Detection and Staging of Chronic Obstructive Pulmonary Disease Using a Computed Tomography-based Weakly Supervised Deep Learning Approach

Jiaxing Sun
Shanghai East Hospital
Ximing Liao
Shanghai East Hospital
Yusheng Yan
First Hospital of Changsha
Xin Zhang
People's Liberation Army Joint Logistic Support Force 960th Hospital
Jian Sun
Shandong Provincial Hospital
Weixiong Tan
Infervision
Baiyun Liu
Infervision
Jiangfen Wu
Infervision
Qian Guo
Shanghai East Hospital
Shaoyong Gao
Shanghai East Hospital
Zhang Li
National University of Defense Technology
Kun Wang (Dr_Wangk@tongji.edu.cn)
Tongji University Affilliated East Hospital: Shanghai East Hospital https://orcid.org/0000-0003-2519-5070
Qiang Li
Shanghai East Hospital

Research

Keywords: chronic obstructive pulmonary disease, computed tomography, spirometry, deep learning, case-finding

Posted Date: July 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-666412/v1

Version of Record: A version of this preprint was published at European Radiology on February 24th, 2022. See the published version at https://doi.org/10.1007/s00330-022-08632-7.

Abstract

Background

Chronic obstructive pulmonary disease (COPD) remains underdiagnosed globally. The coronavirus disease 2019 pandemic has also severely restricted spirometry, the primary tool used for COPD diagnosis and severity evaluation, due to concerns of virus transmission. Computed tomography (CT)-based deep learning (DL) approaches have been suggested as a cost-effective alternative for COPD identification within smokers. The present study aims to develop weakly supervised DL models that utilize CT image data for the automated detection and staging of spirometry-defined COPD among natural population.

Methods

A large, highly heterogenous dataset was established comprising 1393 participants recruited from outpatient, inpatient and physical examination center settings of 4 large public hospitals in China. CT scans, spirometry data, demographic data, and clinical information of each participant were collected for the purpose of model development and evaluation. An attention-based multi-instance learning (MIL) model for COPD detection was trained using CT scans from 837 participants and evaluated using a test set comprised of data from 278 non-overlapping 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 and evaluated using 5-fold cross validation. Spirometry tests were used to diagnose COPD, with stages defined according to the GOLD criteria.

Results

The attention-based MIL model used for COPD detection achieved an area under the receiver operating characteristic curve (AUC) of 0.934 on the test set and 0.866 on the LDCT subset acquired from NLST. The model exhibited high generalizability across distinct scanning devices and slice thicknesses, with an AUC above 0.90. 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, with a Cohen’s weighted Kappa of 0.619 for the assessment of GOLD categorization .

Conclusion

The proposed chest CT-DL approach can automatically identify spirometry-defined COPD and categorize patients according to the GOLD scale, with clinically acceptable performance. As such, this approach may be a powerful novel tool for COPD diagnosis and staging at the population level.

Background

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 therefore 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 estimated of over 40% of COPD patients remain undiagnosed, especially in developing countries (4, 5). Only 12% of individuals with chronic airflow limitations had a previous spirometry-defined COPD diagnosis during the recent screening of 57779 participants in China (6). The coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has severely restricted spirometry tests, which are the primary tool used for COPD case-finding and severity evaluation, due to concerns over droplet and aerosol generation (7, 8). As such, multiple respiratory societies have recommended postponing or limiting spirometry during the COVID-19 pandemic, to prevent virus transmission (9, 10). Prior studies have also reported that COPD can either be misdiagnosed or missed entirely when using spirometry alone (4, 5, 11). Thus, alternative, safe, and effective strategies are in urgent need to provide accurate detection and evaluation of COPD for optimal clinical decision making.

In the past few years, several studies using both qualitative and quantitative imaging techniques have demonstrated the utility of computed tomography (CT) imaging in assessing patients with COPD (12, 13). Typical CT features, such as lung parenchyma, airways, pulmonary vasculature, and the chest wall provide valuable insights in evaluating lung function, categorizing disease severity, and predicting outcomes for patients with COPD (14–16). Thus, CT-based imaging could lead to improvements in COPD detection and evaluation (17, 18). 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 usage for large-scale COPD diagnosis.

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 (19, 20). Sophisticated and subtle image patterns (at distinct spatial scales) have been learned by trained models and used to discriminate diseases without any human guidance (21). As a result, the advantages of a DL strategy for improving the accuracy and efficiency of human COPD detection, and bolstering human knowledge of COPD subtypes, have in principle been established (22–25). It is worth mentioning that DL models proposed in most previous studies were developed on open public datasets primarily composed of current or former smokers instead of the 'real-world' settings. For example, Gonzàlez et al. 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% (26). Later, Lisa et al. developed a novel residual network in the detection of COPD among smokers screened for lung cancer and achieved an area under the receiver operating characteristic curve (AUC) of more than 88% (27). Thus, it remains largely unknown whether this approach could be applied to real-world screening scenarios with clinically acceptable performance.

