network was able to correctly grade 76.4% of COPD patients in the test set (423/553) using the GOLD scale.

**Conclusions** The proposed chest CT-DL approach can automatically identify spirometry-defined COPD and categorize patients according to the GOLD scale. As such, this approach may be an effective case-finding tool for COPD diagnosis and staging.

## Key Points

- *Chronic obstructive pulmonary disease is underdiagnosed globally, particularly in developing countries.*
- *The proposed chest computed tomography (CT)–based deep learning (DL) approaches could accurately identify spirometry-defined COPD and categorize patients according to the GOLD scale.*
- *The chest CT-DL approach may be an alternative case-finding tool for COPD identification and evaluation.*

---

Jiaxing Sun, Ximing Liao and Yusheng Yan contributed equally to this work.

---

✉ Kun Wang  
Dr\_Wangk@tongji.edu.cn

✉ Qiang Li  
liqressh1962@163.com

<sup>1</sup> Department of Pulmonary and Critical Care Medicine, Shanghai East Hospital, Tongji University School of Medicine, No. 150 Jimo Road, Pudong, Shanghai, China

<sup>2</sup> Department of Pulmonary and Critical Care Medicine, The Affiliated Hospital of Qingdao University, Qingdao, China

<sup>3</sup> Department of Pulmonary and Critical Care Medicine, Changsha First Hospital, Changsha, China

<sup>4</sup> Department of Pulmonary and Critical Care Medicine, People's Liberation Army Joint Logistic Support Force 920th Hospital, Kunming, Yunnan, China

<sup>5</sup> Department of Pulmonary and Critical Care Medicine, Shandong Provincial Hospital, Jinan, China

<sup>6</sup> Infervision, Beijing, China

<sup>7</sup> College of Aerospace Science and Engineering, National University of Defense Technology, Changsha, China

**Keywords** Chronic obstructive pulmonary disease · Tomography, X-ray computed · Spirometry · Deep learning · Mass screening

### Abbreviations

%LAA-950	The percentage of lung volume less than or equal to −950 Hounsfield units
95% CI	95% confidence interval
AI	Artificial intelligence
AUC	Area under the receiver operating characteristic curve
BMI	Body mass index
COPD	Chronic obstructive pulmonary disease
CT	Computed tomography
DL	Deep learning
FEV1	Forced expiratory volume in 1 second
FVC	Forced vital capacity
IQR	Interquartile range
LDCT	Low-dose computed tomography
MIL	Multi-instance learning
NA	Not applicable
NLST	National Lung Screening Trial
NPV	Negative predictive value
PPV	Positive predictive value
SD	Standard deviation
Yrs	Years

### Introduction

Chronic obstructive pulmonary disease (COPD) is a worldwide public health challenge, due to its high prevalence and long-term effects on related disabilities and mortality [1, 2]. The accurate diagnosis of COPD is crucial for the timely initiation of appropriate therapeutic intervention to improve the patient's quality of life and reduce the risk of future exacerbation [3]. Previous studies have reported that an estimate of over 40% of COPD patients remain undiagnosed, particularly in developing countries [4, 5]. According to Wang et al [6], only 12% of individuals with chronic airflow limitations had a previous spirometry-defined COPD diagnosis during the recent screening of 57,779 participants in China. Furthermore, prior studies have reported that COPD can either be misdiagnosed or missed entirely when using spirometry alone [4, 5, 7]. Therefore, alternative strategies are urgently needed to provide accurate detection and evaluation of COPD for optimal clinical decision-making.

With the growing use of computed tomography (CT) for screening of lung cancer, there is an opportunity to use these scans to identify patients with COPD, with subsequent confirmation by spirometry. In the past few years, growing evidence has shown that CT-derived qualitative and quantitative imaging analyses have potential in COPD diagnosis and stratification [8, 9]. Typical CT features, such as lung parenchyma,

airways, pulmonary vasculature, and the chest wall, provide valuable insights into evaluating lung function, categorizing disease severity, and predicting outcomes for patients with COPD [10–12]. Thus, CT-based imaging can lead to improvements in COPD detection and evaluation [13, 14]. However, objective CT analysis requires prior knowledge of the anatomical and physiological implications of diseases likely to be associated with certain clinical outcomes. In addition, the conventional manual inspection of CT images is often time-consuming and subjective, which limits its use for large-scale COPD screening.

Recent advances in deep learning (DL)-based artificial intelligence (AI) have enabled the direct interpretation of medical images without relying on specific radiographic features of interest [15, 16]. Sophisticated and subtle image patterns (at distinct spatial scales) have been learned by trained models and used to discriminate diseases without any human guidance [17]. As a result, the advantages of a DL strategy for improving the accuracy and efficiency of human COPD detection, and for bolstering human knowledge of COPD subtypes, have in principle been established [18–21]. For example, González et al [22], using the large cohort COPD genetic epidemiology study (COPDGene), trained a 2D convolutional neural network (CNN) for automated COPD detection in smokers, achieving a c-statistic of 85.6%. Using the same cohort, Hatt et al [23] developed CT-based CNN classifiers for COPD assessment. These classification models were also shown to be predictive in the National Lung Screening Trial (NLST) cohort. Recently, Tang and colleagues [24] proposed a novel residual network in the detection of COPD among smokers screened for lung cancer using low-dose CT (LDCT) and achieved an area under the receiver operating characteristic curve (AUC) of more than 88%. It is worth mentioning that these DL models reported in most previous studies were trained using open public datasets that were primarily composed of current or former smokers. Thus, it remains largely unknown whether this approach could be applied to a heterogeneous dataset that extracted from diverse clinical scenarios, especially among the Chinese population.

In the present study, we recruited 1393 participants from outpatient, inpatient, and physical examination center settings of four large hospitals in China. The dataset were highly heterogeneous which we thought could mimic the screening scenario to a great extent. We developed an attention-based multi-instance learning (MIL) model for COPD detection and a multi-channel 3D residual network for the Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage classification among spirometry-confirmed COPD patients. External validation of the COPD detection model was performed with an LDCT subset acquired from the NLST

cohort, which comprised 620 patients with current or previous smoking history.

