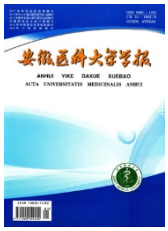

安徽医科大学学报
Acta Universitatis Medicinalis Anhui
ISSN 1000-1492,CN 34-1065/R

《安徽医科大学学报》网络首发论文

题目：基于机器学习和COPD-SQ问卷的COPD风险预测模型构建
作者：陈琳，赵璐娜，周玥，王盼盼，李京坤，张文文，张欣欣，邬超，刘冬
收稿日期：2026-03-22
网络首发日期：2026-05-20
引用格式：陈琳，赵璐娜，周玥，王盼盼，李京坤，张文文，张欣欣，邬超，刘冬．基于机器学习和COPD-SQ问卷的COPD风险预测模型构建[J/OL]．安徽医科大学学报．
https://link.cnki.net/urlid/34.1065.R.20260520.1306.002

基于机器学习和 COPD-SQ 问卷的 COPD 风险预测模型构建

陈　琳[1]，赵璐娜[1]，周　玥[2]，王盼盼[3]，李京坤[2]，张文文[1]，张欣欣[1]，邬　超[1]，刘　冬[1*]

（[1] 石河子大学第一附属医院呼吸与危重症医学科，石河子 832000；[2] 石河子大学临床医学院，石河子 832000；[3] 西南交通大学数学学院，成都 611756）

2026-03-22 接收

基金项目：国家自然科学基金地区基金项目（编号：82560019）；癌症、心脑血管、呼吸和代谢性疾病防治研究国家科技重大专项（编号：2023ZD0506100）；兵团指导科技计划项目（编号：2023ZD019）

作者简介：陈　琳，女，硕士研究生；

刘　冬，男，副主任医师，硕士生导师，通信作者，E-mail：　2322800100@qq.com

摘要　目的 构建并评估多种机器学习模型用于预测个体罹患慢性阻塞性肺疾病（COPD）的风险，为早期筛查和干预提供数据支持。**方法** 选取 823 例研究对象，其中 COPD 高风险组 142 例，低风险组 681 例。收集人口统计学特征、吸烟史、症状（如咳嗽、气短）及慢性阻塞性肺疾病筛查问卷评分等数据。采用 4 种机器学习算法——逻辑回归、随机森林、支持向量机和 XGBoost 构建风险预测模型。采用 5 折交叉验证评估模型性能，评价指标包括准确率、精确率、召回率、F1 分数、受试者工作特征曲线下面积（AUC-ROC）和平均精度（AP）。另，进行了特征重要性分析。**结果** 逻辑回归模型表现出最佳性能（AUC=0.982，AP=0.939），随机森林模型次之（AUC=0.975，AP=0.890）。特征重要性分析显示，吸烟史、呼吸急促症状和体重是关键预测因子。所有模型在识别低风险人群方面均表现出色（精确度>0.93），但在识别高风险人群的能力上存在差异。**结论** 机器学习模型能有效预测 COPD 的高风险人群。逻辑回归模型展现出最优的综合性能，能高效识别 COPD 高危人群，可作为有价值的临床辅助筛查工具。不同模型因其性能特点差异而适用于不同的临床筛查场景，为构建分层、智能化的 COPD 筛查路径提供了具体的决策依据。

关键词 慢性阻塞性肺疾病；风险预测模型；机器学习；逻辑回归；类别不平衡；筛查

中图分类号 R563.9

Construction of a COPD risk prediction model based on machine learning and the COPD-SQ questionnaire

Chen Lin[1], Zhao Luna[1], Zhou Yue[2], Wang Panpan[3], Li Jingkun[2],Zhang Wenwen[1],Zhang Xinxin [1],Wu Chao[1], Liu Dong[1]

[[1]*Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Shihezi University, Shihezi 832000;*[2] *School of Clinical Medicine, Shihezi University, Shihezi 832000;*[3] *School of Mathematics, Southwest Jiaotong University, Chengdu 611756*]

Abstract ***Objective*** This study aims to construct and evaluate various machine learning models for predicting the risk of chronic obstructive pulmonary disease (COPD) in individuals, thereby providing data support for early screening and intervention.***Methods*** A total of 823 subjects were selected for this study, comprising 142 individuals in the high-risk group for COPD and 681 individuals in the low-risk group. Data collected included demographic characteristics, smoking history, symptoms (such as cough and shortness of breath), and scores from the Chronic obstructive pulmonary disease screening questionnaire. Four machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machine, and XGBoost—were utilized to construct risk prediction models. The performance of these models was assessed using 5-fold cross-validation, with evaluation metrics including accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC), and average precision (AP). Furthermore, a feature importance analysis was performed.***Results*** The Logistic Regression model exhibited superior performance, achieving an AUC of 0.982 and an AP of 0.939. This was closely followed by the Random Forest model, which recorded an AUC of 0.975 and an AP of 0.890. Feature importance analysis revealed that smoking history, symptoms of shortness of breath, and body weight were significant predictors. All models demonstrated robust performance in identifying low-risk populations, with precision values exceeding 0.93; however, variations were observed in their efficacy in identifying high-risk populations.***Conclusion*** Machine learning models have proven effective in identifying individuals at high risk for COPD. Among these, the logistic regression model exhibits the best overall performance, efficiently identifying high-risk populations and serving as a valuable clinical auxiliary screening tool. Various models, each with distinct performance characteristics, are suited to different clinical screening scenarios, thereby offering targeted decision-making support for the

establishment of a hierarchical and intelligent COPD screening pathway.