In the present study, we recruited 1393 participants from outpatient, inpatient and physical examination center settings of 4 large hospitals in China. The dataset were highly heterogenous which we thought could mimic the 'real world' scenerio 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 GOLD stage classification among spirometric-confirmed COPD patients. External validation of the COPD detection model was performed with a low-dose CT (LDCT) subset acquired from the National Lung Screening Trial (NLST) cohort, which comprising 620 patients with current or previous smoking history.

Methods

Data Collection

CT image data were retrospectively collected from 1441 participants at outpatient, inpatient and physical examination center settings of 4 large public hospitals in China, including the Affiliated Hospital of Qingdao University, Changsha First Hospital, People's Liberation Army Joint Logistic Support Force 920th Hospital, and Shandong Provincial Hospital. All images were uploaded by principal investigators at each site through the InferScholar research platform (Infervision, Beijing, China). Spirometry data, demographic information, smoking history, clinical indices, and underlying diseases were extracted from electronic medical records using a standardized data collection form. Data collection periods ranged from August 10, 2019 to October 8, 2020. After excluding cases with incomplete clinical data (17 cases), substandard pulmonary function (8 cases), or poor CT image quality (23 cases), a total of 1393 participants were enrolled in the final cohort and randomly divided into a training set (60%), a validation set (20%), and a test set (20%) for subsequent model development (see Fig. 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 (28). 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 thus we could further evaluate the model efficiency among smokers. Detailed NLST subset information are 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) or a forced vital capacity (FVC) ratio of 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 and scanners, representative of clinical routines. Further details regarding image acquisition are provided in **Table S2**.

Data Pre-processing

Since 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 × 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 Fig. 2 and consists of three primary steps: (1) preparation of CT lung instances and bags; (2) feature extraction using a 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 patient labels (COPD vs Non-COPD) used for training the network (29). A weakly supervised approach, multiple instance learning (MIL), was adopted due to the heterogeneous nature of the COPD CT instances (30). MIL has previously been used to examine available CT voxels and facilitate the detection of asymptomatic or subtle lesions during screening (31), while keeping computational costs and memory requirements manageable. In the next step, a deep residual neural network (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 (32). 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 Fig. 3, a lung segmentation algorithm was first applied to raw 3D CT data to create binary lung masks and exclude unrelated information that may cause confusion or reduce learning efficiency. This segmentation algorithm was developed in-house, derived from a signature U-net architecture (33), and implemented in MxNet. We next employed a multi-channel strategy that included raw CT volumes, segmented lung parenchyma, and emphysema features (the percentage of lung volume less than or equal to − 950 HU, %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 curves (ROC) and their confidence interval were determined in accordance with the DeLong methods, to assess the DL model's ability in identifying 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 measurements (%LAA-950), as a reference to prior studies reporting similar outcomes. A 5-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 enable comparison.

Statistical Analysis

Measurement data of the baseline clinical and demographic characteristics with normal distribution were presented as mean ± standard deviation (SD) and with non-normal distributions were presented as the median (M) and upper and lower quartile spacing (IQR). Categorical variables were presented as numbers (%). The Wilcoxon singed-rank or Kruskal-Wallis tests were used for numerical variables, and Fisher exact tests was used for categorical variables. No multivariable analyses were conducted, since 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, Chicago, IL, USA) 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 (34). 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 I, II, III, and IV 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 milletus in the non-COPD group. Among 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.

Table 1
Demographic and clinical characteristics for the development dataset.

Demographic Characteristics	COPD (n = 644)	Non-COPD (n = 749)	P value
Age, Yrs, M (IQR)	62 (22–85)	56 (14–84)	P < 0.001
Sex, %Male (n)	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, % (n)	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, % (n)			
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, % (n)			
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, %(n)	NA	50.20 (376)	NA

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

COPD Detection Performance

We first examined the overall detection performance for the proposed 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 Fig. 4A. When applying the optimal threshold value (a probability of 0.24 determined by Youden 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).

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

Test Set (n = 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 (n = 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 (n = 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 (n = 10)	1.000	NA	1.000	NA	1.000	1.000
25–50 years (n = 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 (n = 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 (n = 20)	1.000	0.889 (0.812, 0.951)	1.000	1.000	0.500	0.667
Subgroup-Apparatus						
GE (n = 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

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.

Test Set (n = 278)	AUC (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)	NPV	PPV	F1 score
SIEMENS (n = 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 (n = 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 (n = 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 (n = 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 (n = 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 (n = 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 (n = 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.

We subsequently evaluated the generalizability of the model among groups categorized by sex, age, CT manufacturer and slice thickness, as we anticipate the model to be applicable in diverse clinical settings. The model exhibited relatively robust performance, with AUC values ranging between 0.874–1.000 (see Table 2 and Fig. 4C). This performance was not affected by modifications to imaging settings or participant demographics. A 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).

For 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 (see Fig. 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 (Grad-CAM) to visualize feature extraction using a heatmap (35). As shown in Fig. 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, implying zero values in the heatmap corresponded to normal regions in the lung. Insights generated from MIL were compared with manual annotations made by experienced respiratory specialists and results indicated 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 spirometric-defined GOLD stage and the differential classification of propose DL model in the pooled dataset. Number of accurate prediction of the GOLD stage were shown in diagonal, with the pooled overall accuracy of 76.4% (423 out of 553) (see Fig. 6). Detailed results in Table 3 showed 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.