## Materials and methods

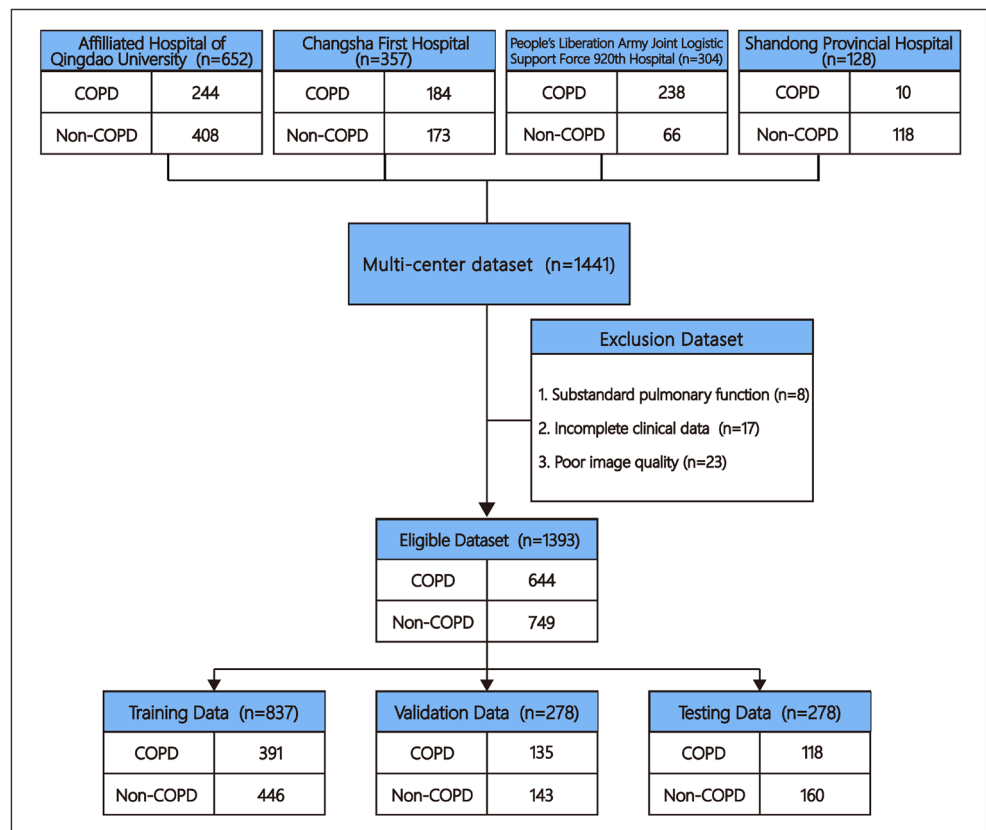
### Data collection

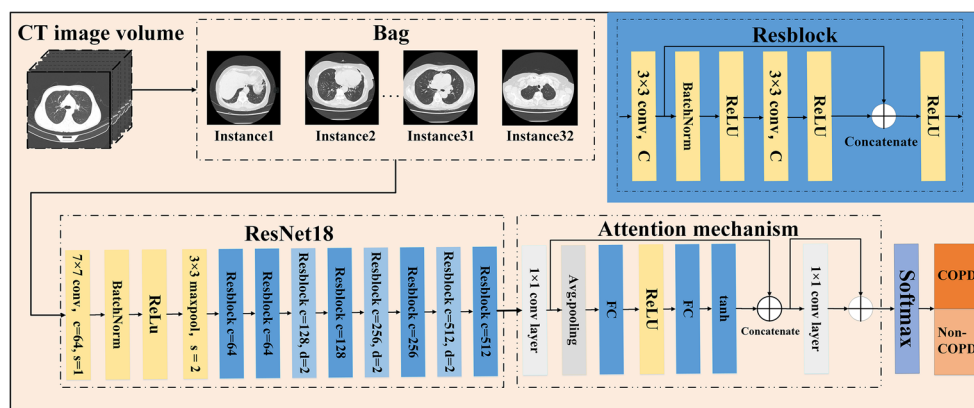
CT image data were retrospectively collected from 1441 participants from outpatient, inpatient, and physical examination center settings of four large public hospitals across China, namely, the Affiliated Hospital of Qingdao University, Changsha First Hospital, People’s Liberation Army Joint Logistic Support Force 920th Hospital, and Shandong Provincial Hospital. The inclusion criteria were (1) participants underwent single inspiratory breath-hold CT scans in the supine position; (2) at least one-time pulmonary function test; and (3) aged 18 years and older and with no history of thoracic surgery. All images were uploaded by principal investigators at each site through the InferScholar research platform (Infervision). Spirometry data, demographic information, smoking history, clinical indices, and underlying diseases were extracted from electronic medical records using a standardized data collection form. The data collection period ranged from August 10, 2019, to October 8, 2020. After excluding cases with incomplete clinical data (17 cases),

substandard pulmonary function (8 cases), and poor CT image quality (23 cases), a total of 1393 participants were enrolled in the final cohort. Eligible subjects were then randomly divided into a training set ( $n = 837$ , 391 with COPD), a validation set ( $n = 278$ , 135 with COPD), and a test set ( $n = 278$ , 118 with COPD) at a ratio of 3:1:1 for subsequent model development (see Figure 1). We further elected to use a random subset of the NLST cohort ( $n = 620$ ) as a means of external validation. The NLST study was conducted by the National Cancer Institute to determine the feasibility of using LDCT for lung cancer screenings and included subjects with spirometry-defined COPD, facilitating an investigation of model efficiency for LDCT and diverse populations [25]. The NLST subset included participants between 55 and 74 years old, with a smoking history of more than 30 pack-years and no self-reported history of lung cancer, thereby allowing us to further evaluate the model efficiency among smokers. Detailed NLST subset information is provided in Table S1. This study was approved by the ethics commissions of all participating hospitals and requirements for written informed consent were waived due to the retrospective nature of the research.