Key words chronic obstructive pulmonary disease ; risk prediction model; machine learning; logistic regression; class imbalance; screening

Fund programs National Natural Science Foundation of China (No. 82560019); National Science and Technology Major Project for Prevention and Treatment of Cancer, Cardiovascular and Cerebrovascular Diseases, Respiratory and Metabolic Diseases (No. 2023ZD0506100); Guiding Science and Technology Plan Project of the Xinjiang Production and Construction Corps (No. 2023ZD019)

Corresponding author Liu Dong,E-mail: 2322800100@qq.com

慢性阻塞性肺疾病(**c**hronic obstructive pulmonary disease,COPD)是一种常见的、以持续气流受限为特征的疾病，是第 4 位全球主要死因的疾病[1]，早期筛查和干预对改善预后至关重要。目前，基于问卷[如慢性阻塞性肺疾病筛查问卷（chronic obstructive pulmonary disease screening questionnaire,COPD-SQ）]的初筛工具虽临床应用简便，但其主观性强、灵敏度有限，难以满足精准风险评估的需求。

近年来，机器学习在医疗预测领域展现出巨大潜力。斯坦福大学等利用三维深度学习使 CT 影像用于肺部结节与肺癌的检测[2]。研究[3]表明，机器学习可以作为深度挖掘 CIN 影响因素的工具。另有研究[4]证明了基于 CT 影像组学的机器学习模型对肺部 pGGN 浸润性的预测具有较高的效能,对患者的治疗方案具有一定的指导价值。为此，该研究旨在整合 COPD-SQ 问卷数据、人口学及生活方式等多维特征，系统构建并比较多种机器学习预测模型，以探索一种性能更优的慢阻肺风险预测工具，为实现早期、智能化的筛查提供新方法。

1 材料与方法

1.1 一般资料

选取石河子大学第一附属医院 2021 年 10 月—2025 年 2 月期间的 823 例受试者作为研究对象，其中 COPD 高风险组 142 例（17.3%），低风险组 681 例（82.7%）。本研究经石河子大学第一附属医院医学伦理委员会审批（伦理审查号：KJ2025-122-01）。

受试者纳入标准：① 完成 COPD-SQ 调查；② 完成肺功能检查；③ 对研究知情并自愿签署知情同意书。高风险组纳入标准：COPD-SQ 总分≥16.5 分。低风险组纳入标准：COPD-SQ 总分<16.5 分。排除标准：① 伴有支气管哮喘、心肌梗死、心力衰竭等其他呼吸系统或心血管

系统疾病者；② 伴有心、肝、肾等脏器功能不全及肿瘤患者；③ 不能理解、配合者；④ 近 3 个月内接受过胸部、腹部及眼科手术；⑤ 近 1 个月内因心脏病住院治疗；⑥ 胸部影像学检查提示肺结核、支气管扩张等因素所致的喘息患者；⑦ 视网膜剥离病史；⑧ 支气管扩张剂过敏；⑨ 精神疾患或认知障碍；⑩ 高位截瘫或可观察到的胸廓畸形；⑪中晚期妊娠患者；⑫不能配合完成 COPD-SQ 调查和肺功能检查，严重神经功能损伤或其他原因无法配合调查者。

1.2 样本量计算

本研究采用 Riley et al[5]提出的临床预测模型样本量计算方法，并使用 Pmsampsize 包在 R 软件(版本 4.3.3)中进行估算。该方法基于四步法，通过同时满足以下 4 个条件确定最小样本量：① 精确估计目标人群的总体结局比例（边际误差 δ≤0.05）；② 控制模型在目标人群中的平均绝对预测误差（mean absolute prediction error, MAPE）；③ 限制过拟合所致预测效应的收缩程度（收缩因子≥0.90）；④ 控制模型拟合的乐观度≤0.05。最终样本量取四步计算结果的最大值。本研究结局为二分类变量，预期高风险人群比例为 15%～20%。保守设定模型的 Cox-Snell R^2（R^2cs）为 max(R^2cs)的 15%。结局比例为 15%时，max(R^2cs)≈0.555，故 R^2cs≈0.083。鉴于本研究对预测精度要求较高，设定 MAPE≤0.04。使用 pmsampsize 包计算，各步骤所需样本量如下：步骤 1（总体比例精确估计）196 例，步骤 2（MAPE≤0.04）772 例，步骤 3（收缩因子≥0.90）203 例，步骤 4（乐观度≤0.05）222 例。取最大值，本研究至少需要 772 例总样本。本研究实际纳入 823 例受试者，包括高风险样本 142 例，基本满足上述样本量要求。

1.3 方法

收集所有受试者的性别、年龄、吸烟史、体质量、身高、身体质量指数（body mass index，BMI）、COPD-SQ 总分及高风险标识等基线资料，同时采用 COPD-SQ 问卷进行调查，记录受试者体质量、身高，并计算 BMI。数据预处理包括数据清洗、特征工程和数值变量标准化处理。采用特定的机器学习算法构建模型，并使用 5 折分层交叉验证评估模型性能，以准确率、精确率、召回率、F1 分数、受试者工作特征曲线下面积和平均精度（average precision，AP）等指标进行评估，同时进行可视化分析。技术路线见图 1。

