DOI:10.16016/j.2097-0927.202403126

基于机器学习的早期胃癌淋巴结转移预测模型构建与验证

孟祥勇,秦嘉怡,陈文生 400038 重庆,陆军军医大学(第三军医大学)第一附属医院消化内科

[摘要] **目的** 通过机器学习构建最佳的早期胃癌淋巴结转移(lymphatic metastasis,LNM)预测模型,并对其性能进行验证。**方法** 收集2015年1月至2022年12月在本院行根治性手术治疗的433例早期胃癌患者的临床资料。以7∶3的比例划分为训练集和验证集。运用LASSO回归筛选变量,并基于多因素Logistic回归确定早期胃癌LNM的独立危险因素。应用类别型特征梯度提升机(categorical boosting,Catboost)、轻量的梯度提升机(light gradient boosting machine,LightGBM)、极端梯度提升机(eXtreme gradient boosting machine,XGboost)、随机森林(random forest,RF)、梯度提升树(gradient boosting machine,GBM)、神经网络(neural networks,NNET)、支持向量机(support vector machine,SVM)、K最临近(K-nearest,KNN)、朴素贝叶斯(naiveBayes,NB)及Logistic回归共10种ML算法构建预测模型。利用准确率、精确率、召回率、F1评分值、灵敏度、特异度、阳性预测率、阴性预测率、Kappa值、受试者工作特征曲线下面积值(AUC)、校准曲线、决策曲线以及精确率-召回率曲线评估和比较模型的预测能力。基于Shaply加性解释(SHapley additive exPlanations,SHAP)方法解释最佳模型中各变量对发生结局影响的贡献度。**结果** 肿瘤浸润深度、淋巴血管浸润和吸烟史是早期胃癌LNM的独立危险因素。Catboost模型具有最佳预测性能,其在训练集中的5个性能指标均优于其他模型[AUC:0.904(95%*CI* 0.868-0.940),F1评分:0.633,Brier评分:0.100,阴性预测率:0.975,Kappa:0.520]。通过计算Catboost的SHAP值发现,肿瘤浸润深度和淋巴血管浸润是预测LNM的2个关键特征变量。**结论** 肿瘤浸润深度为黏膜下层、淋巴血管浸润和吸烟史是早期胃癌LNM的独立危险因素。机器学习可用于预测LNM风险,Catboost模型具有最佳预测性能,并可为临床诊断和治疗决策提供指导。

[关键词] 早期胃癌;淋巴结转移;机器学习;预测模型

[中图法分类号] R319; R73-37; R735.2 [文献标志码] A

Construction and validation of a prediction model for lymph node metastasis in early gastric cancer based on machine learning

MENG Xiangyong, QIN Jiayi, CHEN Wensheng Department of Gastroenterology, First Affiliated Hospital, Army Medical University (Third Military Medical University), Chongqing, 400038, China

[**Abstract**] **Objective** To construct an optimal prediction model for lymph node metastasis (LNM) in early gastric cancer (EGC) using machine learning techniques and assess its predictive performance. **Methods** Clinical data of 433 EGC patients undergoing radical surgery in our hospital from January 2015 to December 2022 were collected. They were divided into a training set and a validation set in a 7∶3 ratio. LASSO regression was used to screen variables and multivariate logistic regression analysis was employed to identify independent risk factors for LNM in the EGC patients. Ten machine learning models were constructed using categorical boosting (Catboost), light gradient boosting machine (LightGBM), extreme gradient boosting machine (XGboost), random forest (RF), gradient boosting machine (GBM), neural networks (NNET), support vector machine (SVM), K nearest (KNN), Naive Bayes (NB) and Logistic regression. The predictive power of the above models was evaluated and compared in terms of accuracy, precision, recall,

[通信作者] 陈文生,E-mail:2303141134@qq.com

2097-0927/Copyright © 2024, 陆军军医大学(第三军医大学)。这篇开放获取文章遵循CC BY许可协议(https://creativecommons.org/licenses/by/4.0/)

F1 score value, sensitivity, specificity, positive predictive rate, negative predictive rate, Kappa value, area value under the receiver operating characteristic curve (AUC), calibration curve, decision curve, and precision-recall curve. Finally, SHAP (SHapley Additive exPlanations) was applied to explain the contribution of each variable in the best model for the prediction outcomes. **Results** Depth of tumor invasion, lymphovascular invasion and smoking history were independent risk factors for LNM in the EGC patients. Catboost model obtained the best predictive performance, and had 5 performance indicators outperforming the other models in the training set, that is, an AUC value of 0.904 (95%*CI* 0.868~0.940), a F1-score of 0.633, a Brier score of 0.100, a negative predictive rate of 0.975, and a Kappa value of 0.520. Finally, calculating the SHAP values of Catboost revealed that the depth of tumor invasion and lymphovascular invasion were two key characteristic variables for predicting LNM. **Conclusion** The depth of tumor invasion of submucosal and lymphovascular invasion and smoking history are independent risk factors for LNM in early gastric cancer. ML can be used to predict LNM risk, the Catboost model has the best predictive performance and can provide guidance for clinical diagnosis and treatment decisions.

[**Key words**] early gastric cancer; lymph node metastasis; machine learning; prediction model

Corresponding author: CHEN Wensheng, E-mail: 2303141134@qq.com

胃癌是我国最常见的恶性肿瘤之一,其致死率在所有癌症死因中位列第三[1]。早期胃癌(early gastric cancer,EGC)指肿瘤侵犯深度局限于黏膜层或黏膜下层,并且与淋巴结转移(lymphatic metastasis,LNM)无关。对于EGC患者,LNM是影响其治疗方案及预后评估的重要因素之一。在临床诊断和治疗中,常使用超声内镜、腹部增强CT或PET-CT以评估淋巴结状态。然而,上述影像学工具并不能完全证实或排除LNM[2]。既往研究报道,术前诊断为LNM的EGC患者中,很大一部分在淋巴结清扫术后的病理分析中证实为阴性,凸显了现有诊断方法的局限性。因此,开发新的诊断工具以评估术前EGC淋巴结状态对于选择合适的治疗方案和确定预后至关重要。

近年来,机器学习(machine learning,ML)算法在疾病预测模型的开发中展现出了巨大潜力。相较于传统的回归方法,ML以其强大的计算能力和自动数据驱动的特性,能够更为准确地预测多个非相关数据集的结果[3-4]。基于此,本研究旨在通过ML构建EGC淋巴结转移的预测模型,并验证其性能,以提供另一种评估方法指导临床诊疗方案的选择。