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.				

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 from multiple scanner manufacturers, slice thickness, and institutes in China. It is also a 'real-world' dataset containing participants recruited from both outpatient, inpatient and physical examination scenario. Implemented with the novel DL networks, our model achieved an AUC of 0.934 for the test group. This DL-based approach also revealed satisfactory robustness across distinct scanner models, and slice thickness 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 comprised 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. To our knowledge, the proposed model offers the best performance for detecting COPD and predicting GOLD stage to date. It is also the first attempt to apply DL-based approaches to COPD case-finding among a natural population in Chinese.

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 or acute exacerbation events (with significant pulmonary structural abnormalities) can often be found among these populations (25, 26). Carpo et al. 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 (27). Results also indicated quantitative CT metrics outperformed spirometry when predicting disease progression and mortality. Thus, CT scans could be used to improve COPD case-finding and evaluation beyond spirometry alone.

The development of artificial intelligence for large-scale data processing has increasingly led to the use of ML-based techniques in establishing a direct link between diagnostic images and disease categorization (14, 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. Previous studies by Gonzàlez and Lisa et al. (23, 24) have explored the application of ML-based methods to CT image analysis for COPD detection and evaluation. The analysis process used in this study differed in terms of patient selection and disease spectrum distribution. Most notably, we adopted a novel attention-based MIL strategy, thus increasing the proportion of lesion character information and achieving high robustness with a relatively limited training set. 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 (ROIs), further improving staging performance (see **Supplemental Appendix 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 both normal spirometry and CT results. This scenario is representative of diverse clinical situations in which COPD could be detected among the general population. 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 in real world settings. While researchers from the COPDgene and ECLIPSE cohorts have reported desired COPD imaging results, it is crucial to further expand this expertise into a Chinese population, as a very small proportion of subjects from these studies were ethnically Chinese (11). 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 since our model was generalized to LDCT in the NLST subset.

The present study does include 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 airflow limitations, such as paraseptal emphysema. This was a result of the relatively objective criteria used for enrollment. However, focusing on this participant group remains practical as they are associated with increased morbidity and mortality rates and could benefit from early detection and proper management (2). Second, the size of our cohort is relatively small compared to other large cohorts, such as COPDgene and ECLIPSE. Though we have adopted strategies to improve the efficiency of detection and staging, performance was unsatisfactory in particular subgroups partly due to the insufficient patients enrolled. We are currently recruiting more participants and hope to optimize our cohort in the future. Third, the ability of DL to detect and stage COPD without specification of clinical or radiographic characteristics could be both a strength and a weakness. The ‘black box’ nature of DL 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.

Conclusion

A highly heterogenous Chinese population consisting of spirometry-defined COPD patients were detected and staged, according to the GOLD scale, using a novel DL technique involving chest CT data. The proposed DL approach achieved clinically acceptable performance and could serve as a powerful tool for COPD case-finding within a general population, providing useful indicators for clinicians and clinically relevant findings that could improve management and follow-up treatment for specific patients.

Abbreviations

COPD
chronic obstructive pulmonary disease
CT
computed tomography
COVID-19
coronavirus disease 2019
SARS-CoV-2
severe acute respiratory syndrome coronavirus 2
LDCT
low-dose computed tomography
MIL
multi-instance learning
DL
deep learning
AI
artificial intelligence
AUC
area under the receiver operating characteristic curve
FEV1
the forced expiratory volume in one second
FVC
forced vital capacity
%LAA-950
the percentage of lung volume less than or equal to − 950 Hounsfield Units
Yrs
years
IQR
inter-quartile range
SD
standard deviation

BMI
body mass index
NA
not applicable
95% CI
95% confidence interval
NPV
negative predictive value
PPV
positive predictive value
NLST
the national lung screening trial

Declarations

Ethics approval and consent to participate

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

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the National Key R&D Program (2018YFC1313700), the National Natural Science Foundation of China (grants no. 81870064 and 82070086), and the "Gaoyuan" project of Pudong Health and Family Planning Commission (PWYgy2018-06).

Author contributions

Kun Wang is responsible for the content of the manuscript, including the data and analysis. Kun Wang, Qiang Li, and Jiaxing Sun conceived and designed the study. Jiaxing Sun, Ximing Liao, Yusheng Yan, Jian Sun, Xin Zhang, Shaoyong Gao, and Qian Guo coordinated collection of the data with technical guidance from Kun Wang. Kun Wang, Weixiong Tan, Jiangfen Wu, and Baiyun Liu analyzed and interpreted the data and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank LetPub (www.letpub.com) for linguistic assistance and pre-submission expert review.

