DOI: 10.12289/j.issn.2097-4345.25390

·临床研究·

整合淋巴细胞亚群与临床特征的机器学习模型在非结核分枝杆菌肺病、肺结核及其他肺部疾病鉴别诊断中的应用与效能评估

王　蕾[1], 曹　婕[1], 刘轾彬[1], 程丽平[1], 吴小翠[2], 孙　勤[1*], 沙　巍[1*]

(1. 同济大学附属上海市肺科医院上海市结核病重点实验室,结核病临床研究中心,上海　200433;
2. 同济大学附属上海市肺科医院检验科,上海　200433)

【摘要】目的　基于淋巴细胞亚群计数指标,利用不同机器学习方法构建诊断模型,区分非结核分枝杆菌肺病(nontuberculous mycobacterial pulmonary disease, NTM-PD)、肺结核(pulmonary tuberculosis, PTB)以及其余常见易混淆的肺部疾病,为早期识别肺部感染性疾病提供科学依据。**方法**　本研究选取 2023 年 1 月—2023 年 12 月同济大学附属上海市肺科医院结核科收治的确诊为活动性结核病、NTM-PD 和其他肺部疾病(肺部炎症性疾病及肺部肿瘤)的患者为研究对象,采用流式细胞技术检测淋巴细胞亚群计数。使用多分类 Logistic 回归、朴素贝叶斯、随机森林、XGBoost 4 种算法进行建模预测,使用贝叶斯和交叉验证方式进行算法超参数优化。从开发集中的差异性分析筛选 $P<0.1$ 变量,使用相关性分析和 Lasso 回归进行特征筛选后进入建模。构建多分类 Logistic 回归、朴素贝叶斯、随机森林、XGBoost 4 种机器学习模型。使用受试者工作特征曲线下面积(area under the receiver operating characteristic curve, AUROC)、平均精度-精确率召回曲线(average precision-precision recall curve, AP-PR)和决策曲线分析(decision curve analysis, DCA)曲线对模型在测试集上的表现进行评价。**结果**　本研究共纳入 1 383 例患者,其中结核组 836 例,NTM 肺病组 254 例,其他组 293 例;以筛选出的人口学信息、合并疾病、淋巴细胞亚群指标为输入变量,以疾病类别为结果变量,成功构建多分类 Logistic 回归、朴素贝叶斯、随机森林、XGBoost 4 种机器学习模型,其中随机森林预测效果最好;模型中变量的重要性排序依次为:身体质量指数(body mass index, BMI)、$CD3^+$T 细胞、$CD16^+56^+$NK 细胞、$CD8^+$T 细胞(细胞毒性 T 细胞)、年龄、%$CD3^+$T 细胞、$CD19^+$B 细胞、$CD4^+$T 细胞(辅助性 T 细胞)、性别、贫血、糖尿病、白细胞减少、低蛋白血症、自身免疫性疾病,其中 BMI 和 $CD3^+$T 细胞、$CD16^+56^+$NK 细胞和 $CD8^+$T 细胞(细胞毒性 T 细胞)贡献度最大。**结论**　本研究构建的机器学习模型通过结合淋巴细胞亚群及临床特征,成功区分了活动性结核、NTM-PD 及其他肺部疾病,为肺部疾病的早期诊断和个性化治疗提供了新的思路和方法。

【关键词】活动性肺结核;淋巴细胞亚群;机器学习

【中图分类号】R52　**【文献标志码】**A　**【文章编号】**2097-4345(2025)06-0848-09

Application and performance of machine learning models integrating lymphocyte subsets and clinical features in: discriminating NTM-PD, pulmonary tuberculosis and other lung diseases

WANG Lei[1], *CAO Jie*[1], *LIU Zhibin*[1], *CHENG Liping*[1], *WU Xiaocui*[2], *SUN Qin*[1*], *SHA Wei*[1*]

(1. Clinical and Research Center for Tuberculosis, Shanghai Key Laboratory of Tuberculosis, Shanghai Pulmonary Hospital, School of Medicine, Tongji University, Shanghai 200433, China; 2. Department of Clinical Laboratory, Shanghai Pulmonary Hospital, School of Medicine, Tongji University, Shanghai 200433, China)

【Abstract】 Objective　Based on lymphocyte subset count indicators, diagnostic models were

收稿日期:2025-08-15　**录用日期:**2025-09-14
基金项目:上海市卫生健康委员会青年项目(20204Y0325)
作者简介:王　蕾(1989—),女,主治医师,硕士研究生,E-mail:17602155688@163.com
通信作者:孙　勤,E-mail:sunqinbongjour@163.com;沙　巍,E-mail:Wei Sha, shfksw@126.com;* 为共同通信作者
©同济大学(开放获取 CC BY-NC-ND 协议)