1 资料与方法

1.1 研究对象

本研究通过本院医学大数据平台,对2015年1月至2022年12月期间行手术治疗的EGC患者的一般资料及病理资料进行回顾性收集。入组标准:①经病理学检查明确诊断为EGC;②接受根治性手术治疗(包括胃切除术联合D2淋巴结清扫);③临床和病理资料完整。排除标准:①术前行新辅助治疗;②术后病理确诊为进展期胃癌;③原位多发肿瘤或合并其他系统恶性肿瘤。

1.2 临床特征

对纳入EGC患者的检验结果、内镜检查和病理报告进行预处理。包括肿瘤标志物(CA125,CA242,CEA)、肿瘤的位置(胃上1/3,胃中1/3,胃下1/3)、大小(≤2 cm,>2~3 cm,>3 cm)、内镜下分型(隆起型,平坦型,凹陷型)、侵犯深度(黏膜层,黏膜下层)、分化程度(高分化,中分化,低分化)、溃疡存在与否、淋巴血管浸润(lymphovascular invasion,LVI)和神经侵犯。同时收集患者的一般资料,包括性别、年龄、体质指数(BMI)以及吸烟史、饮酒史和胃癌家族史。吸烟史定义连续或累积吸烟6个月或以上,吸烟量>20支/d。饮酒史定义男性饮酒者日均酒精摄入量≥61 g,女性饮酒者日均酒精摄入量≥41 g[5]。

1.3 模型的建立

采用LASSO回归分析对纳入的变量进行筛选,通过多因素Logistic回归分析确定EGC合并LNM的独立危险因素。通过对文献报道中不同的ML算法进行综合回顾,并比较其可扩展性、灵活性、准确性和相对快速性,最终选择10种监督式ML算法用于构建预测模型。包括类别型特征梯度提升机(categorical boosting,Catboost)、极端梯度提升机(eXtreme gradient boosting,XGboost)、轻量梯度提升机(light gradient boosting machine,LightGBM)、随机森林(random forest,RF)、梯度提升树(gradient boosting machine,GBM)、支持向量机(support vector machine,SVM)、K最临近(K-nearest neighbor algorithm,KNN)、神经网络(neural network,NNET)、朴素贝叶斯(naiveBeyes,NB)以及

Logistic 回归(Logistic regression,LR)模型。为确保模型性能的最优化,每个模型均通过网格搜索与十折交叉验证来确定最佳超参数。

1.4 模型的评价与验证

在训练集和验证集中分别对 10 种模型的性能进行全面评估。采用准确率、精确率、召回率、F1-score、灵敏度、特异度、Brier score、阳性预测率(PPV)、阴性预测率(NPV)和 Kappa 值共 10 个指标来评价 ML 模型的性能。其中准确率、精确率、召回率、F1-score、灵敏度、特异度、阳性预测率、阴性预测率以及 Kappa 的值越高,表明模型性能越出色。相反,对于 Brier score,评分越接近 0,表示预测模型的准确性越高。一般来说,Brier 评分<0.15 可认为模型具有良好的一致性。

此外,以受试者工作特征(ROC)曲线下面积(AUC)作为主要评价指标,评估 10 个预测模型在区分度上的表现。AUC 值越高,模型的分类性能越出色。同时引入校准曲线计算给定模型的置信区间和预测概率;运用决策曲线分析(decision curve analysis,DCA)量化评估预测模型在临床决策中的实际效用并计算模型的临床“净效益”。利用精确率-召回率(precision recall,PR)曲线评估模型在类别不平衡数据集上的正类预测性能。

1.5 最佳模型的构建与评估

通过对比 10 种机器学习模型,筛选出最佳模型后,再次采用十折交叉验证来评估该最佳模型的区分度能力。具体操作方法如下:①对原始训练集数据进行随机排序;②将训练集拆分为 10 个子集,每次从中抽取 1 个子集作为保留或验证数据集,其余子集作为训练数据集,在训练数据集上拟合模型,并在验证数据集上对其进行评估;③保留评估分数并进行下一组测试;④绘制训练集和验证集的十折交叉验证 ROC 曲线及其平均曲线,直观展示模型区分度。

1.6 SHAP 可视化解释机器学习模型

Shaply 加性解释(SHapley additive exPlanations,SHAP)是用于解释 ML 模型输出结果的一种博弈论方法。SHAP 允许为预测模型中的每个输入特征变量进行赋值,即 SHAP 值。SHAP 值可以解释为特征变量的重要性,反映了特征变量对预测结局的影响。SHAP 值保证了预测结果在不同特征变量之间的公平分布,SHAP 的正值或负值越高,说明某个特征变量越为重要。正值表示该特征变量会增加模型的输出,负值表示模型的输出将会减少,接近于 0 的 SHAP 值表明该变量对预测结果影响较低。根据 SHAP 值构建 SHAP 摘要图可对风险因素进行重要性排序。在 SHAP 摘要图中,每个变量所在的横轴长度表示变量对预测结果的贡献,点的颜色代表变量的大小或类别。

1.7 统计学分析

所有变量均包括在训练集和验证集中,对训练集和验证集之间的基线特征进行比较。对于分类变量采用卡方检验或 Fisher 确切概率法,以频数(%)的方式呈现;对于符合正态分布的连续性变量,采用 t 检验,并表示为 $\bar{x}\pm s$;对不符合正态分布的连续性变量,采用 Mann-Whitney U 检验,并以中位数和四分位数区间展示。所有统计分析和机器学习算法均使用 R 语言(4.1.3 版)。$P<0.05$ 为差异具有统计学意义。

2 结果

2.1 一般特征

本研究共纳入 433 例患者,按 7∶3比例进行数据分割,其中训练集 304 例,验证集 129 例。在训练集中,有 55 例(18.09%)出现 LNM,验证集中有 23 例(17.83%)出现 LNM。两组之间的 LNM 发生率无统计学差异。其余研究变量在训练集与验证集之间均无统计学差异,表明两组数据具备可比性。见表 1。

2.2 LASSO 回归筛选变量

采用 LASSO 回归从纳入指标中筛选变量,分别在 λ 最小值和最小值的 1SE 处绘制垂线(图 1),最终选择 log(λ)= -3.551 时对应的 6 个非零系数的预测变量,包括性别、吸烟史、分化程度、肿瘤浸润深度、LVI 和神经侵犯。