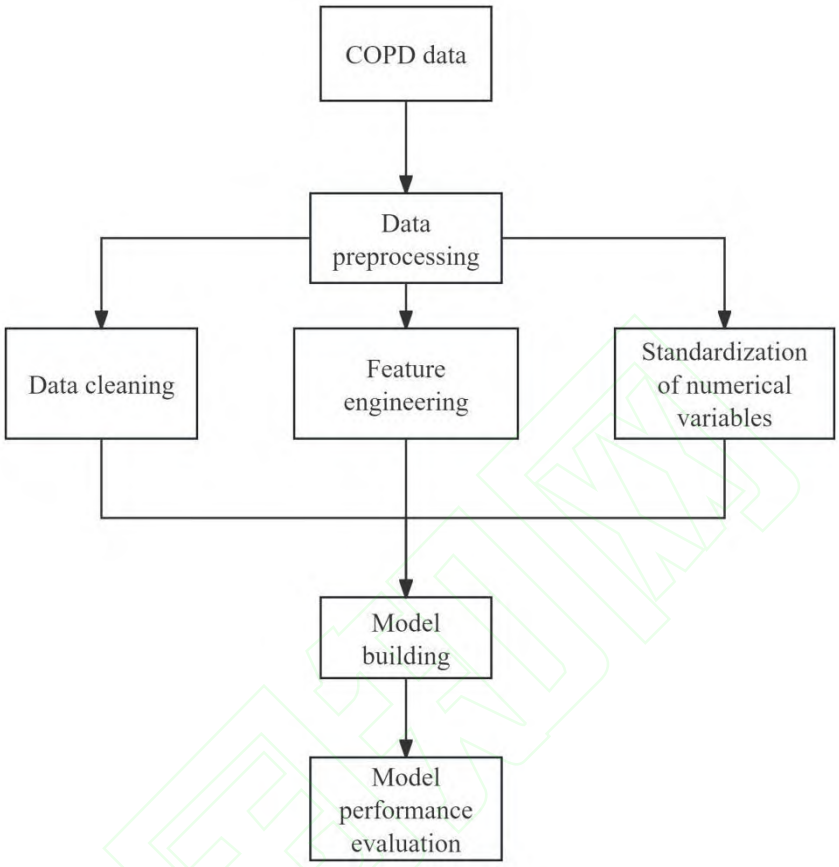

图 1 技术路线图

Fig. 1 Technology roadmap

1.3.1 COPD-SQ 调查

COPD-SQ 由 Maples et al[6]研制，用于帮助研究人员和临床医生确定患者对 COPD 知识的了解程度。该问卷包括 13 项，其中第 1、2、4、6、8、9、10、11 题为正确，第 3、5、7、12、13 题为错误。 COPD-SQ 的弗莱士易读指数分数为 74.7（相当于五年级阅读水平）。

1.3.2 观察指标及评价标准

① 比较高风险组与低风险组受试者的基线资料；② 比较两组受试者的 COPD-SQ 评分和相关特征水平，COPD-SQ 评分为综合风险评分，分值范围 0～28 分，其中 COPD-SQ 评分＜16.5 分为低风险，≥16.5 分为高风险；③ 采用特征重要性分析，分析各特征与 COPD 风险的关系；④ 以 COPD-SQ 评分≥16.5 分作为 COPD 高风险诊断标准，绘制受试者工作特征曲线（receiver

operating characteristic curve，ROC）确定 COPD-SQ 评分及其他特征诊断 COPD 高风险的敏感度、特异度及曲线下面积（area under curve，AUC）。

1.4 统计学处理

1.4.1 统计学分析

采用 SPSS 26.0 统计学软件进行数据处理，对计量资料进行正态性检验，符合正态分布的计量资料用 $\bar{x} \pm s$ 表示，采用 t 检验；计数资料用 n(%)表示，采用 χ^2 检验。采用 Pearson 相关性分析数据间相关性，同时绘制 ROC 曲线评估诊断效能。P<0.05 为差异有统计学意义。

1.4.2 机器学习建模与分析

机器学习模型构建及评估使用 Python3.9.15 及 scikit-learn 库进行，采用 5 折分层交叉验证评估模型性能。为验证模型稳健性，额外采用 3 次重复 5 折交叉验证（即 15 次独立验证），计算性能指标的 $\bar{x} \pm s$，评估模型结果的稳定性。

1.4.3 机器学习算法及超参数优化

1.4.3.1 逻辑回归

逻辑回归是一种基于概率的线性分类模型，通过 Sigmoid 函数将线性回归输出映射至[0,1]区间，实现二分类预测。本研究中，为应对类别不平衡问题（高风险样本占比 14.9%），采用类别权重调整策略（class_weight='balanced'），自动根据样本分布设置权重（低风险样本权重=总样本数/(2×低风险样本数)，高风险样本权重=总样本数/(2×高风险样本数)），降低漏诊风险；同时引入 L2 正则化（penalty='l2'）缓解过拟合，正则化强度参数 C 通过网格搜索优化，搜索范围为[0.01,0.1,1,10,100]；优化器采用拟牛顿法（solver='lbfgs'），迭代次数设置为 1 000 以确保收敛[7]。

1.4.3.2 随机森林

随机森林是集成学习中的 Bagging 算法，通过 Bootstrap 采样构建多个决策树，最终以投票法或平均法输出结果，具有抗过拟合、鲁棒性强的特点。本研究中，关键超参数通过网格搜索优化：决策树数量（n_estimators）搜索范围为[100,200,300,400]，最大树深度（max_depth）搜索范围为[5,10,15,None]，每个节点分裂时考虑的最大特征数（max_features）设置为'auto'（即 sqrt(n_features)），最小样本分裂数（min_samples_split）为[2,5,10]，最小样本叶节点数（min_samples_leaf）为[1,2,4]；为处理类别不平衡，采用 SMOTE 采样（合成少数类过采样技

术）与随机森林结合，在每次 Bootstrap 采样后，对训练集少数类（高风险样本）进行 SMOTE 采样，使训练集类别分布平衡（1:1），提升模型对高风险人群的识别能力[8]。

1.4.3.3 支持向量机（support vector machine ，SVM)

