基于机器学习的呼吸道疾病诊断模型研究

张睿妍[1] 彭铃钦[1] 杨 晶[2]
ZHANG Ruiyan PENG Lingqin YANG Jing

摘 要 由于流行性呼吸道疾病具有传染性强、传播速度快且症状相似的特点，快速准确地鉴别诊断对临床决策和公共卫生防控至关重要。为此构建了基于机器学习的疾病多分类诊断模型，对比研究了 Logistic 回归、K 近邻、朴素贝叶斯、决策树、AdaBoost 和随机森林 6 种模型在疾病多分类任务中的性能。首先，通过 6 种特征选择算法从 300 例数据中初筛特征；随后，利用斯皮尔曼相关系数去除冗余特征，并通过互信息值筛选出 17 个与目标变量最相关的关键特征；最后，结合多种模型进行训练与评估。结果表明，随机森林模型性能最优，精确率达 83%，Micro-AUC 为 0.962 3。该研究为小样本、高维数据的多分类问题和机器学习在疾病诊断领域的应用提供了新的实践参考。

关键词 机器学习；多分类模型；特征选择；随机森林；呼吸道疾病诊断

doi：10.3969/j.issn.1672-9528.2025.08.030

0 引言

呼吸道感染病原体类型各异，但是传播方式与临床症状却很相似，一般通过呼吸道传播，临床症状主要表现为鼻塞、咳嗽、发热、头痛、咽喉痛等[1]。由于病因和治疗方法不同，因此准确的早期诊断至关重要。本文选择新冠感染、甲流、肺炎支原体和百日咳 4 种呼吸道疾病进行研究，是因为它们在临床实践中常被混淆，且早期诊断对患者预后有重要影响。

近年来，人工智能技术在智慧医疗领域逐渐展现出巨大的潜力，在疾病诊断和临床决策支持方面取得了显著进展。本研究聚焦于构建一个基于机器学习的初步多分类诊断模型，旨在通过数据驱动的方法，为呼吸道疾病的识别提供初步的科学依据，助力智慧医疗体系的进一步完善。

1 研究方法与数据构建

1.1 数据来源与变量说明

本文使用的数据主要来源于丁香园病例库、医学慕课病例库、爱爱医病例库，并以 CHARLS 数据库、NHANES 数据库的部分内容作为参考。该数据库包含 300 条样本，每条样本包含 38 个特征，其中有 11 个定性变量，27 个定量变量。目标变量为疾病分类，包含 5 个类别：正常、新冠感染、甲流、肺炎支原体和百日咳。

1. 河南科技大学信息工程学院 河南洛阳 471023
2. 洛阳市青岛路小学 河南洛阳 471003

[基金项目] 河南省大学生创新创业训练计划项目（202410464073）

1.2 项目流程

项目基于机器学习算法构建呼吸道疾病智能诊断模型，通过训练多维度临床病例数据构建分类器，对比 K 近邻、随机森林等算法的综合性能，最终遴选最优模型实现疾病智能分类。 如图 1 所示，系统实施流程包含 4 个阶段：