constructed using different machine learning methods to distinguish non-tuberculous mycobacterial pulmonary disease (NTM-PD), pulmonary tuberculosis (PTB), and other common confounding pulmonary diseases, to provide a scientific basis for the early identification of infectious pulmonary diseases. **Methods** The patients diagnosed with active tuberculosis (ATB), NTM-PD, or other pulmonary diseases (including inflammatory and neoplastic conditions) admitted to the Department of Tuberculosis at Shanghai Pulmonary Hospital from January to December in 2023 were included in this study. Lymphocyte subset counts were measured using flow cytometry. Four machine learning algorithms—multinomial logistic regression, naive Bayes, random forest, and XGBoost—were employed for model development. Hyperparameter tuning was performed using Bayesian optimization and cross-validation. The variables with $P<0.1$ from univariate analysis were selected and further refined via correlation analysis and LASSO for final model input. The models were evaluated using area under the receiver operating characteristic curve (AU-ROC), average precision-precision recall curve (AP-PR), and decision curve analysis (DCA) curves on the test set. **Results** A total of 1 383 patients were included, with 836 cases in the ATB group, 254 in the NTM group, and 293 in the OTHER group. Using selected demographic data, comorbidities, and lymphocyte subset indices as input variables and disease category as the outcome variable, four machine learning models were successfully constructed. Among them, the random forest model demonstrated the best predictive performance; the top contributing variables in the models were body mass index (BMI), $CD3^+$ T cells, $CD16^+56^+$ NK cells, $CD8^+$ T cells (cytotoxic T cells), age, %$CD3^+$T cells, $CD19^+$B cells, $CD4^+$T cells (helper T cells), gender, anemia, diabetes, leukopenia, hypoproteinemia, and autoimmune disease; and BMI, $CD3^+$T cells, $CD16^+56^+$NK cells, and CD^+T cells (cytotoxic T cells) contributed most significantly. **Conclusion** The machine learning models developed in this study successfully differentiated ATB, NTM-PD, and other pulmonary diseases by integrating lymphocyte subset profiles with clinical features. These models provide novel approaches for the early diagnosis and personalized management of pulmonary diseases.

【Key words】 active pulmonary tuberculosis; lymphocyte subsets; machine learning

结核病(tuberculosis, TB)是全球致死率最高的传染性疾病之一[1-2]。尽管近年来抗结核药物和治疗策略不断进步,但其在全球范围内的发病率与死亡率依旧居高不下,构成严重的公共卫生负担[3]。与此同时,非结核分枝杆菌肺病(nontuberculous mycobacterial pulmonary disease, NTM-PD)发病率也呈持续上升趋势,尤其在免疫功能低下人群中更为常见[4],已成为继结核病之后的重要肺部慢性感染性疾病,对临床管理提出了新的挑战[5]。NTM-PD 与活动性结核病(active tuberculosis, ATB)在临床症状和影像学方面高度相似[6-7],在临床上极易混淆。

随着人工智能技术的发展,机器学习在医学影像与临床诊断中的应用日益广泛,为整合多维度数据(包括临床表现、免疫学指标、影像学特征等)提供了新的技术路径。Zhou 等[8]通过提取 CT 影像组学特征,构建了包括随机森林(random forest, RF)、支持向量机(support vector machine, SVM)、逻辑回归(Logistic regression, LR)与极端梯度提升(eXtreme Gradient Boosting, XGBoost)在内的 4 种监督学习分类器,以区分 NTM-PD 与肺结核(pulmonary tuberculosis, PTB),结果表明 SVM 模型在疾病识别方面表现最为稳定。Liu 等[9]分析了 2019 年 1 月—2023 年 6 月住院的 NTM-PD 与 PTB 患者,基于免疫与影像学特征构建逻辑回归诊断模型,其受试者工作特征曲线(receiver operating characteristic curve, ROC)下面积为 0.938,灵敏度和特异度分别达 0.835 与 0.911,显示出良好的诊断能力。另一项基于 892 例患者的大样本研究利用多项实验室指标构建机器学习模型发现,RF 模型可有效识别结核分枝杆菌(*mycobacterium tuberculosis*, Mtb)感染状态,提示此类算法具有临床推广潜力[10]。但需注意的是,已有研究纳入的变量过多,数据处理与临床应用的可行性存在挑战。近年来研

究发现,ATB 与 NTM 感染在诱导宿主免疫应答方面具有显著差异。例如,ATB 患者常伴 $CD4^+$ T 细胞耗竭与自然杀伤细胞(natural killer cell, NK)功能障碍,而 NTM-PD 则可能以 Th17 细胞相关免疫激活为主。2024 年一项前瞻性、单中心研究采用流式细胞术(flow cytometry, FC)与酶联免疫斑点检测(enzyme-linked immunospot assay, ELISA)对 NTM-PD 患者($n=18$)与健康对照($n=22$)进行免疫分型,发现 NTM 患者在纯蛋白衍生物(purified protein derivative, PPD)刺激下,其 $CD4^+/CD8^+$T 细胞共表达 CD25、CD134 的比例显著升高,提示淋巴细胞亚群在其免疫应答中发挥关键作用[11]。尽管淋巴细胞亚群计数已被广泛用于结核病活动性与疗效监测,其在 ATB 与 NTM-PD 鉴别诊断中的潜在价值尚未得到系统评估。

基于上述背景,本研究拟建立一种创新的临床辅助诊断策略:基于流式细胞术获取的淋巴细胞亚群参数,联合患者的人口学特征及合并症信息,构建多分类机器学习模型,实现 ATB、NTM-PD 及其他常见肺部疾病(如肺部炎症与肿瘤)的高效鉴别诊断。该研究有望为复杂肺部感染性疾病的精准诊疗提供理论支持与实践依据。

1　资料与方法

1.1　一般资料

本研究为回顾性病例对照研究,收集 2023 年 1—12 月在同济大学附属上海肺科医院结核科收治的肺部影像异常患者临床资料。采用流式细胞术检测淋巴细胞亚群计数并进行数据分析。纳入标准:(1) 年龄 ≥ 18 岁;(2) 诊断为活动性肺结核(PTB)、NTM-PD 或其他相关肺部疾病(如炎症性肺病和肺肿瘤);(3) 基于病理检查或影像学(如胸片、CT)明确肺部疾病诊断;(4) 具备完整病历资料,便于数据分析和预后追踪。排除标准:(1) 免疫相关性疾病;(2) 接受器官移植或使用免疫抑制剂、长期类固醇药物者;(3) 人类免疫缺陷病毒(human immunodeficiency virus, HIV)感染或艾滋病患者;(4) PTB 合并 NTM 感染;(5) NTM-PD 合并 PTB。本研究获同济大学附属上海肺科医院伦理委员会批准(伦理号:K20-265Y)。