COPD diagnosis was confirmed by forced expiratory volume in 1 second (FEV1) to forced vital capacity (FVC) ratio less than 0.7 after inhalation of bronchodilators. The severity of COPD was graded according to the GOLD standard [2]. CT images were acquired using a range of acquisition protocols

**Fig. 1** A multi-center COPD dataset establishment diagram





**Fig. 2** A flowchart for the proposed COPD detection model. The COPD detection pipeline consisted of three primary steps: (1) preparation of CT lung instances and bags (top left); (2) feature extraction using ResNet18 (bottom left); and (3) an attention mechanism–based classifier (bottom right). The top-right image shows details concerning the ResBlock layer. In 2D ResNet18, we use  $c$ ,  $s$ , and  $d$  to denote number of output channels,

and scanners, representative of clinical routines. Further details regarding image acquisition are provided in Table S2.

## Data preprocessing

Because the CT images were acquired from different vendors with varying scanning parameters, the original data were first adjusted to lung window settings using lower and upper Hounsfield unit (HU) bounds of  $-1500$  and  $600$ , respectively. All images were then resized to a resolution of  $512 \times 512$  pixels using bilinear interpolation and the whole CT volume was normalized.

## Development of the COPD detection model

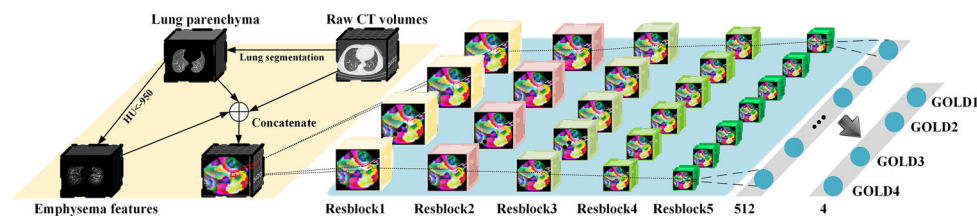
The workflow for the experimental COPD detection model is illustrated in Figure 2 and consists of three primary steps: (1) preparation of CT lung instances and bags; (2) feature extraction using a deep residual neural network, ResNet18; and (3) an attention mechanism–based classifier for COPD detection. Whole CT volumes were divided into multiple parts, with a single axial slice (one instance) being selected from each set and formed into a bag (collection of instances) with defined

strides, and down-sampling factor. “ResBlock,  $c = 128$ ,  $d = 2$ ” denotes a vanilla ResBlock with 128 output channels and a down-sample skip connection that reduces the resolution with a factor of 2 in  $x$ ,  $y$  direction. Conv, convolution; BatchNorm, batch normalization; ReLU, rectified linear unit; Avg, average; FC, fully connected layer

patient labels (COPD vs non-COPD) used for training the network [26]. A weakly supervised approach, MIL, was adopted due to the heterogeneous nature of the COPD CT instances [27]. MIL has previously been used to examine available CT voxels and facilitate the detection of asymptomatic or subtle lesions during screening [28], while keeping computational costs and memory requirements manageable. In the next step, ResNet18 was used for feature extraction, generating a dictionary of visual characteristics from bag instances. Attention mechanisms were further applied to augment the most discriminative features related to COPD, thereby increasing detection accuracy [29]. Finally, the resulting responses were converted into probability values using a softmax classifier. A detailed network architecture and training methodology are provided in Supplemental Appendix 2.

## Development of the COPD staging model

The GOLD stage of confirmed COPD cases was classified by training an end-to-end deep learning model to identify radiographic features suggestive of disease severity. As shown in Figure 3, a lung segmentation algorithm was first applied to raw 3D CT data to create binary lung masks and exclude



**Fig. 3** Illustration of the GOLD stage classification model. Model input is a composite of three channels, including raw CT volumes, segmented lung parenchyma, and emphysema features (CT value  $\leq 950$  HU). Stacked channels were then passed through a 3D ResNet50 network

consisting of five ResBlock layers. A final softmax layer was applied to the output of the fully connected layer to generate four GOLD stage categories. HU, Hounsfield units

unrelated information that might cause confusion or reduce learning efficiency. This segmentation algorithm was developed in-house, derived from a signature U-net architecture [30], and implemented in MxNet. We next employed a multi-channel strategy that included raw CT volumes, segmented lung parenchyma, and emphysema features (percent low-attenuation area less than  $-950$  Hounsfield units, %LAA-950) as model inputs. Stacked channels were concatenated into 3D volumes and passed to a 3D ResNet50 network for post-processing. The proposed 3D ResNet50 consisted of five ResBlock layers capable of processing high-dimensional and complex features for improved prediction outcomes. A final softmax layer was applied to the output of the fully connected layer, to generate four GOLD stage categories. The detailed network architecture and training methodology are provided in Supplemental Appendix 3.

### Model validation

The performance of the proposed attention-based MIL COPD detection model was evaluated using a test set of 278 non-overlapping participants. External validation was further conducted with 620 LDCT scans acquired from the NLST cohort. The receiver operating characteristic (ROC) curves and their confidence interval were determined in accordance with the DeLong methods, to assess the DL model's ability to identify COPD patients from a large heterogeneous dataset. Confusion matrices such as sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and F1 score were determined when applying an optimal threshold selected from the validation set. We also reported the COPD detection accuracy of a common quantitative CT measurement (%LAA-950), as a reference to prior studies reporting similar outcomes. A five-fold cross-validation was used to evaluate the staging performance of the multi-channel COPD staging model. Considering the imbalance in the number of patients within each GOLD stage, micro F1 score and Cohen's weighed kappa were applied to allow comparison.

### Statistical analysis

Measurement data of the baseline clinical and demographic characteristics with normal distribution were presented as mean  $\pm$  standard deviation (SD), and data with non-normal distribution were presented as the median (M) and upper and lower quartile spacing (IQR). Categorical variables were presented as numbers (%). The Wilcoxon signed-rank or Kruskal-Wallis tests were used for numerical variables, and Fisher exact tests were used for categorical variables. No multivariable analyses were conducted, because we deployed each model as an assessment of risk over the entire cohort. Statistical analysis was performed using the IBM SPSS

statistics 20.0 software (SPSS) in the R programming language (version 3.4.0, <http://www.Rproject.org>).

