International Journal of Information Technologies and Systems Approach
Volume 16 • Issue 3

Machine Learning-Assisted Diagnosis Model for Chronic Obstructive Pulmonary Disease

Yongfu Yu, School of Computer Science and Engineering, Central South University, China
Nannan Du, Xiangya Hospital, Central South University, China
Zhongteng Zhang, Central South University, China
Weihong Huang, Central South University, China*
Min Li, Xiangya Hospital, Central South University, China

ABSTRACT

Chronic obstructive pulmonary disease (COPD) is a long-term, irreversible, and progressive respiratory disease that often leads to lung function decline. Pulmonary function tests (PFTs) provide valuable information for diagnosing COPD; however, they are underutilised in clinical practice, with only a subset of test values being used for decision making. The final clinical diagnosis requires combining PFT results with patient information, symptoms, and other tests, such as imaging and blood analysis. This study aims to comprehensively utilise all the testing information in PFTs to assist in the diagnosis of COPD. Various machine learning models, such as logistic regression, support vector machine (SVM), k-nearest neighbour (KNN), random forest, decision tree, and XGBoost, have been employed to establish COPD diagnosis assistance models. The XGBoost model, trained with features extracted by the group LASSO algorithm, achieved the best performance, with an area under the receiver operating characteristic curve (ROC) of 0.90, 88.6% accuracy, and 98.5% sensitivity. This model can assist doctors in the clinical diagnosis and early prediction of COPD.

KEYWORDS

Chronic Obstructive Pulmonary Disease (COPD), Machine Learning, Pulmonary Function Test, Receiver Operating Characteristic Curve (ROC)

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a respiratory disease characterised by airflow limitation (Vestbo et al., 2013). It poses a serious threat to human health with a high incidence. In fact, the disease is one of the most significant public health problems affecting global economic and social development (Alkhathlan et al., 2020; Corlateanu et al., 2020; Halpin et al., 2021). COPD is the third leading cause of death worldwide and the fifth disease causing a substantial social burden,

DOI: 10.4018/IJITSA.324760 *Corresponding Author

with approximately three million deaths each year (Lozano et al., 2012) and more than 300 million people who suffer it (Adeloye et al., 2015). With an aging population, the number of people affected is expected to increase (Lopez-Campos et al., 2016).

Pulmonary function tests (PFTs) are essential tools for diagnosing COPD and assessing the function of the respiratory system (Crapo, 1994). However, interpreting complex, multidimensional, nonlinear, and heterogeneous data in PFT reports can be challenging. Experts rely on international guidelines to identify disease patterns (obstructive, restrictive, mixed, and normal) and grade their severity (Pellegrino et al., 2005; Vogelmeier et al., 2017). The final clinical diagnosis requires combining PFT results with patient information, symptoms, and other tests, such as imaging, blood analysis, and biopsy (Galie et al., 2016; Martinez et al., 2017).

Although PFT reports contain rich clinical information, reliance on human interpretation has limitations. Owing to international diagnostic rules, physicians can only make preliminary judgments based on some of the test values. Furthermore, clinicians may lack experience, leading to misdiagnosis and delayed COPD treatment (Tinkelman et al., 2006).

To improve the accuracy of COPD diagnosis, it is essential to mine hidden clinical information from PFT reports. Utilising the entire report can reduce unnecessary examinations and wasted medical resources. However, few studies have utilised PFT reports to aid in the diagnosis of COPD, likely because of difficulties in acquiring the necessary data. The data contained in PFT reports are physically separated from other hospital data, creating a data island. These data are stored in standalone spirometers; they cannot be integrated and analysed with other data like electronic medical records (EMRs).

Building on the research findings of the authors' laboratory, this study represents the first effort to extract 230,000 PFT reports from stand-alone spirometers at Xiangya Hospital of Central South University, thereby breaking down the data barriers that hindered previous studies. Given the focus of this study on the auxiliary diagnosis of COPD, the authors examined the bronchodilation reports in the PFT reports, totalling 16,012 reports.

The PFT is globally standardised, making it an ideal candidate for developing artificial intelligence (AI) algorithms for assisted diagnosis (Jordan & Mitchelle, 2015; Kononenko, 2001). AI can identify subtle and decisive features that are difficult for humans to detect. Then, the technology can incorporate the features into powerful differential diagnostic algorithms.

This study developed a complete solution for the auxiliary diagnosis of COPD using PDF-format PFT reports. The proposed algorithm comprises six parts: (1) pre-processing PDF report data; (2) matching report ID; (3) handling missing data; (4) data selection; (5) feature selection; and (6) model training. Specifically, the proposed algorithm initially uses text detection and recognition algorithms (Liao et al., 2020; Shi et al., 2017) to process the PDF format and obtain structured data. It then matches the missing patient identifications (IDs) in the PFT reports and handles the missing values in the examination indicators. Subsequently, data inclusion and exclusion are performed to obtain a clean dataset.

The authors used two feature selection algorithms, LASSO (Tibshirani, 1996) and group LASSO (Yuan et al., 2006), to perform data selection and dimensionality reduction on the dataset. The features obtained from the two dimensionality reduction algorithms were then separately used for six machine learning models: (1) logistic regression; (2) support vector machine (SVM); (3) k-nearest neighbour (KNN); (4) random forest; (5) decision tree; and (6) Xgboost (Breiman, 2001; Chang & Lin, 2011; Chen & Guestrin, 2016; Deng et al., 2016; Lavalley, 2008). These models located the best performing model as the authors' final assistive diagnosis model for COPD.

This study also used the Shapley additive explanations (SHAP) (Lundberg & Lee, 2017; Lundberg et al., 2018) algorithm to visualise the importance of each feature when the model makes decisions. The feature importance explanation reveals the machine learning model's preferences when making decisions, thereby providing new recommendations for clinical decision making. The results validate the important features of COPD diagnosis in clinical medicine and provide insights into COPD diagnosis from a big data perspective. The final accuracy rate of the model was 88.6%, with a specificity of 68.9% and a sensitivity of 98.5%, achieving a state-of-the-art effect in COPD diagnosis.

For a clearer presentation of the research, the authors first provide an overview of the study's structure. Section 2 discusses related studies on the auxiliary diagnosis of COPD. Section 3 introduces the core method proposed for diagnosing auxiliary COPD. Section 4 presents the results of the proposed COPD auxiliary diagnostic model. Finally, Section 5 summarises the contributions of this study and discusses the existing limitations.

RELATED WORK

Currently, research on the auxiliary diagnosis of COPD focuses on modelling and analysis of radiographic, bioinformatics, and respiratory function data. Radiographic data-based COPD auxiliary diagnosis is based primarily on computed tomography (CT) and x-rays as research subjects (Bhosale & Patnaik, 2023; Hasenstab et al., 2021; Ho et al., 2021; Tang et al., 2020; Willer et al., 2021; Xu et al., 2020). Machine learning algorithms are used to quantitatively analyse lung structures and extract features to assess COPD severity. Tang et al. (2020) collected multiple low-dose CT scan images from smoking and non-smoking patients. They enhanced the images and extracted several regions of interest, which were combined into a three-channel data structure as input for the constructed deep residual network model to automate COPD diagnosis. Xu et al. (2020) built a network model based on deep convolutional neural network transfer learning and multi-instance learning concepts, using axial divisions of multiple CT images as input for COPD risk prediction. Willer et al. (2021) developed a novel dark-field x-ray system to quantitatively analyse the accuracy of emphysema detection. Bhosale and Patnaik (2023) employed ensemble learning to integrate eight advanced deep learning models and constructed a more powerful ensemble model for the quantitative analysis and diagnosis of x-ray images. However, radiography-based techniques have limitations, as early-stage COPD does not exhibit significant organic changes, making radiographic auxiliary diagnostic methods unsuitable for early COPD diagnosis.

Bioinformatics-based COPD auxiliary diagnosis involves constructing models using gene expression and patient biological characteristic data (Leidy & Malley, 2016; Mostafaei et al., 2018; Spathis & Vlamos, 2019). Mostafei et al. (2018) explored novel genes related to COPD in human airway epithelial cells using classical machine learning algorithms from a genetic perspective. Leidy and Malley (2016) employed random forest algorithms to analyse three large datasets, constructing a COPD auxiliary diagnosis model with features like age, smoking status/history, symptoms, activity limitations, and acute bronchitis exacerbation. Spathis and Vlamos (2019) used clinical features (i.e., age, gender, sputum production, and smoking) as input data. They built a COPD auxiliary diagnosis model using basic machine learning algorithms and ensemble machine learning algorithms. Leidy and Malley (2016) and Spathis and Vlamos (2019) used common clinical features without incorporating standardised features related to respiratory function (e.g., PFT reports), leading to relatively poor interpretability of the model's predictive results and weaker persuasiveness for doctors.

Respiratory function data-based COPD auxiliary diagnosis primarily uses the PFT report data. However, because spirometers are physically isolated from other hospital data, there is minimal research directly analysing PFT reports using machine learning methods. In 2019, Topalovic et al. (2019) proposed a machine learning model for the auxiliary diagnosis of lung diseases based on PFTs. However, the study used only 50 pulmonary function test reports. In addition, the model structure was not disclosed. The accuracy of COPD diagnosis was 72.4%. Owing to the limited number of research samples, undisclosed model, and irreproducibility, the study's influence was slightly limited.

This study is the first to use massive PFT reports for the auxiliary diagnosis of COPD, both at home and abroad. Compared to Topalovic et al. (2019), the accuracy rate in this study improved by 16.2%, reaching 88.6%. This study utilised advanced public machine learning models, such as logistic regression, SVM, decision tree, KNN, random forest, and Xgboost (Breiman, 2001; Chang & Lin, 2011; Chen & Guestrin, 2016; Deng et al., 2016; Lavalley et al., 2008). In addition, the authors conducted an interpretability analysis of the model output to provide effective auxiliary decision making for doctors. The main contributions are as follows:

1. The authors developed a comprehensive solution to process the PDF reports of PFTs and obtain structured data, providing a solution for processing similar medical data.
2. The feature selection algorithm (Tibshirani, 1996; Yuan & Lin, 2006) was applied to reduce the data dimensions of the PFT report and obtain critical information in high-dimensional data. This can improve the accuracy of the model.
3. Using the algorithms of logistic regression, SVM, decision tree, KNN, random forest, and Xgboost (Breiman, 2001; Chang & Lin, 2011; Chen & Guestrin, 2016; Deng et al., 2016; Lavalley, 2008) were used for the first time to construct an auxiliary diagnosis model of COPD based on pulmonary function reports. The model's accuracy of 88.6% is an advancement in the research on the diagnosis of COPD using machine learning algorithms.
4. The SHAP (Lundberg & Lee, 2017; Lundberg et al., 2018) algorithm was used to visualise the model output and verify the critical indicators for the diagnosis of COPD in the past. Furthermore, this study provided a new key indicator for diagnosing COPD.

