Statistics and Application 统计学与应用, 2025, 14(11), 67-76
Published Online November 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.1411311

Hans汉斯

多维度特征交互驱动的肺癌风险分级预测模型构建与临床应用研究

马翎容

曲阜师范大学统计与数据科学学院，山东 济宁

收稿日期：2025年10月6日；录用日期：2025年10月27日；发布日期：2025年11月10日

摘　要

肺癌作为全球癌症死亡的首要诱因，其精准的风险分级与致病机制解析对临床诊疗及早期筛查效率具有重要意义。本研究基于Kaggle平台1000例肺癌患者数据集，涵盖环境暴露、既往病史等24项多维特征，尤其引入灰尘过敏等少被探索因素，系统挖掘特征间交互作用与肺癌风险等级的关联。改变了以往常集中于临床特征等单一维度忽略其他影响的情况。同时选取了五种代表模型，通过可视化分析、数据增强、超参数调优等筛选预测模型，并借助特征重要性排序与决策图增强模型可解释性，旨在筛选出兼顾准确性、鲁棒性和解释性的最优预测模型识别并其关键影响因素，为肺癌早期快速初筛检测提供支持。研究结果显示，除饮酒(0.72)、被动吸烟情况(0.7)等主流因素外，像灰尘过敏(0.71)等少被关注的因素及灰尘过敏与职业危害(0.79)等少研究的交互关系应加以重视。随机森林模型性能最优，准确率达98%，咳血、饮酒和肥胖是模型的三大关键预测因子。本研究构建的高精度预测模型为肺癌早期筛查提供可靠工具，新发现的特征与交互作用为肺癌病因学深入研究与个体化防控提供了新方向。

关键词

肺癌水平预测，机器学习，随机森林，特征重要性，交互作用

Construction and Clinical Application of Lung Cancer Risk Grading Prediction Model Driven by Multi-Dimensional Feature Interaction

Lingrong Ma

School of Statistics and Data Science, Qufu Normal University, Jining Shandong

Received: October 6, 2025; accepted: October 27, 2025; published: November 10, 2025

文章引用: 马翎容. 多维度特征交互驱动的肺癌风险分级预测模型构建与临床应用研究[J]. 统计学与应用, 2025, 14(11): 67-76. DOI: 10.12677/sa.2025.1411311

Abstract

As the leading cause of cancer death worldwide, accurate risk grading and pathogenic mechanism analysis of lung cancer are of great significance for clinical diagnosis and early screening efficiency. This study is based on a dataset of 1000 lung cancer patients on the Kaggle platform, covering 24 multidimensional features such as environmental exposure and past medical history. In particular, less explored factors such as dust allergies are introduced to systematically explore the interaction between features and their association with lung cancer risk levels. Changed the situation of focusing solely on clinical features and ignoring other influences. Five representative models were selected simultaneously, and prediction models were screened through visualization analysis, data augmentation, and hyperparameter tuning. The interpretability of the models was enhanced by feature importance ranking and decision graphs, aiming to identify the optimal prediction model that balances accuracy, robustness, and interpretability, and its key influencing factors, providing support for early rapid screening and detection of lung cancer. The research results show that in addition to mainstream factors such as alcohol consumption (0.72) and passive smoking (0.7), less studied factors such as dust allergy (0.71) and the interaction between dust allergy and occupational hazards (0.79) should be given attention. The random forest model has the best performance, with an accuracy of 98%. Coughing blood, alcohol consumption, and obesity are the three key predictive factors of the model. The high-precision prediction model constructed in this study provides a reliable tool for early screening of lung cancer, and the newly discovered features and interactions provide new directions for in-depth research on the etiology of lung cancer and personalized prevention and control.