2.3 多因素 Logistic 回归筛选独立危险因素

为了进一步控制混杂因素的影响,将 LASSO 回归筛选的 6 个预测变量纳入多因素 LR 分析。结果显示,吸烟史、黏膜下层浸润和 LVI 确定为独立危险因素($P<0.05$),见表 2。

2.4 10 种机器学习模型的建立

基于训练组数据选择 10 种算法 Catboost、LightGBM、XGboost、RF、GBM、NNET、SVM、KNN、NB 及 Logistic 回归,分别构建 EGC 淋巴结转移的预测模型。除 Logistic 回归外,其余 9 种算法均使用网格搜索和十折交叉验证进行调参调优,并在选择最佳超参数后确定最终模型。

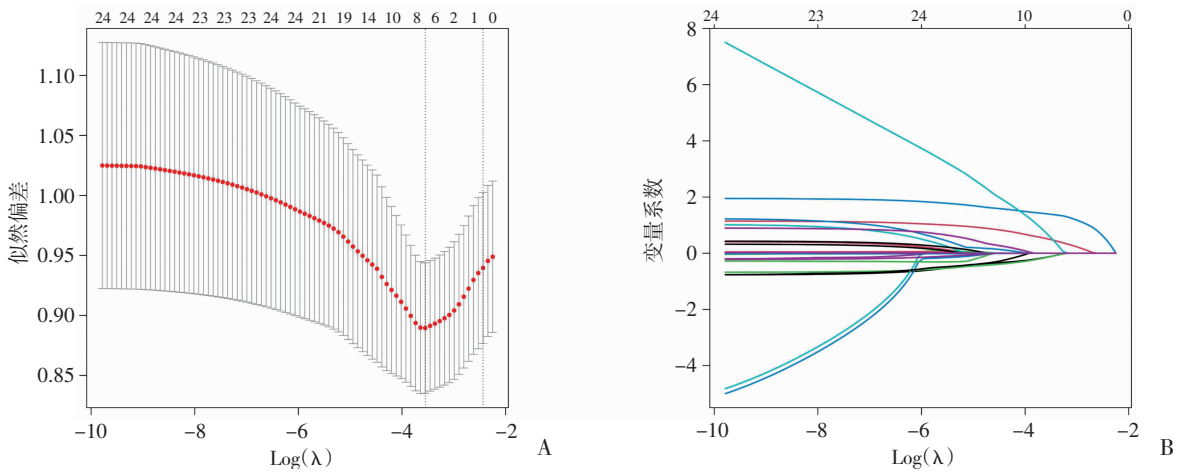
2.5 模型性能验证

基于 LASSO 回归分析筛选的 6 个特征变量构建预测模型,模型均具有良好的稳定性,各模型均无明显的过拟合或欠拟合。在训练集中,Catboost 模型的 4 个性能指标均优于其他模型(F1 评分:0.633,Brier 评分:0.100,阴性预测率:0.975,Kappa:0.520),见表 3。Catboost 模型在训练集(AUC = 0.904)和验证集(AUC=0.806)上的 AUC 值均最高(图 2)。

表 1 训练集与验证集的临床特征比较[$M(P25, P75)$, $n(\%)$, $\bar{x}\pm s$]

变量	总体数据($n=433$)	训练集($n=304$)	验证集($n=129$)	P	$\chi^2/t/U$
年龄	56(49, 64)	56(49, 65)	54(48, 63)	0.126	21 430
性别				0.647	0.209
男	280(64.67)	194(63.82)	86(66.67)		
女	153(35.33)	110(36.18)	43(33.33)		
BMI(kg/m^2)	22.86±2.91	22.80±2.96	22.99±2.79	0.518	−0.648
饮酒史				0.736	0.113
无	238(54.97)	165(54.28)	73(56.59)		
有	195(45.03)	139(45.72)	56(43.41)		
吸烟史				0.390	0.738
无	111(25.64)	82(26.97)	29(22.48)		
有	322(74.36)	222(73.03)	100(77.52)		
高血压				0.789	0.072
无	385(88.91)	269(88.49)	116(89.92)		
有	48(11.09)	35(11.51)	13(10.08)		
糖尿病				0.229	1.446
无	404(93.30)	287(94.41)	117(90.70)		
有	29(6.70)	17(5.59)	12(9.30)		
胃癌家族史				0.875	0.025
无	366(84.53)	258(84.87)	108(83.72)		
有	67(15.47)	46(15.13)	21(16.28)		
肿瘤位置				0.721	Fisher
胃上 1/3	9(2.08)	6(1.97)	3(2.33)		
胃中 1/3	121(27.94)	82(26.97)	39(30.23)		
胃下 1/3	303(69.98)	216(71.05)	87(67.44)		
肿瘤大小(cm)				0.311	2.339
≤2	274(63.28)	188(61.84)	86(66.67)		
>2~3	76(17.55)	52(17.11)	24(18.60)		
>3	83(19.17)	64(21.05)	19(14.73)		
内镜下分型				0.227	2.961
隆起型	37(8.55)	28(9.21)	9(6.98)		
平坦型	141(32.56)	105(34.54)	36(27.91)		
凹陷型	255(58.89)	171(56.25)	84(65.12)		
肿瘤浸润深度				0.277	1.183
黏膜层	226(52.19)	153(50.33)	73(56.59)		
黏膜下层	207(47.81)	151(49.67)	56(43.41)		
淋巴结转移				0.999	0
否	355(81.99)	249(81.91)	106(82.17)		
是	78(18.01)	55(18.09)	23(17.83)		
溃疡				0.135	2.23
无	200(46.19)	148(48.68)	52(40.31)		
有	233(53.81)	156(51.32)	77(59.69)		
分化程度				0.987	0.026
高分化	30(6.93)	21(6.91)	9(6.98)		
中分化	112(25.87)	78(25.66)	34(26.36)		
低分化	291(67.21)	205(67.43)	86(66.67)		
LVI				0.999	0
无	411(94.92)	289(95.07)	122(94.57)		
有	22(5.08)	15(4.93)	7(5.43)		
神经侵犯				0.508	Fisher
无	431(99.54)	303(99.67)	128(99.22)		
有	2(0.46)	1(0.33)	1(0.78)		
CA125				0.767	Fisher
(−)	419(96.77)	293(96.38)	126(97.67)		
(+)	14(3.23)	11(3.62)	3(2.33)		
CA242				0.999	Fisher
(−)	424(97.92)	297(97.70)	127(98.45)		
(+)	9(2.08)	7(2.30)	2(1.55)		
CEA				0.999	0
(−)	411(94.92)	289(95.07)	122(94.57)		
(+)	22(5.08)	15(4.93)	7(5.43)		