References

1. Global regional. and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1736–88. DOI:10.1016/S0140-6736(18)32203-7.
2. Halpin D, Criner GJ, Papi A, et al. 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 2021;203(1):24–36. DOI: 10.1164/rccm.202009-3533SO.
3. Jithoo A, Enright PL, Burney P, et al. Case-finding options for COPD: results from the Burden of Obstructive Lung Disease study. Eur Respir J. 2013;41(3):548–55. DOI:10.1183/09031936.00132011.
4. Lamprecht B, Soriano JB, Studnicka M, et al. Determinants of underdiagnosis of COPD in national and international surveys. Chest. 2015;148(4):971–85. DOI:10.1378/chest.14-2535.
5. Perez-Padilla R, Thirion-Romero I, Guzman N. Underdiagnosis of chronic obstructive pulmonary disease: should smokers be offered routine spirometry tests? Expert Rev Respir Med. 2018;12(2):83–5. DOI:10.1080/17476348.2018.1419868.
6. Wang C, Xu J, Yang L, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study. Lancet. 2018;391(10131):1706–17. DOI:10.1016/S0140-6736(18)30841-9.
7. Helgeson SA, Lim KG, Lee AS, et al. Aerosol Generation during Spirometry. Ann Am Thorac Soc. 2020;17(12):1637–9. DOI:10.1513/AnnalsATS.202005-569RL.
8. Li J, Jing G, Fink JB, et al. Airborne Particulate Concentrations During and After Pulmonary Function Testing. Chest. 2021;159(4):1570–4. DOI:10.1016/j.chest.2020.10.064.
9. Hull JH, Lloyd JK, Cooper BG. Lung function testing in the COVID-19 endemic. Lancet Respir Med. 2020;8(7):666–7. DOI:10.1016/S2213-2600(20)30246-0.
10. Saunders MJ, Haynes JM, McCormack MC, et al. How Local SARS-CoV-2 Prevalence Shapes Pulmonary Function Testing Laboratory Protocols and Practices During the COVID-19 Pandemic. Chest 2021. DOI: 10.1016/j.chest.2021.05.011.

11. Miller MR, Levy ML. Chronic obstructive pulmonary disease: missed diagnosis versus misdiagnosis. BMJ. 2015;351:h3021. DOI:10.1136/bmj.h3021.
12. Estepar RS, Kinney GL, Black-Shinn JL, et al. Computed tomographic measures of pulmonary vascular morphology in smokers and their clinical implications. Am J Respir Crit Care Med. 2013;188(2):231–9. DOI:10.1164/rccm.201301-0162OC.
13. McDonald ML, Diaz AA, Ross JC, et al. Quantitative computed tomography measures of pectoralis muscle area and disease severity in chronic obstructive pulmonary disease. A cross-sectional study. Ann Am Thorac Soc. 2014;11(3):326–34. DOI:10.1513/AnnalsATS.201307-229OC.
14. Bhatt SP, Washko GR, Hoffman EA, et al. 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. 2019;199(3):286–301. DOI:10.1164/rccm.201807-1351SO.
15. Park J, Hobbs BD, Crapo JD, et al. Subtyping COPD by Using Visual and Quantitative CT Imaging Features. Chest. 2020;157(1):47–60. DOI:10.1016/j.chest.2019.06.015.
16. Washko GR, Parraga G. COPD biomarkers and phenotypes: opportunities for better outcomes with precision imaging. Eur Respir J 2018;52(5). DOI:10.1183/13993003.01570-2018.
17. Kauczor HU, Bonomo L, Gaga M, et al. ESR/ERS white paper on lung cancer screening. Eur Radiol. 2015;25(9):2519–31. DOI:10.1007/s00330-015-3697-0.
18. Lathan C, Frank DA. ACP Journal Club. Review: low-dose CT screening reduces lung cancer and mortality in current or former smokers. Ann Intern Med. 2013;159(10):JC3. DOI:10.7326/0003-4819-159-10-201311190-02003.
19. Chassagnon G, Vakalopolou M, Paragios N, et al. Deep learning: definition and perspectives for thoracic imaging. Eur Radiol. 2020;30(4):2021–30. DOI:10.1007/s00330-019-06564-3.
20. Shin HC, Roth HR, Gao M, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging. 2016;35(5):1285–98. DOI:10.1109/TMI.2016.2528162.
21. Kermany DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018;172(5):1122–31.e9. DOI:10.1016/j.cell.2018.02.010.
22. Cho YH, Lee SM, Seo JB, et al. 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. 2018;108:276–82. DOI:10.1016/j.ejrad.2018.09.013.
23. Lynch DA, Moore CM, Wilson C, et al. CT-based Visual Classification of Emphysema: Association with Mortality in the COPDGene Study. Radiology. 2018;288(3):859–66. DOI:10.1148/radiol.2018172294.
24. Peng L, Lin L, Hu H, et al. Classification and Quantification of Emphysema Using a Multi-Scale Residual Network. IEEE J Biomed Health Inform. 2019;23(6):2526–36. DOI:10.1109/JBHI.2018.2890045.
25. Nambu A, Zach J, Schroeder J, et al. 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. 2016;85(11):2144–51. DOI:10.1016/j.ejrad.2016.09.010.