1.2　诊断标准

1.2.1　肺结核　诊断标准按照《中国肺结核病防治工作技术指南》[12]执行。临床症状:咳嗽、咳痰、低热、盗汗、体重减轻等;胸部影像学:胸片或 CT 显示肺部病变(如空洞或浸润影);分枝杆菌分离或分子检测:痰、气管吸取物或支气管灌洗液中通过培养或核酸扩增检测到结核分枝杆菌。

1.2.2　非结核分枝杆菌肺病　参考《非结核分枝杆菌病诊断与治疗指南(2020 年版)》[13]进行判定。临床症状与肺结核相似但通常较轻,呈慢性病程;胸部影像学 CT 可见结节性病变、空洞、肺泡浸润或支气管扩张;微生物学检查痰或支气管灌洗液中分离出 NTM,并结合临床和影像学证据确诊。

1.3　方法

1.3.1　样本制备　取绝对计数管并编号,向管中加入 20 μL 6 色 TBNK 抗体试剂,注意避免触碰底部金属垫片及微球。采用反向加样法加入 50 μL 混匀后的样本,同样避免接触管底金属垫片及微球,涡旋混匀后室温避光孵育 15 min。孵育完成后,向样本管中加入 450 μL 1×FACS Lysing Solution,再次涡旋混匀,避光条件下裂解 15 min。最后于 6 h 内上机检测,使用 FACSCanto Clinical 软件获取至少 2 500 个淋巴细胞进行分析,上机前需充分混匀样本以确保检测准确性。

1.3.2　流式细胞术分析　染色细胞在 FACSAria™ Ⅲ流式细胞分选仪(美国 BD 公司)上分析。数据采集采用 FACSDiva 软件(2.0 版,美国 BD 公司)记录 150 个事件/样本。多参数流式细胞术数据分析使用 FlowJo 软件(version 10.8.1,FlowJo LLC, US)进行。在所有 3 个面板中,首先利用 FSC-A *vs* SSC-A 散点图对细胞进行门控以识别淋巴细胞群,然后通过 SSC-A 与 Zombie Aqua 荧光信号图筛选活细胞,最后利用 FSC-H *vs* FSC-W 散点图排除粘连体,确保最终分析为活的单个淋巴细胞。本研究所采用的流式细胞术门控策略主要参考 BD 标准操作手册中描述的方法。记录细胞群体的数据为细胞百分比或中位荧光强度(MFI),导出为 csv 文件进行统计分析。

1.3.3　各组间淋巴细胞亚群绝对计数　测定 $CD3^+$ T 细胞、$CD4^+$T 细胞、$CD8^+$T 细胞、$CD19^+$B 细胞、$CD16^+56^+$NK 细胞绝对值,使用 FACS 流式细胞仪对经红细胞裂解处理后的外周血样本中的 T 淋巴细胞、NK 细胞、NK 样 T 淋巴细胞和 B 淋巴细胞进行分析。

1.4　统计学处理

采用 R 4.2.1 软件对数据进行统计分析。若比

较变量为连续变量，根据数据分布分别使用 t 检验和 Wilcoxon 秩和检验。若为分类变量，则根据分布分别使用卡方检验和 Fisher 精确概率检验。使用多分类 Logistic 回归、朴素贝叶斯、随机森林、XGBoost 4 种算法进行建模预测，使用贝叶斯和交叉验证方式进行算法超参数优化。从开发集中的差异性分析筛选 $P<0.1$ 变量，使用相关性分析和 lasso 进行特征筛选后进入建模。按照 7∶3 比例随机划分方式将数据集划分为开发集和测试集。以筛选的显著指标为输入变量，以疾病类别为结果变量，构建多分类 Logistic 回归、朴素贝叶斯、随机森林、XGBoost 4 种机器学习模型。使用 AU-ROC、AP-PR 和 DCA 曲线对模型在测试集上的表现进行评价。统计检验均为双侧检验，$P<0.05$ 为差异具有统计学意义。

2　结　　果

2.1　患者基本情况

本研究共纳入 1 383 例患者，其中结核组 836 例，NTM 肺病组 254 例，其他组 293 例；3 组在性别、年龄、身高、体质量、身体质量指数（body mass index，BMI）、糖尿病、贫血、白细胞减少、低蛋白血症、自身免疫性疾病、%$CD3^+$T 细胞、$CD3^+$T 细胞、$CD4^+$T 细胞、$CD8^+$T 细胞、$CD19^+$B 细胞、$CD16^+56^+$NK 细胞整体差异均有统计学意义（$P<0.05$），见表 1。

表 1　患者基本情况对比

Tab. 1　Comparison of basic information in patients among groups　［n(%)，$\bar{x}\pm s$，$M(P_{25},P_{75})$］