## Results

### Demographic and clinical characteristics

A total of 1393 participants were included in the study: 749 spirometry-defined COPD patients and 644 non-COPD participants. The median age of COPD patients was higher than that of non-COPD participants (62 vs 56,  $p < 0.001$ ) and the majority of the COPD cohort was male (76.09%), which is consistent with COPD gender distributions in China [6]. In addition, a higher proportion of smokers (24.53% vs 4.67%,  $p < 0.001$ ), a reduced FEV1 percentage (52.56% vs 103.25%,  $p < 0.001$ ), and a lower average body mass index (BMI) (22.73 vs 24.02,  $p < 0.001$ ) were evident among the COPD patients. The percentages of stage 1, 2, 3, and 4 spirometry-defined COPD subjects on the GOLD scale were 3.73%, 59.63%, 30.28%, and 6.37%, respectively. Cardiovascular disease was the most common comorbidity within the dataset, followed by asthma in the COPD group and diabetes mellitus in the non-COPD group. Of the non-COPD participants, 376 (50.20%) were healthy subjects with normal CT manifestations and clinical assessments. Detailed demographic and clinical characteristics for the participants are provided in Table 1.

### COPD detection performance

We first examined the overall detection performance of the proposed CT-DL model. The present attention-based MIL algorithm correctly determined the presence or absence of COPD in 243 of 278 subjects in the test set, with an AUC of 0.934 (95% CI: 0.903, 0.961), as shown in Figure 4A. When applying the optimal threshold value (a probability of 0.25 determined by Youden [31] in the validation set), we obtained the sensitivity, specificity, NPV, PPV, and F1 score of 0.805, 0.925, 0.888, 0.865, and 0.894, respectively (see Table 2).

We subsequently evaluated the generalizability of the model among groups categorized by sex, age, CT manufacturer, and slice thickness, as we anticipate that the model will be applicable in diverse clinical settings. The model exhibited relatively robust performance, with AUC values ranging between 0.874 and 1.000 (see Table 2 and Figure 4C). This performance was not affected by modifications to imaging settings or participant demographics. The common quantitative CT measurement %LAA-950 was also used as a reference, producing an AUC of 0.708 (95% CI: 0.648, 0.768) for the same test set when detecting COPD using univariate regression analysis (see Table 2).

**Table 1** Demographic and clinical characteristics for the development dataset

Demographic characteristics	COPD (n = 644)	Non-COPD (n = 749)	p value
Age, Yrs, <i>M</i> (IQR)	62 (22–85)	56 (14–84)	$p < 0.001$
Sex, %male ( <i>n</i> )	76.09 (490)	45.93 (344)	$p < 0.001$
BMI, mean (SD)	22.73 (3.78)	24.02 (3.07)	$p < 0.001$
Former or current smokers, % ( <i>n</i> )	24.53 (158)	4.67 (35)	$p < 0.001$
Pack-years, mean (SD)	36.42 (21.33)	32.08 (22.79)	$p = 0.1451$
FEV1% predicted, mean (SD)	52.56 (14.74)	103.25 (12.80)	$p < 0.001$
GOLD stage, % ( <i>n</i> )			
1	3.73 (24)	NA	NA
2	59.63 (384)	NA	NA
3	30.28 (195)	NA	NA
4	6.37 (41)	NA	NA
Underlying diseases, % ( <i>n</i> )			
Cardiovascular disease	45.65 (294)	15.49 (116)	$p < 0.001$
Diabetes mellitus	18.63 (120)	6.54 (49)	$p < 0.001$
Pulmonary nodule	5.59 (36)	10.01 (75)	$p < 0.005$
Asthma	38.66 (249)	0.80 (6)	$p < 0.001$
Bronchiectasis	9.94 (64)	0.13 (1)	$p < 0.001$
Pneumonia	18.94 (122)	2.67 (20)	$p < 0.001$
Others	2.95 (19)	14.55 (109)	$p < 0.001$
Healthy subjects, % ( <i>n</i> )	NA	50.20 (376)	NA

COPD, chronic obstructive pulmonary disease; Yrs, years; *M* (IQR), median, interquartile range; *SD*, standard deviation; *BMI*, body mass index; *FEV1*, forced expiratory volume in 1 second; *NA*, not applicable

For the external validation dataset (NLST), the model showed an AUC of 0.866 (95% CI: 0.805, 0.928), with the sensitivity and specificity of 0.804 and 0.835, using the same threshold. The confusion matrices revealed that 516 of 620 subjects were accurately categorized, with the sensitivity and specificity of 0.804 and 0.835, respectively, using the optimal threshold of the maximized Youden Index (see Figure 4B). Other measurements, including sensitivity, specificity, PPV, NPV, and F1 score, are summarized in Table 2.

### Feature extraction visualization

The lack of transparency in machine learning can be overcome by applying gradient-weighted class activation mapping (Grad-CAM) to visualize feature extraction using a heatmap [16]. As shown in Figure 5, signature lesions related to COPD detection and differential diagnosis, such as emphysema (A), diffuse exudation (B), bronchiectasis (C), and pulmonary mass (D), were manifest as increased values in the Grad-CAM results, while zero values in the heatmap corresponded to normal regions in the lung. Insights generated from the CT-DL model were compared with manual annotations made by experienced respiratory specialists, and the results indicated that the model pays specific attention to these lesions when distinguishing COPD subjects.

### GOLD stage prediction performance

Confusion matrices showed the number of cases between the spirometry-defined GOLD stage and the differential classification of the proposed CT-DL model in the pooled dataset. The number of accurate GOLD stage predictions is shown in diagonal, with a pooled overall accuracy of 76.4% (423 out of 553) (see Figure 6). Detailed results in Table 3 show that the AUC for classifying GOLD stages 1, 2, 3, and 4 were 0.901, 0.903, 0.848, and 0.952, respectively. The model adopted a Cohen's weighted kappa of 0.619, suggesting a strong agreement between predictions and truth labels. Other measurements within each stage, including sensitivity, specificity, and F1 score are summarized in Table 3.