METHOD

Figure 1 provides an overview of the authors' proposed approach for building an auxiliary diagnosis model for COPD. It comprises six key steps: (1) pre-processing the PDF report data; (2) matching the report ID; (3) handling missing data; (4) data selection; (5) feature selection; and (6) model training. This section elaborates on each aspect in detail.

Pre-Processing PDF Report Data

PFT reports, which were obtained from the hospital intranet data centre, comprised real hospital records. Specifically, the authors collected 16,012 reports of patients who underwent bronchodilation examinations at Xiangya Hospital of Central South University between 2011 and 2021 to construct a dataset suitable for the study. PFT reports are typically stored in PDF format. Thus, the numerical data contained within them cannot be easily extracted and processed, presenting significant challenges for data analysis. To address this, the authors proposed a method for extracting PFT test PDF report data based on text detection and recognition techniques. Their work aimed at extracting and saving the data in a usable JSON format for subsequent data analysis and research. The processing flow is illustrated in Figure 2.

Converting PDF to Images

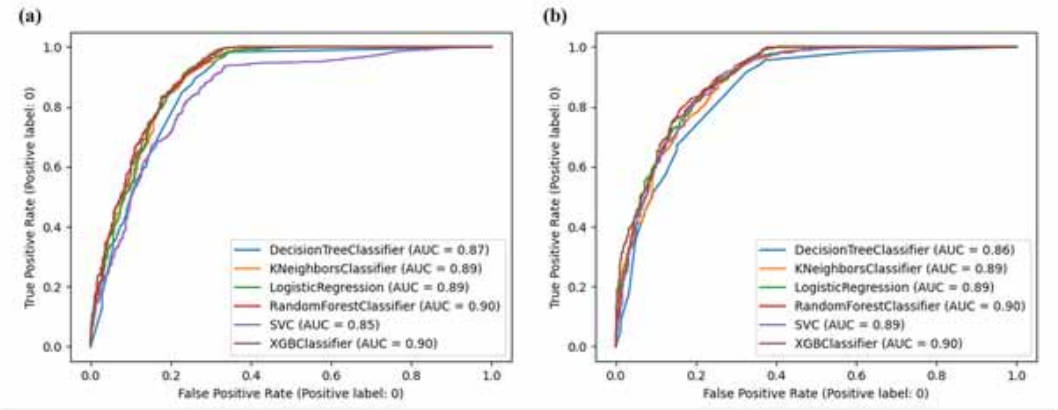
This study utilized the pdf2image Python library to convert a PDF test report into an image format that can be easily used for subsequent text detection and recognition.

Text Detection and Text Recognition

The authors first performed text detection and recognition on images converted from the PDF format of PFT reports. The differentiable binarization (DB) algorithm was used for text detection (Liao et al., 2020); the convolutional recurrent neural network (CRNN) algorithm was used for text recognition (Shi et al., 2017). Specifically, this study used the open-source implementation of PaddleOCR (Li et al., 2022) and the pretrained DB algorithm model and CRNN model built into PaddleOCR for text detection and recognition, respectively. The application of these algorithms and models provided a solid foundation for the research and achieved satisfactory experimental results. Specifically, the study's experiments performed well, as shown in Figure 3.

Figure 1. Overview of the algorithm

Figure 2. PDF report processing workflow



However, sometimes the text detection and recognition algorithm encountered issues, as shown in Figure 4, where the problem is highlighted in a green box. It is evident that there are two numbers within one text-detection box and there are spaces before and after the decimal point of the number. The key issue is correctly recognising the numbers and associating the extracted indicator values with the corresponding feature types. Handling missing values is also a critical issue; text recognition boxes are not continuous, making it difficult to match the columns during alignment. Therefore, this study proposes a text alignment algorithm for data processing.

Text Alignment

The textbox may contain more than one number. In addition, there may be spaces before and after the decimal point of multiple numbers. To address this issue, this study uses a non-capturing group technique in regular

Figure 3. PDF report text detection and recognition results

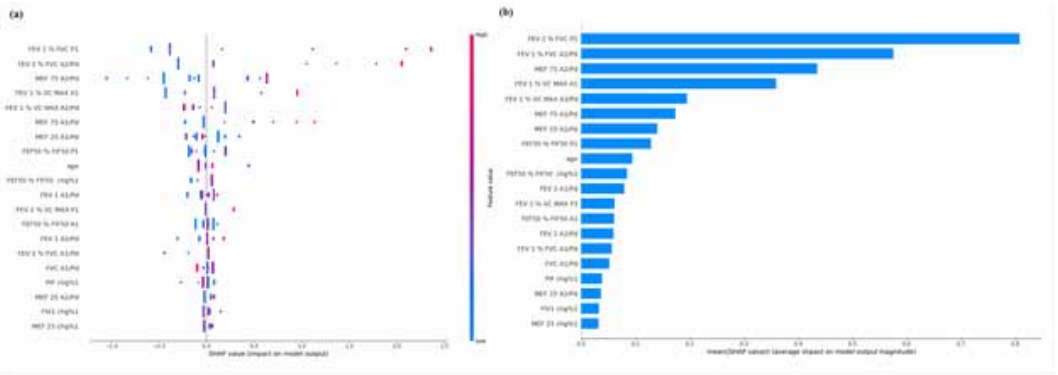


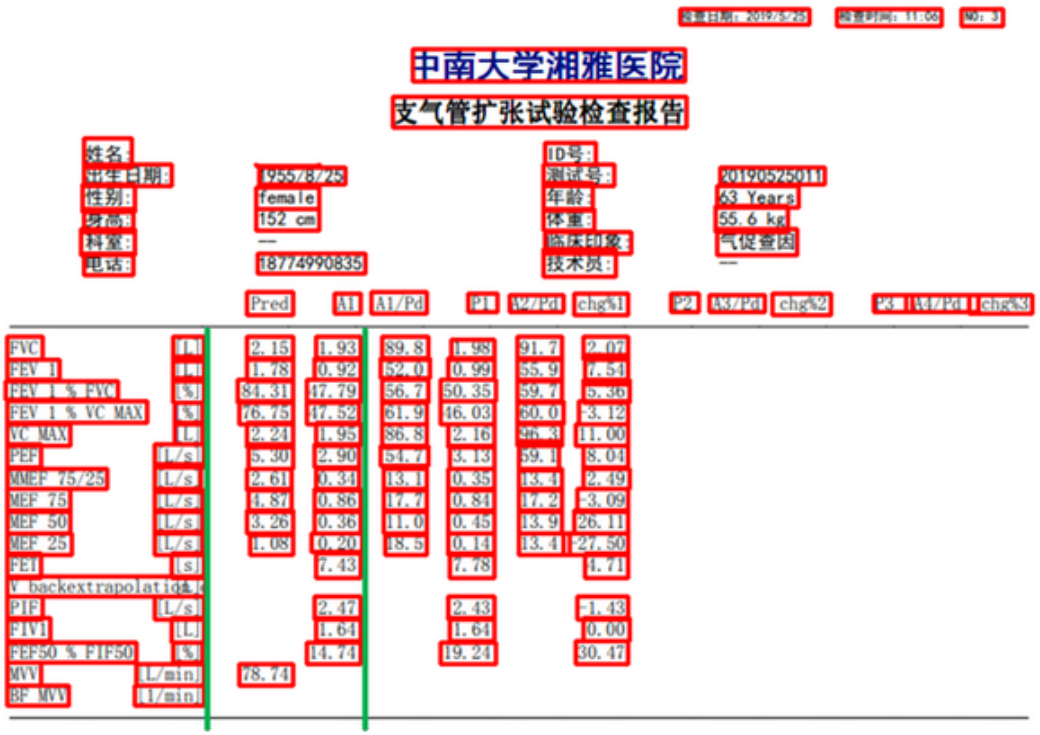
Figure 4. Example of errors in text detection and recognition

expressions for matching. The regular expression is $\text{pattern} = \text{r}'-?\backslash\text{d}+\left(?:\backslash\text{s}*.\backslash\text{s}*\backslash\text{d}+\right)?'$. Here, `$-?$` indicates that the optional negative sign `$-$` can be matched, allowing for matching of positive and negative numbers. `$-d+$` matches one or more digits, thereby matching the integer part. `$?:\backslash s*\backslash.\backslash s*\backslash d+$` is a non-capturing group, where $\left(?:\ldots\right)$ indicates a non-capturing group. `$\backslash s*$` matches zero or more whitespace characters and `$\backslash.$` matches the decimal point. '$\backslash d+$'matches one or more digits, thereby matching the decimal part. The entire expression uses `$?$` to indicate that the decimal part is optional, allowing for matching of both integers and floating-point numbers. There is no need to save the decimal part separately. Therefore, using a non-capturing group can achieve the desired effect of matching the decimal part while saving the entire number as a group in the match result.

The regular expression-matching algorithm is used in line 19 of Algorithm 1 to extract all numbers in the text box and save them in a list referred to as matches_str. To better understand Algorithm 1, it is necessary to clearly show the data format, as shown in Figure 5. The upper part of Figure 5 shows a cropped text detection box; the lower part displays the corresponding data format. In this format, [523.0, 94.0] represents the coordinates of the top-left corner of the text detection box, [842.0, 94.0] the top-right corner, [842.0, 134.0] the bottom-right corner, and [523.0, 134.0] the bottom-left corner. In the tuple (‘中南大学湘雅医院’, 0.9971780776977539), ‘中南大学湘雅医院’ denotes the extracted text information from the text detection box. 0.9971780776977539 indicates the confidence level of the model’s recognition.

