[文章编号] 1671-587X(2026)04-1086-10 DOI:10.13481/j.1671-587X.20260421

基于临床实验室常规检测数据的食管癌机器学习预测模型的构建和验证

李 蕊，郝晓燕，杨 柳，刘家云，何 睦
（空军军医大学第一附属医院检验科，陕西 西安 710032）

[摘 要] **目的：**建立基于临床实验室常规检测数据的食管癌（EC）机器学习（ML）预测模型，阐明关键检测指标在模型预测中的相对贡献。**方法：**采用回顾性病例对照研究，收集2019年1月—2024年12月空军军医大学第一附属医院就诊的2 763例EC患者和3 297名年龄及性别匹配的健康对照（HC）者的临床实验室数据。对原始数据进行缺失值填充、异常值剔除及标准化预处理后，采用递归特征消除（RFE）与最小绝对收缩和选择算子（LASSO）回归联合筛选预测特征，并基于筛选特征分别构建极限梯度提升（XGBoost）、逻辑回归（LR）、轻量级梯度提升机（LightGBM）、随机森林（RF）、自适应提升算法（AdaBoost）和决策树（DT）模型。通过受试者工作特征（ROC）曲线下面积（AUC）、决策曲线分析（DCA）、校准曲线及精确率-召回率（PR）曲线对模型性能进行综合评估。采用沙普利加性解释（SHAP）方法对最优模型进行可解释性分析。**结果：**从67项实验室检测指标中筛选出10项EC风险相关的预测特征，包括癌胚抗原（CEA）、尿管型定量（UCQ）、红细胞分布宽度标准差（RDW-SD）、血小板（PLT）计数、血小板分布宽度（PDW）、碱性磷酸酶（ALP）、尿酸（UA）、肌酐（Cr）、丙氨酸氨基转移酶（ALT）和总蛋白（TP）。在多模型比较中，LightGBM模型在测试集DCA中表现最优，其AUC为0.953［95%置信区间（95%CI）：0.942～0.964］，并在校准曲线中显示出良好的校准度和较高的临床净获益。**结论：**基于临床实验室常规检测数据构建了EC的ML预测模型，并在内部数据集中完成了系统验证；其中LightGBM模型在多模型比较中表现最优，显示出较高的区分能力和良好的验证性能。

[关键词] 食管肿瘤；机器学习；风险预测模型；沙普利加性解释；肿瘤筛查
[中图分类号] R446 [文献标志码] A

Construction and validation of machine learning-based prediction models for esophageal cancer using routine clinical laboratory data

LI Rui, HAO Xiaoyan, YANG Liu, LIU Jiayun, HE Mu
(Department of Clinical Laboratory Medicine, First Affiliated Hospital, Air Force Medical University, Xi'an 710032, China)

ABSTRACT **Objective:** To construct a machine learning (ML)-based prediction model for esophageal

[收稿日期] 2025-11-06 [录用日期] 2026-01-04
[基金项目] 国家重点研发计划项目（2022YFC3602301）
[作者简介] 李 蕊（1988—），女，内蒙古自治区赤峰市人，主管技师，医学硕士，主要从事临床检验诊断学方面的研究。
[通信作者] 何 睦，主管技师（E-mail：394092811@qq.com）

©《吉林大学学报（医学版）》编辑部，开放获取遵循CC BY-NC-ND协议。
© Editorial Board of Journal of Jilin University (Medicine Edition). Open access under CC BY-NC-ND license.