26. Gonzalez G, Ash SY, Vegas-Sanchez-Ferrero G, et al. Disease Staging and Prognosis in Smokers Using Deep Learning in Chest Computed Tomography. Am J Respir Crit Care Med. 2018;197(2):193–203. DOI:10.1164/rccm.201705-0860OC.
27. Tang L, Coxson HO, Lam S, et al. 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. 2020;2(5):e259–67. DOI:10.1016/S2589-7500(20)30064-9.
28. Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395–409. DOI:10.1056/NEJMoa1102873.
29. Xu C, Qi S, Feng J, et al. DCT-MIL: Deep CNN transferred multiple instance learning for COPD identification using CT images. Phys Med Biol. 2020;65(14):145011. DOI:10.1088/1361-6560/ab857d.
30. Yan X, Tao M, Feng Q, et al. Deep learning of feature representation with multiple instance learning for medical image analysis. In: IEEE International Conference on Acoustics. 2014, 2014.
31. Shen Y, Wu N, Phang J, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Med Image Anal. 2021;68:101908. DOI:10.1016/j.media.2020.101908.
32. Ilse M, Tomczak JM, Welling M. Attention-based Deep Multiple Instance Learning. 2018.
33. Alom MZ, Yakopcic C, Hasan M, et al. Recurrent residual U-Net for medical image segmentation. J Med Imaging (Bellingham). 2019;6(1):014006. DOI:10.1117/1.JMI.6.1.014006.
34. Wang C, Xu J, Yang L, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study. Lancet. 2018;391(10131):1706–17. DOI:10.1016/S0140-6736(18)30841-9.
35. Philbrick KA, Yoshida K, Inoue D, et al. What Does Deep Learning See? Insights From a Classifier Trained to Predict Contrast Enhancement Phase From CT Images. AJR Am J Roentgenol. 2018;211(6):1184–93. DOI:10.2214/AJR.18.20331.
36. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9. DOI:10.1038/s41591-018-0316-z.

