Harnessing Artificial Intelligence for Accurate Diagnosis and Radiomics Analysis of Combined Pulmonary Fibrosis and Emphysema: Insights from a Multicenter Cohort Study

Shichen Zhang[1,2#], Hongqiu Wang[3#], Haiyun Tang[4#], Xiaoqin Li[5#], Nian-Wei Wu[6#], Bo Li[7], Xin Chen[5], Kaishi Chen[5], Baosong Xie[5]*, Aiyuan Zhou[8]* and Chunheng Mo[1]*

[1]*The Key Laboratory of Birth Defects and Related Diseases of Women and Children of MOE, State Key Laboratory of Biotherapy, West China Second University Hospital, Sichuan University, Chengdu 610041, China*

[2]*School of Engineering, Westlake University, Hangzhou 310030, China*

[3]*Department of Systems Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China*

[4]*Department of Radiology, Xiangya Hospital, Central South University, Changsha, Hunan, China*

[5]*Department of Respiratory Medicine and Critical Care Medicine, Fujian Provincial Hospital, Fuzhou University Affiliated Provincial Hospital, Fuzhou 350001, China*

[6]*Department of Endocrinology and Metabolism, West China Hospital, Sichuan University, Chengdu 610041, China*

[7]*Department of Radiology, West China Hospital, Sichuan University, Chengdu 610041, China*

[8]*Department of Respiratory Medicine, National Key Clinical Specialty, Branch of National Clinical Research Center for Respiratory Disease, Xiangya Hospital, Central South University, Changsha 410008, China*

[#]These authors have contributed equally to this work.

***Corresponding authors**

Chunheng Mo, Ph.D., Associate Professor, West China Second University Hospital, Sichuan University, Chengdu, 610041, China. E-mail: chunhengmo@gmail.com

Aiyuan Zhou, M.D., Xiangya Hospital, Central South University, Changsha 410008, China. E-mail: aiyuanzhou@126.com

Baosong Xie, M.D. Professor, Fujian Provincial Hospital, Fuzhou University, Fuzhou, Fujian, China. E-mail: baosongx@fjmu.edu.cn

Declarations: The authors declare that they have no competing interests.

Funding: This work was supported by the National Natural Science Foundation of China (82200084), the Natural Science Foundation of Sichuan Province (2023NSFSC1456), the Postdoctoral Science Foundation funded project of Sichuan Province (TB2023047), and the Sichuan University postdoctoral interdisciplinary Innovation Fund (0020404153020).

Authors contributions : S.Z., H.W. and C.M. performed the experiments, data analyses, experimental planning, and wrote the paper. H.T., X.L., N.W., B.L., X.C., K.C., and B.X. performed the experiments. C.M. and A.Z. conceived and designed the work. All authors contributed to the article, read, and approved the final manuscript.

Abstract

Combined Pulmonary Fibrosis and Emphysema (CPFE), recognized as a distinct pulmonary syndrome in 2022, is characterized by unique clinical features and pathogenesis that can lead to respiratory failure and death. However, the diagnosis of CPFE presents significant challenges that impede effective treatment. Here, we assembled a multicenter dataset of three-dimensional (3D) computed tomography (CT) images of patients lungs from multiple hospitals across different provinces in China, including Xiangya Hospital, West China Hospital, and Fujian Provincial Hospital. Utilizing this dataset, we developed CPFENet, a deep learning-assisted diagnostic model specifically designed for CPFE patients. This model accurately classifies patients into CPFE, chronic obstructive pulmonary disease (COPD), and pulmonary fibrosis based on 3D CT data, demonstrating diagnostic performance comparable to that of professional radiologists. Furthermore, we extracted radiomic features from the 3D CT images to generate a CPFE score, a robust and efficient metric for characterizing the presence of CPFE. Using this score, we identified significant differences in CPFE scores between genders. To validate our findings, we retrospectively analyzed the gender distribution of patients across the participating hospitals, corroborating the accuracy of our results. Overall, our study represents the first multicenter systematic investigation of CPFE, providing a diagnostic model and clinical indicators that enable accurate classification and characterization of the syndrome. This research offers a valuable direction and benchmark for future studies on CPFE, potentially facilitating the development of targeted therapeutic strategies and improving patient outcomes.

Keywords: CPFE; deep learning; radiomics; multicenter research

Introduction

Combined Pulmonary Fibrosis and Emphysema (CPFE) is a distinct clinical syndrome characterized by the simultaneous presence of pulmonary fibrosis and emphysema, primarily identified through computed tomography (CT). CPFE patients typically exhibit severe dyspnea, hypoxemia, and a high prevalence of comorbidities, including pulmonary hypertension, contributing to significant morbidity and mortality. Despite its clinical importance, CPFE was only recently recognized as a distinct disease entity in 2022[1–7], reflecting its complex nature and the difficulty of delineating it from other lung conditions, such as chronic obstructive pulmonary disease (COPD) and idiopathic pulmonary fibrosis (IPF)[8–16]. This delayed recognition underscores the challenges in understanding CPFEs pathogenesis, clinical course, and effective diagnostic strategies.[2,17]

The diagnosis of CPFE relies heavily on imaging, where emphysema is identified in the upper lobes and fibrosis in the lower lobes of the lungs [18,19]. However, distinguishing CPFE from other diseases with overlapping imaging and clinical features remains a major challenge, even for

experienced radiologists. Current approaches lack standardization, and the variability in radiological interpretation often contributes to misdiagnoses. Furthermore, most existing studies on CPFE are limited to small cohorts or single-center datasets[20–23]. The absence of a systematic framework for classifying and quantifying CPFE further hampers efforts to improve patient outcomes.[24–26]

Advances in artificial intelligence (AI), particularly deep learning (DL), offer unprecedented opportunities to enhance diagnostic accuracy and standardization in medical imaging[27–29]. DL models have demonstrated remarkable success in identifying complex patterns within radiological data, facilitating the diagnosis of lung diseases such as COPD, IPF, and lung cancer. However, no AI-driven approach has been specifically developed for CPFE, a condition that uniquely combines features of both emphysema and fibrosis.

Here, we present CPFENet, the first deep learning model specifically designed to classify CPFE, COPD, and control cases based on 3D chest CT data. This model was developed and validated using a large, multicenter dataset (2836 cases) encompassing patient samples from Xiangya Hospital, West China Hospital, and Fujian Provincial Hospital, making it one of the most comprehensive studies on CPFE to date. CPFENet achieves superior diagnostic performance (83.3% for accuracy) compared to both junior and senior radiologists (around 72% accuracy for senior doctors and around 67% accuracy for junior doctors), demonstrating its potential to address the challenges of CPFE diagnosis in clinical practice. Beyond classification, we further explore the interpretability of CPFENet by integrating its extracted features with radiomics data derived from lung parenchyma. This integration enables the development of a novel clinical metric, the CPFE score, which provides a quantitative assessment of CPFE severity. Importantly, the CPFE score reveals previously unreported clinical insights, such as gender-based differences in disease presentation, offering new avenues for understanding CPFE pathogenesis and progression.

In summary, this study represents the first multicenter, systematic effort to develop an AI-based diagnostic framework for CPFE. By combining deep learning and radiomics, our work provides a robust tool for the classification and characterization of CPFE, addressing critical gaps in its diagnosis and evaluation. The methods and findings presented here establish a foundation for future research in CPFE and set a benchmark for AI-assisted diagnostic tools in complex lung diseases.

Results

1.1 Patient Characteristics

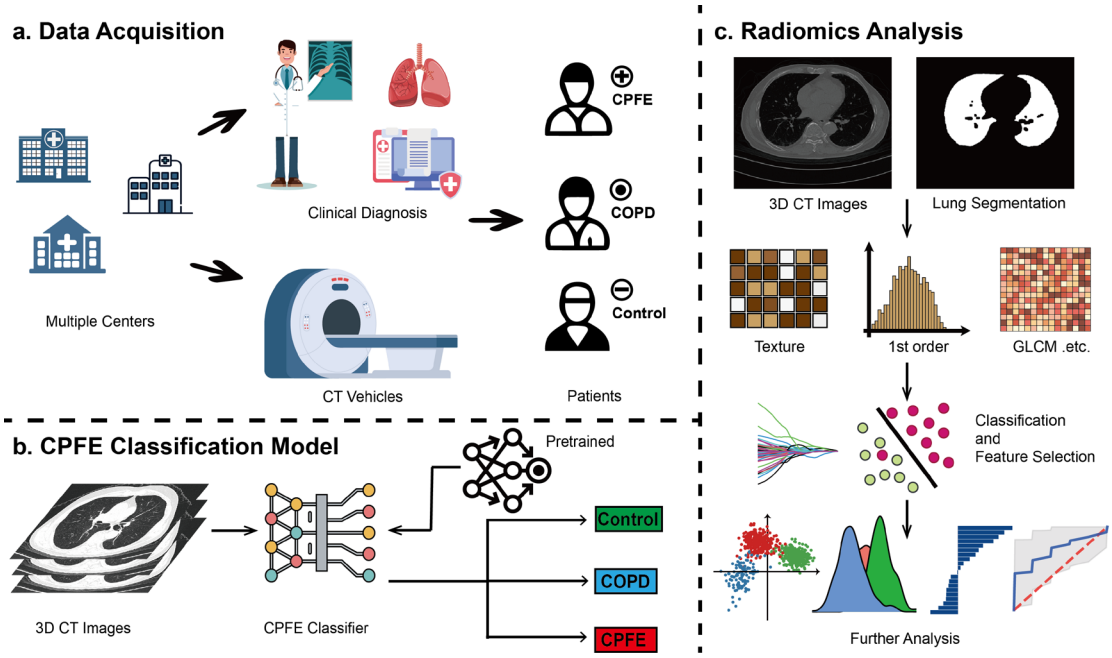


Figure 1: Workflow of the whole study, firstly, in plot a, we acquire our CT data from three clinical centers and we welcome expert doctors to diagnosis the diseases and divided them into three labels including control COPD and CPFE. Next, we design a deep learning model, which is shown in plot b to use artificial intelligence to affiliate the diagnosis. Finally, in plot c, radiomics are employed to construct the index to represent the disease to find relationship between disease and other clinical influencing factors.