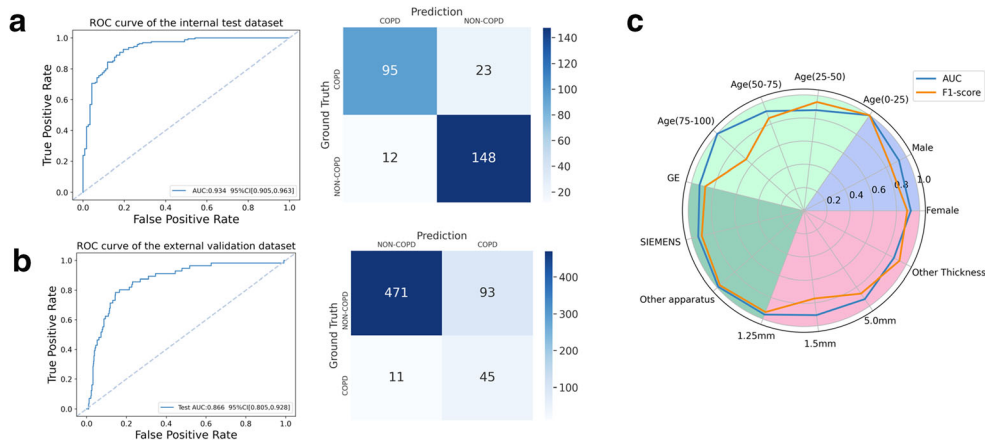
### Discussion

In the present study, an attention-based MIL model was developed to identify spirometry-defined COPD patients using a large and highly heterogeneous collection of CT scans across China. The dataset contained participants recruited from both outpatient and inpatient and physical examination settings. Implemented with the novel DL networks, our model achieved an AUC of 0.934 (95% CI: 0.903, 0.961) in the internal test group of 278 subjects. This DL-based approach

**Table 2** COPD detection performance for the attention-based MIL model

Test set ( <i>n</i> = 278)	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	NPV	PPV	F1 score
Overall performance	0.934 (0.903, 0.961)	0.805 (0.731,0.874)	0.925 (0.881,0.963)	0.888	0.865	0.894
Subgroup—sex						
Female ( <i>n</i> = 115)	0.924 (0.888, 0.963)	0.708 (0.5, 0.88)	0.945 (0.894, 0.989)	0.773	0.925	0.935
Male ( <i>n</i> = 163)	0.929 (0.847, 0.976)	0.829 (0.753, 0.904)	0.899 (0.821, 0.959)	0.918	0.795	0.844
Subgroup—age range						
0–25 years ( <i>n</i> = 10)	1.000	NA	1.000	NA	1.000	1.000
25–50 years ( <i>n</i> = 73)	0.874 (0.733, 0.988)	0.545 (0.222, 0.867)	0.968 (0.919, 1.0)	0.750	0.923	0.945
50–75 years ( <i>n</i> = 175)	0.918 (0.874, 0.953)	0.820 (0.728, 0.893)	0.884 (0.815, 0.947)	0.880	0.826	0.854
75–100 years ( <i>n</i> = 20)	1.000	0.889 (0.812, 0.951)	1.000	1.000	0.500	0.667
Subgroup—apparatus						
GE ( <i>n</i> = 182)	0.929 (0.888, 0.963)	0.918 (0.86, 0.968)	0.798 (0.704, 0.884)	0.841	0.893	0.878
SIEMENS ( <i>n</i> = 66)	0.941 (0.874, 0.988)	0.897 (0.788, 0.976)	0.889 (0.75, 1.0)	0.921	0.857	0.909
Others ( <i>n</i> = 30)	0.989 (0.912, 1)	0.571 (0.391, 0.725)	1.000	1.00	0.884	0.970
Subgroup—slice thickness						
1.25 mm ( <i>n</i> = 164)	0.959 (0.924, 0.988)	0.902 (0.82, 0.969)	0.923 (0.881, 0.978)	0.941	0.887	0.936
1.5 mm ( <i>n</i> = 92)	0.908 (0.844, 0.963)	0.833 (0.731, 0.927)	0.763 (0.625, 0.892)	0.763	0.833	0.763
5 mm ( <i>n</i> = 489)	0.927 (0.796, 0.895)	0.847 (0.926, 0.953)	0.861 (0.818, 0.901)	0.878	0.828	0.869
Others ( <i>n</i> = 52)	0.879 (0.75, 0.983)	0.571 (0.167, 1.0)	0.933 (0.854, 1.0)	0.933	0.571	0.933
Quantitative CT metrics						
%LAA-950	0.708 (0.648,0.768)	0.576 (0.402, 0.715)	0.787 (0.732,0.841)	0.716	0.667	0.618
External validation						
NLST ( <i>n</i> = 620)	0.866 (0.805, 0.928)	0.804 (0.687, 0.907)	0.835 (0.802, 0.86)	0.977	0.326	0.464

AUC, area under the receiver operating characteristic curve; 95% CI, 95% confidence interval; NPV, negative predictive value; PPV, positive predictive value; %LAA-950, percentage of lung volume less than or equal to -950 Hounsfield units; NLST, National Lung Screening Trial



**Fig. 4** COPD detection performance using the attention-based MIL model. Receiver operating characteristic curves and confusion matrices for the MIL model during prediction of COPD in the internal test set. (b) COPD detection performance for the external validation set (NLST). (c) Model

