REVIEW

Using machine learning for early detection of chronic obstructive pulmonary disease: a narrative review

Xueting Shen[1] and Huanbing Liu[1,2*]

Abstract

Chronic obstructive pulmonary disease (COPD) is a prevalent respiratory disease and ranks third in global mortality rates, imposing a significant burden on patients and society. This review looks at recent research, both domestically and abroad, on the application of machine learning (ML) for early COPD screening. The review discusses **the practical application**, **key optimization points**, **and prospects** of ML techniques in early COPD screening. The aim is to establish a scientific foundation and reference framework for future research and the development of screening strategies.

Keywords Chronic obstructive pulmonary disease, Machine learning, Early screening

Chronic obstructive pulmonary disease (COPD) is a common, preventable, and treatable condition that contributes significantly to the global burden of chronic noncommunicable diseases [1]. It has profound implications for both mortality and morbidity. In 2019, COPD affected approximately 391.9 million individuals globally and caused over 3 million fatalities [2, 3]. Projections indicate that by 2060, COPD and related diseases will cause over 5.4 million deaths annually, underscoring the escalating disease burden [4, 5]. In China, COPD prevalence and mortality rates are on the rise. Survey data revealed that COPD prevalence among individuals aged over 40 increased from 8.2% in 2002–2004 to 13.7% by 2015—a ten-year surge of 67% [6]. For those over 60, the prevalence has surpassed 27% [7]. China also experiences a sub-stantial mortality rate from COPD, with 876,300 deaths, or 29.86% of the global COPD mortality, as reported in 2016 [8]. As a leading cause of death and contributor to disability-adjusted life years in China, COPD, alongside chronic ailments like hypertension and diabetes, presents a significant public health challenge and disease burden.

Despite the high disease burden of COPD attracting widespread international attention, early diagnosis still faces significant challenges. The subtlety of the initial symptoms of COPD extends the average time from the first symptoms to diagnosis to 3.6±4 years, a delay which contributes to the prevalence of missed diagnoses [9]. A study from the UK found that 85% of patients with COPD missed the opportunity for early diagnosis in the five years before their diagnosis [10]. Research in China demonstrated that among diagnosed COPD patients, a mere 35.1% had previously been identified with related conditions, including emphysema, asthma, or bronchitis [6]. It is estimated that 70-80% of adult COPD patients remain undiagnosed [11, 12]. Undiagnosed patients with COPD face higher health risks than individuals without COPD, including a significant increase in acute exacerbations,

*Correspondence:
Huanbing Liu
liuhuanbing6911@sina.com
[1]Department of General Medicine, The First Affiliated Hospital of Nanchang University, Nanchang 330000, China
[2]Department of General Practice, The First Affiliated Hospital of Nanchang University, Nanchang 330000, China

pneumonia, respiratory system cause of death, and all-cause mortality if timely intervention is not provided [13]. Notably, the rate of decline in the Forced Expiratory Volume in one second (FEV_1) among patients with mild COPD is significantly higher than that in the high-risk population, highlighting the importance of early intervention in slowing disease progression [14, 15].

Machine learning (ML) is a branch of artificial intelligence and computer science, that offers unprecedented possibilities for the diagnosis, treatment, and management of diseases, marking a technological revolution in the medical field. The emergence of ML holds promise in enhancing the accuracy and efficiency of early COPD diagnosis, presenting novel approaches in dealing with the current challenges of early COPD detection. Thus, we review the recent literature on the application of ML in early COPD screening, both domestically and internationally. It summarises the **practical applications**, **key optimization points**, **and prospects** for development, providing a reference for the initiation of related research and the formulation of screening strategies.

Conventional methods and early COPD screening's limitations

Pulmonary function tests (PFTs) involve multidimensional assessments of lung volume and ventilatory function, providing an all-encompassing evaluation of the pulmonary gas exchange capacity [16]. PFTs are now considered the "gold standard" for COPD diagnosis and are essential for assessing the severity, course, response to treatment, and prognosis of the illness [1]. However, the application of spirometry encounters persistent issues, manifesting as a deficiency in healthcare professionals' expertise with PFT interpretation, prohibitive costs of spirometry apparatus, and the economic burden on patients regarding test affordability—conditions that are exacerbated in under-resourced primary healthcare contexts [13, 17–20]. The suboptimal accessibility of PFTs has increasingly become a significant obstacle in diagnosing COPD, highlighting the necessity for additional resources and efforts to investigate and establish straightforward, cost-effective COPD screening methodologies that will enhance the precision of early COPD detection.

Currently, primary screening for COPD usually adopts a two-tiered screening protocol for COPD, comprising questionnaires followed by pulmonary function tests (PFTs). This framework designates screening questionnaires as a preliminary filter to single out individuals at high risk, who are then subjected to spirometric evaluation to ascertain COPD diagnoses. The 'questionnaire plus PFT' strategy offers a more straightforward and expedient alternative to PFTs alone, with the added benefit of cost-effectiveness. Nevertheless, extant primary care guidelines have yet to delineate the optimal questionnaire for this purpose, presenting clinicians at the grassroots level with a challenge in decision-making. Moreover, current questionnaires predominantly account for physiological indices (age, body mass index), lifestyle habits (smoking status, usage of coal and biomass fuels), family history, and symptoms related to the respiratory tract [21, 22] While relevant, these parameters represent a limited scope and fail to encompass the multifaceted risk factors implicated in the development of COPD.