Between April 2021 and March 2024, we conducted a multicenter study, collecting imaging data from Xiangya Hospital of Central South University, Fujian Provincial Hospital, and West China Hospital of Sichuan University. Following the exclusion of patients with other lung diseases, such as lung cancer, comorbid thoracic conditions, a history of thoracic surgery, and those with significant image artifacts, a total of 2836 patients were included in the study cohort.

We engaged a panel of expert radiologists and pulmonologists to meticulously review the CT images and other relevant clinical data of these patients, thereby rendering a comprehensive diagnosis and assigning diagnostic labels (as depicted in Figure 1a). Subsequent to the acquisition of the 3D CT dataset for CPFE from the three distinct medical centers, a consensus panel of physicians was convened to establish a standardized diagnostic protocol. This protocol was predicated on a holistic evaluation of patients lung function, blood test results, and imaging data.

Given the relative rarity of CPFE and the lack of a singular diagnostic criterion, we harmonized the diagnostic criteria for both COPD and CPFE. Consequently, a total of 405 CT datasets corresponding to CPFE, 1175 cases of COPD, and 1256 control datasets were meticulously curated for the subsequent analysis (as illustrated in Figure 2). This rigorous data curation process ensured the integrity and relevance of the datasets, thereby facilitating a robust

investigation into the diagnostic nuances of CPFE and COPD.

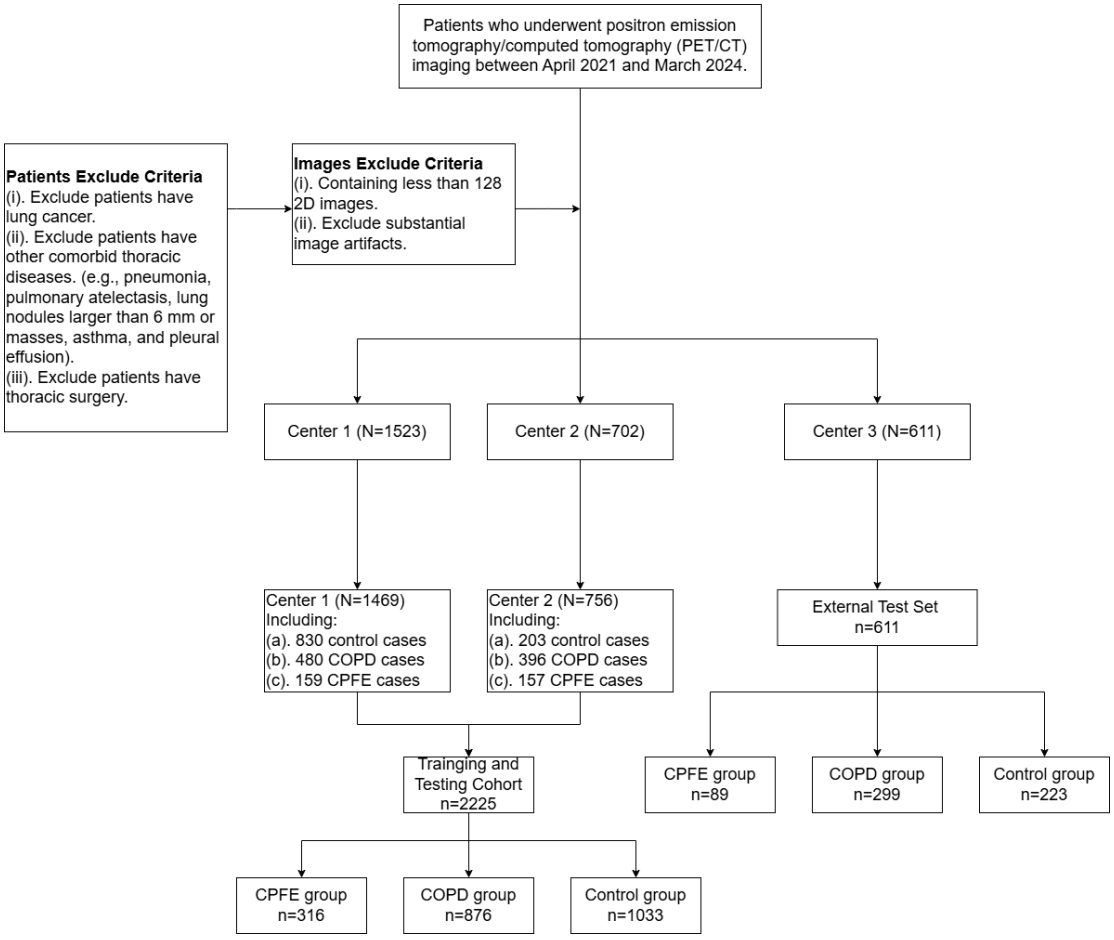


Figure 2: Patients selection flowchart. COPD, Chronic Obstructive Pulmonary disease; CPFE: Combined Pulmonary Fibrosis and Emphysema; PET/CT, Positron Emission Tomography / Computed Tomography

1.2 Auxiliary diagnosis deep learning model-CPFENet

Despite the involvement of experts in assigning accurate labels to the CT data, inherent variabilities in treatment protocols and disease pathogenesis, along with the predominant reliance on physicians subjective interpretation of CT images for hospital-based diagnoses, have given rise to inconsistencies in disease identification. This issue is particularly pronounced in less technologically advanced medical facilities and among less experienced physicians. Furthermore, the paucity of large-scale datasets poses a significant impediment to conducting comprehensive studies on CPFE. Against this backdrop, we endeavored to develop a deep learning model to aid clinicians in the diagnosis of CPFE and to distinguish it from COPD.

Specifically, we devised a deep learning model that harnesses 3D CT data, with a focus on discerning both global and local lesion characteristics pertinent to COPD and CPFE (as

illustrated in Figure 1b). The model, termed CPFENet, capitalizes on the feature extraction prowess of Convolutional Neural Networks (CNNs)[30] and incorporates a self-attention mechanism to meticulously scrutinize each pixel region. In its architecture, we utilized a multi-layer Swin Transformer [31] block as the transformer component and a residual cascade to constitute the CNN module. These modules process 3D CT data and amalgamate their extracted features via a sophisticated fusion layer. For the training regimen, we adopted a transfer learning approach, wherein individual model modules were pre-trained prior to fine-tuning their parameters within the context of the fully assembled model (as delineated in Figure 3). This methodological framework was designed to optimize the models performance and generalizability, thereby enhancing its diagnostic efficacy in the clinical setting.

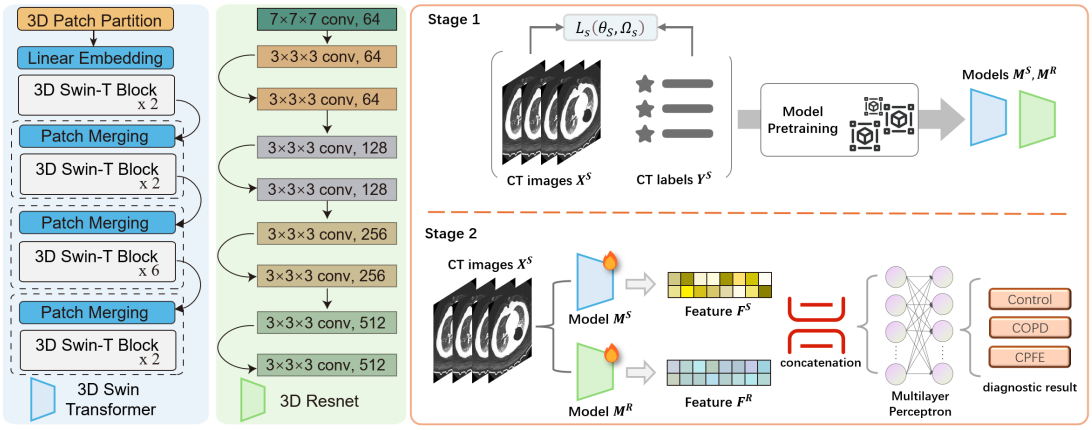


Figure 3: the model overview of the CPFENet. The model employed transferring learning to catch the local and global information by utilizing transformer and CNN module in the 3D CT images and fuse the features to diagnosis.

Following the training of the CPFENet model utilizing the designated training dataset, we sought to evaluate its performance in classifying COPD and CPFE relative to other established deep-learning models. To this end, we constructed a series of benchmark models, encompassing various configurations of 3D ResNet to represent CNN-based architectures and the Swin Transformer to exemplify transformer-based models. Upon analyzing the results derived from both internal and external validation datasets, our CPFENet model demonstrated superior performance across multiple evaluation metrics.

Specifically, for the test dataset, the Swin Transformer model attained an accuracy of approximately 80%, while the ResNet models achieved accuracies ranging from 85% to 87%. In contrast, our CPFENet model achieved an accuracy of 90%. In the external validation dataset, CPFENets performance was further underscored by its AUC and PR values, which surpassed those of the comparator models, reaching 0.94 and 0.9, respectively (as depicted in Figures 4a and 4b, supplementary figure 1a). The accuracy of CPFENet was 0.82, outperforming the 0.78 achieved by ResNet and the 0.71 by Swin-Transformer (as shown in Figure 4c). Collectively,

these findings indicate that our work substantially enhances the overall classification accuracy and diagnostic precision for CPFE when compared to other baseline models.