CA125(+)、CA242(+)、CEA(+)指超过正常范围上限



A:通过十折交叉验证方法确定 LASSO 模型中最佳 λ 值的过程;B:纳入临床变量的 LASSO 系数曲线,不同色线表示特征变量的相应系数

图 1 运用 LASSO 回归模型进行预测变量的筛选

表 2 多因素 Logistic 回归分析筛选独立危险因素

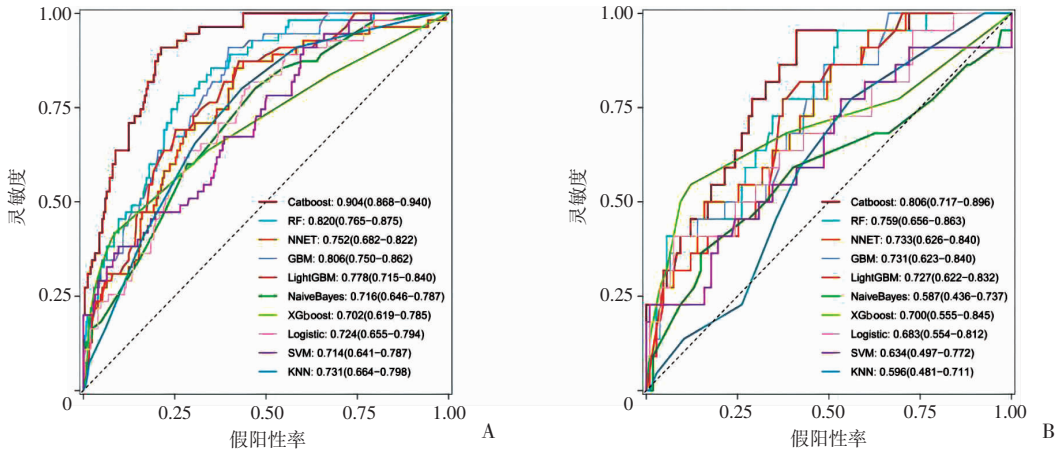
变量	*OR*	95%*CI*	*P*	*Std. Err*	*Wald*	*B*
性别						
男	1.000					
女	1.414	0.704~2.841	0.330	0.356	0.950	0.347
吸烟史						
无	1.000					
有	2.413	0.939~6.201	0.047	0.481	3.347	0.881
分化程度						
高分化	1.000					
中分化	3.755	0.445~31.661	0.224	1.088	1.479	1.323
低分化	4.329	0.546~34.32	0.165	1.056	1.924	1.465
肿瘤浸润深度						
黏膜层	1.000					
黏膜下层	2.455	1.243~4.849	0.010	0.347	6.684	0.898
LVI						
无	1.000					
有	4.634	1.435~14.961	0.010	0.598	6.575	1.533
神经侵犯						
无	1.000					
有	3 864 694.694	0~Inf	0.986	882.744	0.000	15.167

引入校准曲线评估模型预测概率与实际发生概率之间的一致性、DCA 曲线评价预测模型的临床效应以及使用 PR 曲线评估模型在类别不平衡数据集上的正类预测性能。结果显示,Catboost 模型在验证集上的校准曲线具有更高的预测准确性(图 3)。DCA 评估结果表明 Catboost 模型具有更好的临床适用性,当阈值概率在 1%~84%范围内,模型具有临床正收益(图 4)。PR 曲线显示,尽管 Catboost 模型在预测正类结局方面的性能表现不是最佳的,但相较于 XGboost、NNET、KNN 等模型,其预测性能仍表现出较高水平(图 5)。综上所述,Catboost 模型在整体性能上优于其他测试对象,因此选择 Catboost 模型作为最佳预测模型。同时,Catboost 模型自带变量的重要性排序依次为分化程度、肿瘤浸润深度、吸烟史、LVI、性别和神经侵犯(图 6)。

表 3 10 种机器学习模型的评价比较

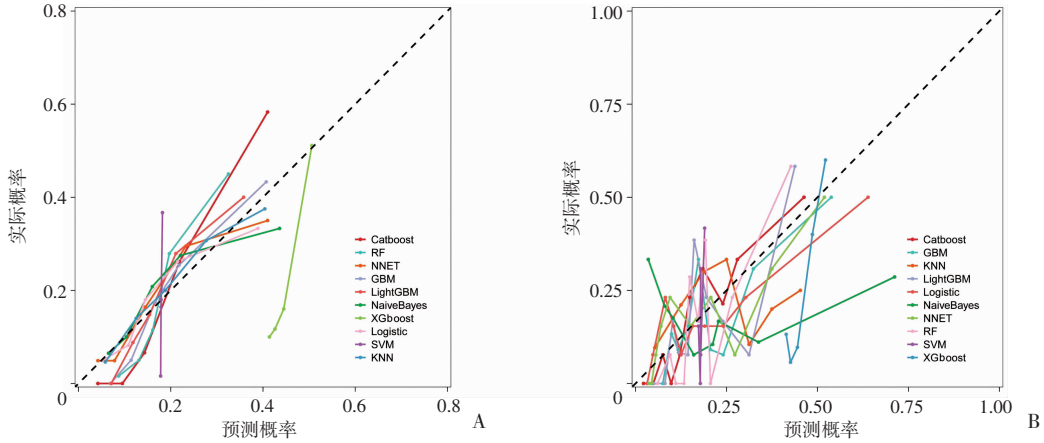
模型	AUC(95% *CI*)	准确率	精确率	召回率	F1 分数	灵敏度
训练集						
Catboost	0.904(0.868~0.940)	0.809	0.485	0.909	0.633	0.909
GBM	0.806(0.750~0.862)	0.658	0.336	0.909	0.490	0.909
KNN	0.731(0.664~0.798)	0.609	0.289	0.800	0.425	0.800
LightGBM	0.778(0.715~0.840)	0.628	0.312	0.873	0.459	0.873
Logistic	0.724(0.655~0.794)	0.595	0.285	0.818	0.423	0.818
NNET	0.752(0.682~0.822)	0.635	0.313	0.855	0.459	0.855
NB	0.716(0.646~0.787)	0.579	0.273	0.800	0.407	0.800
RF	0.820(0.765~0.875)	0.747	0.398	0.782	0.528	0.782
SVM	0.714(0.641~0.787)	0.490	0.250	0.909	0.392	0.909
XGboost	0.702(0.619~0.785)	0.822	0.511	0.418	0.460	0.418
验证集						
Catboost	0.806(0.717~0.896)	0.736	0.350	0.636	0.452	0.636
GBM	0.731(0.623~0.840)	0.620	0.263	0.682	0.380	0.682
KNN	0.596(0.481~0.711)	0.581	0.224	0.591	0.325	0.591
LightGBM	0.727(0.622~0.832)	0.620	0.286	0.818	0.424	0.818
Logistic	0.683(0.554~0.812)	0.566	0.234	0.682	0.349	0.682
NNET	0.733(0.626~0.840)	0.605	0.262	0.727	0.386	0.727
NB	0.587(0.436~0.737)	0.519	0.233	0.739	0.354	0.739
RF	0.759(0.656~0.863)	0.682	0.298	0.636	0.406	0.636
SVM	0.634(0.497~0.772)	0.512	0.219	0.727	0.337	0.727
XGboost	0.700(0.555~0.845)	0.822	0.480	0.545	0.511	0.545