Peak expiratory flow (PEF) is a rapid, convenient, and economical method for assessing pulmonary function and the degree of airway constriction. It indicates lung function status by measuring the maximum speed of expiration [23]. Studies have demonstrated a strong correlation between PEF and FEV_1 measured using spirometry [24]。As an essential tool for monitoring and screening COPD and other respiratory diseases, peak flow meters are compact, low-cost, easy to operate, portable, and capable of delivering quick results, assisting physicians in identifying potential COPD patients or high-risk individuals promptly. PEF is particularly critical for early COPD screening, especially in environments with limited resources [13]. Research evidence suggests that combining PEF measurements with screening questionnaires can effectively enhance the screening efficiency for COPD [25–27]. This could be a result of PEF's ability to measure airway blockage objectively, improving COPD diagnosis capabilities. Nelson et al. recommended a three-tier approach (risk factor questionnaire, PEF, and spirometry) to enhance the sensitivity and specificity of diagnosing moderate to severe COPD [28]. However, there is no standard formula for estimating PEF in China, which limits the widespread use of PEF [29]. Nevertheless, PEF can be used to initially identify patients with abnormal lung function and recommend further pulmonary function tests to confirm the diagnosis.

The application of machine learning (ML) in early screening for COPD
The introduction of ML
ML is a subfield of artificial intelligence (AI) that provides data-driven tools to support and optimize decision-making processes [30]. ML employs algorithms and statistical models to identify patterns within data, improving the performance of computer systems based on accumulated experience and reducing dependency on pre-programmed instructions. ML has been widely applied in various branches of the medical field, including medical image processing, genomics, drug discovery, and patient management, due to its ability to handle complex non-linear relationships between predictive variables and generate novel outcomes [31].

Compared with traditional models, ML has a significant advantage in its ability to handle data. Traditional models usually require rigorous data preprocessing and make assumptions based on linear relationships and variable independence. In contrast, ML can analyze and mine complex patterns and relationships in large and diverse datasets, maintaining robust performance even in cases of incomplete data or noise, and providing more accurate risk predictions [30, 32]. Appropriate training enables ML models to integrate clinical, physiological, imaging, and demographic data, providing a comprehensive perspective for assessing disease risk. This facilitates the development of personalized screening, diagnostic strategies, and interventions [33–35]. The advantages of ML highlight its significant role in improving the quality of medical decision-making, promoting innovation in medical research and practice, and its continuously growing potential for application in the medical field.

The process of ML comprises several steps, including data collection, preprocessing, feature engineering, model selection, training, evaluation, optimization, and deployment [36]. Data collection and preprocessing are foundational steps for building the model, while feature engineering and model selection are critical and distinctive steps for constructing personalized models. Feature engineering involves extracting useful features from raw data to enhance the model's performance. Model selection is the process of choosing the most appropriate algorithm for a specific issue. Machine learning plays an increasingly important role in medical decision-making, disease diagnosis, and the formulation of treatment strategies. It also has notable application value and potential, particularly in the early screening of COPD. It has the potential to achieve more precise and personalized early diagnostics, thereby improving treatment outcomes and enhancing patients' quality of life.

Feature variables for machine learning models in early COPD screening

Feature engineering involves extracting, selecting, and constructing features from raw data to aid the model's learning process. It requires harnessing the raw data and translating it into a form that the model can comprehend. Specialized medical knowledge is necessary to accurately identify characteristics closely associated with early COPD diagnosis.

The progression of COPD is caused by the dynamic, cumulative, and repetitive interplay of multiple risk factors that either damage the lungs or affect their development and aging processes [37]. Early screening and risk assessment for COPD traditionally rely on direct clinical observation, pulmonary function tests, and evaluations of known risk factors. Clinical observation is primarily focused on respiratory symptoms, including coughing, expectoration, wheezing, dyspnea, and shortness of breath [38]. Risk factors include genetic predisposition (such as a family history of respiratory diseases and relevant genetic data), environmental factors (such as smoking and exposure to air pollution), and life course events (such as being underweight, a history of chronic cough during childhood, and a lower level of education) [7]. Although these methods have practical value, they are limited in handling multi-source data and capturing complex disease patterns. To overcome these limitations, ML can be used to construct early screening models for COPD. This approach takes advantage of existing data resources, thoroughly analyzes critical information hidden in high-dimensional data, identifies early characteristics and patterns of COPD, and enhances the accuracy and reliability of screening. Additionally, automated screening processes can be established across multiple devices and in the cloud, enabling remote screening and monitoring. This improves the accessibility of medical services and reduces the workload on healthcare professionals.