To ascertain the models ability to discern salient features within the 3D CT images, we implemented a class activation mapping (CAM) process (as illustrated in Figure 4d). The resultant CAM visualizations revealed that our model accurately localized the lesion areas associated with COPD and CPFE, thereby validating its capacity to identify critical diagnostic features within the imaging data. This capability is instrumental in facilitating accurate and timely diagnoses, which is of paramount importance in the clinical management of these pulmonary conditions.

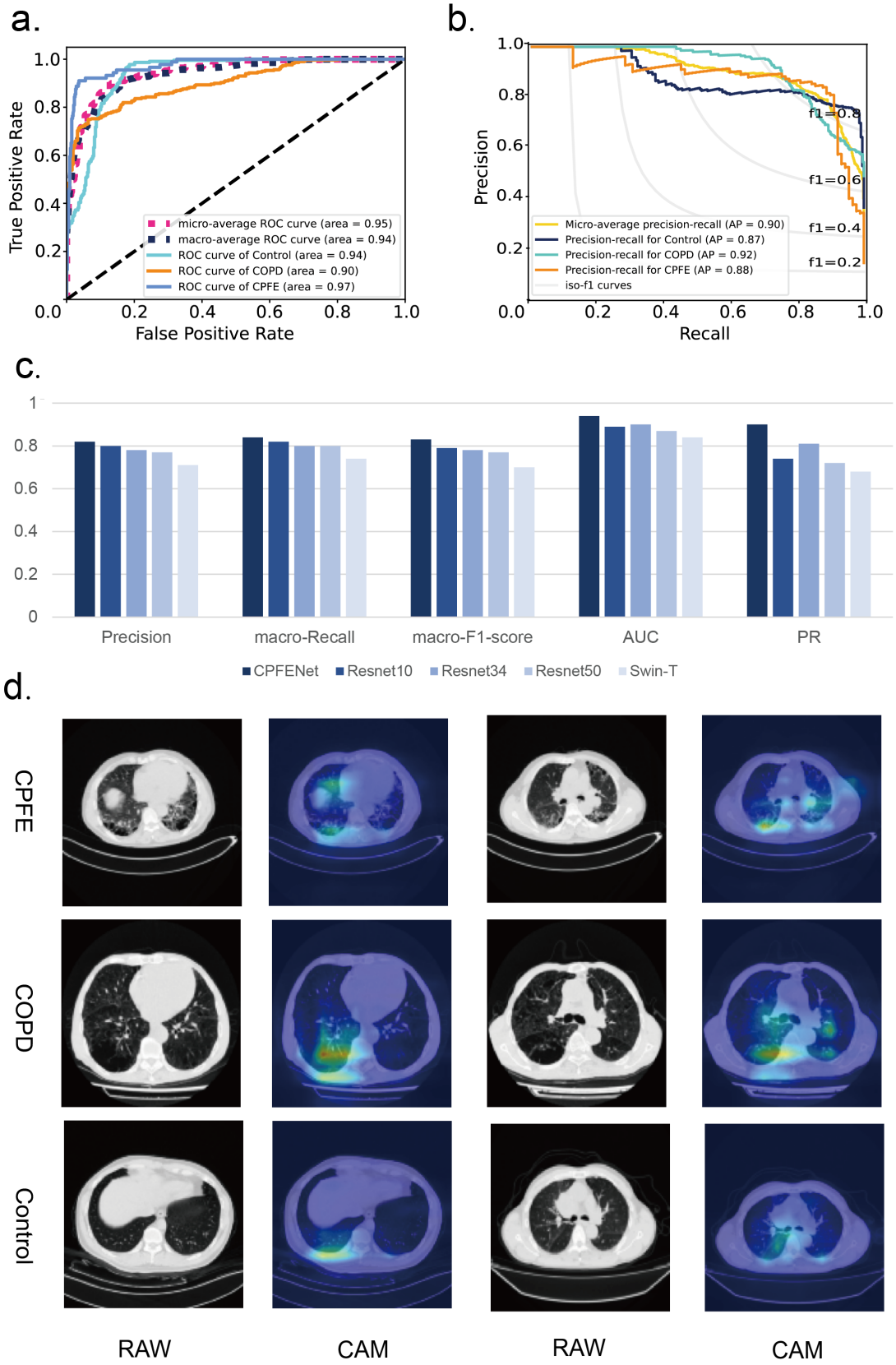


Figure 4: the performance result of the CPFE model. To be specific, the subfigure a and b shows the ROC curve and PR curve of the CPFENet. The subfigure c shows the performance comparing to other baseline models, concrete evaluation indicators include precision, macro-recall, macro-F1 score, AUC

and PR results. The subfigure d shows the visualization CAM plots of the model intermediate layer parameters accompanying with the raw CT images.

To ascertain the clinical viability of our model, we engaged a panel of experts from West China Hospital and Xiangya Hospital to conduct comparative analyses. A random selection of 180 CT images from the external validation cohort was subjected to diagnostic evaluation. Physicians were stratified into two groups: senior (n=2) and junior (n=3) doctors, based on their hierarchical positions and cumulative experience within the imaging department. Our analyses revealed a significant correlation between the success rates of CPFE and COPD diagnoses and the level of expertise of the radiologists. Junior radiologists demonstrated an accuracy rate of approximately 67%, whereas senior radiologists achieved an accuracy rate of 72%. In stark contrast, CPFENet exhibited an accuracy rate of 83.3%, thereby surpassing the diagnostic proficiency of physicians across all experience levels and affirming the models robust stability and precision. The detailed diagnostic outcomes for each physician are presented in Supplementary Table 1, and the confusion matrix delineating the models diagnostic results is depicted in Supplementary Figure 1c.

In summary, CPFENet demonstrates a marked capacity to accurately discern CPFE from 3D CT data, representing a substantial enhancement over traditional physician-based diagnoses in clinical contexts. We posit that CPFENet holds considerable promise for augmenting practical clinical applications and may serve as a transformative catalyst in the evolution of CPFE diagnostics, potentially revolutionizing the approach to disease identification and management in this domain.

1.3 CPFE score acquisition based on radiomics features

Indeed, in conjunction with leveraging 3D CT scans to facilitate deep learning - assisted diagnosis, our objective extends to conducting an exhaustive analysis of the imaging characteristics of CPFE and their correlation with pertinent clinical indicators. Our deep learning model has instilled confidence in the feasibility of diagnosing CPFE through the analysis of 3D CT images, thereby prompting our conviction that the extraction of image - based features can aid in the construction of a comprehensive scoring system for the disease. Consequently, we aspire to quantify the disease progression by employing radiomic features, with the ultimate aim of defining a CPFE - specific clinical metric, herein designated as the CPFE score.

However, acknowledging the interpretability challenges inherent to deep learning models and the non - uniform distribution of features within such data, we have elected to utilize radiomic features for a more robust and interpretable index evaluation. To this end, we harnessed data from the Xiangya Hospital Center, integrating clinical indicators and patient - specific

information to segment lung parenchyma and extract radiomic features. From the 780 3D CT images analyzed, we amassed over 1300 radiomic feature data points.

Due to the noise features and other problems in radiomics, the use of unsupervised clustering methods including PCA, TSNE and UMAP (as shown in supplementary figure 2,3,4) will produce poor results. We utilized Linear Discriminant Analysis (LDA)[32] for visualization, we observed that COPD cases were positioned as an intermediary group between the control and CPFE cohorts within the latent space (as depicted in Figure 5a). This observation aligns with the established medical understanding of disease progression and underscores the capacity of radiomic features to effectively represent the disease spectrum. Such findings not only validate the utility of radiomic features in disease characterization but also highlight their potential as a valuable adjunct in the clinical assessment of CPFE.