cancer（EC）using routine clinical laboratory data, and to clarify the relative contribution of key laboratory indicators to model prediction. **Methods**: A retrospective case-control study was conducted. The clinical laboratory data were collected from 2 763 patients with EC and 3 297 age-and sex-matched healthy controls（HC）treated at the First Affiliated Hospital of Air Force Medical University between January 2019 and December 2024. After missing value imputation, outlier removal, and data normalization, recursive feature elimination（RFE）combined with least absolute shrinkage and selection operator（LASSO）regression was used to screen predictive features. Based on the selected features, extreme gradient boosting（XGBoost）, logistic regression（LR）, light gradient boosting machine（LightGBM）, random forest（RF）, adaptive boosting（AdaBoost）, and decision tree（DT）models were constructed. Model performance was comprehensively evaluated using the area under the receiver operating characteristic（ROC）curve（AUC）, decision curve analysis（DCA）, calibration curves, and precision-recall（PR）curves. SHapley Additive exPlanations（SHAP）were applied to interpret the optimal model. **Results**: From 67 routine laboratory indicators, 10 EC-related predictive features were identified, including carcinoembryonic antigen（CEA）, urinary cast quantity（UCQ）, red cell distribution width-standard deviation（RDW-SD）, platelet（PLT）count, platelet distribution width（PDW）, alkaline phosphatase（ALP）, uric acid（UA）, creatinine（Cr）, alanine aminotransferase（ALT）, and total protein（TP）. Among the compared models, the LightGBM model achieved the best performance in the test cohort, with an AUC of 0.953［95% confidence interval（95%CI）: 0.942－0.964］, and demonstrated good calibration and higher clinical net benefit in calibration and DCA. **Conclusion**: The ML-based prediction models for EC was developed using routine clinical laboratory data and systematically validated in an internal dataset. Among the compared models, the LightGBM model showed the best performance, with high discriminative ability and good validation performance.

KEYWORDS　Esophageal neoplasm; Machine learning; Risk prediction model; Shapley additive explanation; Tumor screening

食管癌（esophageal cancer，EC）是一种发病机制复杂的消化道肿瘤，其发生发展涉及遗传背景、生活方式及环境暴露等多重因素[1]。由于EC早期阶段缺乏特异性临床症状，EC患者在确诊时常处于中晚期，整体预后仍不理想[2]。因此，探讨可用于早期风险评估和预测的临床指标，一直是EC基础研究和临床实践中的重要方向。近年来，围绕EC早期检测的研究逐步从传统形态学评估拓展至分子标志物和多维数据分析领域。目前可以利用循环微小RNA-29c-3p（microRNA-29c-3p，miR-29c-3p）及特定组合[3-5]、表观遗传生物标志物（DNA甲基化、组蛋白修饰及非编码RNAs）[6-8]及代谢异常相关信号（循环血浆挥发性有机化合物）[9]等对EC进行风险评估，但上述方法多依赖高通量或专用检测平台，成本较高，且尚缺乏成熟的临床转化路径。临床实验室常规检测指标具有获取便捷、检测成本低和覆盖人群广等优势，其在肿瘤风险评估中的潜在价值逐渐受到关注。然而，单一实验室指标往往缺乏足够的特异性，其与EC风险之间的复杂非线性关系难以通过传统统计方法加以充分识别。机器学习（machine learning，ML）技术凭借其处理高维度复杂数据的独特优势，已逐步渗透到疾病诊疗的全流程，涵盖风险预测、治疗决策和治疗反应预测等环节[10]。已有研究[11]尝试将ML方法应用于EC的影像分析、病理判读或分子数据挖掘中，显示出较传统方法更优良的预测性能。但目前基于临床实验室常规检测数据构建EC预测模型的研究仍然有限，且多数工作侧重模型性能评估，对模型内部决策机制及特征贡献的解释不足，限制了其在临床实践中的可理解性和推广应用。

基于上述背景，本研究利用临床实验室常规检测数据，构建并比较多种机器学习模型用于EC风险预测，在系统评估模型预测性能的基础上，引入沙普利加性解释（shapley additive explanation，SHAP）方法，对模型中关键特征的相对贡献进行量化分析，旨在为基于常规实验室数据的EC风险预测提供一种兼顾预测性能与可解释性的分析框架，并为后续相关研究提供方法学参考。