To correctly locate a text selection box in its corresponding row and column, the text alignment algorithm first processes a list of text recognition boxes into rows and columns. Specifically, this study uses the right upper corner coordinates of each text recognition box as a reference point. When the difference in the y-coordinates between two boxes is greater than eight, it is considered that they belong to different rows, as shown in lines 3 and 7 of Algorithm 1. Each row of the text recognition boxes is stored in a

Figure 5. Data format for text detection and recognition results



separate list. The entire report is stored using multiple lists. The entire process of handling text-detection boxes by dividing them into rows is shown in lines 1-10 of Algorithm 1. After completing row alignment, column alignment is required. Column alignment is a critical operation that is essential for data accuracy.

For the bronchial dilation test reports, there were six test values: (1) Pred; (2) A1; (3) A1/Pd; (4) P1; (5) A2/Pd; and (6) chg%1 (see Figure 3). At present, the focus is on aligning the feature columns. It is crucial to accurately align the data in the text boxes with the data types for the authors' subsequent research. To simplify the algorithm, this study shows the core operation of aligning the clinical features into columns. Specifically, line 13 of Algorithm 1 describes the process of feature alignment. If the second text detection box in a row ends with ']', it indicates that the row contains clinical features.

Regarding the column alignment, the authors' approach determines the position of the current text recognition box as the n^{th} element. This determines the corresponding test value. However, it was found that the sizes of various text boxes differ. Thus, relying solely on their approximate positions to determine which column they correspond to is difficult because it is a challenge to ascertain the specific coordinate range. Furthermore, the sizes of text recognition boxes vary, leading to changes in their approximate position coordinates.

Therefore, this study adopts a different approach, looking for fixed points in the text recognition box for use as references. Specifically, as shown by the first green vertical line in Figure 3, the right endpoint of this text box was fixed. Thus, this article uses the first green line as a first reference point. The distance of each text recognition box relative to the first reference point can then be calculated to determine the features to which it belongs.

Which point in the text box should be selected for the calculation? As indicated by the second green vertical line in Figure 3, the horizontal coordinates of the right endpoints are the same for the text recognition boxes in the same column. This study, therefore, used the right endpoint of the text recognition box and right endpoint of the first vertical line as benchmarks. The distance between them was used as a standard to measure the column to which the text belonged. Specifically, the number of text boxes relative to the first reference box is calculated as $num = \left(x - 18\right) / 40$, where x is the horizontal coordinate of the upper-right endpoint of the text box and 40 represents the difference in the horizontal coordinate between the right endpoint of the previous text box and current text box's right endpoint. The 18 refers to the distance between the right endpoint of the first reference box and its adjacent text box's right endpoint (this is 18 times greater than the normal 40). Therefore, for convenience of calculation, x was first subtracted by 18.

The variable *num* represents the index of the last feature that corresponds to the number in the current text box. A text box may contain multiple numbers; therefore, it should loop backwards from the obtained index *num* and assign each number in the text box to the corresponding feature type. This process corresponds to lines 11-24 of Algorithm 1:

```
Algorithm 1: Text Alignment Algorithm based on Non-capturing
Grouping Regular Technique
Input: Text detection and recognition results set S
Output: Json Result
1. line_list = [], total_list = [] // Initialize the list of rows
and the whole report data list
2. for data in S:
3.  if len(line_list)>1 and abs(line_list[-1][0][0][1] - data[0]
[0][1]) > 8:
4.        total_list.append(line_list)
5.        line_list=[]
6.        line_list.append(data)
7.  elif len(line_list) == 0 or abs(line_list[-1][0][0][1] -
data[0][0][1]) <= 8:
8.        line_list.append(data)
```

```
9. if line_list is not None:
10.  total_list.append(line_list)
11. result = {'data': {}} // define result dictionary
12. for line_list in total_list: // perform column alignment
13.  if len(line_list) >= 2 and line_list[1].endsWith("]"): //
locate the rows where the feature values occur
14.       feature_name = "_".join(line_list[0].split()) // defines
the name of the feature to handle
15.       start_right = line_list[1][0][1][0]
16.       if len(line_list) == 2:
17.            continue
18.       for line in line_list[2:]:
19.            matches_str = match(line[1][0].strip()) // regular
extraction
20.            dis = line[0][1][0] - start_right
21.            num = round((dis-18)/40)
22.            start_index = num - len(matches_str)
23.            for index, match_str in enumerate(matches_str):
24.                 result['data'][feature_name][feature_
type[start_index + index]] = match_str
```

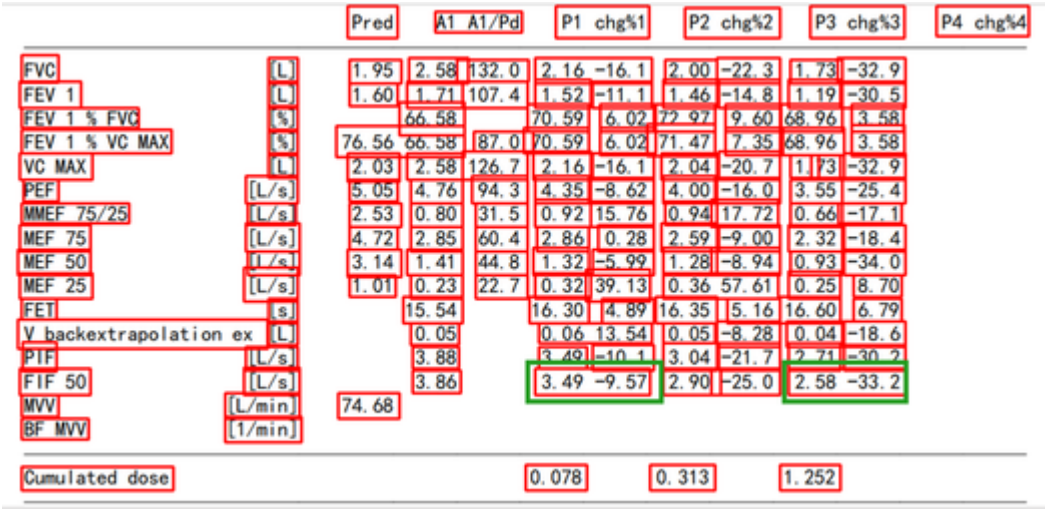
Structured Text Output

After processing using Algorithm 1, the PFT report in the PDF format is transformed into a JSON format that is easy to handle (see Figure 6).

Matching Report's Patient ID Module

The patient ID serves as a unique identifier in the EMR system, allowing the patient's diagnostic data to be linked to their ID. However, some PFT reports lack patient ID. Thus, it is impossible to obtain a corresponding diagnostic label from the EMR. To solve this problem, this study extracts payment information from the hospital charging system for each PFT report, including the report name, payment

Figure 6. Structured reporting data



time, patient name, birth date, and sex. This information is then matched to the corresponding data in the PFT report to obtain the patient ID for each report.

Handling Missing Data

Imputing missing values is a crucial step in data pre-processing. To better understand this process, it is necessary to provide an overview of the bronchodilation test. The test measures the changes in the degree of bronchiectasis before and after using a specific dose of bronchodilator drugs. It then determines whether there is persistent spasm and irreversible airway obstruction. Each report includes 17 test values corresponding to FVC, FEV1, ..., and BF MVV in Figure 3. Each test value has six types of values, including the predicted value, pre-drug test value before bronchodilator inhalation, ratio of the pre-drug test value to the predicted value, post-drug test value after bronchodilator inhalation, ratio of the post-drug test value to the predicted value, and rate of change in the post-drug test value to the pre-drug test value (corresponding to Pred, A1, A1/Pd, P1, A2/Pd, and chg%1 in Figure 3).

When filling in missing values, in this study, the authors first filled in the missing values of the FEV1%FVC Pred indicator using the formula. This represents the ratio of forced expiratory volume in one second to forced vital capacity, which is commonly referred to as the one-second rate. This indicator is crucial for diagnosing asthma and COPD severity. Its predicted value is related to height, calculated using $FEV1\%FVC\ \ Pred = 90.6043 - 0.0414*height$. Then, the authors calculated the missing values for the FEV1%FVC A1/Pd and FEV1%FVC A2/Pd indicators based on their predicted values via their pre-drug values, post-drug values, and predicted values, respectively. See equations (1) and (2).

$$FEV1\%FVC\,A1\,/\,Pd = \left(FEV1\%FVC\,A1\right) / \left(FEV1\%FVC\,Pred\right) \tag{1}$$

$$FEV1\%FVC\,A2\,/\,Pd = \left(FEV1\%FVC\,P1\right) / \left(FEV1\%FVC\,Pred\right) \tag{2}$$

Next, the authors removed indicator values whose missing rate was greater than one-third. They used the missForest algorithm to fill in the remaining indicator values (Stekhoven & Buhlmann, 2012). After completing the data processing, 78 dimensional indicators remained.

Data Selection

Data filtering is crucial in scientific research because it determines the performance of a model. In this study, 16,012 bronchodilation reports were collected from a dataset. The data filtering process is illustrated in Figure 7.

In the screening process for reports labelled as COPD, the authors found that some diagnoses in hospital outpatient clinics were inaccurate, such as COPD pending investigation and COPD waiting to be discharged. After consulting experts from the Xiangya Hospital, Central South University's Department of Respiratory Medicine, a decision was made to exclude data on pending diagnosis, pending investigation, and suspected COPD from the outpatient diagnosis. All inpatient diagnostic data were retained. For each report, if there was a diagnosis of COPD in the EMR two weeks before or after the report's check time, it was labelled 0. This indicated that the diagnosis label was COPD. A total of 803 outpatient reports were excluded; 3,453 reports labelled as COPD were included.

Figure 7. Inclusion and exclusion of PFT reports

In the screening process for reports labelled normal, experts from the department were consulted. They noted that there are cases in which the diagnosis is not entirely written in the outpatient diagnosis. Thus, even if COPD is not mentioned in the outpatient diagnosis, it does not rule out that the patient has COPD. If there was no diagnosis of COPD on the EMR two weeks before or after the report's check time, the report was labelled as normal (defined as 1). To ensure the accuracy of the report label, outpatient reports were excluded when the label was normal. Only inpatient reports were included. A total of 10,038 reports were excluded; 1,718 reports labelled as normal were included.