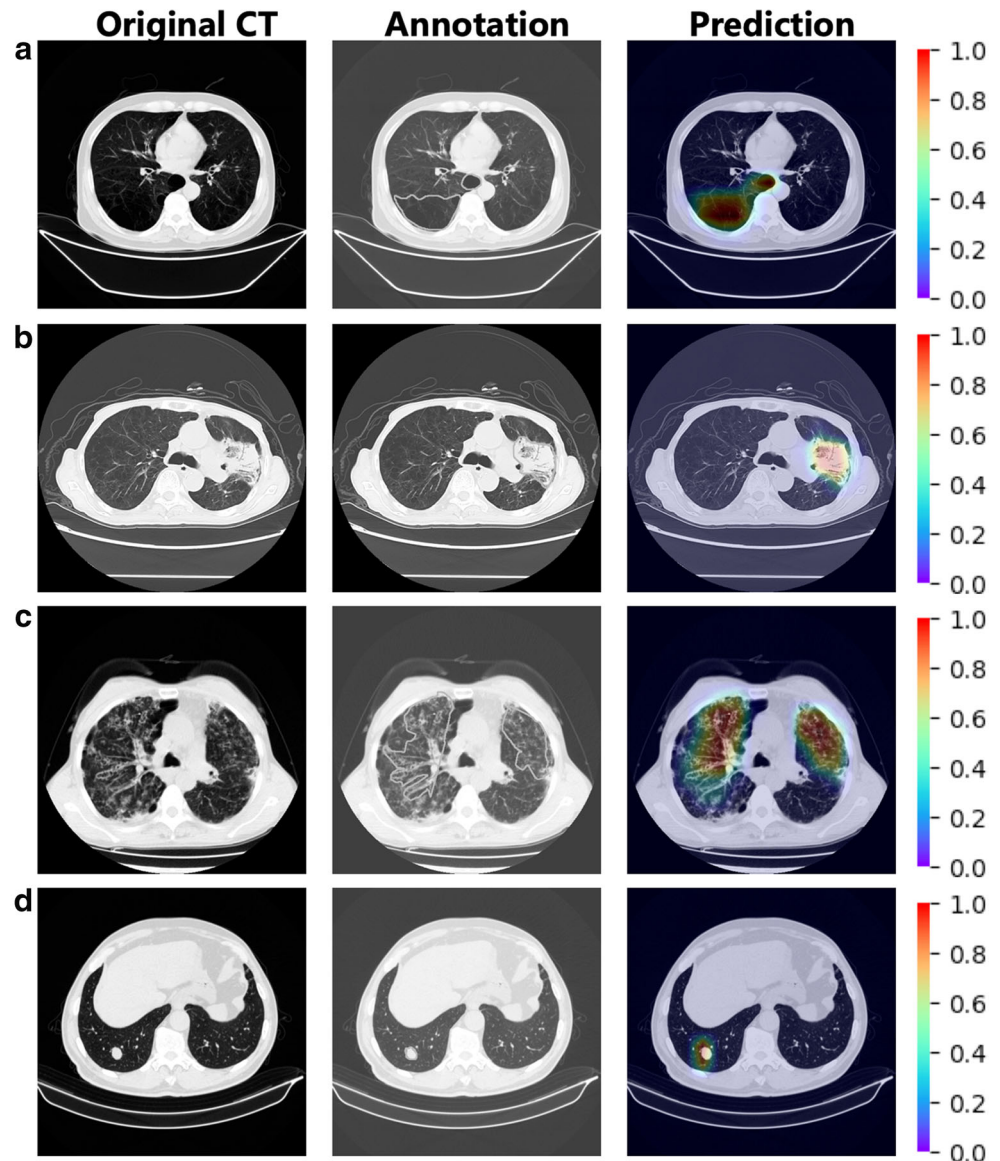
generalizability among groups categorized by sex, age, CT manufacturer, and slice thickness. ROC, receiver operating characteristic curve; AUC, area under the receiver operating characteristic curve; 95% CI, 95% confidence interval; NLST, National Lung Screening Trial

also revealed satisfactory robustness across distinct scanner models, and slice thickness was employed to reconstruct CT scans, with AUC of 0.8 and above. The generalizability of the model was externally validated using a separate dataset collected from a large cohort consisting of LDCT scans (NLST), with the AUC of 0.866 (95% CI: 0.805, 0.928). A multi-channel 3D ResNet50 network was further trained to predict GOLD stages for confirmed COPD patients, achieving an accuracy above 0.8 for every stage. The proposed model offers the requisite performance for detecting COPD and predicting GOLD stage, and it is also the first attempt to apply DL-based approaches to COPD case-finding among the Chinese populace.

Although the heterogeneous pathological nature of COPD has been understood for decades, patients are currently diagnosed primarily by spirometry, a history of

exposure (smoking or other environmental factors), and respiratory symptoms at the time of presentation. Over the last few years, it has become evident that patients without spirometry abnormalities who experience COPD-like respiratory symptoms and acute exacerbation events (with significant pulmonary structural abnormalities) can often be found among these populations [32–34]. Crapo et al [35] presented an analysis of baseline phenotyping and a 5-year longitudinal progression for the COPDGene study, demonstrating that spirometry criteria alone were insufficient to characterize COPD participants among current and former heavy smokers. Results also indicated that quantitative CT metrics outperformed spirometry when predicting disease progression and mortality. Accordingly, CT scans could be used to improve COPD case-finding and evaluation beyond spirometry alone.

**Fig. 5** Visualization of features extracted by the MIL model during COPD detection. The first column shows the original CT axial slices. The second column shows manual annotations of emphysema (a), diffuse exudation (b), bronchiectasis (c), and pulmonary mass (d), identified by experienced respiratory specialists. The third column shows predictions made by the MIL model. The MIL model correctly predicted abnormalities that were highlighted on attention heatmaps. The color depth of the heatmaps represented the possibility of predicted lesions by the model





**Table 3** GOLD stage prediction performance for the multi-channel 3D residual network applied to the test set

Class	Sensitivity	Specificity	F1 score	AUC (95% CI)
GOLD1	0.474	0.994	0.581	0.901 (0.808, 0.994)
GOLD2	0.853	0.815	0.798	0.903 (0.874, 0.932)
GOLD3	0.748	0.811	0.754	0.848 (0.814, 0.882)
GOLD4	0.629	0.986	0.727	0.952 (0.917, 0.987)
Micro Avg.	0.765	0.922	0.765	0.912 (0.882, 0.941)

AUC, area under the receiver operating characteristic curve; 95% CI, 95% confidence interval