模型	特异度	Brier 评分	PPV	NPV	Kappa
训练集					
Catboost	0.787	0.100	0.485	0.975	0.520
GBM	0.602	0.122	0.336	0.968	0.307
KNN	0.566	0.134	0.289	0.928	0.217
LightGBM	0.574	0.127	0.312	0.953	0.263
Logistic	0.546	0.131	0.285	0.932	0.211
NNET	0.586	0.130	0.313	0.948	0.264
NB	0.530	0.134	0.273	0.923	0.189
RF	0.739	0.123	0.398	0.939	0.379
SVM	0.398	0.147	0.250	0.952	0.151
XGboost	0.912	0.207	0.511	0.876	0.355
验证集					
Catboost	0.757	0.112	0.350	0.910	0.297
GBM	0.607	0.126	0.263	0.903	0.177
KNN	0.579	0.147	0.224	0.873	0.103
LightGBM	0.579	0.129	0.286	0.939	0.228
Logistic	0.542	0.126	0.234	0.892	0.127
NNET	0.579	0.129	0.262	0.912	0.180
NB	0.472	0.121	0.233	0.893	0.114
RF	0.692	0.120	0.298	0.902	0.226
SVM	0.467	0.141	0.219	0.893	0.101
XGboost	0.879	0.206	0.480	0.904	0.402



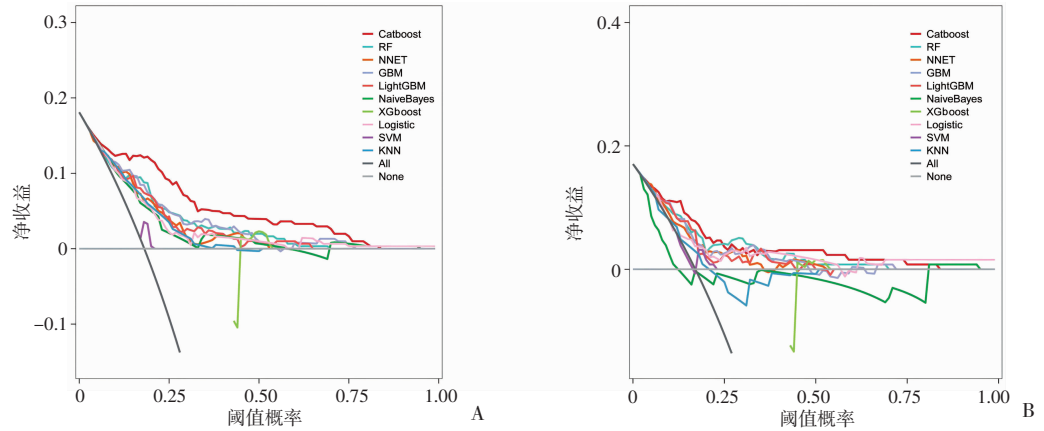
Catboost:类别型特征梯度提升机;LightGBM:轻量的梯度提升机;XGboost:极端梯度提升机;RF:随机森林;GBM:梯度提升树;NNET:神经网络;SVM:支持向量机;KNN:K 最临近

图 2 10 种机器学习模型在训练集(A)和验证集(B)上的受试者工作特征(ROC)曲线



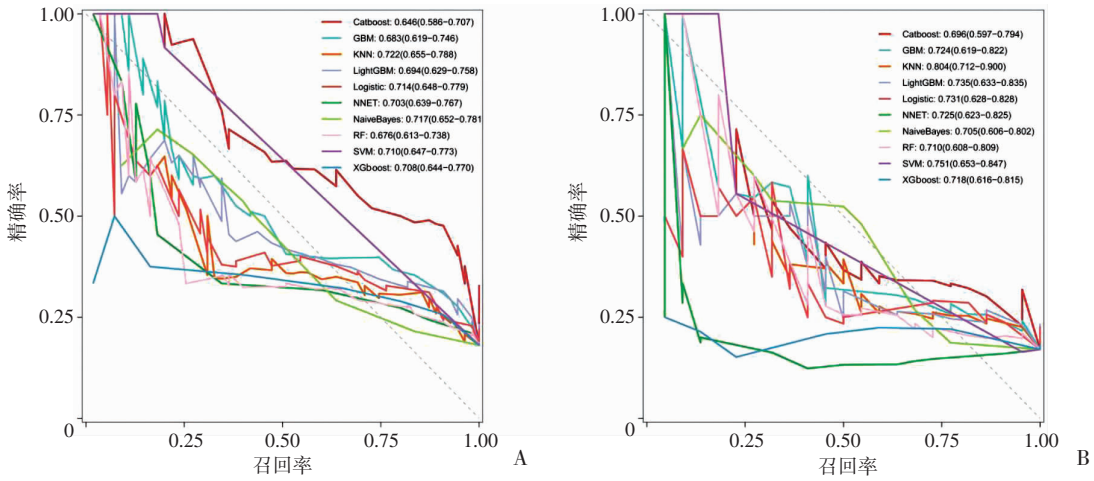
Catboost:类别型特征梯度提升机;LightGBM:轻量的梯度提升机;XGboost:极端梯度提升机;RF:随机森林;GBM:梯度提升树;NNET:神经网络;SVM:支持向量机;KNN:K 最临近;对角虚线表示理想状态下预测概率与实际概率相等