变量	结核组（n=836）	NTM 肺病组（n=254）	其他组（n=293）	整体（n=1 383）
年龄/岁	59.0(48.0,66.0)*	60.0(55.0,69.0)#	59.0(50.0,62.0)	59.0(50.0,66.0)
性别				
女	329(39.4)*#	173(68.1)#	166(56.7)	668(48.3)
男	507(60.6)	81(31.9)	127(43.3)	715(51.7)
身高/cm	159.2(153.2,166.3)#	159.7(152.5,166.0)#	161.6(155.6,167.9)	159.8(153.6,166.5)
体质量/kg	63.1(55.4,70.9)*	66.5(59.1,74.4)	65.0(58.5,72.0)	64.2(56.5,72.0)
BMI/(kg·m^{-2})	23.6(22.4,24.8)*	24.4(23.0,25.7)#	23.4(22.6,24.2)	23.7(22.6,24.8)
BCG 接种	769(92.0)	231(90.9)	268(91.5)	1 268(91.7)
糖尿病	108(12.9)*#	16(6.3)#	0	124(9.0)
贫血	168(20.1)#	47(18.5)#	2(0.7)	217(15.7)
白细胞减少	56(6.7)#	14(5.5)#	0	70(5.1)
低蛋白血症	41(4.9)#	6(2.4)#	0	47(3.4)
自身免疫性疾病	5(0.6)*	6(2.4)#	0	11(0.8)
肺外结核	242(28.9)*#	2(0.8)	0	244(17.6)
肺外结核部位				
关节	8(3.3)	/	/	8(3.3)
其他	18(7.4)	/	/	18(7.4)
消化道	3(1.2)	/	/	3(1.2)
淋巴结	47(19.4)	1(50.0)	/	48(19.7)
生殖系统	5(2.1)	/	/	5(2.0)
胸膜	110(45.5)	/	/	110(45.1)
脑膜	21(8.7)	/	/	21(8.6)
腹膜	8(3.3)	1(50.0)	/	9(3.7)
骨	22(9.1)	/	/	22(9.0)
$CD3^+$T 细胞	895.0(666.3,1 121.8)#	892.2(676.3,1 101.7)#	1078.5(910.4,1 357.2)	925.0(705.2,1 192.6)
%$CD3^+$T 细胞	69.6±8.3	68.6±8.2#	70.9±7.2	69.7±8.1
$CD4^+$T 细胞	540.9(428.8,631.5)#	540.9(432.0,621.1)#	607.4(529.2,726.5)	540.9(447.1,657.7)
%$CD4^+$T 细胞	41.3±8.4	41.1±7.9	41.9±7.3	41.4±8.1
$CD8^+$T 细胞	346.8(275.9,425.9)#	346.8(273.5,402.3)#	366.7(304.9,513.4)	346.8(278.4,446.2)
%$CD8^+$T 细胞	24.6(19.6,29.4)	24.2(19.4,29.1)	24.6(20.0,29.5)	24.6(19.6,29.4)
CD4/CD8	1.8(1.6,2.3)	1.8(1.6,2.4)	1.8(1.6,2.2)	1.8(1.6,2.3)
$CD19^+$B 细胞	170.9(132.3,193.7)#	170.9(130.7,203.0)#	179.6(152.2,244.9)	170.9(136.7,209.1)
%$CD19^+$B 细胞	12.6(8.9,16.7)	13.0(9.6,17.7)	12.8(9.9,15.8)	12.7(9.3,16.5)
$CD16^+56^+$NK 细胞	190.4(142.9,233.4)#	190.4(140.3,251.4)#	194.8(169.6,267.8)	190.4(146.5,247.3)
%$CD16^+56^+$NK 细胞	14.7(11.0,20.0)	14.7(11.9,20.8)	14.7(11.7,18.9)	14.7(11.3,19.9)

与 NTM 肺病组相比，*$P<0.05$；与其他组相比，#$P<0.05$；表中“/”为未测定

2.2　变量筛选

使用多分类 Logistic 回归、朴素贝叶斯、随机森林、XGBoost 4 种算法进行建模预测,使用贝叶斯和交叉验证方式进行算法超参数优化。从开发集中的差异性分析筛选 $P<0.1$ 特征变量进入该部分分析。分析纳入变量的相关性,相关性较大的只保留最有临床意义的一个,见图 1。在经过差异性分析和相关性分析变量筛选后,符合条件的使用 LASSO+交叉验证(5 折)进行变量筛选。图 2 为 LASSO 筛选变量组合以及误分率的关系。*Y* 轴为误分率,*X* 为 lambda 函数值。从图中可以看到,随着 lambda 增大,变量数量在减少,而误分率在增大。图中左侧第一条虚线表示最佳 lambda 值,在该点变量数量较少,而误分率最低。第二条虚线为最佳 lambda 1se 处位置,在该点变量数量进一步减少,而误分率相比前者无显著下降。图 3 为以 ATB 为例,预测患者各变量系数路径图,*Y* 轴为系数大小,*X* 轴为 lambda 大小,可见随着 lambda 增大,各变量系数在逐渐减小。最后共计筛选了 14 个特征进入建模,包括 BMI、$CD3^+$T 细胞、$CD16^+56^+$NK 细胞、$CD8^+$T 细胞(细胞毒性 T 细胞)、年龄、%$CD3^+$T 细胞、$CD19^+$B 细胞、$CD4^+$T 细胞(辅助性 T 细胞)、性别、贫血、糖尿病、白细胞减少、低蛋白血症、自身免疫性疾病。

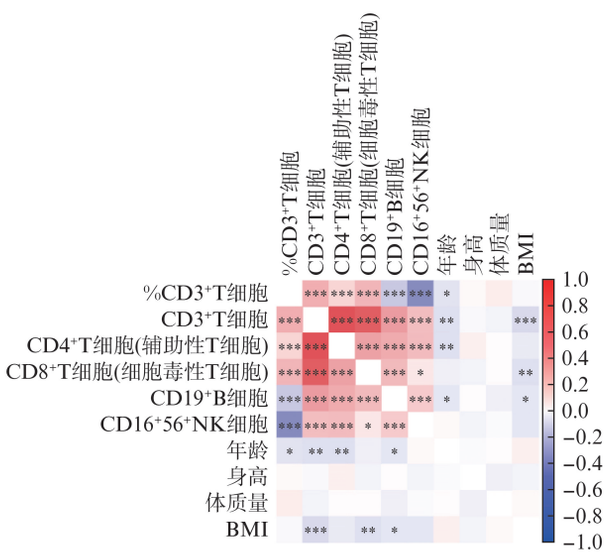

图 1　相关性分析

Fig. 1　Correlation analysis

2.3　建立预测模型

按照 7:3 比例根据疾病类别随机划分方式将数据集划分为开发集和测试集。训练集 968 例、验证集 462 例。训练集和验证集患者人口学信息、合并疾病、淋巴细胞亚群等方面均无统计学差异。以筛选出的人口学信息、合并疾病、淋巴细胞亚群指标为输入变量,以疾病类别为结果变量,成功构建多分类 Logistic 回归、朴素贝叶斯、随机森林、XGBoost 4 种机器学习模型。