（1）数据采集：通过医疗信息平台收集实验室指标数据，构建病例库。

（2）特征工程：进行数据处理，采用多种特征提取算法筛选出关键特征。

（3）模型训练：并行训练随机森林等分类器，计算每个分类器的分类指标结果。

（4）模型选择：对比分类指标结果得到性能最优的模型。项目方案通过机器学习实现从数据到诊断模型的完整闭环，在保证模型可解释性的同时提升诊断准确率与效率。

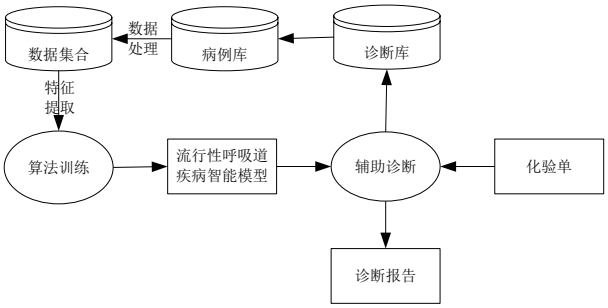

图 1 基于机器学习的呼吸道疾病分类诊断过程

2 特征工程

2.1 数据预处理

数据预处理过程中，使用标签编码将目标变量（中文值）

转换为数值标签。对缺失值进行处理时，定量变量采用均值填充法，定性变量使用众数填充法。数据集按 7:3 的比例划分为训练集和测试集，并使用分层抽样保持各类别比例一致。定性变量通过 OneHotEncoder 进行独热编码，定量变量使用 RobustScaler 进行标准化，达到缩小异常值的影响和统一量纲的目的 [2]。

2.2 特征选择算法

2.2.1 RFECV+SVM

SVM-RFE 的改进版本，SVM-RFE 的核心思想是通过递归消除特征，逐步剔除对 SVM 模型贡献最小的特征。RFECV 在其基础上增加了交叉验证（CV）来提高模型性能 [3]。SVM 作为分类器，适用于处理非线性、高维度的小样本数据，具有良好的泛化能力。

2.2.2 Relief-F

Relief-F 是对经典 Relief 算法的扩展，能够处理多分类问题 [4]。在特征选择中，Relief-F 通过评估特征在邻近样本中的区分能力来确定其重要性，适用于处理非线性关系和噪声数据。

2.2.3 Null-Importance+GBDT

先用树模型计算特征重要性，得到特征原始重要性分布，然后将标签随机打乱 *n* 次，得到特征随机重要性排序，综合比较两者的偏离度是否显著，实现特征的评估和筛选 [5]。GBDT 通过加法模型（基函数的线性组合）捕捉非线性交互，并通过不断减小残差来进行数据分类 [6]。

2.2.4 Distance correlation coefficient

一种能够检测线性和非线性相关性的统计方法，基于特征空间和目标空间中样本之间的距离计算相关性。其值范围为 [0, 1]，0 表示完全独立，1 表示完全依赖。

2.2.5 Random Forest

通过构建多棵决策树评估特征重要性，使用 Gini 重要性或 Permutation Importance 衡量特征对分类的贡献。它能处理高维数据和非线性关系，对过拟合具有较强鲁棒性，适合小样本特征选择。

2.2.6 Kolmogorov–Smirnov test

通过比较两个样本的累积分布函数（cumulative distribution function, CDF）之间的最大差异评估分布差异 [7]。传统 K-S 检验仅支持二分类，本研究通过逐对比较类别并取最大 K-S 统计量，将其扩展至多分类。

2.3 特征提取流程

首先，每种算法从 38 个特征中选择排名前 20 的特征。由于 Relief-F，Null-Importance+GBDT，Random Forest 随机性较另外 3 种更强，故用它们分别进行 10 次特征选择，剩余算法只进行 1 次特征选择，汇总每次得到的排名前 20 特征，统计出现频次大于等于 7 的特征，构成最终排名前 20 特征。各个特征算法初步筛选出的重要性前 20 特征对比图如图 2 所示。

图 2 各特征算法特征重要性排序图对比

随后，对所有算法的结果进行汇总，计算出现频次大于等于 3 次的特征，得到一个共识特征集合 [8]。

为避免特征冗余，利用斯皮尔曼相关系数（spearman's rank correlation coefficient）从集合中筛选出高度相关的特征。各特征间的相关系数热力图如图 3 所示。