Keywords

Lung Cancer Level Prediction, Machine Learning, Random Forest, Feature Importance, Interaction

1. 引言

1.1. 研究背景

肺癌已连续多年居全球恶性肿瘤死亡率榜首。据世界卫生组织最新数据显示，2023 年全球新发肺癌病例约 250 万例，死亡病例约 180 万例，5 年生存率仍低于 20%。我国肺癌发病与死亡负担尤为沉重，根据国家癌症中心 2023 年发布的数据，肺癌年新发病例数占所有恶性肿瘤的 20.3%，年死亡病例数占恶性肿瘤死亡总数的 21.68%，发病率和死亡率呈持续上升趋势。临床研究表明，早期肺癌患者(I 期)经规范治疗后 5 年生存率可超过 75%，而晚期(IV 期)患者 5 年生存率仍不足 10%，凸显出“早筛早诊早治”在肺癌防控中的关键作用。

国家高度重视肺癌防治工作，近年来密集出台多项政策，推动肺癌筛查与早诊早治能力的提升。2024 年国家卫生健康委印发的《肺癌筛查与早诊早治方案(2024 年版)》中明确推荐对肺癌高风险人群开展低剂量螺旋 CT (LDCT)筛查，以提高早期诊断率、降低死亡率。同时，《健康中国行动——癌症防治行动实施方案(2023~2030 年)》强调应加大癌症防控科技攻关，构建更加精准的癌症风险预测与干预体系。当前临床多依赖 LDCT，辐射、成本高，且多单(少)维度评估，漏诊率较高。因此，开展高精度肺癌风险分

级预测与致病因素解析研究，不仅是响应国家健康战略的具体实践，更为优化筛查资源分配、提高肺癌早期诊断率提供了科学依据。

1.2. 文献综述

围绕肺癌预测模型的构建与优化，现有研究主要集中在特征分析、预测模型构建两个方面，分述如下：

1) 肺癌患者的特征研究

肺癌临床特征维度多样，但致病因素多聚焦某类或几类没有综合研究、缺少区分度，且少研究因素的相互作用。在生活方式上，刘宝珠和李晓艺(2020)对 84 例病例分析显示，非小细胞肺癌占比达 83.3%，其中腺癌患者吸烟史占比 67% [1]；在症状上，钟德光(2010)针对 40 岁以下群体发现，72%患者以咳嗽、胸痛为首发症状，低分化癌比例达 58% [2]；遗传层面，夏银川等(2021)调查 418 例患者证实，一级亲属患癌史使个体风险提升 1.8 倍(95% CI 1.2~2.6) [3]。环境暴露研究中，张幸(2021)强调石棉职业暴露人群肺癌发病率超普通人群 5~8 倍[4]；Grzywa-Celińska 等(2020)证实 PM2.5 年均浓度每升高 10 μg/m^3，肺癌死亡率增加 8% [5]。

2) 相关预测模型的研究现状

存在多种模型分别从不同范式切入研究，但未涉及综合预测优良的模型。李秀芹等(2022)融合临床指标与基因数据，将肺癌生存预测准确率提升至 78% [6]；陈睿琳等(2023)纳入饮食、运动等 12 项生活行为变量，构建模型 AUC 达 0.79 [7]。机器学习应用中，孟丹等(2020)优化支持向量机实现肺癌文本分类准确率 83% [8]；蓝潞杭等(2021)的随机森林模型在并发症预测中取得 0.82 的特异性[9]。国际研究中，Barsasella 等(2021)通过多因素回归预测糖尿病合并高血压患者住院时长(RMSE = 2.3 天) [10]；Marshall 等(2022)基于支气管微生物组差异建立早期预警模型(敏感性 71%) [11]。

1.3. 研究意义

本研究多维度剖析肺癌风险影响因素，针对致病因素聚焦单一类别，综合研究不足的问题，整合了 24 项多维特征，挖掘潜在影响因素、关键因素以及各因素间的区分度；针对交互作用研究匮乏，挖掘因素间可能存在的相互作用，为机制研究提供新方向。

同时，针对综合性能优良的模型缺失的问题，筛选了高精准预测模型。最优模型可整合多维特征，补充传统诊断盲区，助力临床早期精准筛查。构建的多维度数据处理、特征选择、多模型比较框架，可为后续研究提供流程参考。

1.4. 主要研究内容

1.4.1. 多维度剖析肺癌影响因素

整合新兴因素、生活习惯、临床症状、环境暴露既往病史等多类变量。可视化各因素与肺癌水平的关联强度，识别关键影响因子，揭示因素间可能存在的协同效应，发现既往研究未涉及的潜在关联。

1.4.2. 筛选肺癌水平预测模型

通过相关性分析、小提琴图、回归分析等进行指标筛选，数据增强后对选取的五种模型进行参数调优。采用多指标评估模型性能，筛选最优模型，利用双重机制揭示模型关键影响因素。