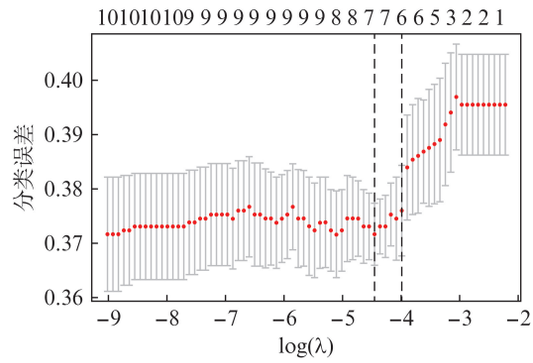

图 2　交叉验证误差曲线图

Fig. 2　Cross-validation error curve

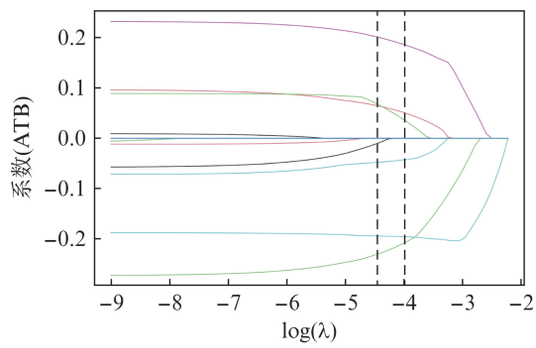

图 3　系数路径图(因变量为 ATB)

Fig. 3　Coefficient path plot(dependent variable: ATB)

2.4　模型性能评估

4 种模型在测试集各个指标上随机森林效果最佳。结局为 3 分类疾病诊断,测试集中随机森林预测疾病和真实疾病的混淆矩阵见图 4。*Y* 轴为预测疾病类型,*X* 轴为真实疾病,图例为患者数量大小。以左上角 252 表示测试集中有 364 例患者患有 ATB,而随机森林模型正确预测 252 例。ATB 组预测灵敏度 0.692 3,特异度 0.724 5,NTM 组预测灵敏度 0.818 2,特异度 0.847 7,其他组灵敏度 0.631 6,特异度 0.870 5,见表 2。随机森林模型在测试集上 3 种疾病的受试者工作特征曲线下面积(area under the receiver operating characteristic curve, AUROC)见图 5。图中可以看出,模型对 3 种疾病有不错的识别能力。其中,识别其他组疾病能力最强。随机森林模型在测试集上 3 种疾病的平均精度-精

确率召回曲线（average precision-precision recall curve，AP-PR）见图 6。从该图中可以看出 3 种疾病分布并不均衡，ATB 患者更多，而 NTM 和其他组患者数量较少，故而这两类的 PR 曲线相比 ATB 稍差，但是平均精准度显示其有不错的判别能力。随机森林模型在测试集中识别 ATB 临床效用评估见图 7，从图中可以看出，该模型在阈值 0~1 的范围内其临床效用均明显大于 Treat All 和 Treat None，显示出较大的临床效用。随机森林在测试集上 ATB 的一致性评价见图 8，反映模型预测风险与实际发生风险的一致程度与一致性。从图中可以看出校准曲线几乎和左下到右上的对角线重叠，显示出训练出的随机森林模型有较高的一致性。

2.5　模型解释

模型中变量的重要性排序依次为：BMI、$CD3^+$ T 细胞、$CD16^+56^+$ NK 细胞、$CD8^+$ T 细胞、年龄、%$CD3^+$T 细胞、$CD19^+$B 细胞、$CD4^+$T 细胞、性别、贫血、糖尿病、白细胞减少、低蛋白血症、自身免疫性疾病，其中 BMI 和 $CD3^+$T 细胞、$CD16^+56^+$NK 细胞和 $CD8^+$T 细胞贡献度最大。

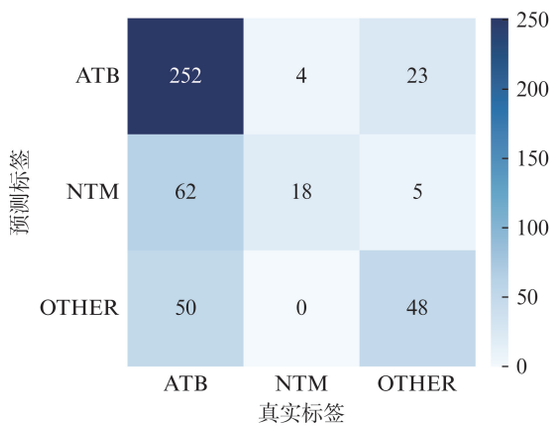

图 4　混淆矩阵热力图（测试集）

Fig. 4　Confusion matrix heatmap (test set)

表 2　灵敏度和特异度

Tab. 2　Sensitivity and specificity

类别	TP	FN	FP	TN	灵敏度	特异度
ATB	252	112	27	71	0.692 3	0.724 5
NTM	18	4	67	373	0.818 2	0.847 7
其他疾病	48	28	50	336	0.631 6	0.870 5