Figures

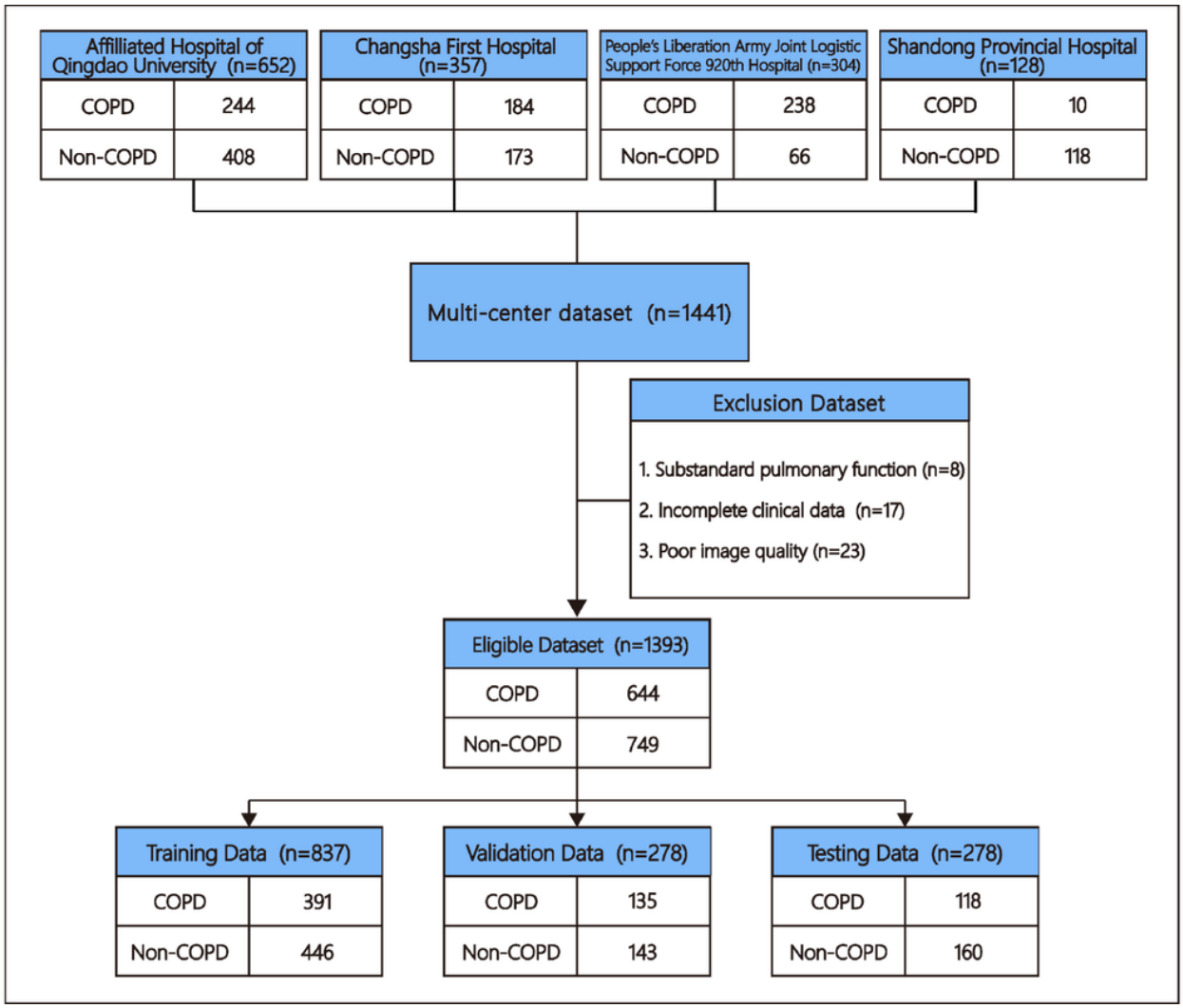

Figure 1

A multi-center COPD dataset establishment diagram.

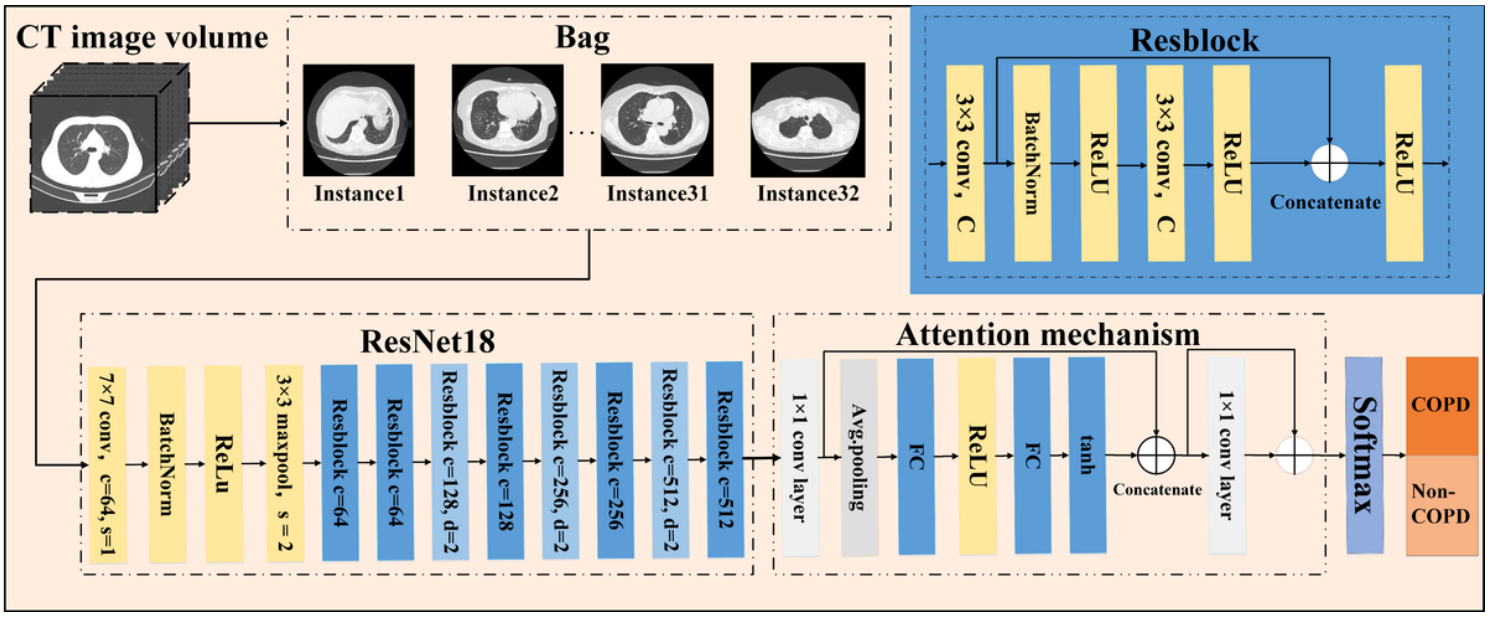

Figure 2

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, 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.

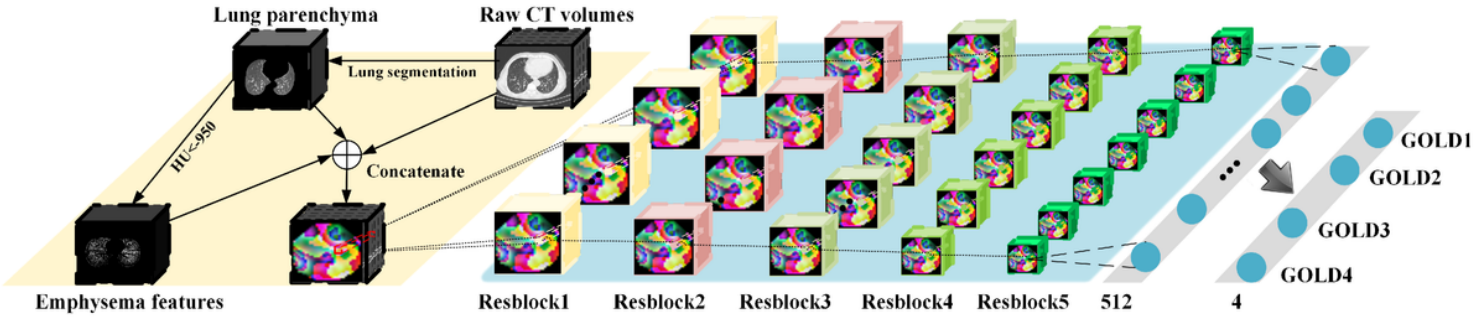

Figure 3

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 ≤ 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.