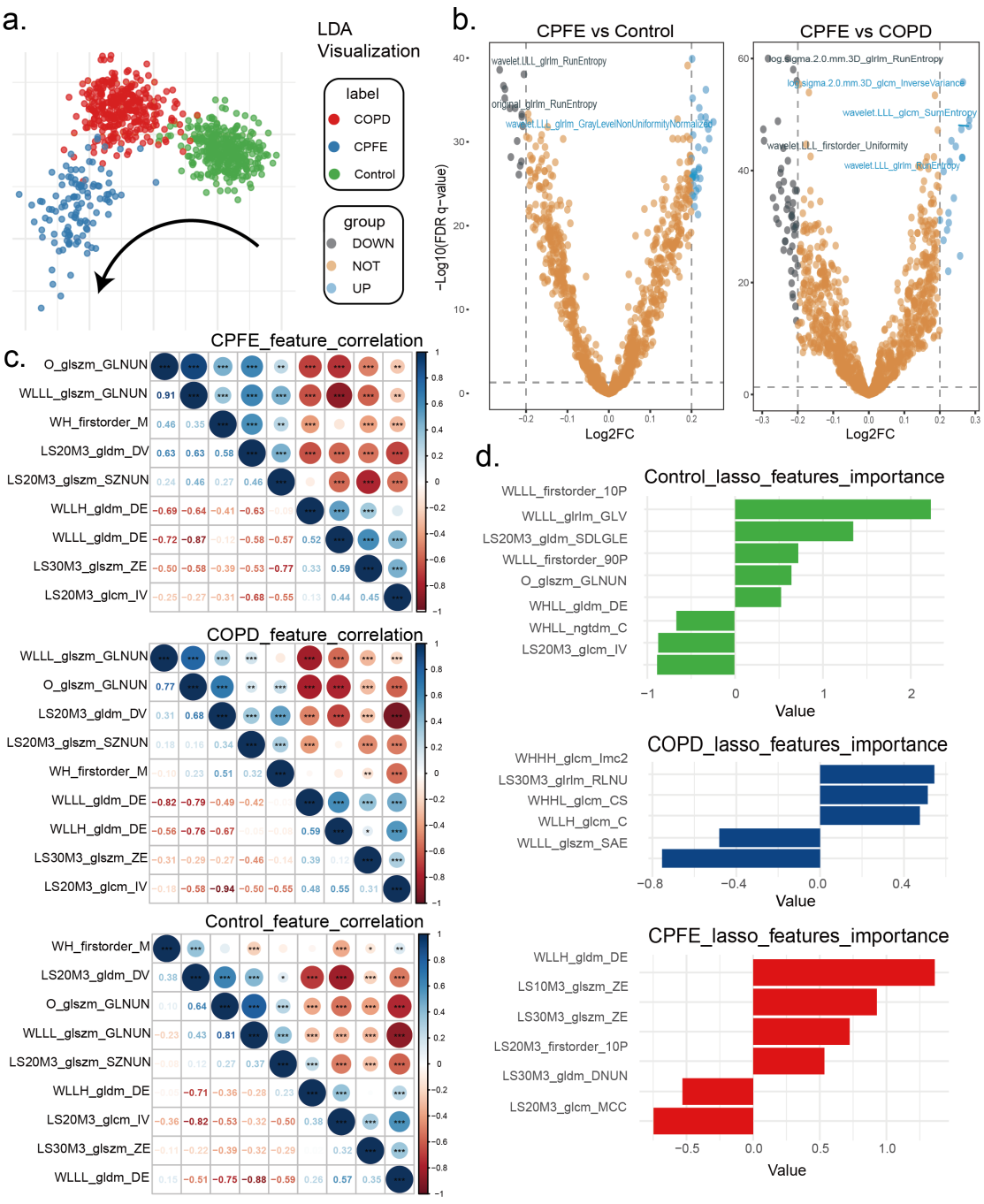


Figure 5: the radiomics analyze result and important radiomics feature selection. Plot a shows the LDA visualization of all radiomics features. Plot b shows the volcano plot of the differentially expression features in CPFE vs Control group and CPFE vs COPD group. Plot c represent the correlation between the selected radiomics features in the CPFE, COPD and Control group. Plot d show the feature importances calculated by the lasso model in the Control, COPD and CPFE groups.

Subsequently, the identification of salient radiomic features became imperative for advancing our research. Following the initial step of screening for differentially expressed features (as illustrated in Figure 5b), we employed Lasso regression for feature selection (as depicted in Figure 5d). This rigorous analytical process led to the identification of nine radiomic features

that were deemed critical for the construction of the CPFE score. A correlation analysis of these features across the three groups revealed distinct correlation patterns, which provided a robust foundation for the definition of the CPFE score (as shown in Figure 5c).

Consequent to the radiomic feature analysis, we selected the Support Vector Machine (SVM)[33] model, which exhibited the highest covariance, as the basis for the CPFE score. This model demonstrated an impressive Area Under the Curve (AUC) of 97% (as presented in Figure 6a), a metric that underscores the high degree of complementarity among the radiomic features and their collective ability to capture the imaging - based commonalities characteristic of CPFE. The contribution graph of the various features to the score further substantiated the necessity and efficacy of employing diverse feature extraction methodologies (as shown in Figure 6b).

Upon examining the distribution of CPFE scores within both the training and test datasets, a consistent pattern emerged wherein the COPD group exhibited higher scores than the control group (as illustrated in Figure 6c). This observation is in concordance with the prevailing hypothesis that CPFE may evolve directly from COPD, thereby lending scientific credence to the validity of the CPFE score as a diagnostic and prognostic tool.

To further unravel the underlying mechanisms of the disease and to enhance our understanding of its progression, we embarked on a comprehensive joint analysis incorporating the CPFE score with other pertinent clinical and radiological indicators. This integrative approach is anticipated to provide deeper insights into the pathophysiology of CPFE, potentially paving the way for more targeted and effective therapeutic interventions.

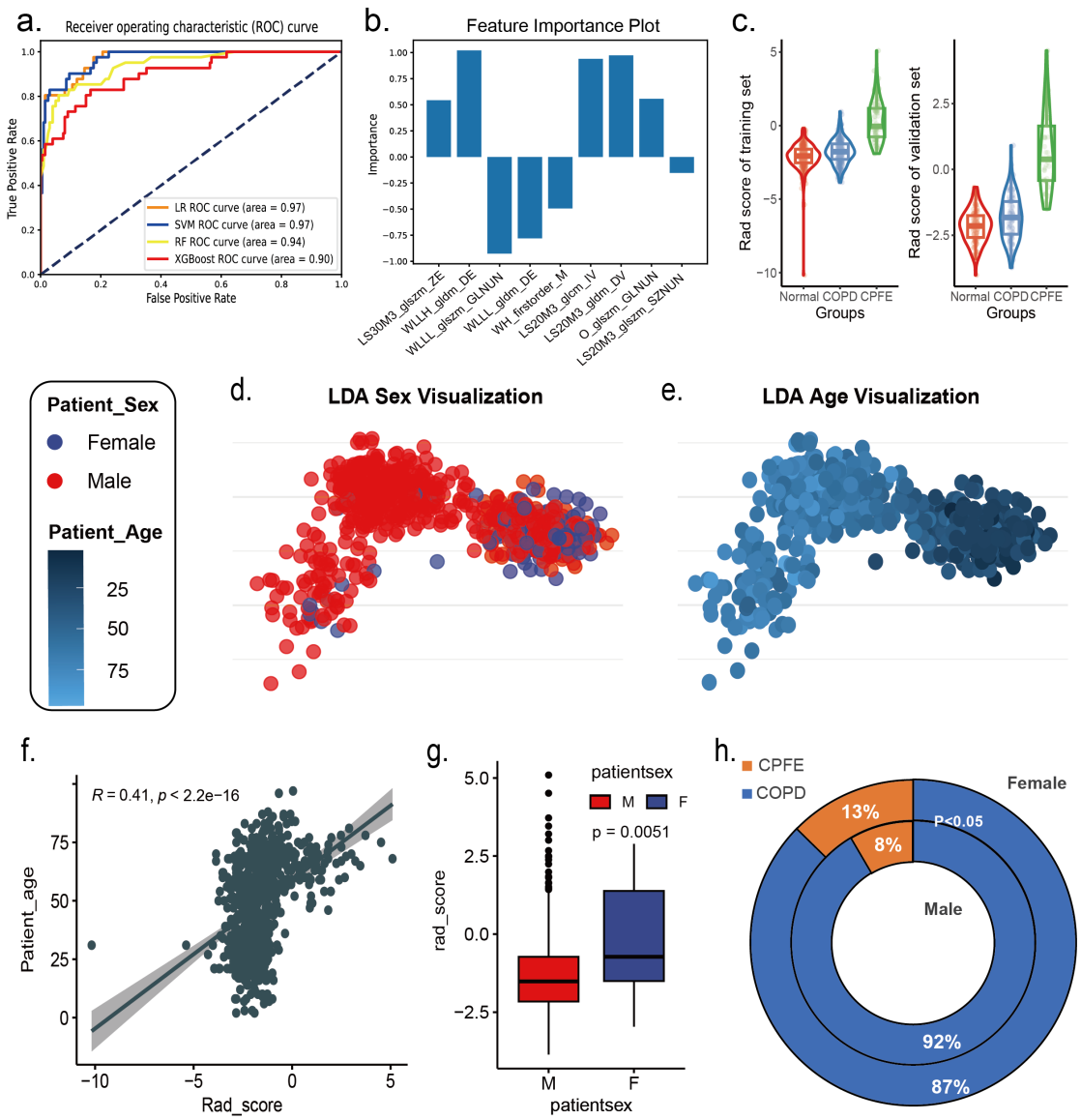


Figure 6: the calculation of the CPFE score and clinical index correlation calculate. Plot a shows the ROC curve of different machine learning models in the selection of the CPFE score. Plot b shows different importance of the radiomics in the calculation of CPFE score. Plot c is the violin plot of the distribution of the calculated CPFE score in the training set (left) and validation set (right). Plot d is the LDA dimensionality reduction plot with the sex label. Plot e is the LDA dimensionality reduction plot with the age label. Plot f is the correlation plot between the age and the CPFE score. Plot g is the boxplot of the CPFE score in different sex group. Plot h is the circle plot of the diagnosis result in the different sex patients of the hospital.

1.4 Clinical indicators related to CPFE disease

The CPFE score, which is derived from radiomic features, has exhibited robust reliability and scientific validity, as evidenced by its high predictive accuracy and consistent distribution across diverse patient groups. In an effort to discern the underlying patterns of CPFE disease,

we conducted an analysis of patient age and gender. Utilizing dimensionality reduction plots (as depicted in Figures 6d and 6e), we observed that different patient cohorts may exhibit discernible patterns in these demographic indicators. Consequently, we are intrigued by the prospect of correlating the CPFE score with clinically relevant parameters to glean novel insights into CPFE from an indicator - based perspective.

Our preliminary analysis centered on the correlation between the CPFE score and patient age, which revealed a moderate correlation with a Pearson correlation coefficient of 0.41 (as shown in Figure 6f). Considering that the prevalence of COPD is notably higher among middle - aged and elderly individuals, and based on comprehensive statistical analyses encompassing dimensionality reduction and T - tests, we did not observe any significant differences between the COPD and CPFE groups ($P > 0.05$). Thus, the correlation with age primarily corroborates the demographic characteristics of COPD in older populations.