1.5. 本文的创新点

① 多维度数的关联挖掘。

本研究综合更全面的因素并纳入了新兴因素，可综合对比因素间的重要程度。展现因素交互效应，

发现未涉及的可存在的潜在关联并探索了各因素间的区分度。

② 多模型优化的预测体系构建。

筛选了兼具准确性与鲁棒性的最优模型。利用双重解释机制，提高模型的可解释性并识别模型的关键影响因素。

2. 模型原理

2.1. 多元 Logistic 回归模型

多元 Logistic 回归模型适用于多分类问题，通过为每个类别构建线性组合并利用 Softmax 函数将输出转化为概率分布，所有类别概率之和为 1。模型以最小化交叉熵损失为目标进行参数优化。其核心公式如下：

Softmax 函数：

$$P\left(y=ix;\theta\right)=\frac{\mathrm{e}^{i\theta^{T}x}}{\sum_{j=1}^{K}\mathrm{e}^{j\theta^{T}x}} \tag{1}$$

其中，K 为类别数，$\theta_i=\left[\theta_{i0},\theta_{i1},\cdots,\theta_{in}\right]$ 为第 i 类对应的参数向量。

交叉熵函数为：

$$L=-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}y_{ik}\log\left(\hat{y}_{ik}\right) \tag{2}$$

本研究通过定义包含 solver、C、max_iter 的超参数网格，利用 GridSearchCV 进行 5 折交叉验证，筛选逻辑回归模型的最优超参数组合，以优化模型性能并控制过拟合。

2.2. 随机森林模型

随机森林(Random Forest)是一种集成学习算法，通过构建多棵决策树并综合其预测结果，以提高模型的准确性和鲁棒性。其核心机制包括 Bootstrap 采样(有放回抽样)和随机特征选择，使得每棵树基于不同的数据子集和特征子集进行训练，增加模型多样性，减少过拟合风险。

在本研究中，随机森林用于处理 24 维特征的多分类任务，通过基尼不纯度(Gini Impurity)作为节点分裂准则，其计算公式为：

$$\mathrm{Gini}=1-\sum_{k=1}^{K}p_k^2 \tag{3}$$

其中 p_k 为第 k 类样本在节点中的比例。我们实现该模型，关键超参数包括 n_estimators、max_depth 和 min_samples_split，并通过网格搜索进行优化。

随机森林不仅能提供高精度预测，还能输出特征重要性排序，为模型解释提供支持，特别适用于高维医学数据的分类问题

2.3. 支持向量机模型

支持向量机是一种基于结构风险最小化原则的分类算法，其核心思想是通过寻找最优超平面实现类别间隔最大化，具有较强的泛化能力。本研究针对肺癌风险分类问题，采用 SVM 处理多分类任务，并使用核函数处理线性不可分情况。

核函数将原始特征映射到高维空间，使数据线性可分。本研究综合比较了以下核函数：

① 线性核(Linear Kernel)：

$$K\left(x_i,x_j\right)=i_T^x x_j \tag{4}$$

② 径向基函数核：

$$K\left(x_i,x_j\right)=\exp\left(-\gamma\left|x_i-x_j\right|^2\right) \tag{5}$$

其中 γ 是核系数。

③ 多项式核：

$$K\left(x_i,x_j\right)=\left(\gamma i_T^x x_j+r\right)^d \tag{6}$$

其中 γ 、 r 和 d 是参数。

通过网格搜索优化惩罚系数 C 及核函数 kernel，以提升模型在肺癌风险分级任务中的性能。

2.4. 高斯朴素贝叶斯模型

高斯朴素贝叶斯是一种基于贝叶斯定理与特征条件独立性假设的分类算法。该模型假设各类别下特征服从高斯分布，通过估计每个类别下特征的均值与方差构建概率模型。分类过程中，依据贝叶斯公式计算后验概率，并选择具有最大后验概率的类别作为预测结果。其核心公式如下：

$$P\left(y\mid X\right)=\frac{P\left(X\mid y\right)\cdot P\left(y\right)}{P\left(X\right)} \tag{7}$$

其中 $P\left(y\mid X\right)$ 为后验概率， $P\left(X\mid y\right)$ 为似然概率。

在本研究中，我们假设所有特征均服从高斯分布，因此似然概率的计算基于高斯概率密度函数：

$$P\left(X_i\mid y\right)=\frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{\left(X_i-\mu_y\right)^2}{2\sigma_y^2}\right) \tag{8}$$