ML can process high-resolution imaging data, identifying key features in pulmonary images such as evidence of emphysema, changes in airway wall thickness, and functional small airway disease [39]. These subtle changes can be detected non-invasively, aiding in the early diagnosis and phenotypic assessment of COPD. Chest X-rays are an inexpensive imaging option that provides readily available images. However, their ability for longitudinal monitoring or precise assessment of disease areas is limited, which makes them less useful in primary COPD screening [40]. Computed Tomography (CT) scans can accurately identify early structural damage in COPD by detecting abnormalities in the airway tree and lung field morphology, quantifying emphysema artifacts in detail, observing low-density areas, and precisely assessing changes in the airway and pulmonary vascular systems [39–42]. Studies have shown that the severity of emphysema on CT is highly correlated with the degree of lung parenchymal destruction seen pathologically [43, 44]. However, the routine use of CT scanning for COPD screening and diagnosis is limited due to the risks associated with radiation exposure [45]. Consequently, researchers are exploring the use of low-dose CT and other non-radiation imaging techniques like Magnetic Resonance Imaging (MRI) and Electrical Impedance Tomography (EIT) for safe and efficient early screening of COPD. Despite its potential, MRI has not been widely used in COPD diagnosis due to low proton density in lung tissue and rapid signal decay [46]. With technological advancements, particularly through the introduction of hyperpolarized gases or contrast agents to quantitatively evaluate lung function, MRI has shown potential for high-resolution visualization of ventilation,

perfusion, and airflow changes [47]. Additionally, EIT allows for non-invasive quantification of regional lung function changes through 3D reconstruction and analysis of impedance data [45].

In addition to high-resolution imaging data, ML can also analyze other types of data, such as audio, to identify COPD. For example, electronic microphones or stethoscopes can record lung auscultation sounds and breathing sounds. The resulting audio can then be analyzed as a time-series signal to extract information about time, frequency, and energy. This allows for the recording of crucial characteristics associated with the respiratory sounds of COPD [48–50]. Impulse Oscillometry (IOS) is a method that measures respiratory impedance reference values by generating rectangular electromagnetic impulses using an external generator during a subject's calm breathing [51]. Applying ML to analyze audio data from iOS can help quantify the lung's response to different pressure frequencies, aiding in the diagnosis of COPD [52, 53]. Moreover, the analysis of respiratory muscle activity is also crucial. The combination of muscles that function independently for breathing is diverse, and abnormal working patterns of these muscles may indicate COPD. ML analysis of electromyography data from the sternocleidomastoid muscle can reveal such abnormal signals. This is very important for the early identification and treatment of COPD [54]. Siddiqui and colleagues utilized an Ultra-Wideband (UWB) radar wireless sensing system to collect time-frequency spectral characteristics related to respiration from a spatial position 1.5 m away from the patient. These characteristics were then used to construct an ML model for COPD [55, 56].

However, the features mentioned above require medical personnel to actively collect data, which to some extent increases the workload of healthcare professionals. To tackle this issue, some researchers have attempted to extract data from electronic medical records and other health information management systems, using ML to construct knowledge graphs or "expert systems" for COPD [57, 58]. The approach centers on using medical data resources and employing efficient data mining and analysis techniques to automatically identify key information and patterns related to COPD. This method can process and analyze large volumes of complex medical data with reduced reliance on medical professionals, thereby enhancing the accuracy and efficiency of early COPD diagnosis. Orchard and colleagues' research has shown how ML methods can integrate remote monitoring and weather data to provide more refined and personalized COPD risk management strategies [59]. Although the research did not specifically target early screening for COPD, it innovatively used non-traditional predictive variables, such as meteorological data. This not only broadens the diversity of data sources but also showcases the considerable capability of machine learning in synthesizing and analyzing large amounts of heterogeneous data. Comparing traditional scoring algorithms with machine learning methods, the research found the latter to be more effective in predicting hospital admission needs for COPD patients and decision-making regarding corticosteroid use. This highlights the potential value of machine learning in developing precise and personalized early screening tools for COPD. The findings indicate that ML technology could be further utilized to process and analyze multimodal data, including clinical, imaging, environmental, and lifestyle data, as a potential direction for future COPD management. ML technology could improve the accuracy of diagnosis and risk assessment, as well as foster more personalized treatment strategies. This could ultimately lead to more effective management and care for COPD patients.

ML models for early screening of COPD

ML techniques can be categorized into four main types based on their reliance on manually annotated data during the learning process: supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is a widely applied method that depends on a labeled training dataset to learn the inherent rules and associations within the data. This training dataset includes input data and their corresponding output labels, allowing supervised learning to effectively predict the labels of new, unknown data. In the medical field, especially in precise diagnostics and risk stratification, supervised learning plays a crucial role. Supervised learning can be classified into two main types: classification and regression tasks. Classification tasks aim to predict discrete label values, also known as 'output', while regression tasks predict continuous label values. In early screening for COPD, the focus is on determining whether a person already has COPD or belongs to a high-risk group. Pulmonary function indicators, such as FEV_1, can be used to predict COPD. However, it is important to follow authoritative guidelines that provide clear standards for the diagnosis and grading of COPD. In order to identify earlier patients, a regression model may be used. However, it may be necessary to perform a more detailed analysis on the lung function indicators to determine more sensitive thresholds, rather than just using the fixed standard of FEV_1 to Forced Vital Capacity (FVC) ratio. Therefore, for the current early screening model for COPD, a supervised learning classification model is more suitable than a regression model because it can directly predict discrete output labels.

Unsupervised learning is a method of discovering hidden structures and patterns in data without labeled outputs. In the context of early COPD screening, unsupervised learning methods (like K-means clustering or

hierarchical clustering) can serve as tools for initial data exploration, identifying patient groups with similar characteristics to uncover potential high-risk groups. Semi-supervised learning combines the features of supervised and unsupervised learning by training models with a small amount of labeled data and a large amount of unlabeled data. This approach enhances data utilization and reduces the need for costly large-scale data annotation. In COPD screening, a preliminary model can predict the labels of unlabeled data, and these predicted results can be used as new training data to optimize model performance through iterative processes. However, in the medical field, the diagnostic gold standard is typically adopted for medical models due to the importance of the accuracy and reliability of model results. Extreme caution must be exercised when directly using model-predicted results for diagnosis. Reinforcement learning learns the optimal decision-making strategy through interaction with the environment, without relying on explicit labels or target outputs. Given its complex environment modeling and feedback mechanisms, its application in clinical diagnosis and optimizing treatment plans is still in the exploratory stage, and its direct application in early COPD screening remains somewhat limited. Figure 1 illustrates the classification of these ML algorithms.