图 3 10 种机器学习模型在训练集(A)和验证集(B)上的校准曲线



Catboost:类别型特征梯度提升机;LightGBM:轻量的梯度提升机;XGboost:极端梯度提升机;RF:随机森林;GBM:梯度提升树;NNET:神经网络;SVM:支持向量机;KNN:K 最临近;浅灰实线代表所有患者未采取干预措施时的临床效益;深灰实线表示所有患者采取干预措施时的临床效益

图 4 10 种机器学习模型在训练集(A)和验证集(B)上的 DCA 曲线



Catboost:类别型特征梯度提升机;LightGBM:轻量的梯度提升机;XGboost:极端梯度提升机;RF:随机森林;GBM:梯度提升树;NNET:神经网络;SVM:支持向量机;KNN:K 最临近

图 5 10 种机器学习模型在训练集(A)和验证集(B)上的 PR 曲线

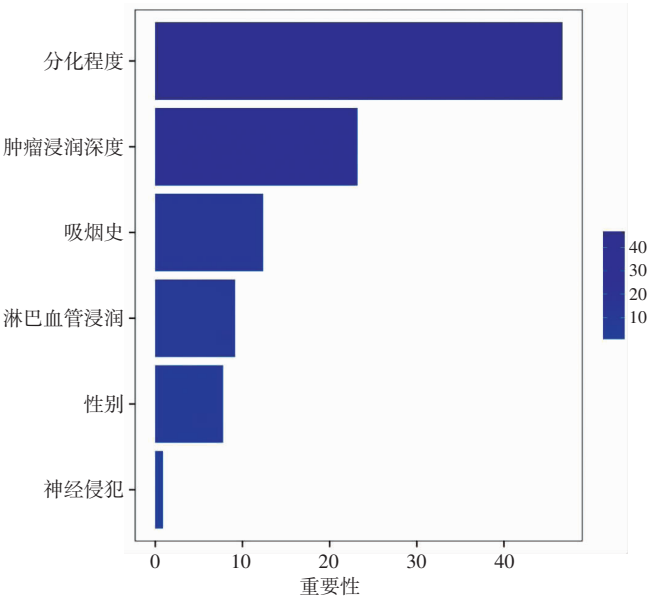


图 6 Catboost 模型变量重要性排序

2.6 最佳模型的再次评估

基于训练集数据作为数据集对 Catboost 模型进行再次评估和十折交叉验证。结果显示,Catboost 模型在训练集的平均 AUC 为 0.915(95% *CI* 0.909~0.920),验证集的平均 AUC 为 0.675(95% *CI* 0.600~0.750)。模型在训练集上表现出色,但验证集上的表现略有不足,这可能与研究纳入的样本量偏少有关。综合来看,Catboost 模型的拟合较为成功(图 7)。

2.7 最佳机器学习模型的 SHAP 解释

为进一步直观解释所选变量对 Catboost 模型的影响,采用 SHAP 进行说明(图 8)。结果表明,SHAP 摘要图中影响 LNM 的重要性排序依次为肿瘤浸润深度、LVI、吸烟史、性别、分化程度和神经侵犯。以肿瘤浸润深度为例来说明 SHAP 值对结果的影响:相较于黏膜内癌,黏膜下层浸润的 EGC 发生 LNM 风险更高。

3 讨论

内镜黏膜下剥离术已成为治疗 EGC 的主要方法。根据日本胃癌协会指南,分化良好且无溃疡的黏膜内癌是内镜治疗的绝对指征,对于存在 LNM 风险的 EGC 患者,仍推荐进行根治性手术[6]。因此,术前精确评估 EGC 患者的 LNM 风险对选择合适的治疗方法至关重要。然而,当前超声内镜、CT 等影像学检查在判断 EGC 淋巴结转移方面的敏感性和特异性尚不理想[7-8]。

多项研究表明,EGC 中 LNM 的临床病理危险因素包括年龄、性别、浸润深度、组织学类型和 LVI[9-10]。在本研究中,基于 LASSO 一致性回归系数,这些临床特征亦被证实为 LNM 的危险因素。与既往研究有所不同的是,在本研究中吸烟史和神经侵犯同样纳入了危险因素,并通过多因素分析确定了吸烟史、黏膜下层浸润和 LVI 为独立的危险因素。值得注意的是,尽管以往较少有将吸烟史纳入到 EGC 淋巴结转移风险预测模型中的报道,但国际癌症研究机构在 2012 年便得出结论:存在充分证据表明吸烟与胃癌之间存在直接关系[11]。神经侵犯表明肿瘤细胞沿神经束浸润,代表着肿瘤细胞更具侵袭性的生物学行为。相关文献显示[12-13]神经侵犯是EGC淋巴结转移的独立危险因素

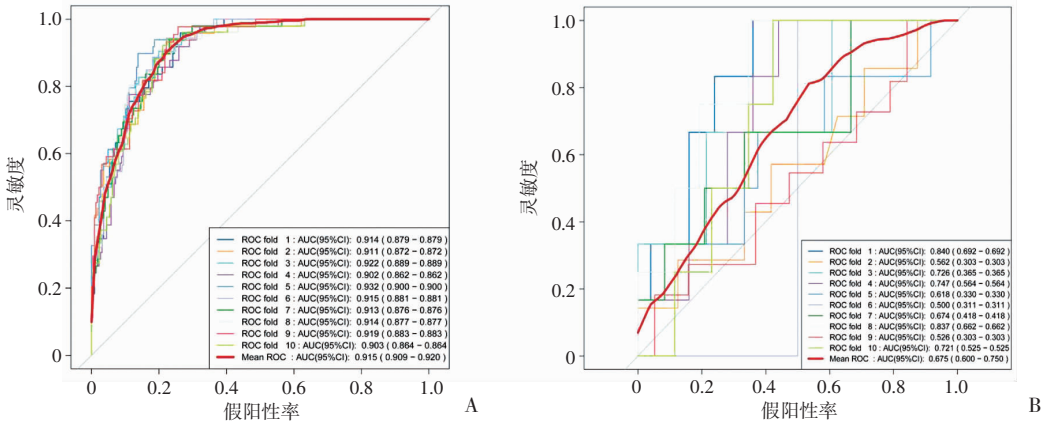
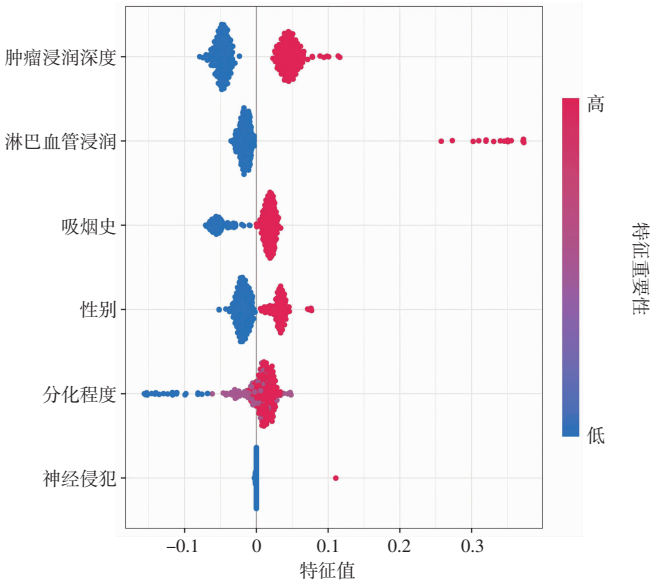