由于特征条件独立性假设，联合似然概率可表示为：

$$P\left(X\mid y\right)=\prod_{i=1}^{n}P\left(X_i\mid y\right) \tag{9}$$

最终通过对数变换避免数值下溢，后验概率的对数形式为：

$$\log P\left(y\mid X\right)=\sum_{i=1}^{n}\log P\left(X_i\mid y\right)+\log P\left(y\right) \tag{10}$$

本研究使用高斯朴素贝叶斯处理多分类问题，通过网格搜索优化平滑参数，以提升模型对肺癌风险等级的预测能力。

2.5. K 近邻模型

K 近邻是一种基于实例的监督学习算法，其核心思想是通过度量待分类样本与训练集中各样本的距离，选取距离最近的 K 个样本，并依据这 K 个近邻的类别投票决定待分类样本的类别。KNN 无需显式训练过程，直接利用已有数据进行预测，适用于多类别分类问题，且对数据分布假设要求较低。

在本研究中，采用欧式距离作为距离度量方式，公式为：

$$d\left(x,y\right)=\sum_{i=1}^{n}\left|x_i-y_i\right| \tag{11}$$

其中 x, y 是两个样本，x_i 和 y_i 分别是它们在第 i 个特征上的值。K 值通过网格搜索与交叉验证进行优化，最终确定最优 K 值用于模型训练与评估。

3. 数据与方法

3.1. 数据来源与预处理

本研究采用 Kaggle 平台的公开肺癌数据集，包含患者症状的多维数据。数据的表现形式除肺癌水平列为 object，其余均为 int。肺癌水平分为“低”“中”“高”三个等级标签集，以此区分病情严重程度。