However, a gender - specific analysis indicated that CPFE scores were significantly elevated in females diagnosed with COPD and CPFE. This suggests that CPFE may be more prone to development and have a higher incidence rate among women with emphysema, and that gender may exert an influence on the progression of CPFE in females with COPD. To substantiate this hypothesis, we performed a statistical analysis of the sample population and COPD patients at Xiangya Hospital. We randomly selected 1000 patients diagnosed with COPD and CPFE, and the analysis revealed significant gender - based differences in the development of CPFE, with 13% of female emphysema patients and 8% of male emphysema patients being affected. Statistical testing confirmed that there were significant differences in the likelihood of CPFE among affected patients of different genders (as shown in Figure 6h).

This analysis implies that the CPFE score can serve as a reflective metric of the impact of CPFE to a certain extent, and the potential role of gender in the pathogenesis and progression of CPFE merits further investigation. These findings underscore the importance of considering gender as a factor in the clinical assessment and management of CPFE, potentially guiding future research and therapeutic strategies.

Methods

2.1 Data acquisition and preprocessing

This retrospective and multicenter study was approved by the Medical Ethics Committee of Xiangya Hospital of Central South University, the Medical Ethics Committee of Fujian Provincial Hospital, and the Medical Ethics Committee of West China Hospital of Sichuan University, in accordance with the guidelines of the Declaration of Helsinki, with the requirement for informed consent waived. The training and internal test datasets were sourced from patients at Xiangya Hospital of Central South University and Fujian Provincial Hospital,

while the external test dataset comprised patients from West China Hospital of Sichuan University.

Following consultations with radiologists, we considered including 2836 patients who underwent lung CT examinations between April 2021 and March 2024 in this study, based on the cohort data from the three centers.

The imaging manifestations of COPD include emphysema and chronic bronchitis. Emphysema is characterized by low - density areas devoid of visible walls and is categorized into three subtypes: centrilobular, paraseptal, and panlobular. Chronic bronchitis is typified by bronchial wall thickening.

The imaging manifestations of CPFE are characterized by the presence of emphysema in the upper lung zones and fibrosis in the lower lung zones. Interstitial fibrosis is defined by areas of increased parenchymal attenuation, presenting as reticular opacities and/or ground - glass opacities, variable degrees of honeycombing, and/or traction - induced bronchiectasis.

Ethical approval for this study was granted by the institutional review boards of Xiangya Hospital of Central South University, Fujian Provincial Hospital, and West China Hospital of Sichuan University, and the study was conducted in strict adherence to the ethical standards set forth in the 1964 Declaration of Helsinki. Given the retrospective design of the study, the need for informed consent was waived.

For the preprocessing of the 3D CT data, the SimpleITK (sitk) Python package was utilized. The ResampleImageFilter method was employed to uniformly resample the image voxels to a size of [1.0, 1.0, 1.0] mm^3, and the CT image direction was standardized. The window width was set to 1300 Hounsfield units (HU) and the window level to -450 HU. Subsequently, the image grayscale values were normalized, and the image dimensions were uniformly resized to 128 × 128 × 128 pixels to serve as the standardized input for the model.

2.2 CPFENet Model construction

In order to solve the characteristics of CPFE disease such as difficult diagnosis, easy confusion and low frequency, we designed a multi-stage deep learning-assisted diagnosis model CPFENet that combines CNN and Transformer framework using transfer learning knowledge. Due to the difficulty of distinguishing between common COPD diseases and CPFE, we divided the data into three groups: Control, COPD and CPFE, aiming to learn the difference information between COPD and CPFE in the classification of CPFE in the model, so as to achieve accurate segmentation of CPFE. In terms of data, we split the training and test samples according to the ratio of 7:3 according to random sampling to meet the subsequent model training.

In terms of the specific architecture of the model, since the commonly used deep learning

frameworks for medical image processing at this stage are mainly divided into two mainstream models: CNN-based[30] and Transformer-based[31], in order to more effectively capture the local effective features and global features of 3D lung image CT, we designed a classification network CPFENet specifically for CPFE diseases, which considers the characteristics of CPFE diseases in the presence of local alveoli and the presence of fibrotic parts in other lung regions. The model was constructed by combining the capture of local information by CNN and the learning ability of Transformer to global features. Specifically, the preprocessed 3D CT data needs to be passed through two sub-modules with ResNet and swin transformer as the main framework to extract the features of different points of interest, and then fuse the learned features and add them to the linear layer for accurate classification. We chose ResNet34[30] as the framework for the first part feature extraction. For another part, we first performed a 3D patch partition for 3D image segmentation, and then used embedding to encode the position information and put it into four layers of 3D swin-T layers, in which the patch merging module between each layer integrates different position information and inputs it into the swin transformer module, and the number of blocks in different layers is set to 2, 2, 6, 2 respectively. Based on the features of the two modules, the three-classification tasks of CPFE, COPD and Control were carried out by concating and then putting them into the classification layer.

In terms of model training, we first used the parameters pre-trained for ImageNet to train the first stage of the model on the 3D ResNet and 3D Swin-Transformer frameworks to learn different disease features from different perspectives. After training, the model parameters were migrated to the CPFENet model framework for the second stage of tuning to obtain the final accurate classification. In terms of hyperparameter settings, we have a learning rate of 1e-5 for the first and second phases, an epoch set to 200, and a batch size of 32.

2.3 Radiological features extraction

In order to perform radiomics studies on CPFE, we first need to perform lung parenchymal segmentation by 3D CT, here we use the SimpleITK[34–36] tool to segment the lung parenchyma by threshold segmentation combined with seed generation. After excluding the missegmented samples, the relevant lung parenchymal radiomics data were extracted by pyradiomics[37]. We mainly extracted the original features, first-order statistical features and second-order statistical features of 3D CT data, and the second-order statistical features were mainly extracted through log and wavelet filters. In the end, we extracted 1302 radiomics features for subsequent analysis.

2.4 CPFE related radiological characteristics selection

In order to be able to study the imaging features associated with CPFE, we hope to be able to extract CPFE-related radiomics features in two ways. The first method is based on the lasso algorithm, we select the appropriate lambda value for the cross-validation method of the

standardized radiomics data, and use the λ value to perform the lasso regression for the classification task between the control, COPD and CPFE of the sample, and select the part of the regression shrinkage coefficient greater than 0 as the set screened out by the first method, which helps us to select the features with anti-noise ability and easy to calculate. and facilitate the calculation of subsequent CPFE scores. In the second method, the differential expression characteristics of Control vs CPFE and COPD vs CPFE were screened, and the threshold of log foldchange was 0.2 and the pvalue was less than 0.05. In the end, nine relevant traits were screened and applied to the subsequent calculation of the CPFE score.

2.5 CPFE score calculation

In the computation of CPFE scores, we utilized nine selected indicators. To specifically target CPFE, we stratified the radiomics data into CPFE and control groups and employed machine learning algorithms, including linear regression, support vector machine (SVM)[33], random forest[38], and XGBoost[39], for classification. We selected the model with the highest accuracy for the determination of the CPFE score. Initially, we randomly assigned data to training and test sets in a 7:3 ratio and performed a grid search for hyperparameter optimization in each model. During training, we implemented five-fold cross-validation. Ultimately, we selected the SVM as the machine learning model and adopted its output as the CPFE score metric.

2.6 Clinical value assessment

Upon deriving the CPFE scores, we collected the corresponding age and gender data for each score and performed a T-test on these discrete variables to assess significant differences in CPFE scores across patient subgroups. For continuous variables, we computed the Pearson correlation coefficient to identify clinical indicators correlated with the severity of CPFE, thereby informing clinical diagnosis of the disease.

Discussion

CPFE poses a significant challenge in clinical diagnostics, distinguished by its chronicity, poor prognosis, and frequent co - morbidities, such as pulmonary hypertension and lung cancer[40–44]. Since the initial description of CPFE in 1948, the medical community has accrued a more profound understanding of this intricate syndrome[3,24]. This condition is characterized by marginal reductions in lung volume and profound impairments in pulmonary diffusion capacity, with pathological features encompassing emphysematous changes, inflammatory responses, and fibrotic processes[45–47]. Despite these advances, the diagnosis of CPFE remains reliant on the expertise of radiologists, highlighting the urgent need for the development of innovative diagnostic adjuncts. In this study, we utilized deep learning and radiomics to conduct a comprehensive analysis of CPFE from 3D CT data, with the objective of enhancing diagnostic

precision and efficacy.

Within the scope of this investigation, we amalgamated CPFE 3D CT datasets from three disparate centers, thereby establishing a standardized diagnostic protocol and constructing a deep learning model, CPFENet. CPFENet integrates the feature extraction capabilities and self - attention mechanisms of Convolutional Neural Networks (CNNs), in conjunction with multi - layer Swin Transformer blocks and residual cascades, to markedly enhance CPFE diagnostic accuracy through transfer learning and subsequent fine - tuning[48,49]. Our model outperformed established benchmarks, including different types of 3D ResNet and Swin Transformer, achieving a 90% accuracy rate and surpassing the diagnostic outcomes of physicians across all levels. In the external validation set, our model attained an accuracy rate of 83%, compared to the accuracy rate of approximately 67% for junior doctors and 72% for experienced doctors, thereby affirming CPFENets robustness and reliability.

Furthermore, we delineated the CPFE score, a clinical metric specifically tailored to CPFE that delineates disease trajectory. By incorporating radiomics features, we extracted over 1,800 feature data points from 780 3D CT images, culminating in the identification of 9 pivotal features. The SVM model, with its high covariance outcomes, substantiates the synergistic potential of radiomics in capturing CPFEs imaging characteristics. Distributional analysis of CPFE scores revealed elevated COPD scores relative to control groups in both training and testing cohorts, thereby validating the scientific underpinnings of CPFE scores.