The final pulmonary function report dataset included 5,171 reports. This included 3,453 COPD reports and 1,718 normal reports.

Feature Selection

Previous studies have found that only a small portion of these features are related to the data labels that need to be predicted (Li et al., 2022). Most features are simply noise variables that can negatively affect model training and response speed. To reduce the data dimensionality, two types of sparse feature selection models were employed. These include the individual sparse feature selection and group sparse feature selection. The individual model independently evaluates the importance of each feature. It does not consider the combined effect of different features. The group model considers the joint effect of different feature combinations.

To ensure the comprehensiveness of the current research, representative algorithms from both methods were selected for feature screening. Specifically, the LASSO algorithm (Tibshirani, 1996) was used for individual sparse feature selection. The group LASSO algorithm (Yuan & Lin, 2006) was employed for group-sparse feature selection.

LASSO is a linear regression method that uses L1 regularisation (Tibshirani, 1996). Using L1 regularisation makes some of the learned feature weights zero to achieve sparsity and feature selection. The group LASSO algorithm, an extension of the LASSO algorithm, applies feature grouping prior to feature selection. This approach considers the interaction effects between features. In the actual implementation, open-source implementations of the two algorithms were adopted using the sklearn and group-LASSO packages in Python.

Model Training

This study utilised features extracted by the LASSO and group LASSO algorithms as inputs to train six machine-learning models (Breiman, 2001; Chang & Lin, 2011; Chen & Guestrin, 2016; Deng et al., 2016; Lavalley, 2008). The performance of each model was evaluated. Then, the best-performing model was selected as the final model to assist in the diagnosis of COPD. The training and test sets were split at a ratio of 7:3. The performance of the model was tested using the test set.

The following metrics were used to test the validity of the model: accuracy; recall; precision; sensitivity; specificity; FPR; NPV; F1score; and AUC. COPD was defined as a positive class and normal as a negative class.

- **Accuracy:** This represents the accuracy of model classification.
- **Recall:** This represents the recall ability of positive samples and reflects the proportion of positive samples correctly judged as positive. Its calculation formula is the same as that for sensitivity.
- **Precision:** This represents the proportion of positive samples classified correctly among the classified positive samples.
- **Sensitivity:** This represents the predictive ability of a positive class. The higher the sensitivity, the lower the probability of a missed diagnosis.
- **Specificity:** This represents the predictive power of a negative class. The higher the specificity, the higher the probability of diagnosis and the lower the probability of misdiagnosis.

- **FPR:** This represents the false positive rate, which indicates the ratio of negative samples predicted as positive to the number of negative examples. The lower the false positive rate, the better the effect of the model.
- **NPV:** This represents the precision of negative samples (the proportion of negative samples predicted to be negative).
- **F1_Score:** This score is an indicator used to measure the accuracy of a binary classification model in statistics. It considers the precision and recall of the model. In addition, it can be regarded as a weighted average of precision and recall.
- **AUC:** This is the area under the receiver operating characteristic curve (ROC), a performance indicator used to measure the model's performance.

The authors define TP as true positive (the number of positive samples predicted to positive class), FP as false positive (the number of negative samples predicted to positive class), TN as true negative (the number of positive samples predicted to negative class), and FN as false negative (the number of negative samples predicted to negative class). The above evaluation metrics are expressed as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$specificity = \frac{TN}{TN + FP} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

$$NPV = \frac{TN}{FN + TN} \tag{9}$$

$$F1\_score = \frac{2*Recall*Precision}{Recall + Precision} \tag{10}$$

EXPERIMENT RESULTS

Dataset Baseline Characteristics

The study's dataset comprised 5,171 PFT reports, each of which included patient characteristics and test values. Table 1 presents a detailed summary of the characteristics of these reports, dividing them into three groups: (1) overall PFT reports (N = 5,171); (2) reports labelled as COPD (n = 3,453); and (3) reports labelled as normal (n = 1,718). The table also displays the median, first quartile, and third quartile of the features of each group. In addition, the authors conducted significance tests for each feature of COPD. They found that 15 features had no significant effect on COPD ($p > 0.05$). As shown in Table 1, these features mainly pertain to the chg%1 test value, which measures the rate of change in the post-drug test value compared to the pre-drug test value. It is speculated that this type of feature may carry relatively little information regarding the patient's lung function because it only reflects the changes in the test value rather than the patient's overall condition.

Table 1. Characteristics of the PFT report in the dataset

Variables	Total N = 5171	COPD n = 3453	Normal n = 1718	p-Value
Basic feature				
Age, median[Q1,Q3]	64.0 [57.0,64.0]	65.0 [59.0,70.0]	62.0[52.0,69.0]	<0.001
Gender				<0.001
Male	4077	2975	1102	
Female	1094	478	616	
Weight, median[Q1,Q3]	57.0 [49.5,65.0]	56.0 [49.0, 64.0]	58.0 [50.0,66.5]	<0.001
Height, median[Q1,Q3]	160.0 [155.0,165.0]	160.0 [156.0,165.0]	160.0 [154.0, 165.0]	<0.001
Feature Prediction Value				
FVC, median[Q1,Q3]	3.07 [2.64,3.42]	3.08 [2.73,3.41]	3.04 [2.40,3.46]	<0.001
FEV1, median[Q1,Q3]	2.39 [2.04,2.69]	2.39 [2.09,2.68]	2.38 [1.94,2.74]	0.88
FEV1%FVC, median[Q1,Q3]	83.98 [83.77,84.19]	83.98 [83.77,84.15]	83.98 [83.77,84.23]	<0.001
FEV1%VC MAX, median[Q1,Q3]	75.15 [74.07,76.75]	74.97 [74.07,76.18]	75.80 [74.43,77.70]	<0.001
VC MAX, median[Q1,Q3]	3.16 [2.72,3.54]	3.17 [2.80,3.52]	3.12 [2.48,3.58]	0.0018
PEF, median[Q1,Q3]	6.96 [6.31,7.44]	6.98 [6.50,7.42]	6.84 [5.67,7.48]	<0.001
MMEF 75/25, median[Q1,Q3]	2.86 [2.55,3.14]	2.84 [2.55,3.11]	2.91 [2.55,3.23]	<0.001
MEF 75, median[Q1,Q3]	6.24 [5.71,6.60]	6.26 [5.88,6.59]	6.15 [5.16,6.64]	<0.001
MEF 50, median[Q1,Q3]	3.59 [3.30,3.88]	3.58 [3.31,3.85]	3.61 [3.28,3.93]	<0.001
MEF 25, median[Q1,Q3]	1.10 [0.91,1.33]	1.07 [0.89,1.28]	1.17 [0.95,1.42]	<0.001
MVV, median[Q1,Q3]	96.24 [86.63,105.09]	96.43 [88.10,104.65]	95.80 [82.99,105.97]	0.062
Before Bronchodilator Inhalation				
FVC, median[Q1,Q3]	2.41 [1.91,2.91]	2.43 [1.96,2.93]	2.37 [1.80,2.90]	0.0062
FEV1, median[Q1,Q3]	1.05 [0.77,1.44]	0.91 [0.69,1.22]	1.42 [1.08,1.79]	<0.001
FEV1%FVC, median[Q1,Q3]	44.29 [35.18,55.44]	39.45 [32.26, 46.81]	63.59 [49.99,78.50]	<0.001
FEV1%VC MAX, median[Q1,Q3]	44.10 [35.07,55.05]	39.29 [32.13, 46.65]	63.35 [49.87,77.27]	<0.001
VC MAX, median[Q1,Q3]	2.42 [1.92,2.93]	2.44 [1.97, 2.94]	2.37 [1.81,2.90]	0.004
PEF, median[Q1,Q3]	3.27 [2.36,4.40]	2.89 [2.17,3.80]	4.27 [3.25,5.57]	<0.001
MMEF 75/25, median[Q1,Q3]	0.35 [0.23,0.58]	0.29 [0.21,0.41]	0.70 [0.41, 1.28]	<0.001
MEF 75, median[Q1,Q3]	1.01 [0.60,1.84]	0.77 [0.51,1.20]	2.33 [1.37,4.08]	<0.001
MEF 50, median[Q1,Q3]	0.44 [0.28,0.76]	0.34 [0.24,0.51]	0.97 [0.55,1.74]	<0.001
MEF 25, median[Q1,Q3]	0.18 [0.13,0.27]	0.16 [0.12,0.21]	0.29 [0.18,0.55]	<0.001
FET, median[Q1,Q3]	11.41 [8.56,14.75]	12.41 [9.77,15.40]	8.97 [7.01,12.48]	<0.001
PIF, median[Q1,Q3]	2.32 [1.63,3.22]	2.26 [1.57,3.14]	2.47 [1.74,3.34]	<0.001
FIV1, median[Q1,Q3]	1.53 [1.14,2.00]	1.52 [1.12,1.96]	1.57 [1.20,2.08]	<0.001
FEF50%FIF50, median[Q1,Q3]	24.10 [16.19,41.79]	19.57 [14.36,28.04]	45.07 [26.20,82.18]	<0.001
MVV, median[Q1,Q3]	44.47 [31.77,59.04]	36.42 [26.76,49.37]	49.75 [36.80,64.16]	<0.001
BF MVV, median[Q1,Q3]	68.33 [53.10,81.63]	63.10 [47.79, 77.53]	71.22 [58.03,83.67]	<0.001
Value Before Inhalation of Bronchodilator/Predictive Value				
FVC, median[Q1,Q3]	81.90 [68.00,94.40]	82.20 [68.20,94.80]	81.25 [67.68,93.13]	0.161
FEV1, median[Q1,Q3]	47.10 [34.30,61.20]	40.55 [30.50,52.78]	61.70 [49.80,71.70]	<0.001
FEV1%FVC, median[Q1,Q3]	52.90 [42.00,66.00]	47.00 [38.40,55.95]	75.90 [59.68,93.53]	<0.001