3.1.1. 数据清洗与转换

为保障数据质量与分析一致性，对原始数据进行了以下处理：

① 无关特征剔除：移除如 Patient Id 等与分析目标无关的标识符变量。

② 数据标准化：将列名统一转换为小写并用下划线连接，提升数据可读性与程序处理一致性。

③ 类型转换：将肺癌严重程度(level)的分类变量(High, Medium, Low)转换为有序数值变量(2, 1, 0)，以适配后续建模要求。

④ 缺失值与异常值处理：经检查，数据集中无缺失值；采用箱线图与 IQR 方法进行异常值检测，未发现显著异常值，表明数据质量较高，无需进一步处理。

3.1.2. 探索性数据分析与特征选择

为深入理解特征与肺癌严重程度的关系及其分布特点，我们进行了多角度可视化分析：

① 相关性分析：

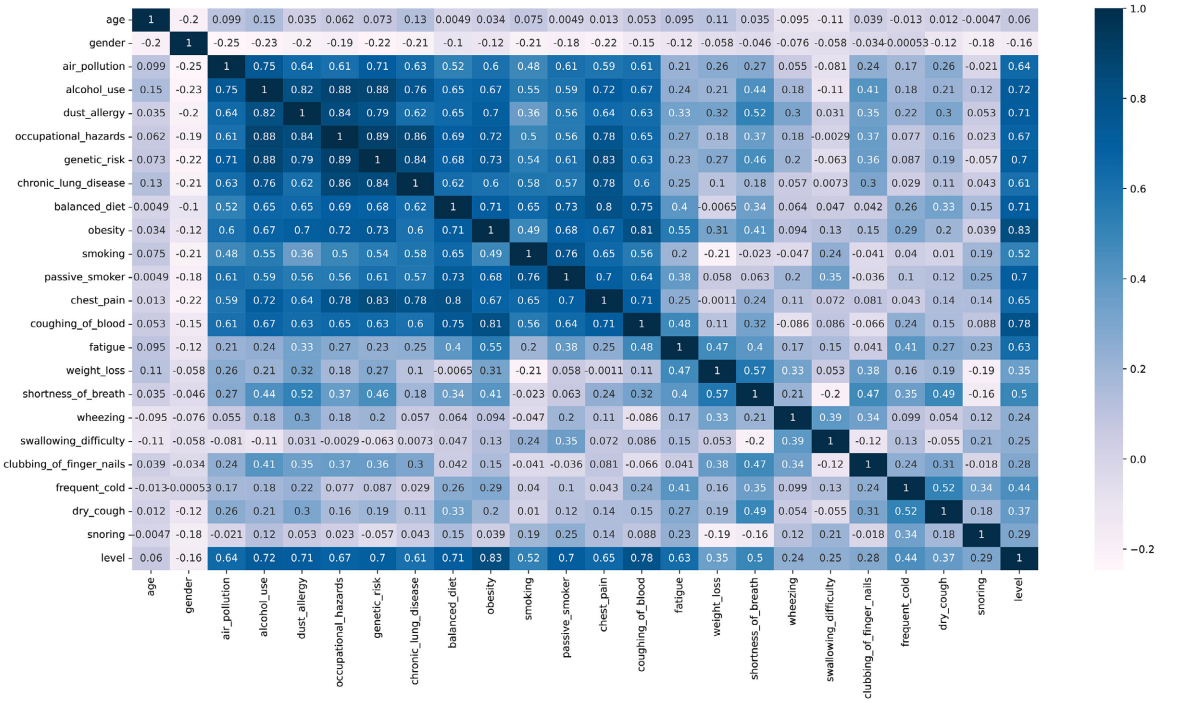

Figure 1. Heatmap of correlations between variables
图 1. 变量间的相关性热力图

见图 1 识别出与肺癌水平显著相关的特征，如灰尘过敏(0.71)、饮酒(0.72)、咳血(0.78)等，并发现特

征间可能存在较强交互作用(如灰尘过敏与职业危害相关性达 0.84)，灰尘过敏作为新兴风险因素需重点关注。

② 回归图与小提琴图：

直观展示特征与肺癌水平的线性与分布关系，确认部分特征(如吸烟、空气污染)在不同风险等级下分布差异显著，区分性强；而年龄、均衡饮食等因素区分能力有限。

③ 分布分析：

通过变量的直方及核密度估计曲线图。计验证了关键特征的数值分布特性，如吸烟程度($\mu = 3.9$, $\sigma = 2.5$)呈较宽分布，提示其影响显著。

基于上述分析，为最大限度保留特征交互信息与潜在模式，本研究未进行特征筛选，而是构建“宽特征空间”纳入全部特征，以支持后续复杂模型学习多维度交互效应。

3.1.3. 数据增强与标准化

为进一步增强模型表达能力，进行了如下特征工程操作。

构造交互项：如灰尘过敏与空气污染的乘积项，以捕捉环境因素的协同效应。

数据增强：针对类别不平衡问题，采用 SMOTE 过采样方法，增强少数类样本的表示，提升模型对罕见类别的识别能力。SMOTE 优于欠采样方法，因其能保留原始样本分布信息，避免信息损失。

数据标准化：对所有连续变量进行 Z-score 标准化，以消除量纲差异，提升模型收敛速度与稳定性。

3.1.4. 数据集的划分

采用分层抽样策略，将数据按 7:2:1 比例划分为训练集、验证集与测试集，确保各子集中类别比例与总体一致，避免因随机划分引入偏差。

3.2. 模型构建与训练

本研究系统比较了五种经典机器学习模型，旨在筛选出兼具预测性能与解释性的肺癌风险分级模型。

3.2.1. 模型选择与超参数优化

所选模型包括：多元逻辑回归、随机森林、支持向量机、K 近邻和高斯朴素贝叶斯。选择这些模型是基于其在不同数据结构下的优势互补性：逻辑回归具有良好的解释性；随机森林能捕捉非线性关系与交互效应；SVM 适用于高维数据；KNN 与朴素贝叶斯则作为基准模型参与对比。

为最大化模型性能，采用网格搜索与 5 折交叉验证进行超参数调优。选择 5 折交叉验证是为了在有限数据下实现偏差与方差的平衡，避免过拟合的同时充分利用样本信息。各模型调优参数如下。