Our research also revealed correlations between CPFE and patient age and sex, with a pronounced elevation in CPFE scores among women affected by COPD and CPFE, suggesting that gender may be a determinant in CPFEs progression. These insights offer novel perspectives on CPFEs etiology, genetic influences, therapeutic approaches, and preventative strategies. The identification of sex - based disparities is particularly significant, as it may unravel distinct pathological mechanisms of CPFE across genders, which is crucial for the formulation of targeted therapeutics.

Although constrained by data limitations and an absence of follow - up data, our study lays the groundwork for preliminary inter - populational disparity analyses in CPFE. We aspire to conduct sex - specific sequencing of CPFE biopsy samples to elucidate the role of gender in CPFE and COPD, and to establish comprehensive follow - up datasets to systematically investigate the impact of lifestyle factors, such as smoking[50–52], on CPFE. Additionally, we intend to explore the correlation between CPFE scores and a spectrum of clinical indicators, including pulmonary function test results and blood biomarkers, to achieve a holistic understanding of CPFEs pathophysiology.

In summary, CPFENet and CPFE scores represent novel diagnostic and investigative tools for CPFE, setting the stage for future research endeavors. We are confident that, with technological

progress and data accumulation, CPFENet and CPFE scores will assume increasingly pivotal roles in CPFEs diagnosis, treatment, and prevention. These tools will enable a more nuanced comprehension of CPFEs complexities, inform more precise treatment strategies for patients, and ultimately, enhance patient outcomes and prognoses.

Reference

1. Lin, H. & Jiang, S. Combined pulmonary fibrosis and emphysema (CPFE): An entity different from emphysema or pulmonary fibrosis alone. *J Thorac Dis* (2015) doi:10.3978/j.issn.2072-1439.2015.04.17.

2. Nemoto, M., Koo, C. W., Scanlon, P. D. & Ryu, J. H. Combined Pulmonary Fibrosis and Emphysema: A Narrative Review. *Mayo Clinic Proceedings* Preprint at https://doi.org/10.1016/j.mayocp.2023.05.002 (2023).

3. Cottin, V. *et al.* Syndrome of Combined Pulmonary Fibrosis and Emphysema An Official ATS/ERS/JRS/ALAT Research Statement. *Am J Respir Crit Care Med* (2022) doi:10.1164/rccm.202206-1041ST.

4. Gredic, M. *et al.* Combined Pulmonary Fibrosis and Emphysema: When Scylla and Charybdis Ally. *Cells* Preprint at https://doi.org/10.3390/cells12091278 (2023).

5. Zhou, J. & Jiang, Y. Meta-analysis: Clinical features and treatments of lung cancer in combined pulmonary fibrosis and emphysema. *Sarcoidosis Vasculitis and Diffuse Lung Diseases* Preprint at https://doi.org/10.36141/svdld.v40i4.14433 (2023).

6. Chen, Q., Liu, P., Zhou, H., Kong, H. & Xie, W. An increased risk of lung cancer in combined pulmonary fibrosis and emphysema patients with usual interstitial pneumonia compared with patients with idiopathic pulmonary fibrosis alone: a systematic review and meta-analysis. *Ther Adv Respir Dis* (2021) doi:10.1177/17534666211017050.

7. Nathan, S. D. *et al.* Impact of lung morphology on clinical outcomes with riociguat in patients with pulmonary hypertension and idiopathic interstitial pneumonia: A post hoc subgroup analysis of the RISE-IIP study. *Journal of Heart and Lung Transplantation* (2021) doi:10.1016/j.healun.2021.02.006.

8. Yan, M., Xu, S., Wang, H., Dong, S. & Mo, C. Ferroptosis in chronic obstructive pulmonary disease: From cellular mechanisms to therapeutic applications. *Chinese Medical Journal* Preprint at https://doi.org/10.1097/CM9.0000000000003079 (2024).

9. Zhang, J. *et al.* The role of cGAS-STING signaling in pulmonary fibrosis and its therapeutic potential. *Frontiers in Immunology* Preprint at https://doi.org/10.3389/fimmu.2023.1273248 (2023).

10. Ma, J., Li, G., Wang, H. & Mo, C. Comprehensive review of potential drugs with anti-pulmonary fibrosis properties. *Biomedicine and Pharmacotherapy* Preprint at https://doi.org/10.1016/j.biopha.2024.116282 (2024).

11. Chen, Y. *et al.* Lung regeneration: diverse cell types and the therapeutic potential. *MedComm* Preprint at https://doi.org/10.1002/mco2.494 (2024).

12. Zhang, X. *et al.* Single-cell RNA sequencing analysis of lung cells in COVID-19 patients with diabetes, hypertension, and comorbid diabetes-hypertension. *Front Endocrinol (Lausanne)* (2023) doi:10.3389/fendo.2023.1258646.

13. Mo, C. *et al.* Dopaminylation of endothelial TPI1 suppresses ferroptotic angiocrine signals to promote lung regeneration over fibrosis. *Cell Metab* (2024) doi:10.1016/j.cmet.2024.07.008.

14. Zhang, S. *et al.* Machine learning identified MDK score has prognostic value for idiopathic pulmonary fibrosis based on integrated bulk and single cell expression data. *Front Genet* (2023) doi:10.3389/fgene.2023.1246983.

15. Zhou, A. *et al.* Loss of interferon regulatory factor-1 prevents lung fibrosis by upregulation of pon1 expression. *Respir Res* (2024) doi:10.1186/s12931-024-02987-9.

16. Zantah, M. *et al.* Acute exacerbations of COPD versus IPF in patients with combined pulmonary fibrosis and emphysema. *Respir Res* (2020) doi:10.1186/s12931-020-01432-x.

17. Hage, R., Gautschi, F., Steinack, C. & Schuurmans, M. M. Combined pulmonary fibrosis and emphysema (CPFE) clinical features and management. *International Journal of COPD* Preprint at https://doi.org/10.2147/COPD.S286360 (2021).

18. Made Agustya Darma Putra Wesnawa. Diagnosis and treatment of Combined Pulmonary Fibrosis and Emphysema (CPFE). *World Journal of Advanced Research and Reviews* (2023) doi:10.30574/wjarr.2023.18.3.1147.

19. Wang, J. M., Araki, T., Cottin, V., Han, M. K. & Oldham, J. M. Quantitative Imaging Methods in Combined Pulmonary Fibrosis and Emphysema. *Chest* (2024) doi:10.1016/j.chest.2024.08.007.

20. Schmidt, S. L. *et al.* Pulmonary function measures predict mortality differently in IPF versus combined pulmonary fibrosis and emphysema. *European Respiratory Journal* (2011) doi:10.1183/09031936.00114010.

21. Robledo, G. C., Hernández, M. Y. J., Lucas, S. A. G. & Delgado, F. R. C. Combined Pulmonary Fibrosis and Emphysema With Pulmonary Hypertension: Cases Report. *Current Problems in Cardiology* Preprint at https://doi.org/10.1016/j.cpcardiol.2021.100856 (2022).

22. Chae, K. J. *et al.* Prevalence and progression of combined pulmonary fibrosis and emphysema

in asymptomatic smokers: A case-control study. *Eur Radiol* (2015) doi:10.1007/s00330-015-3617-3.

23. Malli, F. *et al.* Combined pulmonary fibrosis and emphysema characteristics in a Greek cohort. *ERJ Open Res* (2019) doi:10.1183/23120541.00014-2018.

24. Seeger, W. *et al.* Pulmonary hypertension in chronic lung diseases. in *Journal of the American College of Cardiology* (2013). doi:10.1016/j.jacc.2013.10.036.

25. Wang, J. M., Araki, T., Cottin, V., Han, M. K. & Oldham, J. M. Quantitative Imaging Methods in Combined Pulmonary Fibrosis and Emphysema. *Chest* (2024) doi:10.1016/j.chest.2024.08.007.

26. Hage, R., Gautschi, F., Steinack, C. & Schuurmans, M. M. Combined pulmonary fibrosis and emphysema (CPFE) clinical features and management. *International Journal of COPD* Preprint at https://doi.org/10.2147/COPD.S286360 (2021).

27. Wang, H. *et al.* Dual-Reference Source-Free Active Domain Adaptation for Nasopharyngeal Carcinoma Tumor Segmentation across Multiple Hospitals. *IEEE Trans Med Imaging* (2024) doi:10.1109/TMI.2024.3412923.

28. Wang, H. *et al.* Video-Instrument Synergistic Network for Referring Video Instrument Segmentation in Robotic Surgery. *IEEE Trans Med Imaging* (2024) doi:10.1109/TMI.2024.3426953.

29. Jacob, J. *et al.* Likelihood of pulmonary hypertension in patients with idiopathic pulmonary fibrosis and emphysema. *Respirology* (2018) doi:10.1111/resp.13231.

30. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2016). doi:10.1109/CVPR.2016.90.

31. Liu, Z. *et al.* Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. in *Proceedings of the IEEE International Conference on Computer Vision* (2021). doi:10.1109/ICCV48922.2021.00986.

32. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning Deep Features for Discriminative Localization. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2016). doi:10.1109/CVPR.2016.319.

33. Schölkopf, B. SVMs - A practical consequence of learning theory. *IEEE Intelligent Systems and Their Applications* (1998) doi:10.1109/5254.708428.

34. Beare, R., Lowekamp, B. & Yaniv, Z. Image segmentation, registration and characterization in R with simpleITK. *J Stat Softw* (2018) doi:10.18637/jss.v086.i08.