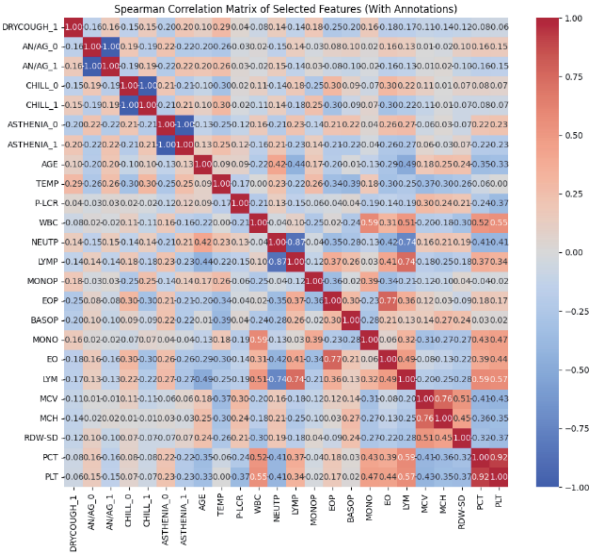

图 3 斯皮尔曼相关系数热力图

对得到的特征对，进一步计算其中单个特征和目标变量的互信息值。互信息（mutual information, MI）是用来衡量两个变量之间依赖关系的指标。互信息值较高的特征表示其与目标变量之间有更多的信息共享，可以作为预测模型中的重要特征，于是剔除特征对中互信息值较低的特征。

2.4 最终特征集

最终筛选出 17 个关键特征，如表 1 所示。

表 1 最终特征分类表

定性变量	干咳、味觉或嗅觉失灵、畏寒、乏力
定量变量	年龄、体温、白细胞计数、中性粒细胞百分比、淋巴细胞绝对值、大血小板比率、血小板计数、嗜碱性粒细胞百分比、嗜酸性粒细胞绝对值、单核细胞百分比、单核细胞绝对值、平均红细胞体积、红细胞分布宽度 SD

3 基于机器学习的分类算法

流行性呼吸道疾病诊断是一个多分类任务，需区分甲流、新冠肺炎等多种类型的疾病。本项目旨在基于机器学习，通过对比分类算法的性能，筛选出对流行性呼吸道疾病数据最具适应性的诊断模型。在本段中，介绍 Logistic 回归、K 近邻、朴素贝叶斯、决策树 4 种经典算法以及随机森林、AdaBoost 两种集成算法的相关知识。

3.1 Logistic 回归

Logistic 回归是一种分类算法，通过 Sigmoid 函数将线性组合特征映射到 (0,1) 区间。其核心公式为：

$$P(x)=\frac{1}{1+e^{-\omega^T\cdot x}} \tag{1}$$

适用于二分类或多分类任务。多分类逻辑回归模型可通过二分类逻辑回归模型推广得到，假设因变量 Y 取值为 $\{1,2,\cdots,k\}$，那么多分类逻辑回归模型的条件概率计算公式为：

$$P(Y=K\mid x)=\frac{1}{1+\sum_{k=1}^{K-1}\exp(w_k\bullet x)} \tag{2}$$

式中：$x\in\mathbf{R}^{n+1}$，$w_k\in\mathbf{R}^{n+1}$。

其原理基于最大似然估计优化参数，支持正则化防止过拟合。优点是可解释性强、计算高效，但对非线性关系建模能力有限。

3.2 K 近邻（KNN）

KNN 是一种基于实例的非参数算法，通过计算样本间的距离，选取最近的 K 个邻居，根据多数投票确定类别。其中衡量距离的方式有很多种，假设 $\boldsymbol{X}$ 是 n 维实数特征空间 $\mathbf{R}^n$，x_1 和 x_2 是 $\boldsymbol{X}$ 中任意两点，其中 $x_i, x_j\in\boldsymbol{X}$，$x_i=(x_i^{(1)}, x_i^{(2)},\cdots, x_i^{(n)})$，$x_j=(x_j^{(1)}, x_j^{(2)},\cdots, x_j^{(n)})$，$x_i$ 和 x_j 的 L_p 距离定义为：

$$L_p(x_i,x_j)=\left(\sum_{l=1}^{n}\mid x_i^{(l)}-x_j^{(l)}\mid^p\right)^{\frac{1}{p}},\text{其中}p\geq 1 \tag{3}$$