Evaluation of ML models for early screening of COPD

In early COPD screening, machine learning models are evaluated using two types of indicators: those based on the confusion matrix and those based on model prediction probability values. Confusion matrix is a two-dimensional grid that categorizes results as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) based on the model's predictions compared to the actual situation. Indicators such as sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), F1 score, and Matthews correlation coefficient are calculated from these four basic elements. These indicators assess the model's predictive performance in positive and negative classification tasks comprehensively. In screening trials, it is crucial to balance sensitivity and specificity. Sensitivity measures a screening test's ability to accurately identify patients with the disease, while specificity reflects its ability to accurately exclude non-patients. Ideally, high sensitivity can minimize false-negative results, reducing misdiagnoses. When the disease is in the early screening stage, sensitivity is often increased at the expense of specificity to increase potential screening value [60].

Evaluation indicators based on model prediction probability values, such as the Receiver Operating Characteristic Curve (ROC), the Area Under the ROC Curve (AUROC), Precision-Recall Curves, and the Area Under
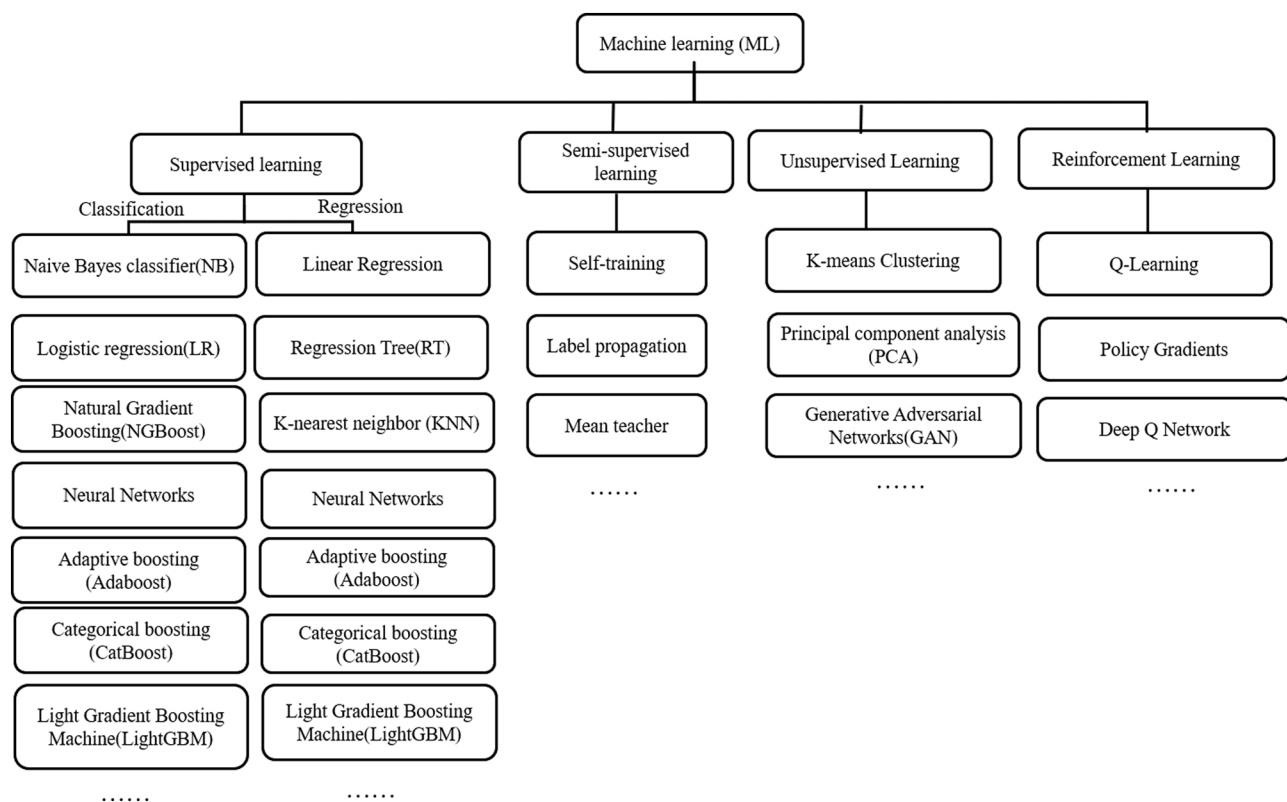

Fig. 1 Classification of ML Algorithms

the Precision-Recall Curve (AUPRC), provide methods to assess the performance of a model under different thresholds. This differs from indicators based on the confusion matrix, which evaluates the model's categorical outputs at a fixed threshold. The AUROC is a widely used indicator for measuring the overall performance of diagnostic methods, especially suitable for comparing different diagnostic methods and determining the best diagnostic boundaries [61]. However, since the ROC curve is mainly based on sensitivity and specificity, it does not reflect the impact of the proportion of positive samples. This can potentially result in a large number of false positives, especially in scenarios where disease prevalence is low (less than 5%) [62]. To provide a more accurate performance assessment for datasets with imbalanced proportions of positive and negative samples, the AUPRC offers an alternative perspective by emphasizing the precision and recall rate of positive case predictions. Consequently, when evaluating a model's ability to predict rare events, the AUPRC is more effective than the AUROC.