图 7 Catboost 模型在训练集(A)和验证集(B)上的十折交叉验证的 ROC 曲线



SHAP 值越高,代表发生 LNM 的概率越高;颜色展示特征变量的大小或类别(红色区域代表高特征值,蓝色区域代表低特征值)

图 8 Catboost 模型 6 个特征变量的 SHAP 摘要图

之一,并且与患者预后密切相关。此外,LVI 被视为 LNM 的危险因素和淋巴结微转移的指标[14]。大多数研究评估了胃癌中 LVI 的作用,并显示出其与 LNM 之间存在密切的相关性。本研究中,女性发生 LNM 的风险概率是男性的 1.414 倍,这可能与女性内源性雌激素的水平有关[15],然而目前尚无定论。MILHOMEM 等[16]发现未分化肿瘤和黏膜下浸润与 EGC 淋巴结转移密切相关,这一发现与本研究结论相符。低分化和深部浸润的肿瘤具备足够的营养支持,癌细胞易于侵入周围组织、毛细血管和淋巴管道。因此,它们具有更快生长和转移的潜力。

为了更准确地预测 LNM 风险并筛选出最佳预测模型,考虑到 Logistic 回归在筛选变量时无法保证自变量之间是否存在共线性,而 LASSO 回归方法在很大程度上降低了变量间的多重共线性。因此,本研究基于 LASSO 回归筛选的 6 个特征变量,采用 ML 算法构建了 10 个预测模型并进行了比较。通过准确率、精确率、召回率、F1 评分、灵敏度、特异度、阳性预测率、阴性预测率、Brier 评分、Kappa 值、ROC 曲线、校准曲线、DCA 曲线和 PR 曲线检验模型的性能。结果显示,Catboost 模型在训练集中的 4 个性能指标均优于其他模型(F1 评分:0.633,Brier 评分:0.100,阴性预测率:0.975,Kappa:0.520),模型在训练集(AUC=0.904)和验证集(AUC=0.806)上的 AUC 值均最高。此外,Catboost 的 AUC 值也高于以往研究中所构建的模型[17-18]。校准曲线证明 Catboost 模型在验证集上具有更高的预测准确性;DCA 曲线则表明该模型具有更好的临床适用性,当阈值概率在 1%~84%范围内,模型呈现出临床正收益;PR 曲线则展示了 Catboost 模型在正类预测方面同样具有较好的表现。经过初步确认 Catboost 模型为最佳预测模型后,进一步采用十折交叉验证评估模型的区分度能力。结果显示,在训练集和验证集上,Catboost 模型均取得了较好的平均 AUC 值(训练集 AUC=0.915,验证集 AUC=0.675)。尽管 Catboost 模型在验证集上的表现稍显不足,这可能与研究纳入的样本量较少有关。因此,本研究认定 Catboost 模型是预测 EGC 患者 LNM 的最佳模型。

Catboost 即分类提升,是基于梯度提升决策树的集成 ML 算法,通过有序提升和在训练阶段对分类特征进行灵活处理,改进了经典的梯度提升算法[19]。该算法具有易于使用、快速训练以及防止过拟合等优点,

并且减少了对大量超参数调整的需求[20]。在POSTNIKOV等[21]的研究中,采用Catboost算法对临床数据进行训练,以预测肺结核患者的耐药性。该研究表明,Catboost算法在处理高度分散的医疗数据时具有很高的实用性。Catboost算法是当前处理临床数据时的一个优选方法,其自带可视化工具,有助于解释模型中变量的重要性。在本研究中,Catboost模型的自带重要性排序依次为分化程度、肿瘤浸润深度、吸烟、LVI、性别和神经侵犯。然而,Catboost模型的重要性排序仅能反映变量整体的影响,无法表达变量在特定类别下的作用。因此,本研究引入SHAP方法来解释Catboost模型中变量的重要性和贡献值。SHAP摘要图结果显示,肿瘤浸润深度和LVI是预测LNM最重要的2个特征变量,这与之前的研究报道相一致[14],同时从侧面证实了模型预测的准确度。

本研究存在的局限性:①使用回顾性队列构建模型可能存在选择偏倚;相比之下,前瞻性数据集更适合用于提高模型的预测能力;②本研究中所有临床数据均来自检查报告,因此信息偏差是无法避免的;③本研究所采用ML模型中的部分变量是基于手术后的病理结果,对于评估LNM风险来说,几个主要变量(如LVI、浸润深度)无法仅通过内镜检查获得,需要结合内镜下活检进行综合判断;④本研究为单中心设计,研究在有限样本量下进行,缺乏外部验证,研究结果需要在其他机构进行验证以确保结果的可靠性。

综上所述,本研究基于LASSO回归筛选的6个特征变量,构建了10个机器学习模型用于预测EGC中的LNM。结果表明,ML可用于预测EGC中的LNM,提高预测的准确性。其中Catboost模型具有最佳的综合预测能力。

利益冲突 所有作者声明无利益冲突

作者贡献声明 孟祥勇:数据统计分析,文章撰写;秦嘉怡:数据核对;陈文生:研究指导、论文审阅

参考文献:

[1] 郑荣寿, 陈茹, 韩冰峰, 等. 2022年中国恶性肿瘤流行情况分析[J]. 中华肿瘤杂志, 2024, 46(3): 221-231. DOI: 10.3760/cma.j.cn112152-20240119-00035.
ZHENG R S, CHEN R, HAN B F, *et al*. Analysis on the prevalence of malignant tumors in China in 2022[J]. Chin J Oncol, 2024, 46(3): 221-231. DOI: 10.3760/cma.j.cn112152-20240119-00035.