无需显式训练，但预测时需遍历全部数据，计算复杂度高，易受维度灾难影响，适合低维小数据，K 值需调优以平衡噪声与泛化能力。

3.3 朴素贝叶斯

朴素贝叶斯法是一种专门用于处理分类问题的算法，基于贝叶斯定理与特征条件独立假设，计算后验概率进行分类。朴素贝叶斯分类的基本公式为：

$$P(Y=c_k\mid X=x)=\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)},k=1,2,...,K \tag{4}$$

该模型计算高效、适合高维稀疏数据，对小样本和噪声数据表现稳健，但独立性假设在现实中常不成立，导致模型偏差。

3.4 决策树

决策树是一种可应用于分类与回归的基本方法，通过递归分割特征空间构建树形结构，以信息增益或 Gini 指数为分类准则实现。本项目基于 scikit-learn 库实现，其中的决策树实现是基于 CART 的优化版本，通过 Gini 指数选择特征。对集合 D，k 是集合中类的个数，C_k 是 D 中属于第 k 类的样本子集，$|C_k|$ 表示属于类 C_k 的样本个数，则 Gini 指数为：

$$\text{Gini}(D)=1-\sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2 \tag{5}$$

模型直观易解释，对缺失值和异常值鲁棒，但易过拟合，需剪枝或集成提升泛化性，适用于非线性、多模态数据场景。

3.5 随机森林

随机森林模型是一种基于 Bagging 集成学习的多分类算法，通过集成学习思想构建多个决策树，每棵树基于对训练数据的有效性抽样和随机特征选择进行训练，最终通过所有决策树的预测结果进行多数投票，将得票最多的类别作为待测样本的分类结果 [9]。该框架通过数据加特征的双重随机性与集成决策，有效平衡偏差与方差，尤其适合高维、非线性医学数据的复杂分类任务。

3.6 AdaBoost

AdaBoost 算法属于 Boosting 框架，是通过迭代训练多个弱分类器，并动态调整样本权重与分类器权重，最终将其加权组合为强分类器。其核心在于动态调整样本权重，预测其权重为：

$$\alpha_j=\eta\log\frac{1-r_j}{r_j} \tag{6}$$

初始时所有样本权重相同，每轮训练后增加分类错误样本的权重，迫使后续分类器更关注难例；同时根据每个弱分类器的错误率计算其投票权重，准确率高的分类器在最终集成中占更高比重，最后得到大多数加权投票的类就是预测器

给出的预测类。

$$\hat{y}(x)=\arg\max_{k}\sum_{\substack{j=1\\ \hat{y}(x)=k}}^{N}\alpha_j \tag{7}$$

式中：N 是预测器的数量。

scikit-learn 库中使用的是 Adaboost 的一个多分类版本，叫作 SAMME（基于多类指数损失函数的逐步添加模型）。

4 模型选择

4.1 模型的评估指标

4.1.1 多分类混淆矩阵

混淆矩阵是评估多分类模型性能的直观工具，其结构为 $N\times N$ 矩阵（N 为类别总数）。矩阵的行表示样本的真实类别，列对应模型的预测结果，每个元素反映真实类别为 i 但被预测为 j 的样本数量。对于具有 k 个类别的分类任务，混淆矩阵 $\boldsymbol{C}$ 可表示为：

$$\boldsymbol{C}=\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kk} \end{bmatrix} \tag{8}$$

基于此矩阵可得到多项评估指标，用公式分别表示为：

（1）准确率：表示模型正确预测的样本占总样本的比例。

$$\text{accuracy}=\frac{\sum_{i=1}^{k}c_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k}c_{ij}} \tag{9}$$

（2）精确率：表示预测为某类的样本中实际属于该类的比例。

$$\text{precision}=\frac{c_{ii}}{\sum_{j=1}^{k}c_{ji}} \tag{10}$$

（3）召回率：表示某类样本中被模型正确预测的比例。

$$\text{recall}=\frac{c_{ii}}{\sum_{j=1}^{k}c_{ij}} \tag{11}$$

（4）F_1 分数：表示精确率和召回率的调和均值，可衡量综合评估模型的性能。

$$F_1=\frac{2\,\text{precision}\bullet\text{recall}}{\text{precision}+\text{recall}} \tag{12}$$