SVM 通过在特征空间中寻找最优超平面实现分类，适用于高维数据，通过核函数可处理非线性问题。本研究中，首先通过网格搜索选择最优核函数，候选核函数包括线性核（kernel='linear'）、多项式核（kernel='poly'）和径向基核（kernel='rbf'）；针对最优核函数（本研究最终确定为径向基核），进一步优化关键参数：惩罚系数 C（搜索范围[0.1,1,10,100]）、核函数参数 gamma（搜索范围[0.001,0.01,0.1,1,10]）；类别不平衡处理采用加权 SVM 策略（class_weight='balanced'），同时引入概率输出（probability=True），通过 Platt 缩放将分类结果转换为风险概率，便于临床应用；为提升训练效率，设置缓存大小（cache_size）为 200MB，迭代次数（max_iter）为-1（无限制，直至收敛）[9]。

1.4.3.4 XGBoost

XGBoost 是基于梯度提升决策树的集成算法，通过梯度下降优化损失函数，引入正则化项与稀疏感知算法，具有训练速度快、预测精度高的优势。本研究中，超参数优化采用网格搜索结合 5 折交叉验证：学习率（learning_rate）搜索范围为[0.01,0.05,0.1,0.2]，决策树数量（n_estimators）为[100,200,300]，最大树深度（max_depth）为[3,5,7,9]，最小样本权重和（min_child_weight）为[1,2,3]，正则化参数 lambda（L2 正则）为[0,1,5,10]、alpha（L1 正则）为[0,0.1,1]，子样本比例（subsample）为[0.7,0.8,0.9,1.0]，列采样比例（colsample_bytree）为[0.7,0.8,0.9,1.0]；类别不平衡处理采用 scale_pos_weight 参数（设置为低风险样本数/高风险样本数≈5.7），自动调整正负样本的损失权重；缺失值处理采用 XGBoost 内置策略，自动学习缺失值的分裂方向，无需额外填充篇[10]。

1.4.4 超参数优化流程

所有模型的超参数优化均采用网格搜索实现，以 AUC-ROC 为优化目标函数，5 折分层交叉验证为评估方式，通过遍历预设参数组合，筛选出最优参数组合用于最终模型训练。优化流程如下：① 划分训练集与验证集（基于 5 折分层交叉验证框架）；② 对每个参数组合，在训练集上训练模型；③ 在验证集上计算 AUC-ROC；④选择 AUC-ROC 最高的参数组合作为最优参数；⑤ 使用最优参数在全量训练数据上训练最终模型。

2 结果

2.1 患者人口学特征情况

所调查的 823 份有效问卷中，患者的平均年龄（58.3±9.7）岁，最小 26 岁，最大 87 岁，72%的受试者年龄≥53 岁，48%集中在 53~65 岁区间，这符合 COPD 高发年龄段特征。

从患者的 BMI 来看，平均 BMI（24.6±3.2）kg/m²，处于正常体质量上限，23.5%受试者 BMI<24.0 kg/m²，属于正常偏低范围，52%处于 24.0 ~ <28.0 kg/m²，属于正常至超重临界，24.5%的受试者 BMI≥28.0 kg/m²，属于超重或肥胖范围。

在生活方式方面，从不吸烟者占 58.9%，现吸烟者占 16.8%，其中重度吸烟（吸烟指数≥400）者比例为 16.8%，呼吸道症状方面，频繁咳嗽（18.5%）相对常见，而气促发生率较低（8.4%）。而在 823 例受试者中，有 9.2%暴露于生物燃料烹饪，仅有 3.8%的受试者有肺部疾病史。根据 COPD-SQ 评分，平均（10.8±5.9）分，其中 17.3%（142 例）的受试者被判定为 COPD 高风险人群。见表 1。

表 1 问卷数据特征

Tab. 1 Characteristics of questionnaire data

Variable	Number of people	Proportion (%)
Age (years)		
<53	230	27.9
53-65	396	48.1
>65	197	23.9
BMI (kg/m²)		
<18.5 (Under weight)	12	1.5
18.5- <24.0 (Normal weight)	181	22.0
24.0- <28.0 (Normal to borderline overweight)	428	52.0
≥28.0 (Overweight/obesity)	202	24.5
Body weight		
mean body weight (kg, $\bar{x} \pm s$)	67.2±11.5	—
Body height		
mean height (m, $\bar{x} \pm s$)	1.64±0.08	—
Smoking history		
Never smoked	485	58.9
Former smoker	200	24.3

Variable	Number of people	Proportion (%)
Current smoker	138	16.8
Heavy smoking (≥20 pack-years)	138	16.8
Exposure to biomass fuel for cooking		
Exposure	76	9.2
Unexposed	747	90.8
Respiratory symptoms		
Frequent cough	152	18.5
Anhelation	69	8.4
Sputum production	98	11.9
Wheezing	45	5.5
History of lung disease		
History of lung disease	31	3.8
No history of lung disease	792	96.2
COPD-SQ score		
Average score	10.8±5.9	—
Median	10	—
High-risk (≥16.5 points)	142	17.3
Low-risk (<16.5 points)	681	82.7

2.2 样本特征

在 COPD-SQ 评分中，所有 823 例样本 16 项特征均完整，无缺失值，总分平均分为(10.8±5.9)分，范围 1~28 分，中位数 10 分，显著低于高风险阈值 16.5 分，且 75%受试者≤15 分。得分的分布右偏（偏度=0.42）,表明存在少量极高风险个体。最终，823 例样本中 17.3%（142/823）达到高风险标准，验证了样本的类别不平衡问题（高风险:低风险≈1:4.8）。