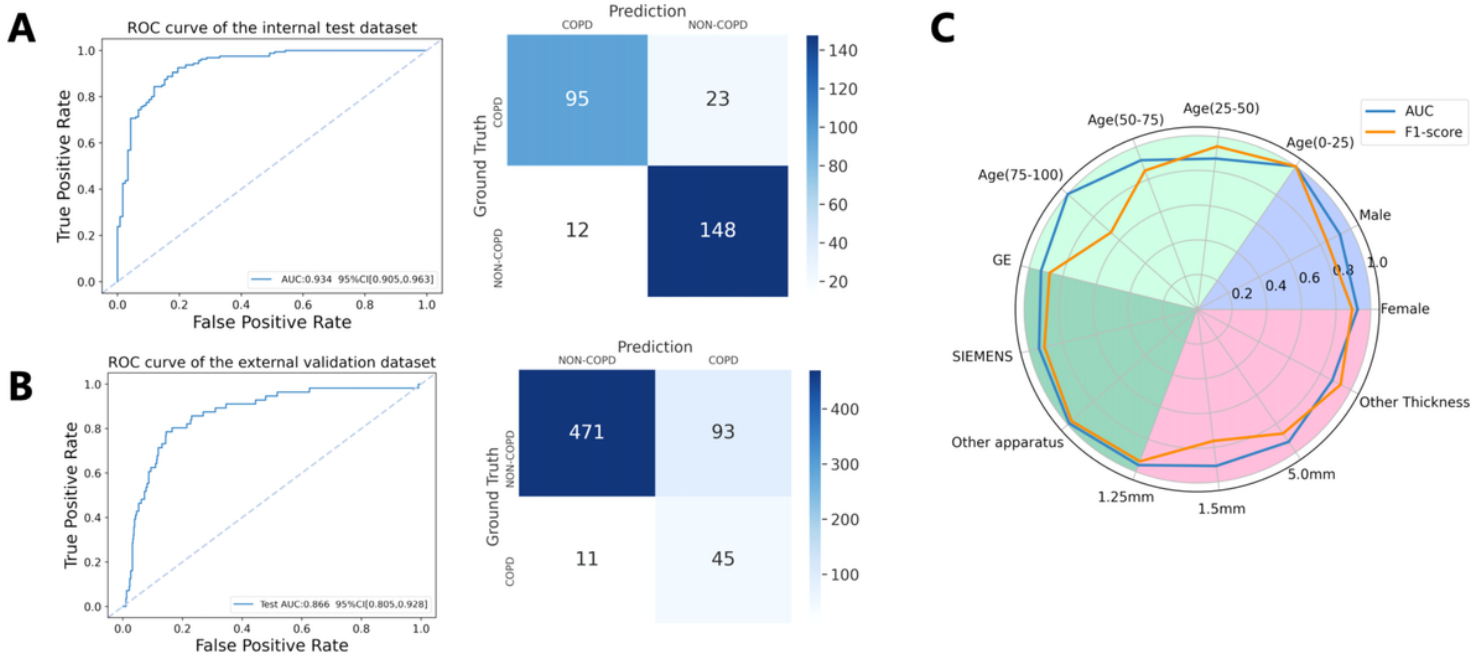

Figure 4

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.