[2] CARDOSO R, COBURN N, SEEVARATNAM R, *et al*. A systematic review and meta-analysis of the utility of EUS for preoperative staging for gastric cancer[J]. Gastric Cancer, 2012, 15(Suppl 1): S19-S26. DOI: 10.1007/s10120-011-0115-4.

[3] OLIVEIRA A L. Biotechnology, big data and artificial intelligence[J]. Biotechnol J, 2019, 14(8): e1800613. DOI: 10.1002/biot.201800613.

[4] MIRZA B, WANG W, WANG J, *et al*. Machine learning and integrative analysis of biomedical big data[J]. Genes, 2019, 10(2): 87. DOI: 10.3390/genes10020087.

[5] WHO. International guide for monitoring alcohol consumption and related harm[M]. Geneva: WHO, 2000: 54.

[6] Japanese Gastric Cancer Association. Japanese gastric cancer treatment guidelines 2018(5th edition)[J]. Gastric Cancer, 2021, 24(1): 1-21. DOI: 10.1007/s10120-020-01042-y.

[7] SAITO T, KUROKAWA Y, TAKIGUCHI S, *et al*. Accuracy of multidetector-row CT in diagnosing lymph node metastasis in patients with gastric cancer[J]. Eur Radiol, 2015, 25(2): 368-374. DOI: 10.1007/s00330-014-3373-9.

[8] MOCELLIN S, PASQUALI S. Diagnostic accuracy of endoscopic ultrasonography (EUS) for the preoperative locoregional staging of primary gastric cancer[J]. Cochrane Database Syst Rev, 2015, 2015(2): CD009944. DOI: 10.1002/14651858.CD009944.pub2.

[9] ZHANG M Q, DING C, XU L, *et al*. A nomogram to predict risk of lymph node metastasis in early gastric cancer[J]. Sci Rep, 2021, 11(1): 22873. DOI: 10.1038/s41598-021-02305-z.

[10] WEI J T, ZHANG Y N, LIU Y Q, *et al*. Construction and validation of a risk-scoring model that preoperatively predicts lymph node metastasis in early gastric cancer patients[J]. Ann Surg Oncol, 2021, 28(11): 6665-6672. DOI: 10.1245/s10434-021-09867-2.

[11] ROTA M, POSSENTI I, VALSASSINA V, *et al*. Dose-response association between cigarette smoking and gastric cancer risk: a systematic review and meta-analysis[J]. Gastric Cancer, 2024, 27(2): 197-209. DOI: 10.1007/s10120-023-01459-1.

[12] 李园园, 吕民, 李苓. 早期胃癌患者淋巴结转移的影响因素及其生存情况[J]. 中国卫生工程学, 2023, 22(4): 494-496. DOI: 10.19937/j.issn.1671-4199.2023.04.019.
LI Y Y, LYU M, LI L. Influencing factors and survival status of lymph node metastasis in patients with early gastric cancer[J]. Chin J Public Health Eng, 2023, 22(4): 494-496. DOI: 10.19937/j.issn.1671-4199.2023.04.019.

[13] 刘庆伟, 李勇, 檀碧波, 等. 早期胃癌患者临床病理特征与淋巴结转移因素分析[J]. 中华普通外科杂志,

2022, 37(4): 255-259. DOI: 10.3760/cma.j.cn113855-20210831-00525.
LIU Q W, LI Y, TAN B B, *et al*. Clinicopathological characteristics and lymph node metastasis in patients with early gastric cancer[J]. Chin J Gen Surg, 2022, 37(4): 255-259. DOI: 10.3760/cma.j.cn113855-20210831-00525.

[14] ZENG Q W, LI H, ZHU Y Y, *et al*. Development and validation of a predictive model combining clinical, radiomics, and deep transfer learning features for lymph node metastasis in early gastric cancer[J]. Front Med, 2022, 9: 986437. DOI: 10.3389/fmed.2022.986437.

[15] RYU E S, CHANG S J, AN J, *et al*. Sex-specific differences in risk factors of lymph node metastasis in patients with early gastric cancer[J]. PLoS One, 2019, 14(10): e0224019. DOI: 10.1371/journal.pone.0224019.

[16] MILHOMEM L M, MILHOMEM-CARDOSO D M, DA MOTA O M, *et al*. Risk of lymph node metastasis in early gastric cancer and indications for endoscopic resection: is it worth applying the east rules to the west?[J]. Surg Endosc, 2021, 35(8): 4380-4388. DOI: 10.1007/s00464-020-07932-7.

[17] YANG T, MARTINEZ-USEROS J, LIU J W, *et al*. A retrospective analysis based on multiple machine learning models to predict lymph node metastasis in early gastric cancer[J]. Front Oncol, 2022, 12: 1023110. DOI: 10.3389/fonc.2022.1023110.

[18] NA J E, LEE Y C, KIM T J, *et al*. Machine learning model to stratify the risk of lymph node metastasis for early gastric cancer: a single-center cohort study[J]. Cancers, 2022, 14(5): 1121. DOI: 10.3390/cancers14051121.

[19] PROKHORENKOVA L, GUSEV G, VOROBEV A, *et al*. CatBoost: unbiased boosting with categorical features[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018: 6639-6649. DOI: 10.48550/arXiv.1706.09516.

[20] SAFAEI N, SAFAEI B, SEYEDEKRAMI S, *et al*. E-CatBoost: an efficient machine learning framework for predicting ICU mortality using the eICU Collaborative Research Database[J]. PLoS One, 2022, 17(5): e0262895. DOI: 10.1371/journal.pone.0262895.

[21] POSTNIKOV E B, ESMEDLJAEVA D A, LAVROVA A I. A CatBoost machine learning for prognosis of pathogen's drug resistance in pulmonary tuberculosis[C]//2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech). IEEE, 2020: 86-87. DOI: 10.1109/LifeTech48969.2020.1570619054.

(**收稿**:2024-03-31;**修回**:2024-06-29)
(编辑 邓强庭)