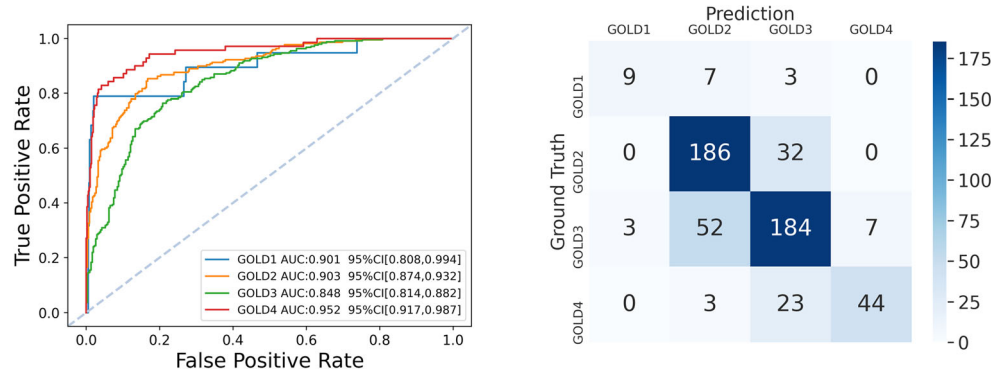
The development of AI for large-scale data processing has increasingly led to the use of DL-based techniques in establishing a direct link between diagnostic images and disease categorization [10, 36]. This approach overcomes the limitations of conventional manual CT image inspection, such as inter/intra-observer variability and heavy workloads. It also bypasses the requirement of prior knowledge of radiographic features, which is required for quantitative CT analysis. González et al [22] have trained CNN models using CT images alone to detect and stage COPD among 7983 COPDGene participants, and achieved an accuracy of 0.773 for 1000 non-overlapping participants. In addition, Hatt et al [23] implemented CNN models with accuracy of 0.777 in COPDGene cohort and 0.762 in the NLST cohort. Due to the constrained graphical processing units’ capability, both studies adopted a strategy of extracting a subset of CT slices to build up a single montage for an individual instead of using the full images. The spatial heterogeneity nature of COPD has then prompted researchers to adopt an MIL strategy in COPD detection, as only the label of COPD or non-COPD is needed in case-finding scenarios. Bruijne et al [37] reported that the use of MIL combined with support vector machine (MI-SVM) could facilitate the detection of COPD by automatically identifying emphysema regions on CT images. Xu et al [26] also proposed a deep CNN transferred multiple instance learning (DCT-MIL) model to identify COPD using CT images from 280 participants including 190 COPD patients, with an accuracy of 99.29% and AUC of 0.983. The analysis process used

in the present study differed from that used in the aforementioned studies in terms of patient selection and disease spectrum distribution. Meanwhile, we adopted a novel attention-based MIL strategy that improved the proportion of lesion character information without increasing the computational power and costs. A multi-channel 3D ResNet50 network allowed the model to extract spatial information between slices and identify abnormal images exhibiting relatively small regions of interest, further improving staging performance (see Supplemental Appendices 2–4).

This study offers several clinical benefits. The deep learning model was trained using subjects recruited from both respiratory clinics and health management centers, thus including participants with normal spirometry as well as CT results. This scenario is representative of diverse clinical situations in which COPD patients could be detected. Previous attempts using DL algorithms for COPD detection have mostly been trained using cohorts enrolling former and current smokers, which may not truly reflect case-finding settings. While researchers from the COPDGene and ECLIPSE cohorts have reported desired COPD imaging results, it is crucial to further extend this expertise into a Chinese population, because a very small percentage of subjects from the aforementioned studies were ethnically Chinese. Furthermore, the increased use of LDCT for pulmonary nodule assessment and lung cancer screenings has created an opportunity to apply the present model to COPD detection, with subsequent confirmation using spirometry. This is particularly relevant, as our model was generalized to LDCT in the NLST subset.

The present study does have some limitations. First, spirometry was used to diagnose COPD instead of symptoms or radiographs, which may prevent our algorithm from being generalized to the detection of COPD in patients without air-flow limitations, such as para-septal emphysema. This was a result of the relatively objective criteria used for enrollment. Second, the size of our cohort is relatively small compared with other larger cohorts, and the inclusion criteria may lead to some undesired biases. Third, a limited number of subjects with GOLD 1 and 4 were enrolled, which might constrain the model’s stratification capacity and resulted in the discrepancy

**Fig. 6** Confusion matrices of five-fold cross-validation performed on the multi-channel 3D ResNet50 network in classifying the GOLD stage. AUC, area under the receiver operating characteristic curve; 95% CI, 95% confidence interval



of staging efficiency between groups. Fourth, the NLST dataset is considered as a poor validation cohort given its predominate portion of smokers with high risk of lung cancer. To improve the efficiency of detection and staging, we are currently recruiting more participants and aim to optimize our cohort in the future. Last but certainly not least, the ability of DL to detect and stage COPD without specification of clinical or radiographic characteristics could be considered both a strength and a weakness. The “black box” nature of the DL model may severely limit its utility in clinical situations, as it does not provide sufficient information to clinicians concerning its decision-making process. Future work is urgently needed to elucidate the decision path.

In conclusion, we developed and validated a CT-DL approach for detecting and staging of spirometry-defined COPD via a highly heterogeneous Chinese population cohort. The proposed model approach achieved the desired performance and could serve as a powerful tool for COPD case-finding, providing useful indicators for clinicians and clinically relevant findings that could improve management and follow-up treatment for specific patients.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00330-022-08632-7>.

**Acknowledgements** We thank LetPub ([www.letpub.com](http://www.letpub.com)) for linguistic assistance and pre-submission expert review.

**Funding** This work was supported by the National Key R&D Program (2018YFC1313700), the National Natural Science Foundation of China (grant nos. 82100089, 81870064, and 82070086), and the “Gaoyuan” project of Pudong Health and Family Planning Commission (PWYgy2018-06).

## Declarations

**Guarantor** The scientific guarantor of this publication is professor Qiang Li, the head of the Department of Pulmonary and Critical Care Medicine of Shanghai East Hospital, Tongji University School of Medicine.

**Conflict of interest** The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

**Informed consent** The requirement for written informed consent was waived due to the retrospective nature of the study.

**Ethical approval** This study was approved by the ethics commissions of all participating hospitals, including the Affiliated Hospital of Qingdao University, Changsha First Hospital, People’s Liberation Army Joint Logistic Support Force 920th Hospital, and Shandong Provincial Hospital.