4.1.2 ROC 曲线和 AUC 值

受试者工作特征曲线（ROC 曲线）是评估分类模型综合性能的重要工具，通过映射真阳性率与假阳性率的关系，展示模型在不同阈值下的性能表现 [10]。ROC 曲线图中横轴代表负类误判率，纵轴代表正类识别能力，曲线凸起程度体现分类边界优化效果。

曲线下面积（AUC）是对 ROC 曲线性能的量化指标，它衡量模型区分正负类的能力。AUC 值具有阈值无关性，避免了单一阈值评估的片面性。AUC=0.5 为随机猜测，AUC=1.0 为理想分类器，值越高表明模型预测排序越可靠。

4.2 评估结果

将收集的 38 个特征实验室数据利用多种特征选择方法选出 17 个特征，将 70% 的数据集当作训练集，剩余的 30% 的数据集当作测试集，使用前文介绍的 4 种经典单一算法和两种集成算法对呼吸道疾病进行分类训练。经过上述 6 种算法的训练后在验证集上进行预测，得到每个算法的评价指标如表 2 所示。

表 2 算法分类指标结果显示表

Logistic 回归				
类别	Precision	Recall	F_1-score	AUC
0	0.67	0.44	0.53	0.779
1	0.52	0.78	0.62	0.891
2	0.69	0.61	0.65	0.863
3	0.94	0.94	0.94	0.985
4	0.94	0.89	0.91	0.987
Accuracy	—	—	0.73	0.901
Macro Avg	0.75	0.73	0.73	0.901
Weighted Avg	0.75	0.73	0.73	0.901
K 近邻				
类别	Precision	Recall	F_1-score	AUC
0	0.71	0.56	0.62	0.789 0
1	0.60	0.67	0.63	0.871 1
2	0.56	0.56	0.56	0.811 0
3	0.89	0.94	0.92	0.997 3
4	0.79	0.83	0.81	0.978 4
Accuracy	—	—	0.71	0.889 4
Macro Avg	0.71	0.71	0.71	0.889 4
Weighted Avg	0.71	0.71	0.71	0.889 4
朴素贝叶斯				
类别	Precision	Recall	F_1-score	AUC
0	0.73	0.44	0.55	0.778 2
1	0.61	0.78	0.68	0.888 9
2	0.58	0.39	0.47	0.833 3
3	0.69	1.00	0.82	1.000 0
4	0.78	0.78	0.78	0.864 2
Accuracy	—	—	0.68	0.873 0
Macro Avg	0.68	0.68	0.66	0.873 0
Weighted Avg	0.68	0.68	0.66	0.873 0
决策树				
类别	Precision	Recall	F_1-score	AUC
0	0.64	0.39	0.48	0.666 7
1	0.68	0.83	0.75	0.868 1
2	0.56	0.83	0.67	0.833 3
3	0.88	0.83	0.86	0.902 8

表 2（续）

决策树				
类别	Precision	Recall	F_1-score	AUC
4	0.92	0.67	0.77	0.826 4
Accuracy	—	—	0.71	0.819 4
Macro Avg	0.74	0.71	0.71	0.819 4
Weighted Avg	0.74	0.71	0.71	0.819 4
随机森林				
类别	Precision	Recall	F_1-score	AUC
0	0.77	0.56	0.65	0.923 2
1	0.73	0.89	0.80	0.962 6
2	0.74	0.78	0.76	0.941 0
3	0.89	0.94	0.92	0.997 3
4	1.00	0.94	0.97	0.987 3
Accuracy	—	—	0.82	0.962 3
Macro Avg	0.83	0.82	0.82	0.962 3
Weighted Avg	0.83	0.82	0.82	0.962 3
AdaBoost				
类别	Precision	Recall	F_1-score	AUC
0	0.71	0.56	0.62	0.904 3
1	0.68	0.83	0.75	0.934 4
2	0.67	0.78	0.72	0.937 5
3	0.94	0.89	0.91	0.998 5
4	1.00	0.89	0.94	0.994 6
Accuracy	—	—	0.79	0.953 9
Macro Avg	0.80	0.79	0.79	0.953 9
Weighted Avg	0.80	0.79	0.79	0.953 9