样本分布特性与 COPD 临床特征高度吻合：连续变量（年龄、BMI 等）基本呈正态分布，分类变量（如气促，发生率 8.4%）呈偏态分布，符合临床罕见症状特征。这种类别不平衡性（1:4.8）对模型识别高风险个体构成挑战，需通过阈值优化策略提升召回率。

本研究 823 例样本以中老年（72%≥53 岁）、超重/肥胖（24.5% BMI≥28.0 kg/m²）、较高吸烟暴露率（16.8%重度吸烟）人群为主，契合慢阻肺典型高危人群特征[11]。

2.3 模型性能比较

2.3.1 ROC 曲线分析

4 种模型 AUC 均>0.97，且 ROC 曲线均十分靠近左上角。其中逻辑回归的 AUC 最高，AUC 值为 0.982，区分能力略优于随机森林、SVM 和 XGBoost 模型。见图 2。

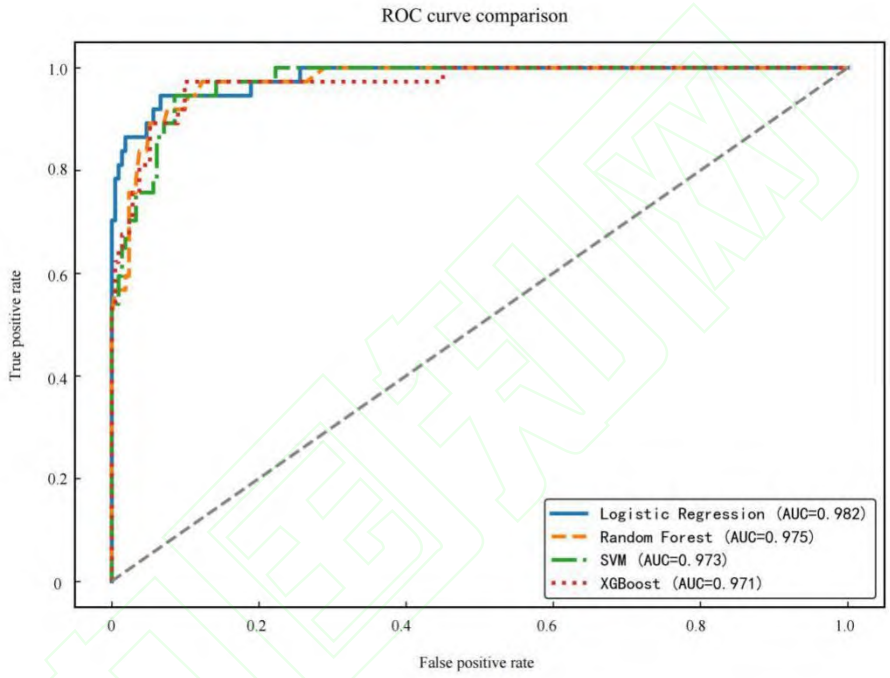

图 2 4 种机器学习模型预测 COPD 风险的 ROC 曲线比较

Fig. 2 Comparison of ROC curves of four machine learning models in predicting the risk of COPD

于 AP 值与 PR 曲线的模型对比分析基 如表 2 所示，逻辑回归 AP 值为 0.939，随机森林的 AP 值为 0.890，SVM 的 AP 值为 0.884，XGBoos 的 AP 值为 0.896。此外，如图 3 所示，不同模型在精确率与召回率之间呈现出明显的权衡关系：逻辑回归和 SVM 的召回率均>90%，但是精确率均<70%，呈现高召回率、低精确率的特点；随机森林虽精确率达 91%，召回率仅为 57%，预测偏保守；XGBoest 则在精确率（77%）与召回率（81%）之间取得较好平衡。见图 3。

表 2 五折交叉验证结果

Tab. 2 Results of five-fold cross-validation

Model	Accuracy	Precision (high-risk)	Recall (high-risk)	F1 score (high-risk)	AUC	AP
Logistic Regression	0.93	0.70	0.95	0.80	0.982	0.939
Random Forest	0.93	0.91	0.57	0.70	0.975	0.890
SVM	0.92	0.67	0.92	0.77	0.973	0.884
XGBoost	0.94	0.77	0.81	0.79	0.971	0.896

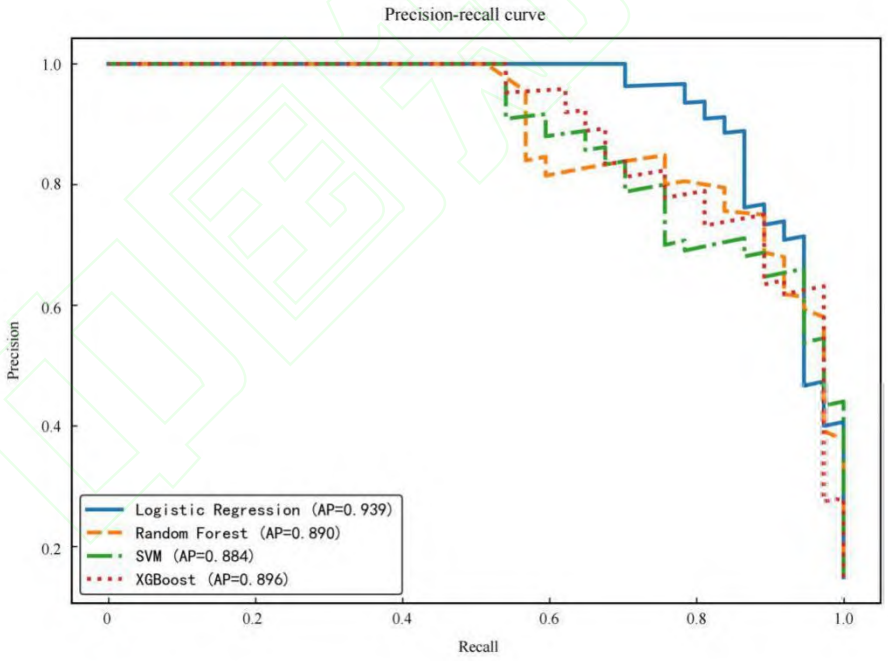

图 3 4 种机器学习模型预测慢阻肺风险的精确率-召回率曲线

Fig. 3 Precision-recall curves of four machine learning models in predicting the risk of chronic obstructive pulmonary disease