continued on following page

Table 1. Continued

Variables	Total N = 5171	COPD n = 3453	Normal n = 1718	p-Value
FEV1%VC MAX, median[Q1,Q3]	58.50 [46.60,73.10]	52.40 [43.00,62.30]	83.10 [66.10,100.00]	<0.001
VC MAX, median[Q1,Q3]	79.35 [66.00,91.50]	79.80 [66.20,91.83]	78.60 [65.60,90.50]	0.091
PEF, median[Q1,Q3]	48.50 [35.80,64.18]	42.50 [32.70,54.80]	65.10 [51.60,82.10]	<0.001
MMEF 75/25, median[Q1,Q3]	12.60 [8.40,20.00]	10.20 [7.40,14.70]	24.40 [14.80,43.53]	<0.001
MEF 75, median[Q1,Q3]	16.90 [9.90,30.00]	12.80 [8.40,19.63]	39.30 [22.85,69.60]	<0.001
MEF 50, median[Q1,Q3]	12.30 [7.80,20.40]	9.70 [6.80,14.30]	26.70 [15.60,47.40]	<0.001
MEF 25, median[Q1,Q3]	16.70 [12.20,25.30]	14.65 [11.30,19.80]	24.90 [15.40,47.13]	<0.001
MVV, median[Q1,Q3]	47.45 [34.70,59.60]	38.60 [28.90,49.95]	53.20 [41.30,64.60]	<0.001
After Bronchodilator Inhalation				
FVC, median[Q1,Q3]	2.59 [2.05,3.10]	2.64 [2.11,3.12]	2.51 [1.89,3.04]	<0.001
FEV1, median[Q1,Q3]	1.17 [0.85,1.56]	1.03 [0.77,1.35]	1.53 [1.19,1.92]	<0.001
FEV1%FVC, median[Q1,Q3]	46.21 [36.21,56.60]	40.99 [33.43,48.65]	65.61 [53.00,79.36]	<0.001
FEV1%VC MAX, median[Q1,Q3]	46.09 [36.14,56.43]	40.84 [33.35,48.51]	65.44 [52.66,78.88]	<0.001
VC MAX, median[Q1,Q3]	2.60 [2.06,3.10]	2.64 [2.12,3.13]	2.51 [1.90,3.04]	<0.001
PEF, median[Q1,Q3]	3.61 [2.66,4.81]	3.23 [2.45,4.20]	4.66 [3.54,5.92]	<0.001
MMEF 75/25, median[Q1,Q3]	0.41 [0.27,0.65]	0.33 [0.24,0.46]	0.82 [0.48,1.51]	<0.001
MEF 75, median[Q1,Q3]	1.19 [0.69,2.10]	0.90 [0.58,1.37]	2.67 [1.65,4.45]	<0.001
MEF 50, median[Q1,Q3]	0.51 [0.32,0.86]	0.40 [0.27,0.58]	1.13 [0.66,2.02]	<0.001
MEF 25, median[Q1,Q3]	0.19 [0.14,0.29]	0.17 [0.13,0.22]	0.32 [0.20,0.60]	<0.001
FET, median[Q1,Q3]	11.41 [8.69,14.72]	12.44 [9.92,15.51]	9.07 [7.00,12.01]	<0.001
PIF, median[Q1,Q3]	2.58 [1.79,3.58]	2.55 [1.76,3.59]	2.64 [1.87,3.57]	0.239
FIV1, median[Q1,Q3]	1.64 [1.23,2.12]	1.62 [1.24,2.09]	1.66 [1.21,2.15]	0.019
FEF50%FIF50, median[Q1,Q3]	25.07 [16.56,45.07]	19.80 [14.51, 28.61]	51.29 [29.09, 93.08]	<0.001
MVV, median[Q1,Q3]	62.77 [39.60,85.12]	25.85 [18.24,43.29]	67.49 [44.70,89.43]	<0.001
BF MVV, median[Q1,Q3]	71.01 [55.06,83.52]	50.11 [34.94,74.76]	73.23 [61.37,83.87]	<0.001
Value After Inhalation of Bronchodilator/Predictive Value				
FVC, median[Q1,Q3]	87.70 [73.60,100.50]	88.40 [74.60,101.00]	85.80 [71.10,99.50]	<0.001
FEV1, median[Q1,Q3]	52.20 [38.00,66.90]	45.50 [34.20,58.10]	67.15 [55.40,78.73]	<0.001
FEV1%FVC, median[Q1,Q3]	55.00 [43.00,67.40]	49.00 [39.98,58.00]	78.00 [63.10,94.40]	<0.001
FEV1%VC MAX, median[Q1,Q3]	61.40 [48.20,74.90]	54.30 [44.40,64.70]	86.45 [69.80,102.20]	<0.001
VC MAX, median[Q1,Q3]	84.90 [71.40,97.40]	85.60 [72.53,97.98]	83.10 [68.90,96.50]	<0.001
PEF, median[Q1,Q3]	54.40 [40.20,70.30]	47.90 [36.50,60.50]	71.80 [57.40,87.40]	<0.001
MMEF 75/25, median[Q1,Q3]	14.60 [9.70,22.80]	12.00 [8.50,16.50]	28.65 [17.50,50.80]	<0.001
MEF 75, median[Q1,Q3]	19.70 [11.60,34.30]	14.70 [9.70,22.50]	45.30 [27.70,77.70]	<0.001
MEF 50, median[Q1,Q3]	14.40 [9.00, 23.80]	11.30 [7.80,16.30]	31.30 [18.60,57.00]	<0.001
MEF 25, median[Q1,Q3]	18.40 [13.20,27.50]	16.00 [12.20,21.30]	28.20 [18.20,50.35]	<0.001
MVV, median[Q1,Q3]	69.00 [48.70,92.60]	29.30 [20.78,50.43]	72.50 [52.50,93.95]	<0.001
Change Rate of Values Before and After Inhalation of Bronchodilator				
FVC, median[Q1,Q3]	5.53 [-0.35,13.69]	6.27 [0.00,14.71]	4.06 [-1.14,10.90]	<0.001
FEV1, median[Q1,Q3]	8.88 [2.72,17.55]	9.79 [3.27,18.76]	7.11 [1.91,14.31]	0.004

continued on following page

Table 1. Continued

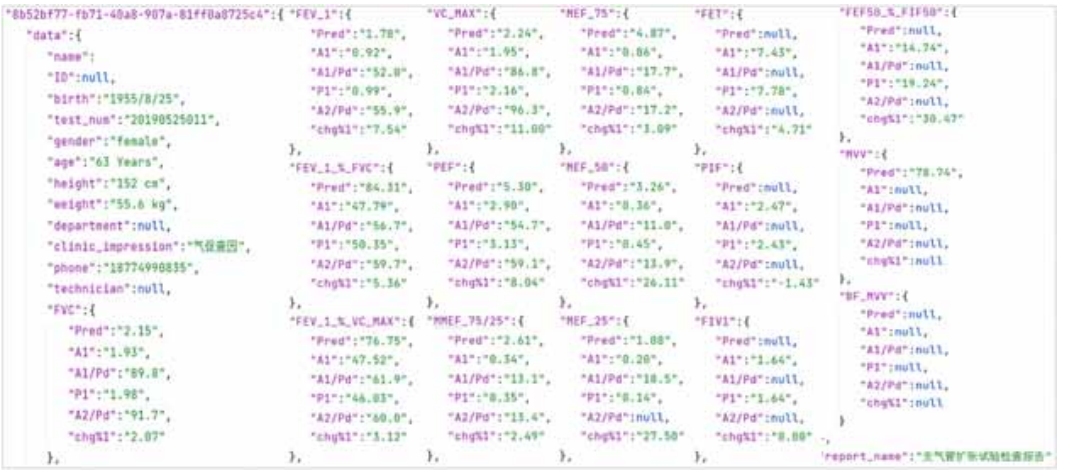
Variables	Total N = 5171	COPD n = 3453	Normal n = 1718	p-Value
FEV1%FVC, median[Q1,Q3]	3.33 [-2.63,9.72]	3.51 [-2.94,10.20]	2.90 [-2.08,8.71]	0.531
FEV1%VC MAX, median[Q1,Q3]	3.34 [-2.61,9.89]	3.48 [-2.95,10.35]	3.00 [-1.95,9.06]	0.456
VC MAX, median[Q1,Q3]	5.52 [-0.47,13.80]	6.34 [-0.07,14.77]	4.04 [-1.25,11.21]	0.102
PEF, median[Q1,Q3]	9.99 [2.06,19.61]	10.73 [2.69,20.51]	8.31 [1.03,17.81]	0.512
MMEF 75/25, median[Q1,Q3]	14.24 [0.02,31.59]	14.46 [0.97,30.47]	13.59 [-2.03,34.65]	0.022
MEF 75, median[Q1,Q3]	14.37 [0.95,30.12]	14.89 [1.08,30.81]	13.31 [0.64,28.81]	0.412
MEF 50, median[Q1,Q3]	15.00 [0.00,33.33]	15.00 [0.00,33.09]	14.92 [-1.60,34.33]	0.051
MEF 25, median[Q1,Q3]	10.66 [-6.68,32.80]	10.00 [-4.91,30.00]	12.39 [-10.24,40.00]	<0.001
FET, median[Q1,Q3]	1.22 [-12.66,15.84]	1.39 [-11.45,14.67]	0.66 [-16.94,20.27]	<0.001
PIF, median[Q1,Q3]	10.41 [-6.50,31.99]	11.39 [-5.37,33.16]	7.56 [-10.09,29.07]	0.077
FIV1, median[Q1,Q3]	6.69 [-5.27,21.31]	7.60 [-4.17,21.98]	4.36 [-8.43,19.57]	0.454
FEF50%FIF50, median[Q1,Q3]	4.35 [18.11,34.08]	2.72 [-18.82,31.57]	7.39 [-16.62,40.70]	0.002
MVV, median[Q1,Q3]	13.92 [-6.68,48.78]	5.50 [-17.70,16.13]	14.15 [-6.68,49.35]	0.542
BF MVV, median[Q1,Q3]	6.99 [-13.32,26.64]	-3.29 [-33.49,22.00]	7.32 [-12.78,27.19]	0.204
COPD, chronic obstructive pulmonary disease; Normal, no chronic obstructive pulmonary disease.				