35. Lowekamp, B. C., Chen, D. T., Ibáñez, L. & Blezek, D. The design of simpleITK. *Front Neuroinform* (2013) doi:10.3389/fninf.2013.00045.

36. Yaniv, Z., Lowekamp, B. C., Johnson, H. J. & Beare, R. SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. *J Digit Imaging* (2018) doi:10.1007/s10278-017-0037-8.

37. Van Griethuysen, J. J. M. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Res* (2017) doi:10.1158/0008-5472.CAN-17-0339.

38. Rigatti, S. J. Random Forest. *J Insur Med* (2017) doi:10.17849/insm-47-01-31-39.1.

39. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016). doi:10.1145/2939672.2939785.

40. Champtiaux, N. *et al.* Combined pulmonary fibrosis and emphysema in systemic sclerosis: A syndrome associated with heavy morbidity and mortality. *Semin Arthritis Rheum* (2019) doi:10.1016/j.semarthrit.2018.10.011.

41. Girard, N. *et al.* Lung cancer in combined pulmonary fibrosis and emphysema: A series of 47 western patients. *Journal of Thoracic Oncology* (2014) doi:10.1097/JTO.0000000000000209.

42. Koo, H. J., Do, K. H., Lee, J. B., Alblushi, S. & Lee, S. M. Lung cancer in combined pulmonary fibrosis and emphysema: A systematic review and meta-analysis. *PLoS One* (2016) doi:10.1371/journal.pone.0161437.

43. Ariani, A. *et al.* Overall mortality in combined pulmonary fibrosis and emphysema related to systemic sclerosis. *RMD Open* (2019) doi:10.1136/rmdopen-2018-000820.

44. Cantor, J. The Potential Role of Cigarette Smoke, Elastic Fibers, and Secondary Lung Injury in the Transition of Pulmonary Emphysema to Combined Pulmonary Fibrosis and Emphysema. *Int J Mol Sci* (2024) doi:10.3390/ijms252111793.

45. Zhao, A. *et al.* Mortality surrogates in combined pulmonary fibrosis and emphysema. *European Respiratory Journal* (2024) doi:10.1183/13993003.00127-2023.

46. Jacob, J. *et al.* Functional associations of pleuroparenchymal fibroelastosis and emphysema with hypersensitivity pneumonitis. *Respir Med* (2018) doi:10.1016/j.rmed.2018.03.031.

47. Antoniou, K. M. *et al.* Combined Pulmonary Fibrosis and Emphysema in Scleroderma-Related Lung Disease Has a Major Confounding Effect on Lung Physiology and Screening for Pulmonary Hypertension. *Arthritis and Rheumatology* (2016) doi:10.1002/art.39528.

48. Vaswani, A. *et al.* Attention is all you need. in *Advances in Neural Information Processing Systems* (2017).

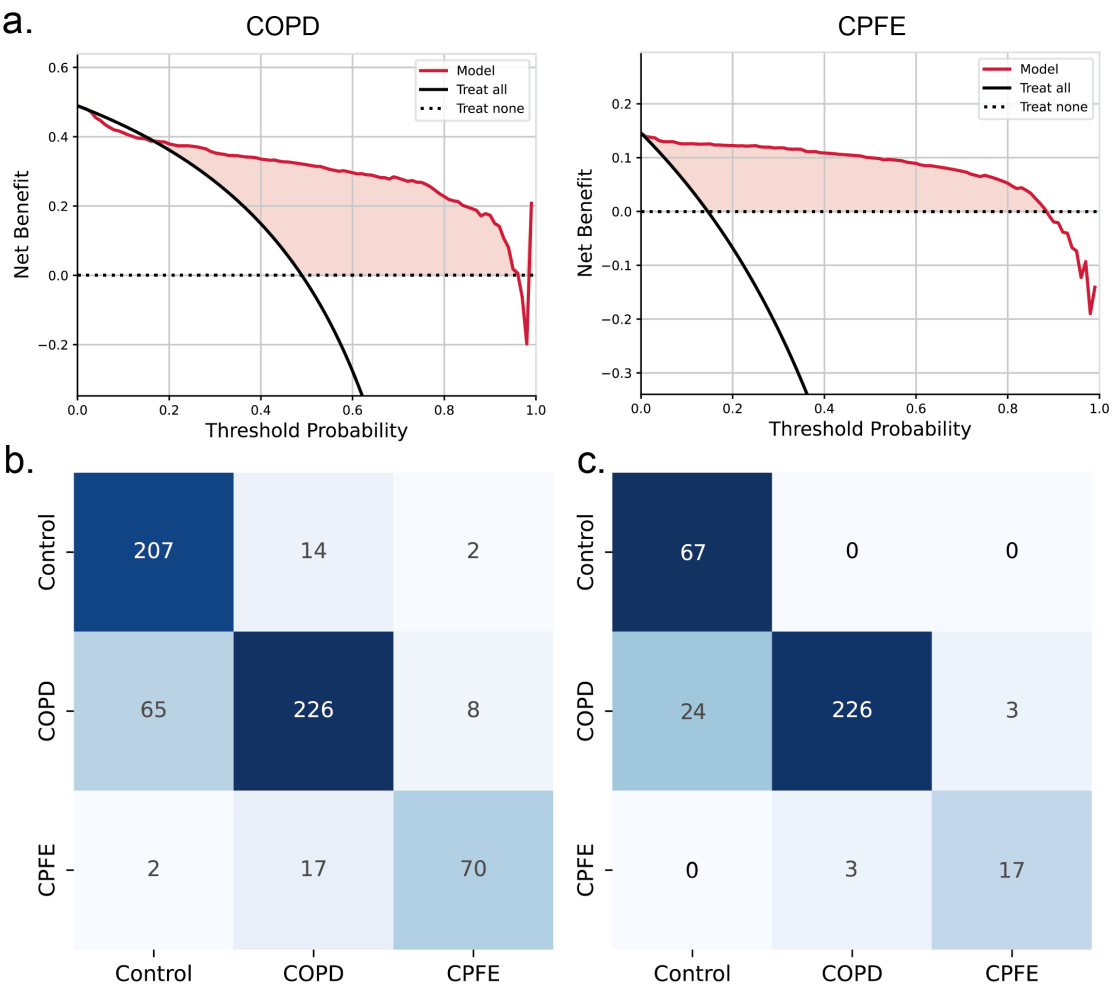
49. Saxena, A. An Introduction to Convolutional Neural Networks. *Int J Res Appl Sci Eng Technol* (2022) doi:10.22214/ijraset.2022.47789.

50. Hagmeyer, L. & Randerath, W. Tabakrauch-assoziierte interstitielle lungenerkrankungen. *Deutsches Arzteblatt International* Preprint at https://doi.org/10.3238/arztebl.2015.0043 (2015).

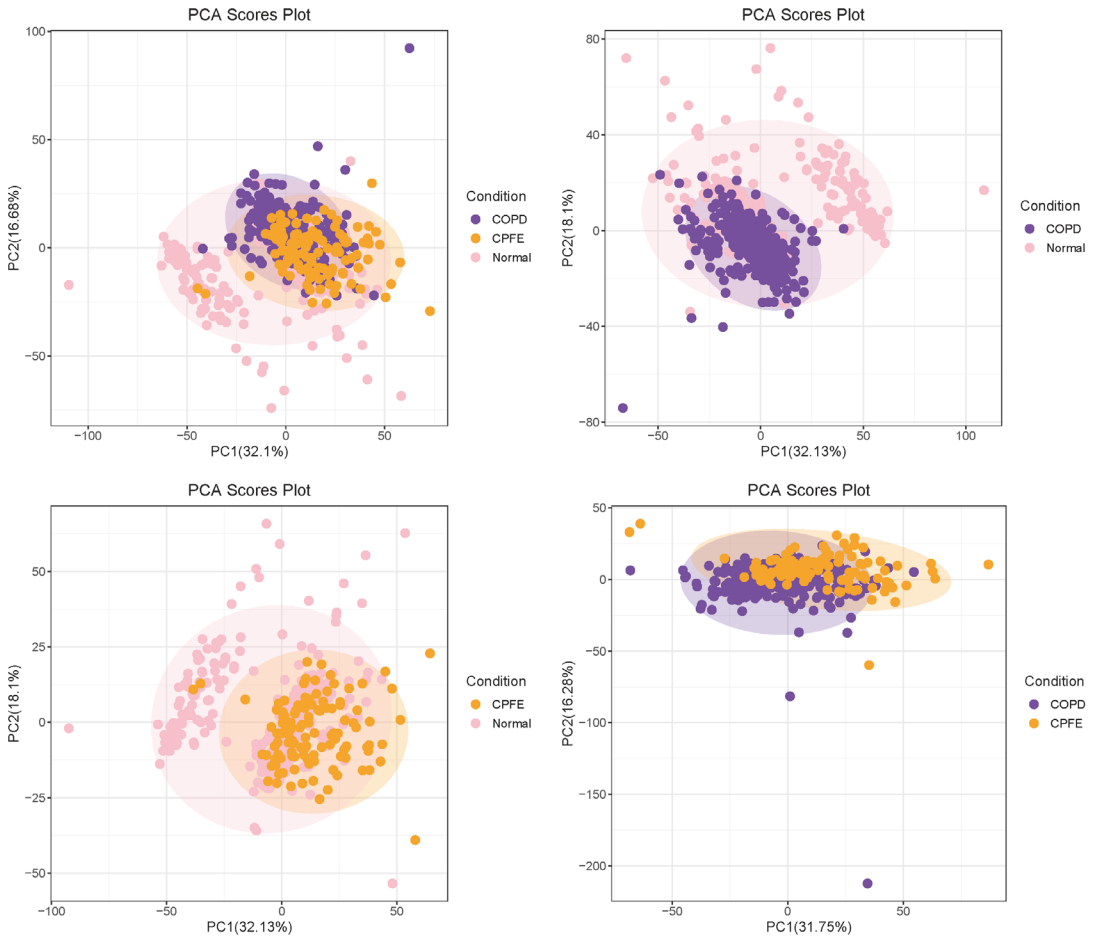
51. I. Papaioannou, A. *et al.* Surfactant Proteins in Smoking-Related Lung Disease. *Curr Top Med Chem* (2015) doi:10.2174/1568026616666150930120640.