2.3.2 基于随机森林的特征重要性排序

基于随机森林模型的特征重要性分析其结果，高风险人群重要性最高的特征为“年龄”，

其后依次为“气促”（0.158）、“经常咳嗽”（0.105）、“不经常咳嗽”（0.100）、“主动吸烟”（0.048）、“体质量”（0.047）、“每日吸烟支数”（0.042）、“身高”（0.039）、“从不吸烟”（0.038）、“BMI”（0.035）、“平地正常行走时感觉气促”（0.030）、“在平地急行或爬小坡时感觉气促”（0.026）、“吸烟总量-少量”（0.025）、“吸烟年数”（0.024）及“吸烟总量-大量”（0.021）。见图 4。

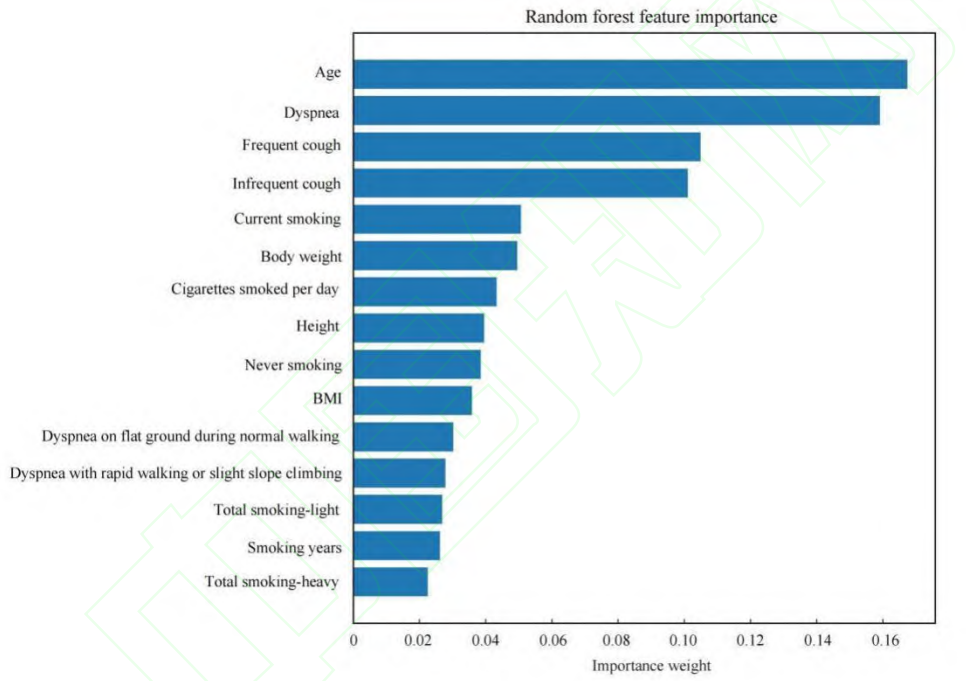

图 4 基于随机森林算法的慢阻肺风险预测特征重要性评估

Fig. 4 Evaluation of the importance of risk prediction features of chronic obstructive pulmonary disease based on the random forest algorithm

2.3.3 阈值优化策略有效提升高风险人群筛查灵敏度

从逻辑回归混淆矩阵图可见，实际值为低风险、预测值为低风险的样本量为 646 例，实际值为低风险、预测值为高风险的样本量为 35 例，实际值为高风险、预测值为低风险的样本量为 7 例，实际值为高风险、预测值为高风险的样本量为 135 例。见图 5。