It's important to note that the performance of a model is significantly influenced by the characteristics of the training dataset. The impact of the dataset is primarily manifested in differences in data sources and processing. COPD is influenced by a variety of factors, including population demographics, genetics, and the environment. The complexity of individual physical responses, variability of the environment, and differences in equipment performance and measurement techniques can all affect data quality and, consequently, the authenticity and accuracy of model predictions. Similarly, differences in preprocessing, cleaning, and annotation methods for the data inputted into the model can impact the model's learning effectiveness and prediction accuracy. Therefore, **the performance of models cannot simply be compared on a horizontal basis**, especially when evaluating the effects of models based on different studies and datasets. This is an issue that requires continuous vigilance in practical applications and research.

Case Study of ML models for early screening of COPD

In previous research on early screening models for COPD, Lin et al. extracted data from public health databases of residents [63]. They compared the effectiveness of 18 machine learning models in distinguishing high-risk COPD groups (with scores greater than 16 on the COPD-SQ scale). As a result, they developed a convenient and effective clinical decision system to assist in case discovery. This improved the ability and efficiency of primary medical institutions in detecting COPD cases. This research stands out in current COPD early screening studies for its utilization of the most diverse models. Among the compared models, the gradient boosting classifier (GBC), CatBoost (Categorical boosting), Light

Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Logistic Regression (LR) performed excellently. Among them, the CatBoost model performed the best, with an AUC of 99.85% and a sensitivity of 94.81%. Wang et al. collected data on COPD through questionnaires. They used Logistic Regression (LR), XGBoost, Generalized Additive Model (GAM), and Random Forest (RF) to create a COPD risk screening questionnaire for high-risk cohorts. The study revealed that GAM had the best screening performance, with an AUC exceeding 0.8 [64]. Similar to Wang et al.'s approach, another researcher combined Support Vector Machine (SVM), Natural Gradient Boosting (NGBoost), and stacking with LR, RF, and XGBoost to identify individuals with a high risk of COPD [65]。. Shigeo et al. compared the effectiveness of the XGBoost and LR models in COPD screening among healthy populations using employee health examination data. The results showed that XGBoost outperformed LR in prediction (AUC>0.95). However, it is important to note that the model was built using data from a single company's employee health examinations, which may limit its generalizability [66].

The analysis of various studies indicates that **gradient-boosting strategies** are effective in early COPD screening. The occurrence could plausibly be attributed to these models' ability to progressively learn from multiple weak predictor models and amalgamate them into a compelling predictive model. Each new model attempts to correct the errors of all previous models, resulting in a highly accurate model. While gradient-boosting strategies have shown excellent performance in multiple studies, considerations for their effectiveness in real-world applications should also include additional factors. These factors include dataset size, the number of features, intervariable relationships, and the time required to train and test the model. Therefore, when selecting a model, we need to consider not only its performance but also have a sufficient understanding of the environment and tasks for which the model is suitable.

Challenges and prospects of ML in early screening of COPD

In the field of early COPD screening, ML exhibits tremendous potential due to its unique advantages. ML allows for the integration of multimodal data, including clinical, radiological, and genomic data, to provide more comprehensive disease information and improve the accuracy of early COPD detection.

Current practices in implementing early-stage COPD screening models often rely solely on single-modal features. That is, features of a single modality (such as imaging, pulmonary auscultation sound, biochemical indicators, etc.) are input separately into a classifier for processing. While this method simplifies the design

and training of the classifier and is easy to understand, it fails to fully utilize information from other modalities, limiting the performance of model classification and prediction. On the other hand, a more comprehensive information profile can be obtained by inputting all modal features, such as pulse oximetry, medical history, genetic data, CT scans, and lung auscultation sounds, into a single classifier. This approach would reveal both shared and unique information among the different modal features, thereby enhancing diagnostic accuracy. Notably, a strategy that capitalizes on the synergistic capabilities of multi-criterion decision-making (MCDM) and multi-classifier fusion (MCF) merits attention [67, 68]. This methodology encompasses multiple objectives, constraint conditions, and interrelated decision criteria, providing superior performance over any single classifier. Utilizing this tactic in constructing an early-stage COPD screening model will help us leverage various predictive technologies, handle multi-type patient data, and build and optimize models with multiple decision-making criteria such as sensitivity, AUROC, and AUPRC. Therefore, it is important not to overlook the significance of MCDM and MCF in early COPD screening. Their involvement requires careful consideration and application in future research.