逻辑回归：优化 solver、正则化系数 C 和最大迭代次数；

随机森林：调整树的数量、最大深度、最小分裂样本数等；

SVM：优化核函数类型、惩罚系数 C 与核参数；

高斯朴素贝叶斯：优化平滑参数 var_smoothing；

KNN：调整近邻数 k 与权重函数。

3.2.2. 模型训练与评估策略

使用训练集进行模型训练，验证集用于超参数选择与早期停止，测试集用于最终性能评估。所有模型均在同一数据划分下训练与测试，确保结果可比性。

3.3. 模型评估

本研究综合采用了多维度指标系统评估了五类模型的预测效能。

3.3.1. 模型预测性能对比

Table 1. Classification performance evaluation of each model
表 1. 各模型分类性能评价表

	Log LOSS	ACC	AUC			加权平均 F1-Score	宏平均 F1-Score
			0	1	2		
多元逻辑回归	0.2279	0.95	0.98	0.90	0.99	0.95	0.95
随机森林	0.4123	0.98	0.99	0.98	0.99	0.98	0.98
GNB	1.6744	0.83	0.95	0.80	0.97	0.83	0.83
KNN	1.1690	0.97	0.98	0.96	0.99	0.97	0.97
SVM	0.1465	0.97	0.98	0.93	1.00	0.97	0.97

见表 1，随机森林模型在大多数指标上表现最为优异，其准确率和加权平均 F1-Score 均为最高，表明其具有出色的整体分类性能。SVM 模型虽在 Log Loss 和 Class 2 的 AUC 上表现极佳，暗示其对高风险类别有极强的区分能力和良好的概率校准，但其宏平均 F1-Score 低于随机森林。

综合来看，随机森林表现最佳。

3.3.2. 模型稳定性分析

结合各模型的学习曲线，随机森林拟合与泛化能力强，训练集与验证集得分收敛于较高水平，几乎无过拟合；多元逻辑回归、KNN 和 SVM 前期均呈现一定程度的过拟合，但随着训练样本量增加，泛化能力逐步提升；高斯朴素贝叶斯模型稳定性和泛化性较差，得分波动较大，这可能源于其特征条件独立性假设在复杂数据上的局限性。

综合对比，随机森林模型在预测性能、鲁棒性和稳定性上综合表现最为优异，故将其确定为最优模型。

3.4. 最优模型解释

随机森林模型不仅性能优异，还具备良好的可解释性。通过分析其特征重要性，发现咳血(0.125)、饮酒情况(0.088)、肥胖(0.085)是前三大关键预测因子，这与临床认知高度吻合(咳血是典型呼吸道严重症状，饮酒与肥胖是已知癌症风险因素)。而性别的特征重要性最低，近乎为 0，表明在该数据集中其对肺癌风险等级的区分能力很弱。为进一步理解模型的决策机制，研究可视化了随机森林中一棵深度为 3 的典型决策树。该子树显示，模型首先根据咳血程度进行初级分裂，随后在分支中进一步依据饮酒情况和肥胖程度进行判断。这直观地揭示了这些关键特征在模型决策路径中的核心作用及其间的交互关系，极大地增强了模型的可信度和透明度，为临床辅助决策提供了清晰的依据。

4. 讨论与结论

4.1. 讨论

本研究通过集成多维度特征与多种机器学习算法，构建了一个高精度的肺癌风险分级预测模型，并探讨了特征间的交互作用。

本研究最重要的发现是，随机森林模型在肺癌风险分级预测中展现出卓越的综合性能，且揭示了咳血、饮酒等关键影响因素。此外还发现了如灰尘过敏与职业危害之间存在强交互作用等，这一新颖发现为肺癌病因学提供了新的视角。

与既往研究相比，本研究的模型更具有通用性。以往模型大多针对具体情况具体选择，本研究采用的随机森林模型可针对所有情况进行快速初筛，这主要得益于对多维度特征更为全面的整合与利用，以及系统的超参数优化策略与模型评估。

对关键特征的机理解释能提升研究的深度。灰尘过敏(相关性 0.71)与肺癌风险的高关联性可能源于：持续的过敏性炎症反应导致肺部组织反复损伤与修复，此过程可能促进细胞增殖与基因突变，从而增加癌变风险。咳血作为重要性最高的特征，是肺部肿瘤侵犯血管的直接临床表现，其预测价值毋庸置疑。在特征解读上，模型中性别重要性近乎为 0，这与现有肺癌流行病学研究可部分呼应：虽《中国肺癌流行病学报告》提及性别差异与吸烟暴露、激素水平相关，但本数据集中可能因性别相关混杂因素(如男女吸烟率差异小、未纳入激素水平等特征)被咳血、饮酒等强预测因子掩盖，此结果与夏银川等(2021)研究中性别未成为肺癌独立风险因子的结论一致，也符合《肺癌筛查与早诊早治方案(2024 年版)》未将性别作为核心筛查指标的临床导向，进一步说明肺癌风险评估中需优先关注病理症状与核心生活暴露因素。