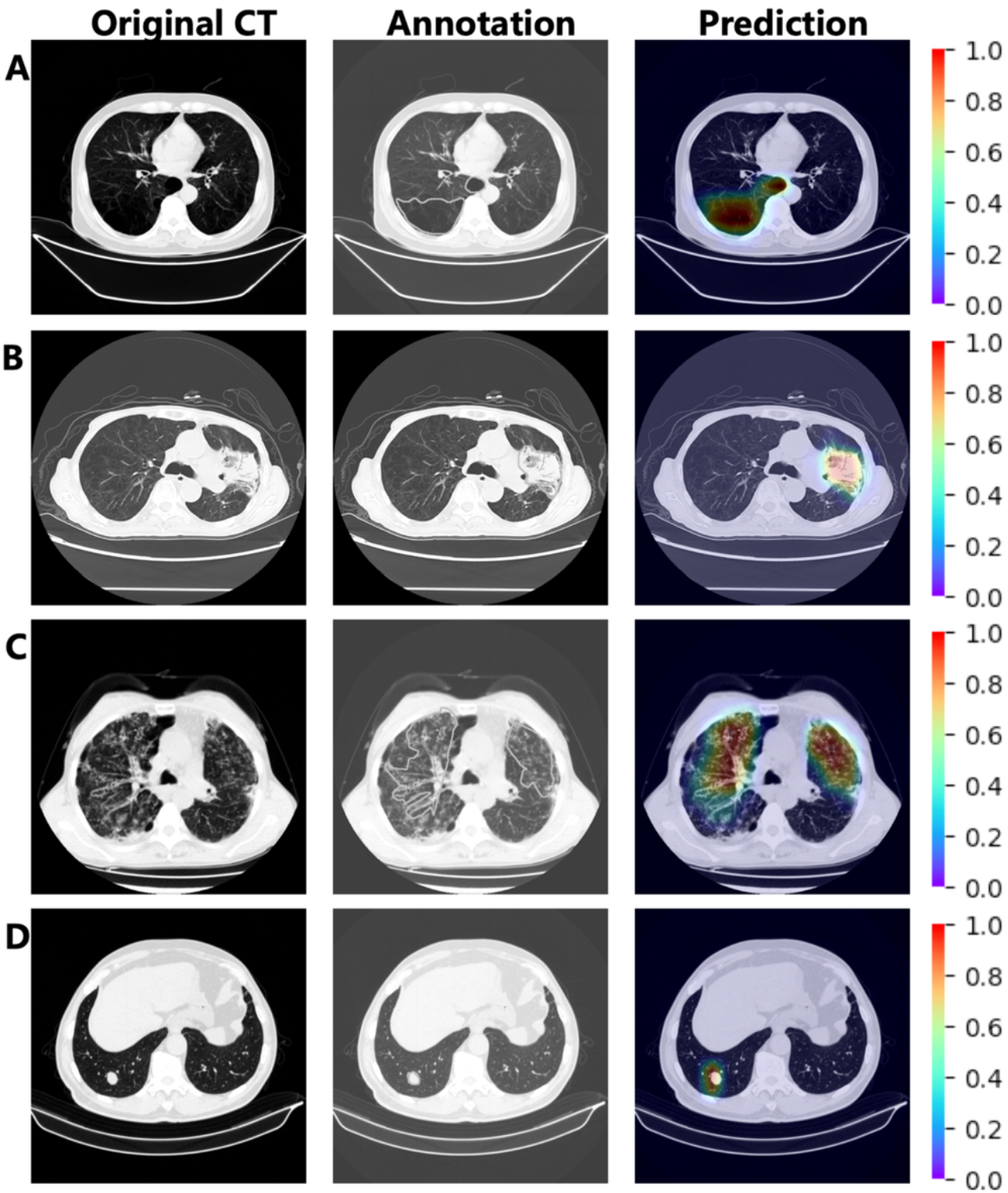

Figure 5

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 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.

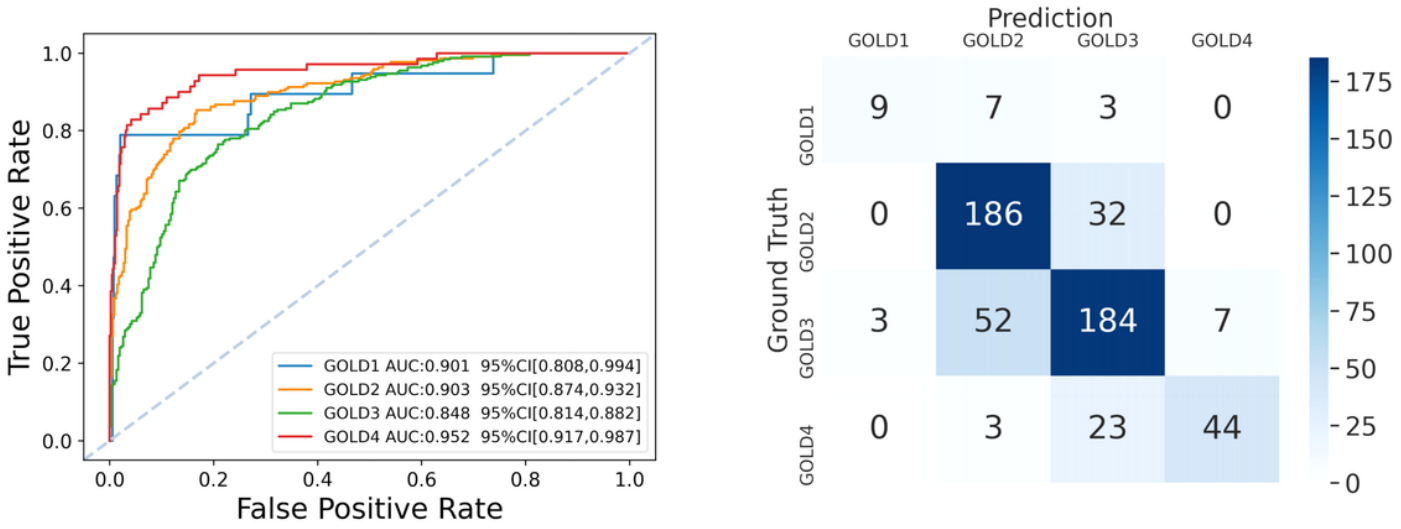

Figure 6

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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- COPDSupplementalAppendix.docx