TP（true positive）：模型正确预测的阳性；FP（false positive）：模型错误预测的阳性；FN（false negative）：模型错误预测的阴性；TN（true negative）：模型正确预测的阴性；灵敏度＝TP／（TP+FN）；特异度＝TN/（TN+FP）

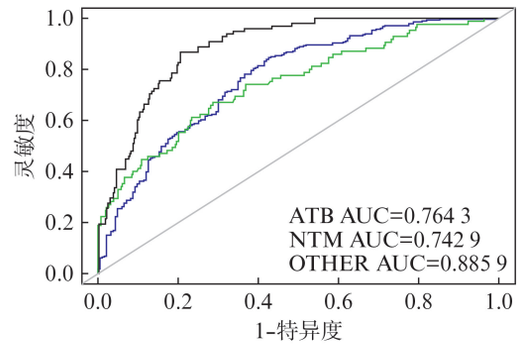

图 5　随机森林在测试集上的 ROC 曲线

Fig. 5　ROC curve of random forest on the test set

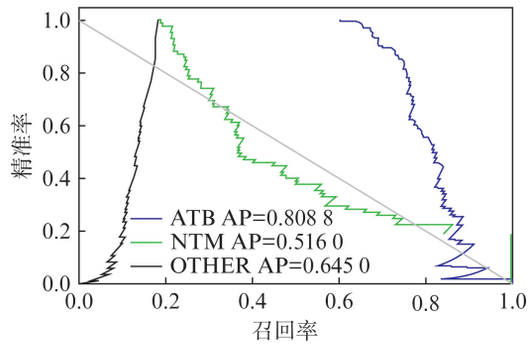

图 6　精准率和召回率分析曲线

Fig. 6　Precision-recall analysis curve

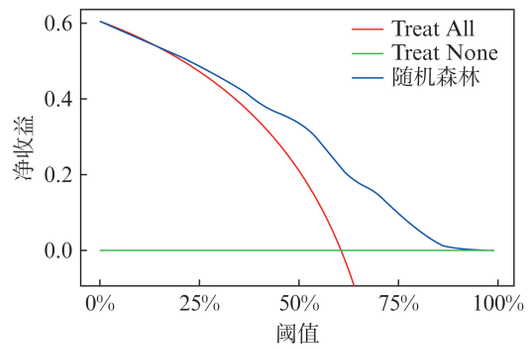

图 7　因变量为 ATB 的随机森林决策分析曲线

Fig. 7　Decision analysis curve of random forest with dependent variable ATB

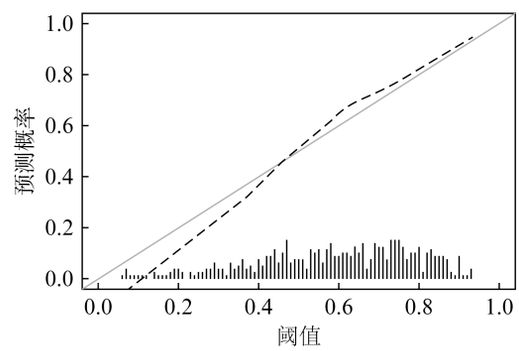

图 8　因变量为 ATB 的校准曲线

Fig. 8　Calibration plot for the dependent variable ATB

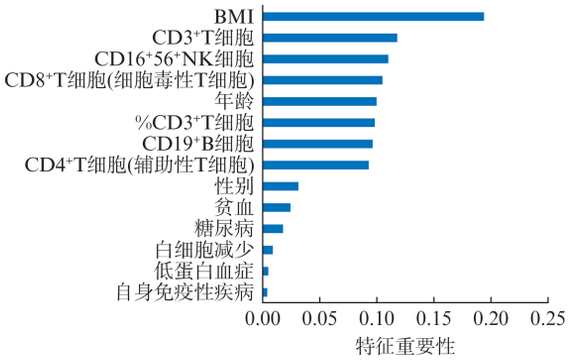

图 9　随机森林特征重要性排序

Fig. 9　Random forest feature importance ranking

3　讨　　论

本研究成功构建了基于机器学习的分类模型,包括多分类 Logistic 回归、朴素贝叶斯、随机森林和 XGBoost,用于对 ATB、NTM-PD 及其他肺部疾病进行分类预测。从整体模型表现来看,本研究构建的多种机器学习模型均展现出较强的分类能力,其中随机森林模型表现最为优异。该模型不仅具备较高的分类准确性与稳定性,能够精准识别 3 类疾病,还展现出良好的一致性。在多变量条件下,随机森林模型表现出良好的适应性与解释能力,能够敏锐捕捉与特定肺部疾病密切相关的免疫细胞群体变化,助力临床医生在疾病早期识别潜在免疫异常。通过精准识别关键免疫指标,该模型有望实现肺部疾病的早期筛查,在疾病尚未出现典型临床症状时及时预警,从而提升医院诊断效率与资源利用率,推动疾病的早期干预与精准治疗。随机森林模型的优势在于其卓越的高维数据处理能力,对缺失值具有较强的鲁棒性,通过集成多个决策树有效降低过拟合风险,显著提高了模型的泛化能力与预测准确性。另一方面,该模型能够输出变量重要性排序,帮助研究人员快速定位对疾病分类影响最大的特征变量,为临床挖掘更具指导意义的潜在生物标志物提供有力支持[14]。