Feature Selection Results

Figure 8 illustrates the results of the feature selection using the LASSO algorithm (Tibshirani et al., 1996). Figure 8a shows the LASSO path plot, which visually displays the changes in the feature coefficients in the LASSO regression as the regularisation parameter lambda varies. The horizontal axis represents the regularisation parameter lambda, the vertical axis represents the feature coefficients, and the coloured curves represent distinct features. The variations in the curves reflect the importance of each feature for the different lambda values. As lambda increases, some feature coefficients shrink to zero, indicating a decreased influence of these features in the model. The vertical black dashed line in Figure 8a represents the optimal regularisation parameter lambda value obtained by the LASSO algorithm (0.012).

Figure 8b presents the importance of the feature under the optimal λ value model, with the X-axis representing the risk coefficient of the feature and the Y-axis representing the feature name.

Figure 8. Feature extraction by the LASSO algorithm



Owing to plot size limitations, features with a risk coefficient of zero were not displayed. Figure 8b shows the 34 high-risk (positive coefficients) and low-risk features (negative coefficients). Notably, the features of FEV1%VC MAX Pred, FEV1%FVC A2/Pd, FEV1 A1/Pd, MEF75 A2/Pd, and height were significant in the results obtained using LASSO (Tibshirani et al., 1996).

The types of indicators will not be repeated here; only the meaning of the indicators appearing for the first time will be described. For instance, FEV1%VC MAX refers to the forced expiratory volume in one second as a percentage of the maximum vital capacity. FEV1 denotes forced expiratory volume in 1s. MEF75 represents the maximum expiratory flow when 25% of forced vital capacity (remaining 75%) is exhaled. The height indicator represented the patient's height at the time of consultation. FEV1%FVC is a known indicator for diagnosing COPD in clinical practice. The sparse feature results obtained also verify the usefulness of this indicator. Furthermore, this demonstrates that the feature dimensionality reduction method effectively removed the interference from redundant features, retained the most relevant features in the sample, and identified new vital features.

Figure 9 shows the results of the group LASSO feature screening (Yuan & Lin, 2006). The X-axis indicates the risk coefficient of each feature; the Y-axis shows the corresponding feature names. Features with a risk coefficient of zero are not shown. There were 26 high-risk (positive coefficient) and low-risk (negative coefficient) features, as shown in Figure 9. Compared with the features selected by LASSO, the results of the group LASSO algorithm partly differed. However, both methods identified FEV1%FVC A2/Pd as important features.

The top-ranked features in Figure 9 were MEF75 A2/Pd, FEV1%FVC A2/Pd, MEF75 A1/Pd, FEV1%FVC P1, and FEV1%FVC A1/Pd. It was observed that the features given high importance by the group LASSO algorithm were consistent with those identified by the LASSO algorithm. However, they differed in terms of feature types. Both types of sparse feature results were employed as the final dataset for PFT report analysis.

Model Training Results

Table 2 presents the evaluation results of the six models trained using the features extracted by LASSO (Tibshirani et al., 1996). As shown in Table 2, XGBoost achieved the highest classification accuracy of 86.5%, which was also the optimal F1_score. While the other evaluation metrics were slightly lower

Figure 9. Feature extraction by group LASSO algorithm

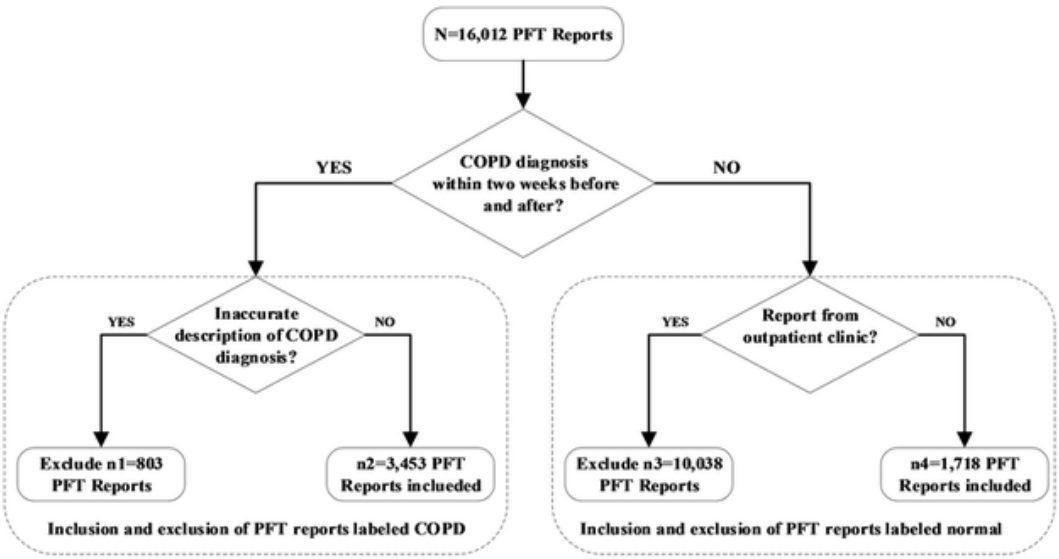


Table 2. Results of models trained with features extracted by the LASSO algorithm

Evaluation Metrics	Accuracy	Recall	Precision	Sensitivity	Specificity	FPR	NPV	F1_Score	AUC
Logistic	0.859	0.950	**0.855**	0.950	**0.676**	**0.324**	0.870	0.900	0.89
SVM	0.860	0.964	0.847	0.964	0.648	0.351	0.900	0.902	0.85
KNN	0.852	**0.984**	0.827	**0.984**	0.586	0.414	**0.947**	0.899	0.89
decision tree	0.843	0.947	0.840	0.947	0.635	0.365	0.856	0.890	0.87
random forest	0.863	0.970	0.847	0.970	0.649	0.351	0.915	0.905	**0.90**
Xgboost	**0.865**	0.970	0.850	0.970	0.654	0.346	0.916	**0.906**	**0.90**

than or equal to the optimal values, the AUC of XGBoost and random forest reached 0.9 (depicted in the ROC curve in Figure 10a). Considering all the evaluation metrics, it is concluded that XGBoost is the best-performing model among those trained with the features selected by LASSO.

Table 3 presents a detailed overview of the evaluation metrics obtained from the model trained using the features selected by group LASSO (Yuan & Lin, 2006). The table shows that Xgboost's classification accuracy is 88.6%, which is 2.1% higher than that of the XGBoost model trained based on the features selected by LASSO. Furthermore, the remaining evaluation metrics of the XGBoost model reached their optimal values. Figure 10b illustrates the ROC curve of the model trained with the features selected by group LASSO. The AUC values of the XGBoost model and the random forest model reached their optimal values of 0.9.

Ultimately, the XGBoost model, trained with the features extracted by group LASSO (Yuan & Lin, 2006), was used as the COPD auxiliary diagnosis model, GL-XGBoost. The accuracy rate of the COPD auxiliary diagnosis model, GL-Xgboost, reached 88.6%. This is 16.2% higher than the 72.4% accuracy of the COPD diagnosis model proposed by Topalovic et al. (2019). This was a significant breakthrough in the field of COPD auxiliary diagnosis because it achieved state-of-the-art performance. The sensitivity of the GL-XGBoost model was 98.5%. It indicated that the model is suitable for large-scale clinical auxiliary screening, can facilitate early detection of COPD, and enables early intervention.

Figure 10. Receiver operating characteristic curve of two type features

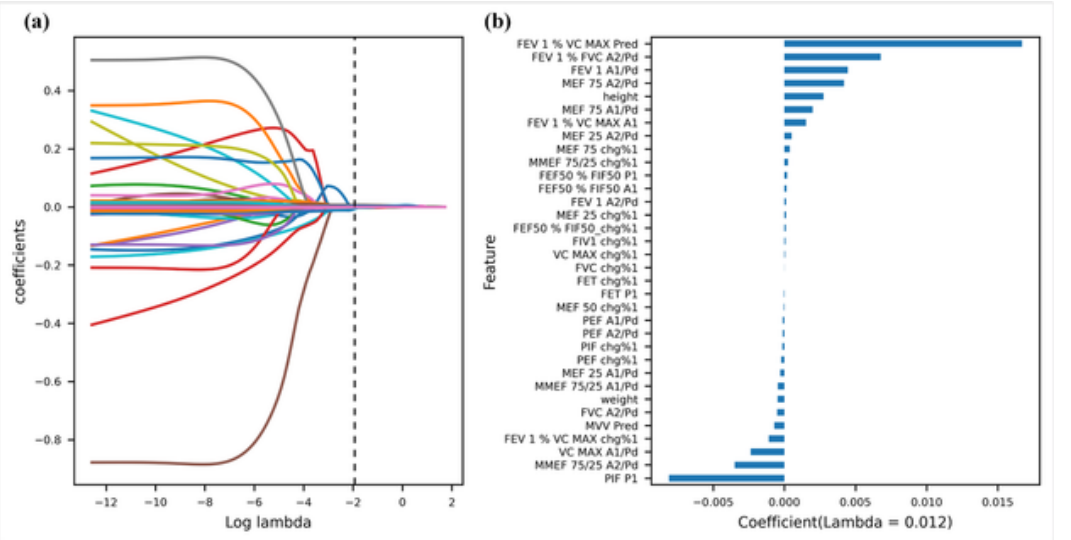


Table 3. Results of models trained with features extracted by the group LASSO algorithm

Evaluation Metrics	Accuracy	Recall	Precision	Sensitivity	Specificity	FPR	NPV	F1_Score	AUC
logistic	0.867	0.936	0.861	0.936	0.685	0.315	0.852	0.903	0.89
SVM	0.756	0.964	0.744	0.964	0.344	0.656	0.829	0.840	0.89
KNN	0.875	0.983	0.852	0.983	0.662	0.338	0.950	0.913	0.89
decision tree	0.866	0.965	0.852	0.965	0.670	0.330	0.906	0.905	0.86
random forest	0.883	0.980	0.855	0.981	0.670	0.330	0.950	0.919	**0.90**
Xgboost	**0.886**	**0.985**	**0.862**	**0.985**	**0.689**	**0.310**	**0.960**	**0.920**	**0.90**

In addition, different feature selection methods can significantly affect model performance. In comparison to the accuracy of the models trained with LASSO, the use of group LASSO increased the accuracy by 2.1%. This indicates that different feature selection methods have different areas of focus. In addition, the COPD auxiliary diagnosis problem is better suited for screening with the group-sparse feature selection algorithm.