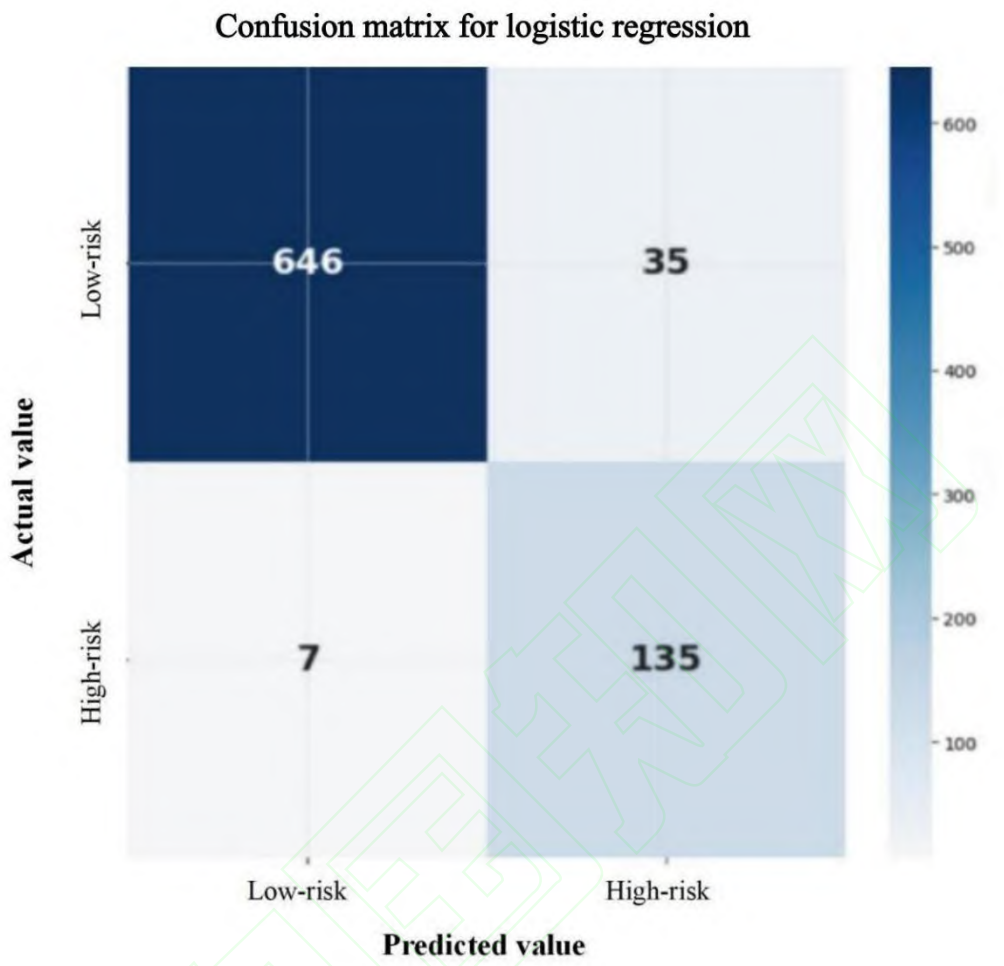

图 5 针对类别不平衡的阈值优化策略在逻辑回归模型中的应用效果

Fig. 5 The application effect of the threshold optimization strategy for class imbalance in the logistic regression model

3 讨论

COPD 作为一种高发的慢性呼吸系统疾病，好发于具有吸烟史的中老年男性群体。其病理进程呈进行性且不可逆，将 COPD 与机器学习相结合，能利用数据驱动的先进方法，从而解决 COPD 在早期诊断、精准分型、病情预测和个性化治疗等方面的传统难题[12]。本研究基于 COPD 筛查问卷数据，系统评估与比较了 4 种机器学习模型在 COPD 风险预测中的性能。

结果表明，4 种模型 AUC 均>0.97，且 ROC 曲线均十分靠近左上角，展现出优异的区分能力[14]。其中，逻辑回归模型的 AP 显著高于其他模型，说明其在保持较高召回率的同时，能够实现更高的精确率，展现出最优的正样本识别能力。特别是在识别高风险人群方面，该模型通过阈值优化后将召回率提升至 95%，显著降低漏诊风险，与临床筛查中对于高敏感性的需求高

度契合，该结论也从 PR 曲线的形态上得到支持：逻辑回归的曲线整体位于其他模型上方，进一步验证其在识别高风险人群任务中的卓越性能，这提示逻辑回归模型可作为基层医疗机构中进行 COPD 初筛的理想工具[13]。

值得注意的是，不同模型在精确率与召回率之间呈现出显著的权衡关系。逻辑回归与 SVM 偏向高召回率，侧重于最大限度发现潜在患者，这种特性使其非常适用于社区或基层医疗机构的初步大规模筛查，其首要目标是最大限度降低漏诊率，宁可误判也需将高危个体纳入后续精查流程；随机森林模型则表现为高精确率，倾向于减少误诊，这有助于在确诊前阶段减少不必要的随访和医疗资源消耗，适用于对筛查阳性人群进行二次验证或风险分层；XGBoost 模型则在两者间取得了较好平衡。这种差异源于模型的内在机制与对数据分布的处理策略，如：逻辑回归可通过调整类别权重以降低漏诊风险，随机森林则因其集成特性而对少数类样本的识别相对谨慎，XGBoost 则通过加权损失函数在一定程度上平衡了类别不平衡的影响。这体现了根据不同临床场景的实际需求选择合适的预测模型的重要性[14]。因此，本研究的结论并非简单地宣布某个模型“最佳”，而是明确了不同模型的适用场景：在强调敏感性的初筛阶段可选用逻辑回归，在强调特异性的确认阶段可考虑随机森林。这种基于性能权衡的、场景化的模型选择策略，是本研究超越简单性能比较、指向临床实践的核心贡献之一。

随机森林模型的特征重要性分析结果增强了模型的可解释性，其识别出的关键预测因子（如吸烟史、气促症状、BMI）与临床认知高度一致，这反映出随机森林模型在特征选择方面具有良好的可解释性与实际意义。而生物燃料暴露等因素影响不显著，增强了模型的可解释性与实际应用价值。一个尤为值得关注的发现是，“气促”症状虽然在总体样本中发生率较低，却在预测模型中具有极高重要性。这提示，即便轻微或偶发的气促，也可能是一个关键的早期预警信号，在筛查实践中应加强对该类非典型症状的主动识别[15-16]。

本研究的主要创新之处在于，研究视角上，首次系统结合 COPD-SQ 问卷与多种机器学习算法，实现了筛查方法从“固定分值阈值”到“动态风险预测”的转变。方法上，针对样本高度不平衡的挑战，实施了以阈值优化为代表的针对性策略，提升了模型对少数类的识别能力。研究发现上，引入 AP 与 AUC 双指标进行评估，系统揭示了不同模型在精确率-召回率上存在鲜明权衡，并联系不同临床场景为模型选择提供了具体依据；此外，通过特征重要性分析，发现了“气促”症状虽发生率低但预测价值高的反直觉临床洞察。