从精确率、召回率、F_1-score 以及 AUC 值综合判断，随机森林模型对呼吸道疾病分类性能表现最为突出。这是因为随机森林模型通过集成多棵决策树进行特征选择与样本加权投票，有效降低了过拟合风险，使得其精确率达到 83%，召回率为 82%，F_1-score 值为 82%，均优于其他模型。特别在 AUC 值指标上，随机森林以 0.962 3 的优异表现逼近理想分类器阈值 1，进一步印证了其强大的类别分类能力。

综上所述，基于集成学习策略的随机森林模型在呼吸道疾病诊断中展现出最佳综合性能，对构建针对流行性呼吸道疾病的多分类诊断模型具有重要应用价值。

5 结语

机器学习辅助诊断已成为近年来计算机科学领域的一个研究焦点，本文对特征提取方法、机器学习中的 3 种经典单一算法和 2 种集成算法，以及常用模型评估指标进行了简要介绍，并整合多种特征选择算法在复杂数据集中提取出关键特征，通过对比多种机器学习模型，验证了随机森林模型在临床数据处理中的优异性能。研究结论为智能辅助诊断系统的开发奠定了基础，后续可探究该模型在真实医疗环境中的部署与应用，以促进精准医疗诊断的进一步发展。

参考文献：

[1] 赵亚虹，胥俊越，李弈，等 .2022 年与 2023 年呼吸道疾病高发月份 7 种常见病原体变化趋势分析 [J]. 标记免疫分析与临床，2024,31(12):2222-2227.

[2] PENG X L, LIU Y L, ZHANG B,et al.A preliminary prediction model of pediatric mycoplasma pneumoniae pneumonia based on routine blood parameters by using machine learning method[J].BMC infectious diseases, 2024, 24(1): 707.

[3] 李楠 . 基于 DE-SVM 的慢性病预测研究 [D]. 太原：山西大学，2022.

[4] ROBNIK-ŠIKONJA M, KONONENKO I.Theoretical and empirical analysis of ReliefF and RReliefF[J].Machine learning, 2003,53(1):23-69.

[5] 曹佳悦，罗冬梅 . 基于 Null Importance 与 GS-LGBM 的糖尿病视网膜病变因素分析与风险预测 [J]. 中国医学物理学杂志，2023,40(8):1033-1038.

[6] 贝壳技术有限公司 . 信息处理方法和装置、电子设备和存储介质：202010886868.1[P].2021-01-08.

[7] 岳健，史秉帅，范寒，等 . 基于多级特征提取框架的风电机组载荷预测方法 [J]. 太阳能学报，2024,45(12):350-359.

[8] 苏畅 . 机器学习算法在冠心病预测中的应用研究 [D]. 桂林：桂林理工大学，2021.

[9] 刘静，汪泓，张磊，等 . 基于高光谱的辣椒叶片氮素含量反演 [J]. 中国农业科学，2025,58(2):252-265.

[10] 北京大学第一医院 . 一种基于机器学习模型的患者睡眠觉醒状态检测方法 :202311201600.X[P].2024-01-05.

【作者简介】

张睿妍（2004—），女，河南洛阳人，本科在读，研究方向：机器学习、人工智能。

彭铃钦（2003—），女，重庆人，本科在读，研究方向：人工智能。

杨晶（1986—），女，河南临颍人，本科，研究方向：大数据分析与人工智能。

（收稿日期：2025-03-22 修回日期：2025-07-31）