Furthermore, the reasons for the different model performances caused by various feature selection algorithms were investigated. Unlike individual sparse feature selection algorithms, group sparse feature selection algorithms consider the combined effects of multiple features and output a list of features ranked in the order of importance. This approach is particularly suitable for medical diagnostic scenarios because a single test value is often inadequate for decision making. Instead, it is necessary to consider multiple test values carefully like the operation of the group sparse feature selection algorithm. Therefore, for complex medical problems, the group-sparse feature selection algorithm is more appropriate for feature screening.

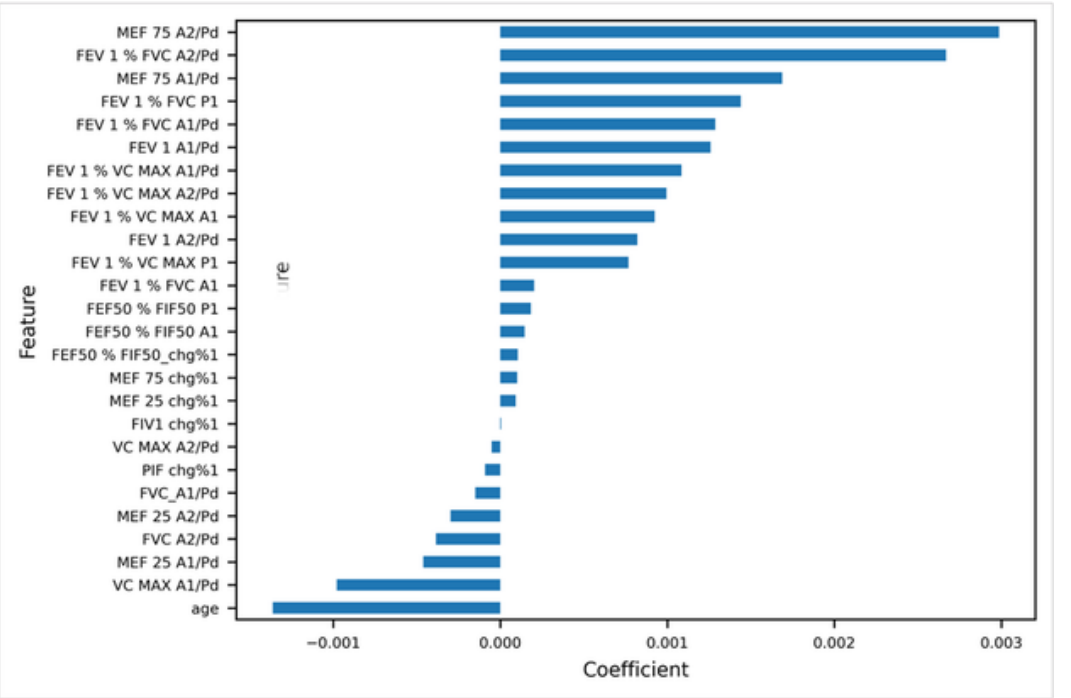
Feature Importance Analysis

This study utilised the framework of SHAP values (Lundberg & Lee, 2017; Lundberg et al., 2018) to rank the feature contributions to COPD predictions by averaging the feature importance estimates (see Figure 11). Figure 11a presents the SHAP summary plot for the top 20 clinical features for COPD prediction. It shows the SHAP values for the most important features of the GL-XGBoost model in the training data. The features in the summary plot (Y-axis) were ordered by the mean absolute SHAP values, representing the importance of the features in driving COPD prediction. The X-axis represents the SHAP value of each sample. The PFT report feature values are coloured according to their relative values in Figure 11a, with red indicating high values and blue indicating low values. Negative SHAP values indicate an increased risk of COPD; positive values indicate trends toward normal ventilation.

Figure 11a shows that FEV 1% FVC P1 was the most important feature for predicting COPD, followed by FEV1% FVC A2/Pd, MEF 75 A2/Pd, FEV 1% VC MAX A1, and FEV 1% VC MAX A2/Pd. FEV1%FVC is a crucial indicator for the clinical diagnosis of COPD. The international diagnostic criteria for COPD approve it (Pellegrino et al., 2005; Vogelmeier et al., 2017). The authors' GL-XGBoost model validates this known feature by focusing on its post-drug value and post-drug ratio to the predicted values. Figure 11a shows the impact of the feature values on the model output. The results are in line with our understanding of the impact of previously known features. Taking the characteristic age as an example, the risk of COPD increases with age. The massive number of older samples in Figure 11a negatively affect the model output, which tends to produce COPD labels.

Furthermore, the model proposes new decision features, such as MEF75 A2/Pd, FEV1%VC MAX A1, and FEV1%VC MAX A2/Pd. The specific meanings of these features were described above and are not repeated here. The discovery of new vital features means that not only previously known clinical features need to be paid attention to, but also newly discovered features should be focused

Figure 11. Feature inspection for GL-Xgboost based on SHAP value



on in the process of clinical diagnosis. Doctors should pay attention to when these characteristics change significantly after bronchodilator inhalation.

Figure 11b shows the ability of each feature to affect the model output sorted by the mean absolute value of SHAP. Figure 11b shows the relative size of the ability of each feature to affect the model output. This demonstrates the importance ranking of features and the decision-making ratio of each feature in the model output more intuitively. Considering the influence of FEV1%FVC P1 on the model output as a benchmark, FEV1%FVC A2/Pd accounted for approximately two-thirds, MEF75 A2/Pd and FEV1%VC MAX A1 accounted for nearly half, and FEV1%VC MAX A2/Pd accounted for slightly less. When making clinical decisions, doctors should pay attention to the suggestions provided by the model.

Moreover, it provides differentiated attention based on the importance ranking of the features. When focusing on a high proportion of decision-making features, even small changes should be carefully considered. Certain tolerance should be given to changes in characteristics that account for a small proportion of decision making.

CONCLUSION

This study presents a novel approach for COPD auxiliary diagnosis using clinical PFTs. The authors developed a comprehensive GL-XGBoost algorithm with exceptional performance, achieving an 88.6% accuracy and 98.5% sensitivity. As the first study to utilise large-scale PFT reports for COPD diagnosis, this study overcomes data acquisition barriers, opens avenues in lung disease analysis, investigates the impact of different feature selection algorithms on model performance, and applies the SHAP algorithm to analyse the importance of GL-Xgboost.

A limitation of this study is found in the need for a larger dataset to validate the transferability and validity of the model. This requires the collection of real data from other hospitals for model verification and improvement. By integrating the model with clinicians' expertise, diagnostic accuracy can be enhanced, COPD misdiagnosis can be reduced, and clinical resources can be saved by avoiding redundant examinations.

FUNDING STATEMENT

This work is supported by Grant No.2022YFC2010200 from the National Key R&D Program of China and the science and technology innovation Program of Hunan Province (No. 2022RC3013).

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

Adeloye, D., Chua, S., Lee, C., Basquill, C., Papana, A., Theodoratou, E., Nair, H., Gasevic, D., Sridhar, D., Campbell, H., Chan, K. Y., Sheikh, A., & Rudan, I. (2015). Global and regional estimates of COPD prevalence: Systematic review and meta-analysis. *Journal of Global Health*, *5*(2), 186–202. doi:10.7189/jogh.05.020415 PMID:26755942

Alkhathlan, B., Greening, N., Harvey-Dunstan, T., & Singh, S. (2020). Acute exacerbation of COPD: A qualitative exploration of the incident symptomatic experience. *The European Respiratory Journal*, *56*(S64), 3025.

Bhosale, Y. H., & Patnaik, K. S. (2023). PulDi-COVID: Chronic obstructive pulmonary (lung) diseases with COVID-19 classification using ensemble deep convolutional neural network from chest X-ray images to minimize severity and mortality rates. *Biomedical Signal Processing and Control*, *81*, 104445. doi:10.1016/j.bspc.2022.104445 PMID:36466567

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27. doi:10.1145/1961189.1961199

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings KDD*, 785–794.

Corlateanu, A., Mendez, Y., Wang, Y., Garnica, R. D., Botnaru, V., & Siafakas, N. (2020). Chronic obstructive pulmonary disease and phenotypes: A state-of-the-art. *Pulmonology*, *26*(2), 95–100. doi:10.1016/j.pulmoe.2019.10.006 PMID:31740261

Crapo, R. O. (1994). Pulmonary-function testing. *The New England Journal of Medicine*, *331*(1), 25–30. doi:10.1056/NEJM199407073310107 PMID:8202099

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing*, *195*(1), 143–148. doi:10.1016/j.neucom.2015.08.112

Galie, N., Humbert, M., Vachiery, J. L., Gibbs, S., Lang, I., Torbicki, A., Simonneau, G., Peacock, A., Noordegraaf, A. V., Beghetti, M., Ghofrani, A., Sanchez, M. A., Hansmann, G., Klepetko, W., Lancellotti, P., Matucci, M., McDonagh, T., Pierard, L. A., Trindade, P. T., & Hoeper, M. et al. (2016). 2015 ESC/ERS guidelines for the diagnosis and treatment of pulmonary hypertension: The joint task force for the diagnosis and treatment of pulmonary hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *European Heart Journal*, *37*(1), 67–119. doi:10.1093/eurheartj/ehv317 PMID:26320113

Halpin, D. M., Criner, G. J., Papi, A., Singh, D., Anzueto, A., Martinez, F. J., Agusti, A. A., & Vogelmeier, C. F. (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. *American Journal of Respiratory and Critical Care Medicine*, *203*(1), 24–36. doi:10.1164/rccm.202009-3533SO PMID:33146552

Hasenstab, K. A., Yuan, N., Retson, T., Conrad, D. J., Kligerman, S., Lynch, D. A., & Hsiao, A. (2021). Automated CT staging of chronic obstructive pulmonary disease severity for predicting disease progression and mortality with a deep learning convolutional neural network. *Radiology: Cardiothoracic Imaging*, *3*(2), e200477. PMID:33969307

Ho, T. T., Kim, T., Kim, W. J., Lee, C. H., Chae, K. J., Bak, S. H., Kwon, S. O., Jin, G. Y., Park, E. K., & Choi, S. (2021). A 3D-CNN model with CT-based parametric response mapping for classifying COPD subjects. *Scientific Reports*, *11*(1), 1–12. doi:10.1038/s41598-020-79336-5 PMID:33420092

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. doi:10.1126/science.aaa8415 PMID:26185243

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, *23*(1), 89–109. doi:10.1016/S0933-3657(01)00077-X PMID:11470218

Lavalley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395–2399. doi:10.1161/CIRCULATIONAHA.106.682658 PMID:18458181

Leidy, N. K., Malley, K. G., Steenrod, A. W., Mannino, D. M., Make, B. J., Bowler, R. P., Thomashow, B. M., Barr, R. G., Rennard, S. I., Houfek, J. F., Yawn, B. P., Han, M. K., Meldrum, C. A., Bacci, E. D., Walsh, J. W., & Martinez, F. (2016). Insight into best variables for COPD case identification: A random forests analysis. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation*, *3*(1), 406–418. doi:10.15326/jcopdf.3.1.2015.0144 PMID:26835508

Li, C., Liu, W., Guo, R., Yin, X., Jiang, K., Du, Y., Du, Y., Zhu, L., Lai, B., Hu, X., Yu, D., & Ma, Y. (2022). *PP-OCRv3: More attempts for the improvement of ultra lightweight OCR system.* arXiv:2109.03144.