52. Samara, K. D., Margaritopoulos, G., Wells, A. U., Siafakas, N. M. & Antoniou, K. M. Smoking and pulmonary fibrosis: novel insights. *Pulm Med* (2011) doi:10.1155/2011/461439.

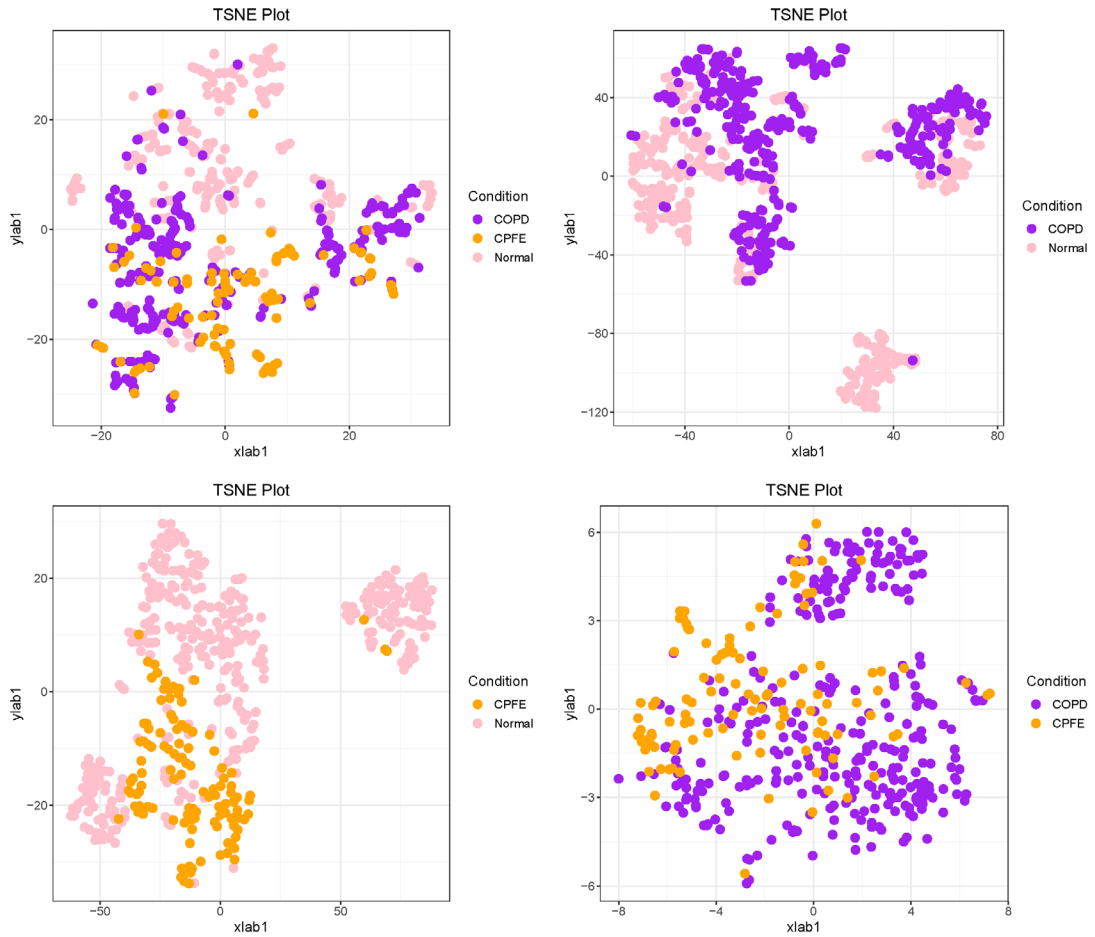
Supplementary Figures



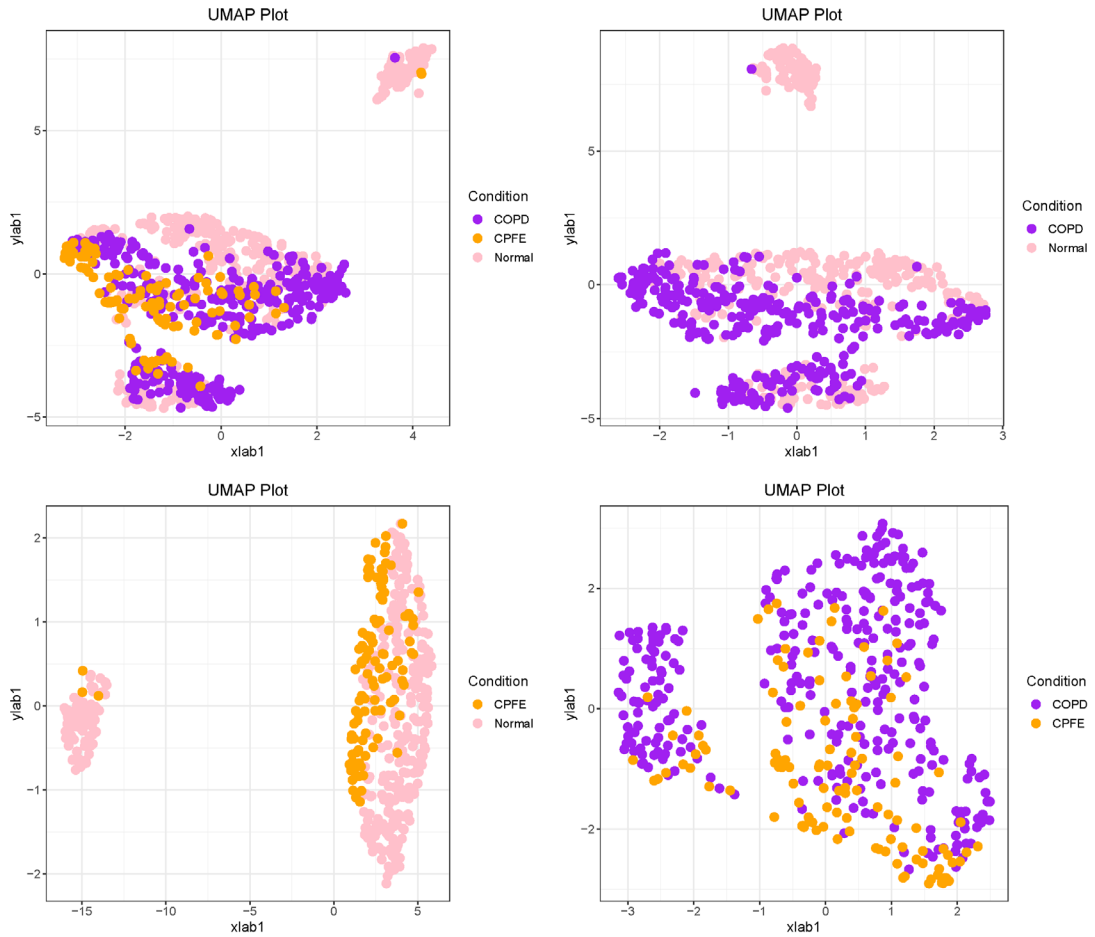
Supplementary figure 1: a: the DCA plot of the CPFENet in COPD and CPFE groups, b: the confusion matrix of the CPFENet for the validation dataset, c: the confusion matirx of the CPFENet for the dataset used into the competetion with doctors.



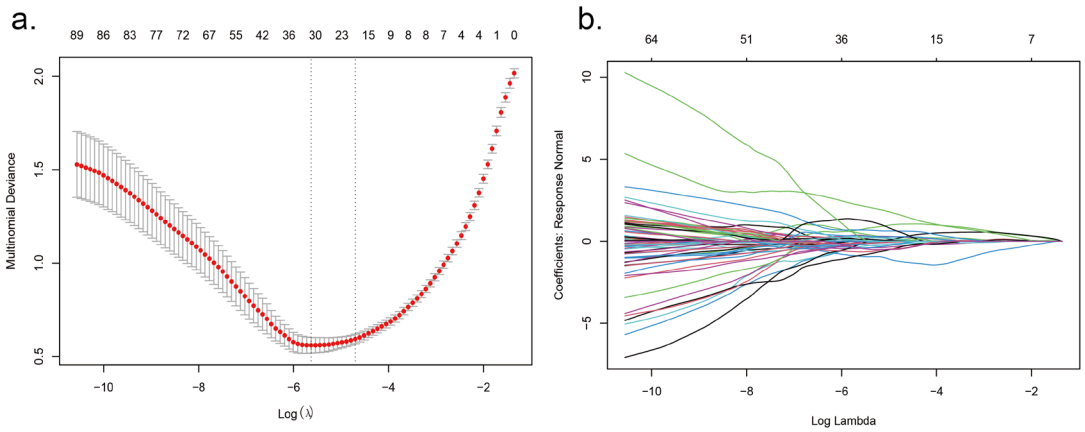
Supplementary figure 2: the PCA plot of different groups with radiomics features.



Supplementary figure 3: the TSNE plot of different groups with radiomics features



Supplementary figure 4: the UMAP plot of different groups with radiomics features.



Supplementary figure 5: a: the lasso multinomial deviance-log(λ) plot of the lasso model selection, b: the lasso regularization path plot of the selected lasso model.