本研究仍存在一定局限性。在研究特征体系中缺乏肺功能等客观生理指标，主要依赖问卷数据，这可能加剧了数据不平衡问题并增加了建模难度。在未来的研究中，将通过整合肺功能检查等客观指标，开展外部验证，进一步提升模型的稳健性与临床适用价值。

综上所述，机器学习能够有效提升 COPD 筛查效率，为个体化预防提供新的思路，本研究显示，通过有针对性的模型选择与调优，能够有效提升筛查效率，并为个体化预防与管理提供新的决策支持工具。本研究构建的预测模型，未来可嵌入临床决策支持系统，助力实现 COPD 的早期识别、规范诊治与长期管理，从而推动呼吸系统慢性病的防控进程。

参考文献

[1] Adeloye D, Song P, Zhu Y, et al. Global, regional, and national prevalence of, and risk factors for, chronic obstructive pulmonary disease (COPD) in 2019: a systematic review and modelling analysis[J]. Lancet Respir Med, 2022, 10(5): 447-58. doi:10.1016/S2213-2600(21)00511-7.

[2] Ardila D, Kiraly A P, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography[J]. Nat Med, 2019, 25(6): 954-61. doi:10.1038/s41591-019-0447-x.

[3] 王志鹏, 张晓刚, 张宏伟, 等. 机器学习在腰椎间盘突出症患者预后预测模型中应用价值的系统评价[J]. 中国组织工程研究, 2026, 30(3): 740-8. doi:10.12307/2026.875.

Wang Z P, Zhang X G, Zhang H W, et al. A systematic review of application value of machine learning to prognostic prediction models for patients with lumbar disc herniation[J]. Chin J Tissue Eng Res, 2026, 30(3): 740-8. doi:10.12307/2026.875.

[4] 张　娜, 杜　静, 郭子泉, 等. 基于 CT 影像组学的机器学习模型可预测肺部纯磨玻璃结节的浸润性[J]. 分子影像学杂志, 2025, 48(5): 614-9. doi:10.12122/j.issn.1674-4500.2025.05.13.

Zhang N, Du J, Guo Z Q, et al. Machine learning models based on CT radiomics can effectively predict invasiveness of pulmonary pure ground-glass nodules[J]. J Mol Imag, 2025, 48(5): 614-9. doi:10.12122/j.issn.1674-4500.2025.05.13.

[5] Riley R D, Ensor J, Snell K I E, et al. Calculating the sample size required for developing a clinical prediction model[J]. BMJ, 2020, 368: m441. doi:10.1136/bmj.m441.

[6] Maples P, Franks A, Ray S, et al. Development and validation of a low-literacy Chronic Obstructive Pulmonary Disease knowledge Questionnaire (COPD-Q)[J]. Patient Educ Couns, 2010, 81(1): 19-22. doi:10.1016/j.pec.2009.11.020.

[7]Kong J, Ha D, Lee J, et al. Network-based machine learning approach to predict immunotherapy response in cancer patients[J].Nat Commun,2022,13(1):3703. doi:10.1038/s41467-022-31535-6.

[8] Ghalwash M A, Abdelrazek S M, Eladawi N H, et al. Enhancing credit card fraud detection using DBSCAN-augmented disjunctive voting ensemble[J]. Sci Rep, 2025, 15: 39754.

doi:10.1038/s41598-025-22960-w.

[9] Mohammadi E, Rastegar M, Jamshidnezhad A, et al. Machine learning improves detection of alpha thalassemia carriers compared to clinical features[J]. Sci Rep, 2025, 15: 36717. doi:10.1038/s41598-025-20605-6.

[10] Im C, Kim W, Kim H. Explainable machine learning for heat-related illness prediction: an XGBoost‐SHAP approach using Korean meteorological data[J]. Bioengineering, 2025, 12(11): 1276. doi:10.3390/bioengineering12111276.

[11] Dey S, Eapen M S, Chia C, et al. Pathogenesis, clinical features of asthma COPD overlap, and therapeutic modalities[J]. Am J Physiol Lung Cell Mol Physiol, 2022, 322(1): L64-83. doi:10.1152/ajplung.00121.2021.

[12] Ko E J, Bae S O, Kang D. A baseline study of interpretable machine learning using GC-MS breath VOCs for classifying asthma, bronchiectasis, and COPD[J]. Sci Rep, 2025, 15: 44392. doi:10.1038/s41598-025-28143-x.

[13] Glyde H M G, Morgan C, Wilkinson T M A, et al. Remote patient monitoring and machine learning in acute exacerbations of chronic obstructive pulmonary disease: dual systematic literature review and narrative synthesis[J]. J Med Internet Res, 2024, 26: e52143. doi:10.2196/52143.

[14] 常敏丽, 由淑萍, 陈晓蝶, 等. 基于机器学习的肺结核肺炎患者判别分析研究[J]. 安徽医科大学学报, 2025, 60(3): 507-14. doi:10.19405/j.cnki.issn1000-1492.2025.03.017.

Chang M L, You S P, Chen X D, et al. Discriminant analysis of pulmonary tuberculosis patients and pneumonia patients based on machine learning[J]. Acta Univ Med Anhui, 2025, 60(3): 507-14. doi:10.19405/j.cnki.issn1000-1492.2025.03.017.

[15] Sobhanan A, Shrinath V, Deshpande S, et al. Beyond smoking: exploring etiotypes of chronic obstructive pulmonary disease (COPD) using the global initiative for chronic obstructive lung disease (GOLD) 2023 classification[J]. Cureus, 2025, 17(10): e94579. doi:10.7759/cureus.94579.

[16] 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. doi:10.1001/jama.2023.0128.