在模型变量重要性排序中,BMI 的重要性位居前列,表明免疫功能状态与营养水平是区分不同肺部疾病的重要因素。BMI 作为反映机体营养状况和整体健康水平的指标,在肺部感染性疾病的识别中具有重要意义。研究发现,BMI 的异常变化与免疫功能紊乱、慢性炎症反应密切相关,尤其在 ATB 和 NTM-PD 患者中表现明显。在本研究中,NTM-PD 组的 BMI 水平最高,其次为 ATB 组,其他肺病组最低。低 BMI 通过蛋白质能量缺乏直接损害胸腺微环境,减少原始 T 细胞输出,导致外周血 $CD3^+$ T 细胞、$CD4^+$T 细胞数量减少,削弱抗结核免疫导致 ATB 发生,同时 Th1 型免疫应答被削弱而促进结核潜伏感染再激活。而过高的 BMI 也容易导致免疫代谢失调,由于内脏脂肪堆积引发慢性低度炎症,持续升高的白细胞介素-6(inter leukin-6, IL-6)和肿瘤坏死因子-α(tumor necrosis factor-α, TNF-α)抑制 $CD16^+56^+$NK 细胞的细胞毒性,NTM 感染的肺泡上皮细胞和巨噬细胞无法被有效清除,促进 NTM 的胞内寄生和扩散[15]。BMI 的分布趋势显著增强了模型对疾病类别的判别能力,表明其在疾病早期识别中的潜在应用价值。

除 BMI 外,$CD3^+$T 细胞、$CD16^+CD56^+$NK 细胞及 $CD8^+$T 细胞对模型的贡献度也较高,进一步强调了淋巴细胞亚群在鉴别 ATB、NTM-PD 与其他肺病中的诊断价值。本研究中,ATB 组与 NTM-PD 组上述 3 类细胞的水平差异不显著,但两者均低于其他肺病组,提示慢性感染状态下的免疫抑制在这两种疾病中均较为突出。既往研究亦有类似发现,PTB 患者的 $CD3^+$T 细胞、$CD16^+CD56^+$NK 细胞、$CD4^+$T 细胞和 $CD19^+$B 细胞水平较健康人显著降低[16-17]。Liu 等[9]的研究也证实,与健康对照组相比,ATB 和 NTM-PD 患者的 T 细胞、B 细胞及 NK 细胞的绝对数量显著减少,提示其细胞免疫和体液免疫功能均受到抑制,反映了淋巴细胞亚群水平与 ATB 和 NTM 疾病之间具有较强关联。具体而言,$CD3^+$T 细胞作为 T 细胞的总标志,在抗感染免疫中发挥核心作用。在 ATB 患者中,$CD3^+$T 细胞呈现出更强的活化状态,分泌干扰素-γ(interferon-γ, IFN-γ)和 TNF-α 等细胞因子以介导免疫反应,同时其活化标志物如人类白细胞抗原-DR(human leukocyte antigen-D-related, HLA-DR)水平显著升高[18]。相比之下,NTM-PD 患者中 $CD3^+$T 细胞数量和功能活性明显下降,表明其免疫应答较弱,可能与免疫逃逸机制有关[19]。此外,研究还发现,$CD3^+$T 细胞的功能状态(如 CD25、CD134 表达)有助于区分 NTM-PD 的进展性与非进展性亚型[11]。NK 细胞在先天免疫中具有关键作用,其通过增强巨噬细胞活性、直接杀伤感染病原体等方式参与抗分枝杆菌感染。研究发现,在活动性结核中,$CD16^+CD56^+$NK 细胞表现出高度活化状态,伴随颗粒酶表达增加以及自然杀伤细胞 p30(natural killer cell p30, NKp30)、自然杀伤细

胞 p46(natural killer cell p46, NKp46)等受体密度升高[20]。而在 NTM-PD 患者中,NK 细胞的 IFN-γ 产生能力和活化状态减弱,提示其功能受损,影响早期免疫控制[21]。$CD8^+$T 细胞(细胞毒性 T 细胞)主要通过穿孔素和颗粒酶杀伤感染细胞,并分泌 IFN-γ 和 TNF-α 激活巨噬细胞,是控制 ATB 感染的重要免疫细胞[22]。然而在 NTM-PD 患者中,$CD8^+$T 细胞(细胞毒性 T 细胞)功能明显下降,表现为 IFNγR1 下调以及 TIM-3 抑制性受体上调,提示其处于免疫耗竭状态[23]。Kudryavtsev 等[24]也指出,$CD8^+$T 细胞(细胞毒性 T 细胞)参与 ATB 的感染控制过程,功能状态的差异可为鉴别肺部疾病提供依据。

本研究构建的基于淋巴细胞亚群的机器学习诊断模型,利用流式细胞检测的常规淋巴细胞亚群数据,整合 BMI、性别年龄等易获取的临床指标,与 Liu 等[9]采用的 ELISA 联合 CT 检测方案、Zhou 等[8]构建的影像组学模型相比,将检测时间更短,检测成本低,避免辐射暴露风险的同时保持了较好的敏感度和特异度,实现了 NTM-PD、PTB 及其他肺部疾病的高效鉴别。然而本研究未纳入影像学指标,也导致在疾病定位和形态学鉴别方面存在一定局限,未来研究可进一步引入更多临床变量、影像学特征及多组学数据,以进一步提升鉴别诊断的准确性。其次本研究为单中心回顾性研究,尽管通过严格的入排标准和内部验证确保了模型的稳健性,但结果仍可能存在选择偏倚。未来需要通过多中心外部验证进一步验证模型的泛化能力,联合不同地域、不同流行特征的医疗机构数据,以评估模型在更广泛人群中的适用性。

综上所述,本研究通过整合淋巴细胞亚群计数和机器学习方法,不仅为 ATB、NTM-PD 与其他肺部疾病的鉴别诊断提供了新的技术路径,也为个体化免疫评估、治疗决策支持及疾病预后监测提供了理论依据和方法支撑。该模型在早期筛查、辅助诊断、免疫干预及资源优化配置等方面具有重要的临床应用潜力。

利益冲突声明　所有作者声明不存在利益冲突。

作者贡献说明　王蕾：研究的构思与设计、论文初稿的撰写以及审阅和编辑；程丽平、刘轾彬：方法学的开发与设计；曹婕：研究资源的提供；吴小翠：实验结果的验证；沙巍、孙勤：论文的审阅与修改；所有作者均阅读并批准了最终稿件。

【参考文献】

[1] JANSSEN S, MURPHY M, UPTON C, et al. Tuberculosis: an update for the clinician[J]. Respirology, 2025,30(3): 196-205.