然而，本研究仍存在若干局限性。首先，数据来源于单一公开数据集，虽经精心处理，仍可能存在未知的选择偏倚，如样本或集中于特定地域、年龄层或疾病亚型，导致模型对异质人群的泛化能力受限，间接影响结论在更广泛临床场景的适用性，未来可进一步整合多源数据，进一步提升预测精度。其次，所有特征均为横断面数据，无法反映动态变化对风险的影响。第三，尽管模型在内部验证中表现优异，但尚未在独立的外部队列中进行验证，其泛化能力有待进一步证实。最后，一些新颖的关联(如打鼾)虽被揭示，但其背后的生物学机制仍需进行深入探索。

4.2. 结论

本文立足于肺癌水平预测研究的刚需，本研究的主要结论与贡献如下：

1) 多维度剖析了肺癌风险影响因素

① 关键影响因素挖掘。发现灰尘过敏等新兴因素及咳血等临床症状与肺癌水平显著相关，拓展肺癌防治关注维度。

② 因素交互作用揭示。发现空气污染与饮酒、灰尘过敏与职业危害等有强交互关系，为解析肺癌复杂发病机制提供全新视角。

③ 因素区分性甄别。证实吸烟、职业危害等因素在不同肺癌等级间区分显著，性别、打鼾等相关性较弱，锚定预测模型核心特征。

2) 筛选了高精准预测模型

① 模型优选：确定随机森林模型为最优预测模型。该模型可整合多维度特征，量化风险评分，助力早期精准筛查。

② 关键特征识别：咳血、饮酒、肥胖成为模型核心驱动因子，实现多维度特征高效整合。

参考文献

[1] 刘宝珠, 李晓艺. 肺癌 84 例临床病理分析[J]. 基层医学论坛, 2020, 24(7): 977-978.

[2] 钟德光. 40 岁以下肺癌临床特征分析[J]. 重庆医学, 2010, 39(13): 1777.

[3] 夏银川, 张冉, 柯晓庆, 等. 418 例肺癌患者家族癌症发病史流行病学调查分析[J]. 公共卫生与预防医学, 2021,

32(1): 121-124.

[4] 张幸. 石棉相关癌症防控的紧迫性不容忽视[J]. 中华劳动卫生职业病杂志, 2021, 39(2): 81-84.

[5] Grzywa-Celińska, A., Krusiński, A. and Milanowski, J. (2020) 'Smoging Kills'—Effects of Air Pollution on Human Respiratory System. *Annals of Agricultural and Environmental Medicine*, **27**, 1-5. https://doi.org/10.26444/aaem/110477

[6] 李秀芹, 李琳, 张慢丽. 基于集成学习的肺癌存活性预测分析[J]. 软件工程, 2022, 25(1): 41-46.

[7] 陈睿琳, 王静茹, 王硕, 唐思琦, 索晨. 大规模人群队列生活行为方式相关的肺癌风险预测模型的构建[J]. 四川大学学报(医学版), 2023, 54(5): 892-898.

[8] 孟丹, 张卫东, 李昌, 王杨, 甄磊. 基于支持向量机的中文极短文本分类模型[J]. 计算机应用研究, 2020, 37(2): 347-350.

[9] 蓝潞杭, 蒋炫东, 王茂峰, 等. 随机森林模型预测急性心肌梗死后急性肾损伤[J]. 中华急诊医学杂志, 2021, 30(4): 491-495.

[10] Barsasella, D., Gupta, S., Malwade, S., Aminin,, Susanti, Y., Tirmadi, B., *et al.* (2021) Predicting Length of Stay and Mortality among Hospitalized Patients with Type 2 Diabetes Mellitus and Hypertension. *International Journal of Medical Informatics*, **154**, Article ID: 104569. https://doi.org/10.1016/j.ijmedinf.2021.104569

[11] Marshall, E.A., Filho, F.S.L., Sin, D.D., Lam, S., Leung, J.M. and Lam, W.L. (2022) Distinct Bronchial Microbiome Precedes Clinical Diagnosis of Lung Cancer. *Molecular Cancer*, **21**, Article No. 68. https://doi.org/10.1186/s12943-022-01544-6