The quality and accessibility of data present another significant challenge in model construction. Despite an abundance of clinical data accumulated within the current healthcare system, the integrity, consistency, and accuracy of this data profoundly impact model performance. Furthermore, restrictions stemming from patient privacy protection and medical confidentiality regulations hinder the full utilization of data of research value, undoubtedly obstructing the model's training and optimization. During the construction of early COPD screening models, relevant information can be gathered from existing databases. However, the final evaluation of the model still relies on the 'gold standard' of pulmonary function diagnosis, which is a post-bronchodilator FEV_1 to FVC ratio of less than 0.7. However, pulmonary function testing is a time-consuming and strenuous process that requires the cooperation of professional medical personnel and patients. This undoubtedly adds to the difficulty of data acquisition and limits further optimization of early COPD screening models. In a real-world environment, pulmonary function tests are typically only conducted on populations with respiratory symptoms or those considered high-risk by physicians, rather than on the general population. This may potentially impact the model results. Due to the non-prominent clinical manifestations of early-stage COPD, relevant medical data may not have been fully captured or may have been overlooked, resulting in a sparse and incomplete data source. To address these challenges, future efforts should prioritize promoting digitization and data sharing in healthcare while ensuring patient privacy and data security. Standardized data collection strategies with unified norms can be implemented to determine the format, granularity, category, indicators, and timestamps of data. This will improve the consistency of data and reduce deviations in data quality.

Furthermore, the interpretability of a model is crucial for its applicability in medical decision-making. A significant challenge, however, is that while many of the current machine learning techniques have the potential to outperform traditional statistical methods in terms of predictive accuracy, their decision-making process is often metaphorically described as a 'black box'. Although an ML model can clarify the relationship between inputs and outputs, understanding its internal processing mechanism, particularly the mutual influence of feature selection and results, remains relatively vague. This presents a significant challenge for decision-makers, as comprehending the basis for model predictions to formulate appropriate treatment plans is essential in clinical applications. Efforts have been made to address this issue. This includes creating models with built-in interpretability mechanisms and implementing algorithms such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to improve model interpretability. These methods often demand significant computational resources and their outcomes are only estimations of the actual model, which poses certain risks. These methods typically require substantial computational resources, and their results are only approximations of the actual model, thus posing certain risks. However, looking to the future, there will be ample opportunities to address this issue due to the increasing convergence of machine learning and medicine.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