[2] 杜姗姗. 肺结核短疗程治疗的研究进展[J]. 同济大学学报(医学版),2024,45(2): 303-309.

[3] NATHAVITHARANA R R, JIJON D F, PAL P, et al. Diagnosing active tuberculosis in primary care[J]. Bmj, 2021,374: n1590.

[4] DAHL V N, MØLHAVE M, FLØE A, et al. Global trends of pulmonary infections with nontuberculous mycobacteria: a systematic review[J]. Int J Infect Dis, 2022,125: 120-131.

[5] PREVOTS D R, MARSHALL J E, WAGNER D, et al. Global epidemiology of nontuberculous mycobacterial pulmonary disease: a review[J]. Clin Chest Med, 2023,44(4): 675-721.

[6] DALEY C L, IACCARINO J M, LANGE C, et al. Treatment of nontuberculous mycobacterial pulmonary disease: an official ATS/ERS/ESCMID/IDSA clinical practice guideline[J]. Eur Respir J, 2020,56(1): e1-e36.

[7] KUMAR K, PONNUSWAMY A, CAPSTICK T G, et al. Non-tuberculous mycobacterial pulmonary disease (NTM-PD): epidemiology, diagnosis and multidisciplinary management[J]. Clin Med(Lond), 2024,24(1): 100017.

[8] ZHOU L, WANG Y, ZHU W, et al. A retrospective study differentiating nontuberculous mycobacterial pulmonary disease from pulmonary tuberculosis on computed tomography using radiomics and machine learning algorithms[J]. Ann Med, 2024,56(1): 2401613.

[9] LIU Q, PAN X, AN H, et al. Building a model for the differential diagnosis of non-tuberculous mycobacterial lung disease and pulmonary tuberculosis: a case-control study based on immunological and radiological features[J]. Int Immunopharmacol, 2023,124(Pt B): 111064.

[10] LUO Y, XUE Y, LIU W, et al. Development of diagnostic algorithm using machine learning for distinguishing between active tuberculosis and latent tuberculosis infection[J]. BMC Infect Dis, 2022,22(1): 965.

[11] MARTY P K, PATHAKUMARI B, COX T M, et al. Multiparameter immunoprofiling for the diagnosis and differentiation of progressive versus nonprogressive

nontuberculous mycobacterial lung disease-a pilot study [J]. PLoS One, 2024,19(4): e0301659.

[12] 中国疾病预防控制中心结核病预防控制中心. 中国结核病防治工作技术指南[M]. 北京: 人民卫生出版社,2021: 388.

[13] 中华医学会结核病学分会. 非结核分枝杆菌病诊断与治疗指南(2020年版)[J]. 中华医学会结核病学分会中华结核和呼吸杂志,2020,43(11): 918-946.

[14] UDDIN S, KHAN A, HOSSAIN M E, et al. Comparing different supervised machine learning algorithms for disease prediction [J]. BMC Med Inform Decis Mak, 2019,19(1): 281.

[15] KANG J Y, HAN K, KIM M K. Severity of underweight affects the development of nontuberculous mycobacterial pulmonary disease; a nationwide longitudinal study[J]. Sci Rep, 2022,12(1): 17180.

[16] MORAIS-PAPINI T F, COELHO-DOS-REIS J G A, WENDLING A P B, et al. Systemic immunological changes in patients with distinct clinical outcomes during *Mycobacterium tuberculosis* infection [J]. Immunobiology, 2017,222(11): 1014-1024.

[17] LIU Q, DU J, AN H, et al. Clinical characteristics of patients with non-tuberculous mycobacterial pulmonary disease: a seven-year follow-up study conducted in a certain tertiary hospital in Beijing[J]. Front Cell Infect Microbiol, 2023,13: 1205225.

[18] AKTAS E, CIFTCI F, BILGIC S, et al. Peripheral immune response in pulmonary tuberculosis[J]. Scand J Immunol, 2009,70(3): 300-308.

[19] KOH J, KIM S, KIM J Y, et al. Immunologic features of nontuberculous mycobacterial pulmonary disease based on spatially resolved whole transcriptomics [J]. BMC Pulm Med, 2024,24(1): 392.

[20] ALBAYRAK N, DIRIX V, AERTS L, et al. Differential expression of maturation and activation markers on NK cells in patients with active and latent tuberculosis[J]. J Leukoc Biol, 2022, 111(5): 1031-1042.

[21] LAI H C, CHANG C J, LIN C S, et al. NK Cell-Derived IFN-γ protects against nontuberculous mycobacterial lung infection [J]. J Immunol, 2018, 201(5): 1478-1490.

[22] LAZAREVIC V, NOLT D, FLYNN J L. Long-term control of Mycobacterium tuberculosis infection is mediated by dynamic immune responses[J]. J Immunol, 2005,175(2): 1107-1117.

[23] RATNATUNGA C N, TUNGATT K, PROIETTI C, et al. Characterizing and correcting immune dysfunction in non-tuberculous mycobacterial disease [J]. Front Immunol, 2022,13: 1047781.

[24] KUDRYAVTSEV I, ZINCHENKO Y, SEREBRIAKOVA M, et al. A key role of $CD8^+$T cells in controlling of tuberculosis infection[J]. Diagnostics(Basel), 2023,13(18): 2961.