1 资料与方法

1.1 研究对象 本研究为回顾性病例对照研究，本院提供了一套真实的和去标识化的临床数据集，包括2019－2024年所有实验室数据。所有数据经匿名化处理并分配唯一识别码，收集内容包括基线资料、诊断信息、食管镜检查和病理报告、实验室检测结果（均为食管镜检查或确诊前15 d内采集）。本研究经本院医学伦理委员会批准（伦理审批号：KY20242119），由于采用去标识化数据进行回顾性分析，伦理委员会批准免除受试者知情同意。EC组纳入标准：①病理确诊的EC患者，年龄18～80岁；②拥有完整的临床数据，包括病史、体格检查、实验室检测和影像学检查结果。排除标准：①并发其他肿瘤病史；②既往接受过抗癌治疗（手术、放化疗和免疫治疗等）；③并发严重基础疾病（如心脑血管疾病、自身免疫性疾病、糖尿病、肝肾功能衰竭和凝血障碍等）。本研究采用频率匹配法，按照约1∶1.2的比例，根据EC组患者的年龄段（每5岁为一个层级）和性别分布，在同期健康体检人群中筛选研究对象作为对照组，以确保组间基础特征的可比性。按照上述标准共纳入EC患者2 763例，健康对照（health control，HC）3 297名。

1.2 数据清洗和标准化 所有患者的检测结果均在食管镜检查或最终诊断前15 d内收集。原始数据预处理步骤：①剔除缺失率＞30％的指标，对多次检测受试者仅保留首次结果；②采用逻辑观察标识符名称与代码（Logical Observation Identifiers Names and Codes，LOINC）系统标准化数据集；③对定量数据进行二值化处理，定性数据采用虚拟编码转换；④使用Python的sklearn. preprocessing包中的MinMaxScaler对数据进行归一化处理，公式如下：$v'=(v-V_{min})/(V_{max}-V_{min})$。其中，v代表原始数据，$V_{min}$代表该指标在数据集中的最小值，$V_{max}$代表该指标在数据集中的最大值，$v'$代表归一化后数值；⑤异常值处理：采用四分位距法与Z-score法联合识别潜在异常值——对非正态分布指标；所有潜在异常值均通过查阅原始病历和咨询临床专家的方式验证合理性，剔除检测误差、标本污染等无效异常值，保留患者病理状态相关的有效异常值；⑥采用随机森林（random forest，RF）算法补充缺失数据。

1.3 预测模型的构建和评估 从所有独立变量中选择特征因子后，将EC患者分为训练集和测试集。应用多种机器学习分类模型进行综合分析，并比较了不同模型训练集和测试集中每个指标的重要性。此外，使用最优模型评估和验证结果，并开发了SHAP解释模型。①筛选特征因子：采用R软件（glmnet 4.1.2）行递归特征消除（recursive feature elimination，RFE）与最小绝对收缩和选择算子（least absolute shrinkage and selection operator，LASSO）交叉验证分析，取两者重叠因子进一步开展回归分析，采用SPSS软件进行差异性比较，最终获得$P<0.05$的特征因子；②数据划分：采用Python（0.22.1）随机数方法根据5∶1的比例随机将数据集分为训练集和测试集；③分类多模型综合分析：采用Python构建极限梯度提升（extreme gradient boosting，XGBoost）、逻辑回归（logistic regression，LR）、轻量级梯度提升机（light gradient boosting machine，LightGBM）、RF、自适应提升算法（adaptive boosting，AdaBoost）和决策树（decision tree，DT）。训练和测试上述参数模型（重复10次样本），分析不同模型训练集和测试集中指标的重要性，并选择最优模型。采用Python绘制受试者工作特征（receiver operating characteristic，ROC）曲线并计算曲线下面积（area under the curve，AUC），用于描述诊断测试工具或预测模型的识别准确性[12]。使用R软件绘制决策曲线分析（decision curve analysis，DCA），通过量化不同阈值下的临床净获益评估模型临床适用性及模型间优劣差异；利用Python sklearn库绘制校准曲线，评估模型预测概率与实际事件发生概率的一致性（校准度），同时绘制精确率-召回率（precision-recall，PR）曲线并计算平均精确率（average precision，AP），该指标在样本分布不平衡场景下可有效补充传统评估方法的不足[13-14]；④模型训练、验证与泛化能力评估方面，使用5折交叉验证训练集，并用测试集进行评估。采用Python绘制学习曲线，用于评估训练集和验证集的模型拟合和稳定性[15]；⑤模型可解释性分析通过Python shap库实现，借助SHAP绘制摘要图和特征重要性图，量化各特征对预测结果的贡献度及正负向作用以明确核心预测因子，同时构建单个样本的SHAP依赖图与力导向图，直观展示特征对个体预测结果的贡献方向与强度，实现预测过程的个体化解释，为临床医生理解预测逻辑提供可视化支持。