References

1. Chronic Obstructive Pulmonary Disease Group of Chinese Thoracic Society, Chronic Obstructive Pulmonary Disease Committee of Chinese Association of Chest Physician. Guidelines for the diagnosis and management of chronic obstructive pulmonary disease (revised version 2021) [J]. Chin J Tuberc Respir Dis. 2021;44(3):170–205.
2. SORIANO JB, KENDRICK P J, PAULSON K R, et al. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the global burden of Disease Study 2017 [J]. Lancet Respiratory Med. 2020;8(6):585–96.
3. CHRISTENSON S A, SMITH B M BAFADHELM, et al. Chronic obstructive pulmonary disease [J]. Lancet. 2022;399(10342):2227–42.
4. HALPIN D M G, CELLI B R, CRINER G J, et al. The GOLD Summit on chronic obstructive pulmonary disease in low- and middle-income countries [J]. Int J Tuberc Lung Dis. 2019;23(11):1131–41.
5. Global Strategy for the Diagnosis. Management, and Prevention of Chronic Obstructive Pulmonary Disease [R], 2020.
6. ZHONG N, WANG C, YAO W, et al. Prevalence of chronic obstructive pulmonary disease in China: a large, population-based survey [J]. Am J Respir Crit Care Med. 2007;176(8):753–60.
7. 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 [J]. Lancet. 2018;391(10131):1706–17.
8. Global regional. National incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of Disease Study 2016 [J]. Lancet. 2017;390(10100):1211–59.
9. CHOI J Y, RHEE CK. Diagnosis and treatment of early chronic obstructive lung disease (COPD) [J]. J Clin Med. 2020;9(11):3426.
10. JONES RC, PRICE D. Opportunities to diagnose chronic obstructive pulmonary disease in routine care in the UK: a retrospective study of a clinical cohort [J]. Lancet Respir Med. 2014;2(4):267–76.
11. MARTINEZ C H, MANNINO D M, JAIMES F A, et al. Undiagnosed obstructive lung disease in the United States. Associated factors and long-term mortality [J]. Ann Am Thorac Soc. 2015;12(12):1788–95.
12. LAMPRECHT B, SORIANO J B, STUDNICKA M, et al. Determinants of underdiagnosis of COPD in national and international surveys [J]. Chest. 2015;148(4):971–85.
13. LIN C H, CHENG S L, CHEN C Z, et al. Current progress of COPD Early detection: key points and novel strategies [J]. Int J Chron Obstruct Pulmon Dis. 2023;18:1511–24.
14. DECRAMER M, CELLI B. Effect of tiotropium on outcomes in patients with moderate chronic obstructive pulmonary disease (UPLIFT): a pre-specified subgroup analysis of a randomised controlled trial [J]. Lancet. 2009;374(9696):1171–8.
15. JENKINS C R, JONES P W, CALVERLEY P M, et al. Efficacy of salmeterol/fluticasone propionate by GOLD stage of chronic obstructive pulmonary disease: analysis from the randomised, placebo-controlled TORCH study [J]. Respir Res. 2009;10(1):59.
16. Ponce MC, Sankari A, Sharma S. Pulmonary function tests. [Updated 2023 Aug 28]. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan.
17. PONCE MC, SANKARI A. SHARMA S. Pulmonary function tests [M]. Treasure Island (FL) ineligible companies.: StatPearls Publishing Copyright © 2023. StatPearls Publishing LLC; 2023.
18. SCHNIEDERS E, UNAL E. WINKLER V, Performance of alternative COPD case-finding tools: a systematic review and meta-analysis [J]. Eur Respir Rev, 2021, 30(160).
19. HANGAARD S, HELLE T, NIELSEN C, et al. Causes of misdiagnosis of chronic obstructive pulmonary disease: a systematic scoping review [J]. Respir Med. 2017;129:63–84.
20. VANJARE N, CHHOWALA S, MADAS S, et al. Use of spirometry among chest physicians and primary care physicians in India [J]. NPJ Prim Care Respir Med. 2016;26:16036.
21. MARTINEZ F J, RACZEK A E, SEIFER F D, et al. Development and initial validation of a self-scored COPD Population Screener Questionnaire (COPD-PS) [J]. COPD. 2008;5(2):85–95.
22. ZHOU Y M, CHEN S Y, TIAN J, et al. Development and validation of a chronic obstructive pulmonary disease screening questionnaire in China [J]. Int J Tuberc Lung Dis. 2013;17(12):1645–51.
23. DEVRIEZE B W, MODI P, GIWA AO. Peak Flow Rate Measurement [M]. StatPearls. Treasure Island (FL) ineligible companies. Disclosure: Pranav Modi declares no relevant financial relationships with ineligible companies. Disclosure: Al Giwa declares no relevant financial relationships with ineligible companies; 2024.
24. HANSEN E F, VESTBO J, PHANARETH K, et al. Peak flow as predictor of overall mortality in asthma and chronic obstructive pulmonary disease [J]. Am J Respir Crit Care Med. 2001;163(3 Pt 1):690–3.
25. SIDDHARTHAN T, POLLARD S L, QUADERI S A, et al. Discriminative accuracy of Chronic Obstructive Pulmonary Disease Screening instruments in 3 low- and Middle-Income Country settings [J]. JAMA. 2022;327(2):151–60.
26. MARTINEZ F J, LEIDY N K MANNINOD, et al. A New Approach for identifying patients with undiagnosed Chronic Obstructive Pulmonary Disease [J]. Am J Respir Crit Care Med. 2017;195(6):748–56.
27. MARTINEZ F J, HAN M K, LOPEZ C, et al. Discriminative accuracy of the CAPTURE Tool for identifying Chronic Obstructive Pulmonary Disease in US Primary Care settings [J]. JAMA. 2023;329(6):490–501.
28. NELSON S B, LAVANGE L M, NIE Y, et al. Questionnaires and pocket spirometers provide an alternative approach for COPD screening in the general population [J]. Chest. 2012;142(2):358–66.
29. Ji C, Xia Y, Dai H, Zhao Z, Liu T, Tong S, Zhang X, Zhao Y. Reference values and related factors for Peak Expiratory Flow in Middle-aged and Elderly Chinese. Front Public Health. 2021;9:706524.
30. COATES J T, DE KONING C. Machine learning-driven critical care decision making [J]. J R Soc Med. 2022;115(6):236–8.
31. SARKER I H. Machine learning: algorithms, real-world applications and research directions [J]. SN Comput Sci. 2021;2(3):160.
32. ALBAHRA S, GORBETT T, ROBERTSON S, et al. Artificial intelligence and machine learning overview in pathology & laboratory medicine: a general review of data preprocessing and basic supervised concepts [J]. Semin Diagn Pathol. 2023;40(2):71–87.
33. ELVAS L B NUNESM. FERREIRA J C, AI-Driven decision support for early detection of cardiac events: unveiling patterns and Predicting Myocardial ischemia [J]. J Pers Med, 2023, 13(9).
34. SANTOSH KC. AI-Driven tools for Coronavirus Outbreak: need of active learning and Cross-population Train/Test models on Multitudinal/Multimodal data [J]. J Med Syst. 2020;44(5):93.
35. KUMAR V V, T R M, IZONIN I, et al. Efficient data preprocessing with Ensemble Machine Learning Technique for the early detection of chronic kidney disease [J]. Appl Sci. 2023;13:2885.
36. SALEHIN I, ISLAM M S, SAHA P, et al. AutoML: a systematic review on automated machine learning with neural architecture search [J]. J Inform Intell. 2024;2(1):52–81.
37. PELLEGRINO D, CASAS-RECASENS S, FANER R, et al. When GETomics meets aging and exercise in COPD [J]. Respir Med. 2023;216:107294.
38. HAN M K, STEENROD A W, BACCI E D, et al. Identifying patients with undiagnosed COPD in primary care settings: insight from screening tools and epidemiologic studies [J]. Chronic Obstr Pulm Dis. 2015;2(2):103–21.
39. WU Y, DU R, FENG J, et al. Deep CNN for COPD identification by Multi-view snapshot integration of 3D airway tree and lung field [J]. Biomed Signal Process Control. 2023;79:104162.
40. WASHKO G R. Diagnostic imaging in COPD [J]. Semin Respir Crit Care Med. 2010;31(3):276–85.
41. PARK J, HOBBS B D, CRAPO JD, et al. Subtyping COPD by using visual and Quantitative CT Imaging Features [J]. Chest. 2020;157(1):47–60.
42. MONDONEDO JR, SATO S, OGUMA T, et al. CT imaging-based low-attenuation Super clusters in three dimensions and the progression of Emphysema [J]. Chest. 2019;155(1):79–87.
43. KUWANO K, MATSUBA K, IKEDA T, et al. The diagnosis of mild emphysema. Correlation of computed tomography and pathology scores [J]. Am Rev Respir Dis. 1990;141(1):169–78.