Li, X., Wang, Y., & Ruiz, R. (2022). A survey on sparse learning models for feature selection. *IEEE Transactions on Cybernetics*, *52*(3), 1642–1660. doi:10.1109/TCYB.2020.2982445 PMID:32386172

Liao, M., Wan, Z., Yao, C., Chen, K., & Bai, X. (2020). Real-time scene text detection with differentiable binarization. *Proceedings AAAI*, 11474–11481. doi:10.1609/aaai.v34i07.6812

Lopez-Campos, J. L., Tan, W., & Soriano, J. B. (2016). Global burden of COPD. *Respirology (Carlton, Vic.)*, *21*(1), 14–23. doi:10.1111/resp.12660 PMID:26494423

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Adair, T., Aggarwal, R., Ahn, S., Alvarado, M., Andrews, K., Anderson, H. R., Atkinson, C., Bennett, D., Berry, R. J., Bhalla, K., Bikbov, B., Bolliger, I., Bucello, C., & Murray, C. J. et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the global burden of disease study 2010. *Lancet*, *380*(9859), 2095–2128. doi:10.1016/S0140-6736(12)61728-0 PMID:23245604

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Proceedings of the Advances in Neural Information Processing Systems*, 4765–4774.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K., Newman, S., Kim, J., & Lee, S. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, *2*(10), 749–760. doi:10.1038/s41551-018-0304-0 PMID:31001455

Martinez, F. J., Chisholm, A., Collard, H. R., Flaherty, K. R., Myers, J., Raghu, G., Walsh, S. L., White, E. S., & Richeldi, L. (2017). The diagnosis of idiopathic pulmonary fibrosis: Current and future approaches. *The Lancet. Respiratory Medicine*, *5*(1), 61–71. doi:10.1016/S2213-2600(16)30325-3 PMID:27932290

Mostafaei, S., Kazemnejad, A., Jamalkandi, S. A., Amirhashchi, S., Donnelly, S. C., Armstrong, M. E., & Doroudian, M. (2018). Identification of novel genes in human airway epithelial cells associated with chronic obstructive pulmonary disease (COPD) using machine-based learning algorithms. *Scientific Reports*, *8*(1), 1–20. doi:10.1038/s41598-018-33986-8 PMID:30361509

Pellegrino, R., Viegi, G., Brusasco, V., Crapo, R. O., Burgos, F., Casaburi, R., Coates, A., Grinten, C. P. M., Gustafsson, P., Hankinson, J., Jensen, R., Johnson, D. C., Maclntyre, N., Mckay, R., Miller, M. R., Navajas, D., Pedersen, O. F., & Wanger, J. (2005). Interpretative strategies for lung function tests. *The European Respiratory Journal*, *26*(5), 948–968. doi:10.1183/09031936.05.00035205 PMID:16264058

Shi, B., Bai, X., & Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2298–2304. doi:10.1109/TPAMI.2016.2646371 PMID:28055850

Spathis, D., & Vlamos, P. (2019). Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Informatics Journal*, *25*(3), 811–827. doi:10.1177/1460458217723169 PMID:28820010

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, *28*(1), 112–118. doi:10.1093/bioinformatics/btr597 PMID:22039212

Tang, L. Y., Coxson, H. O., Lam, S., Leipsic, J., Tam, R. C., & Sin, D. D. (2020). Towards large-scale case-finding: Training and validation of residual networks for detection of chronic obstructive pulmonary disease using low-dose CT. *The Lancet Digital Health*, *2*(5), e259–e267. doi:10.1016/S2589-7500(20)30064-9 PMID:33328058

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, *58*(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Tinkelman, D. G., Price, D. B., Nordyke, R. J., & Halbert, R. J. (2006). Misdiagnosis of COPD and asthma in primary care patients 40 years of age and over. *The Journal of Asthma*, *43*(1), 75–80. doi:10.1080/02770900500448738 PMID:16448970

Topalovic, M., Das, N., Burgel, P. R., Daenen, M., Derom, E., Haenebalcke, C., Janssen, R., Kerstjens, H. A. M., Liistro, G., Louis, R., Ninane, V., Pison, C., Schlesser, M., Vercauter, P., Vogelmeier, C. F., Wouters, E., Wynants, J., & Janssens, W. (2019). Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *The European Respiratory Journal*, *53*(4), 1801660. doi:10.1183/13993003.01660-2018 PMID:30765505

Vestbo, J., Hurd, S. S., Agusti, A. G., Jones, P. W., Vogelmeier, C., Anzueto, A., Barnes, P. J., Fabbri, L. M., Martinez, F. J., Nishimura, M., Stockley, R. A., Sin, D. D., & Rodriguez-Roisin, R. (2013). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American Journal of Respiratory and Critical Care Medicine*, *187*(4), 347–365. doi:10.1164/rccm.201204-0596PP PMID:22878278

Vogelmeier, C. F., Criner, G. J., Martinez, F. J., Anzueto, A., Barnes, P. J., Bourbeau, J., Celli, B. R., Chen, R., Decramer, M., Fabbri, L. M., Frith, P., Halpin, D. M. G., Varela, M. V., Nishimura, M., Roche, N., Rodriguez-Roisin, R., Sin, D. D., Singh, D., Stockley, R., & Agusti, A. et al. (2017). Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. GOLD executive summary. *American Journal of Respiratory and Critical Care Medicine*, *195*(5), 557–582. doi:10.1164/rccm.201701-0218PP PMID:28128970

Willer, K., Fingerle, A. A., Noichl, W., De Marco, F., Frank, M., Urban, T., Schick, R., Gustschin, A., Gleich, B., Herzen, J., Koehler, T., Yaroshenko, A., Pralow, T., Zimmermann, G. S., Renger, B., Sauter, A. P., Pfeiffer, D., Makowski, M. R., Rummeny, E. J., & Pfeiffer, F. et al. (2021). X-ray dark-field chest imaging for detection and quantification of emphysema in patients with chronic obstructive pulmonary disease: A diagnostic accuracy study. *The Lancet Digital Health*, *3*(11), e733–e744. doi:10.1016/S2589-7500(21)00146-1 PMID:34711378

Xu, C., Qi, S., Feng, J., Xia, S., Kang, Y., Yao, Y., & Qian, W. (2020). DCT-MIL: Deep CNN transferred multiple instance learning for COPD identification using CT images. *Physics in Medicine and Biology*, *65*(14), 145011. doi:10.1088/1361-6560/ab857d PMID:32235077

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *68*(1), 49–67. doi:10.1111/j.1467-9868.2005.00532.x

Yongfu Yu received the B.E. degree in computer science from Qingdao University of Science and Technology, Qingdao, China, in 2020. He is currently pursuing M.E. degree in software engineering with the Central South University. His research interests include medical big data processing, machine learning and artificial intelligence assisted medical diagnosis. Nannan Du received the M.B. degree in clinical medicine from Zheng-zhou University, ZhengZhou, China, in 2021. She is currently pursuing M.S. degree in Xiangya Hospital of Central South University. Her research interests include occurrence and evolution of lung cancer and the molecular mechanism of lung cancer metastasis. He received the B.E. degree in software engineering from Central South University, Changsha, China in 2022. He is currently pursuing M.E. degree in software engineering with the Central South Uni-versity. His research interests include machine learning, sports health and medical health. Weihong Huang received the B.Eng. degree in automation and the M.Eng. degree in pattern recognition and smart control from Southeast University, China, in 1995 and 1998, respectively, and the Ph.D. degree in computer science from Nanjing University, China, in 2001. From 2001 to 2002, he was a Postdoctoral Research Fellow with CNRS University Lyon 1, France. From 2002 to 2005, he was a Lecturer with the Department of Computer Science, University of Hull, U.K. From 2005 to 2014, he was a Senior Lecturer with the School of Computer and Information Systems, Kingston University London, U.K. Since 2016, he has been a Professor and the Deputy Director of the Mobile Health Ministry of Education–China Mobile Joint Laboratory, Xiangya Hospital Central South University, China. His research interests include mobile health, artificial intelligence in healthcare, cognitive computing for healthcare, semantic multimedia computing, and knowledge graph applications. Dr. Huang is a Committee Member of the China Hospital Information Management Association, a Standing Committee Member of the Medical and Health Big Data Evaluation and Assurance Board of the Chinese Health Information and Big Data Association, and the Chairman of the Specialized Committee of Information Management of the Hunan Health Management Association.

Min Li is the post doctor of medicine, deputy director of the Department of Respiratory and Critical Care Medicine, Xiangya Hospital, Central South University, and executive deputy director of the Hunan Clinical Medical Research Center for Respiratory Diseases; Her research interests include occurrence and evolution of lung cancer and the molecular mechanism of lung cancer metastasis.