## Methodology

- retrospective
- diagnostic or prognostic study
- multi-center study

## References

1. Disease GBD, Injury I, Prevalence C (2016) Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* 388:1545–1602
2. Halpin DMG, Criner GJ, Papi A et al (2021) Global initiative for the diagnosis, management, and prevention of chronic obstructive lung disease. The 2020 GOLD Science Committee report on COVID-19 and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 203:24–36
3. Jithoo A, Enright PL, Burney P et al (2013) Case-finding options for COPD: results from the Burden of Obstructive Lung Disease study. *Eur Respir J* 41:548–555
4. Perez-Padilla R, Thirion-Romero I, Guzman N (2018) Underdiagnosis of chronic obstructive pulmonary disease: should smokers be offered routine spirometry tests? *Expert Rev Respir Med* 12:83–85
5. Lamprecht B, Soriano JB, Studnicka M et al (2015) Determinants of underdiagnosis of COPD in national and international surveys. *Chest* 148:971–985
6. Wang C, Xu J, Yang L et al (2018) Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study. *Lancet* 391:1706–1717
7. Miller MR, Levy ML (2015) Chronic obstructive pulmonary disease: missed diagnosis versus misdiagnosis. *BMJ* 351:h3021
8. Estepar RS, Kinney GL, Black-Shinn JL et al (2013) Computed tomographic measures of pulmonary vascular morphology in smokers and their clinical implications. *Am J Respir Crit Care Med* 188:231–239
9. McDonald ML, Diaz AA, Ross JC et al (2014) Quantitative computed tomography measures of pectoralis muscle area and disease severity in chronic obstructive pulmonary disease. A cross-sectional study. *Ann Am Thorac Soc* 11:326–334
10. Bhatt SP, Washko GR, Hoffman EA et al (2019) Imaging advances in chronic obstructive pulmonary disease. Insights from the Genetic Epidemiology of Chronic Obstructive Pulmonary Disease (COPDGene) study. *Am J Respir Crit Care Med* 199:286–301
11. Park J, Hobbs BD, Crapo JD et al (2019) Subtyping COPD using visual and quantitative CT features. *Chest*. <https://doi.org/10.1016/j.chest.2019.06.015>
12. Washko GR, Parraga G (2018) COPD biomarkers and phenotypes: opportunities for better outcomes with precision imaging. *Eur Respir J* 52
13. Kauczor HU, Bonomo L, Gaga M et al (2015) ESR/ERS white paper on lung cancer screening. *Eur Radiol* 25:2519–2531
14. Lathan C, Frank DA (2013) ACP Journal Club. Review: Low-dose CT screening reduces lung cancer and mortality in current or former smokers. *Ann Intern Med* 159(JC3)
15. Chassagnon G, Vakalopoulou M, Paragios N, Revel MP (2020) Deep learning: definition and perspectives for thoracic imaging. *Eur Radiol* 30:2021–2030
16. Philbrick KA, Yoshida K, Inoue D et al (2018) What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. *AJR Am J Roentgenol* 211: 1184–1193
17. Kermany DS, Goldbaum M, Cai W et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172:1122–1131 e1129
18. Cho YH, Lee SM, Seo JB et al (2018) Quantitative assessment of pulmonary vascular alterations in chronic obstructive lung disease: associations with pulmonary function test and survival in the KOLD cohort. *Eur J Radiol* 108:276–282

19. Lynch DA, Moore CM, Wilson C et al (2018) CT-based visual classification of emphysema: association with mortality in the COPDGene study. *Radiology* 288:859–866
20. Peng L, Lin L, Hu H et al (2019) Classification and quantification of emphysema using a multi-scale residual network. *IEEE J Biomed Health Inform* 23:2526–2536
21. Nambu A, Zach J, Schroeder J et al (2016) Quantitative computed tomography measurements to evaluate airway disease in chronic obstructive pulmonary disease: relationship to physiological measurements, clinical index and visual assessment of airway disease. *Eur J Radiol* 85:2144–2151
22. Gonzalez G, Ash SY, Vegas-Sanchez-Ferrero G et al (2018) Disease staging and prognosis in smokers using deep learning in chest computed tomography. *Am J Respir Crit Care Med* 197:193–203
23. Hatt CR, Galban CJ, Labaki W, Kazerooni EA, Han ML (2018) Convolutional neural network based COPD and emphysema classifications are predictive of lung cancer diagnosis. *Image Analysis for Moving Organ, Breast, and Thoracic Images*:11040
24. Tang LYW, Coxson HO, Lam S, Leipsic J, Tam RC, Sin DD (2020) Towards large-scale case-finding: training and validation of residual networks for detection of chronic obstructive pulmonary disease using low-dose CT. *Lancet Digit Health* 2:e259–e267
25. National Lung Screening Trial Research T, Aberle DR, Adams AM et al (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 365:395–409
26. Xu C, Qi S, Feng J et al (2020) DCT-MIL: deep CNN transferred multiple instance learning for COPD identification using CT images. *Phys Med Biol* 65:145011
27. Yan X, Tao M, Feng Q, Zhong P, Chang IC (2014) Deep learning of feature representation with multiple instance learning for medical image analysis. *IEEE International Conference on Acoustics*,
28. Shen Y, Wu N, Phang J et al (2021) An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med Image Anal* 68:101908
29. Ilse M, Tomczak JM, Welling M (2018) Attention-based deep multiple instance learning. *The 35th International Conference on Machine Learning*.
30. Alom MZ, Yakopcic C, Hasan M, Taha TM, Asari VK (2019) Recurrent residual U-Net for medical image segmentation. *J Med Imaging (Bellingham)* 6:014006
31. Fluss R, Faraggi D, Reiser B (2005) Estimation of the Youden Index and its associated cutoff point. *Biom J* 47:458–472
32. Young KA, Strand M, Ragland MF et al (2019) Pulmonary subtypes exhibit differential global initiative for chronic obstructive lung disease spirometry stage progression: the COPDGene(R) study. *Chronic Obstr Pulm Dis* 6:414–429
33. Kinney GL, Santorico SA, Young KA et al (2018) Identification of chronic obstructive pulmonary disease axes that predict all-cause mortality: the COPDGene study. *Am J Epidemiol* 187:2109–2116
34. Woodruff PG, Barr RG, Bleecker E et al (2016) Clinical significance of symptoms in smokers with preserved pulmonary function. *N Engl J Med* 374:1811–1821
35. Lowe KE, Regan EA, Anzueto A et al (2019) COPDGene((R)) 2019: redefining the diagnosis of chronic obstructive pulmonary disease. *Chronic Obstr Pulm Dis* 6:384–399
36. Esteva A, Robicquet A, Ramsundar B et al (2019) A guide to deep learning in healthcare. *Nat Med* 25:24–29
37. Pino Pena I, Cheplygina V, Paschaloudi S et al (2018) Automatic emphysema detection using weakly labeled HRCT lung images. *PLoS One* 13:e0205397

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.