44. Takahashi M, Fukuoka J, Nitta N, Takazakura R, Nagatani Y, Murakami Y, et al. Imaging of pulmonary emphysema: a pictorial review. Int J Chron Obstruct Pulmon Dis. 2008;3(2):193–204. https://doi.org/10.2147/copd.s2639.

45. JUNG T. VIJ N. Early diagnosis and real-time monitoring of Regional Lung function changes to Prevent Chronic Obstructive Pulmonary Disease progression to severe emphysema [J]. J Clin Med, 2021, 10(24).

46. BIEDERER J, BEER M, HIRSCH W, et al. MRI of the lung (2/3). Why … when … how? [J]. Insights Imaging. 2012;3(4):355–71.

47. SVERZELLATI N, MOLINARI F, PIRRONTI T, et al. New insights on COPD imaging via CT and MRI [J]. Int J Chron Obstruct Pulmon Dis. 2007;2(3):301–12.

48. ALTAN G, KUTLU Y. ALLAHVERDI N. Deep Learning on Computerized Analysis of Chronic Obstructive Pulmonary Disease [J]. IEEE J Biomed Health Inf, 2019.

49. MORILLO D S, LEON JIMENEZ A, MORENO SA. Computer-aided diagnosis of pneumonia in patients with chronic obstructive pulmonary disease [J]. J Am Med Inf Assoc. 2013;20(e1):e111–7.

50. HAIDER N S, SINGH B K, PERIYASAMY R, et al. Respiratory sound based classification of Chronic Obstructive Pulmonary Disease: a risk Stratification Approach in Machine Learning paradigm [J]. J Med Syst. 2019;43(8):255.

51. Pulmonary Function Group, Respiratory Branch of Chinese Pediatric Society of Chinese Medical Association Editorial Board of Chinese Journal of Applied Clinical Pediatrics. Series guidelines for pediatric pulmonary function(part III): impulse oscillometry [J]. Chin J Appl Clin Pediatr, 2016, (11): 821–5.

52. AMARAL JL, LOPES A J, JANSEN JM, et al. Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease [J]. Comput Methods Programs Biomed. 2012;105(3):183–93.

53. LIPWORTH B J JABBALS. What can we learn about COPD from impulse oscillometry? [J]. Respir Med. 2018;139:106–9.

54. KANWADE A, BAIRAGI VK. Classification of COPD and normal lung airways using feature extraction of electromyographic signals [J]. J King Saud Univ - Comput Inform Sci. 2019;31(4):506–13.

55. SIDDIQUI H U R, SALEEM A A, BASHIR I, et al. Respiration-based COPD detection using UWB Radar Incorporation with Machine Learning [J]. Electronics. 2022;11(18):2875.

56. SIDDIQUI H U RAZAA, SALEEM A A et al. An Approach to Detect Chronic Obstructive Pulmonary Disease using UWB Radar-based temporal and spectral features [J]. Diagnostics (Basel), 2023, 13(6).

57. FANG Y, WANG H, WANG L, et al. Diagnosis of COPD based on a knowledge graph and Integrated Model [J]. IEEE Access. 2019;7:46004–13.

58. FENG Y, WANG Y, ZENG C, et al. Artificial Intelligence and Machine Learning in Chronic Airway diseases: Focus on Asthma and Chronic Obstructive Pulmonary Disease [J]. Int J Med Sci. 2021;18(13):2871–89.

59. ORCHARD P, AGAKOVA A, PINNOCK H, et al. Improving prediction of risk of Hospital Admission in Chronic Obstructive Pulmonary Disease: application of machine learning to Telemonitoring Data [J]. J Med Internet Res. 2018;20(9):e263.

60. MAXIM LD, NIEBO R, UTELL M J. Screening tests: a review with examples [J]. Inhal Toxicol. 2014;26(13):811–28.

61. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol. 2022;75(1):25–36. https://doi.org/10.4097/kja.21209.

62. OZENNE B, SUBTIL F, MAUCORT-BOULCH D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases [J]. J Clin Epidemiol. 2015;68(8):855–9.

63. LIN X, LEI Y, CHEN J, et al. A case-finding clinical decision support system to identify subjects with chronic obstructive Pulmonary Disease based on Public Health Data [J]. Tsinghua Sci Technol. 2023;28(3):525–40.

64. WANG X, HE H, XU L, et al. Developing and validating a chronic obstructive pulmonary disease quick screening questionnaire using statistical learning models [J]. Chron Respir Dis. 2022;19:14799731221116585.

65. WANG X, REN H, REN J, et al. Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data [J]. Comput Methods Programs Biomed. 2023;230:107340.

66. MURO S, ISHIDA M, HORIE Y, et al. Machine learning methods for the diagnosis of Chronic Obstructive Pulmonary Disease in healthy subjects: retrospective observational cohort study [J]. JMIR Med Inf. 2021;9(7):e24796.

67. HE Q, LI X. KIM D, Feasibility study of a multi-criteria decision-making based hierarchical model for multi-modality feature and multi-classifier fusion: applications in medical prognosis prediction [J]. Inform Fusion, 2019, 55.

68. CAI G, HUANG F, GAO Y, et al. Artificial intelligence-based models enabling accurate diagnosis of ovarian cancer using laboratory tests in China: a multi-centre, retrospective cohort study [J]. Lancet Digit Health; 2024.

Publisher's note