1.4　统计学分析　采用SPSS25.0软件、R软件（3.6.1）和Python软件（3.4.3）进行统计学分析。符合正态分布的观察指标以$\bar{x}\pm s$表示，组间比较采用独立样本t检验；不符合正态分布的检以中位数（四分位数间距）表示，组间比较采用Mann-Whitney U检验；分类变量以例数（百分比）表示，组间比较采用χ^2检验。以$P<0.05$为差异有统计学意义。

2　结　果

2.1　基线分析　本研究按照标准共计纳入EC患者2 763例，HC 3 297名。采用随机数字表法按照5∶1比例将总体样本划分为训练集（$n=5\,050$）与测试集（$n=1\,010$）。EC患者和HC者在年龄、性别构成、吸烟史和饮酒史方面比较差异均无统计学意义（$P>0.05$），提示具有可比性。训练集和测试集研究对象在上述基线特征及主要实验室指标分布方面比较差异亦无统计学意义（$P>0.05$）。见表1。

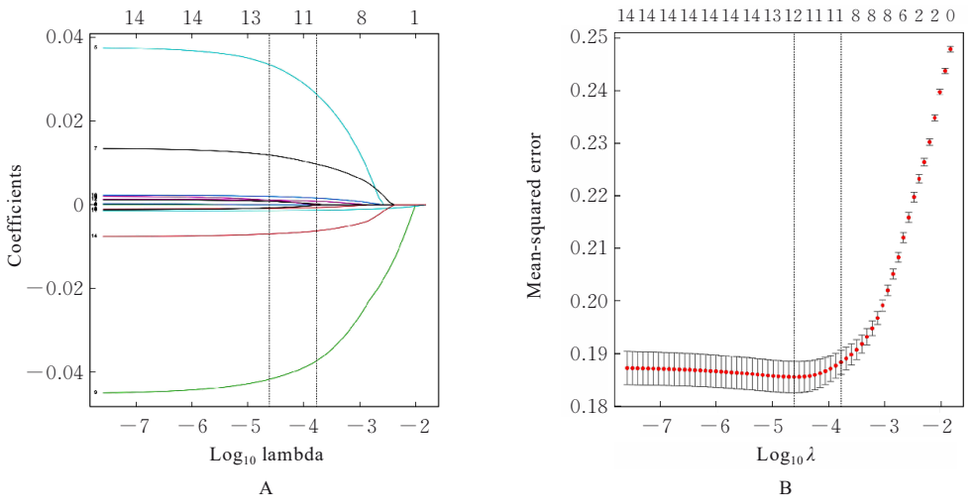
表1　训练集和测试集研究对象的基线特征

Tab.1　Baseline characteristics of subjects in training set and test set

Variable	All subjects	Training set	Test set	P
Diagnosis [n(%)]				0.972 4
HC	3 297(54.41)	2 747(83.32)	550(16.68)	
EC	2 763(45.59)	2 303(83.35)	460(16.65)	
HC				
Age ($\bar{x}\pm s$, year)	57.25±13.85	57.81±13.98	56.64±9.05	0.099 8
Gender [n(%)]				0.495 0
Male	1 800(54.60)	1 507(54.86)	293(53.27)	
Female	1 497(45.40)	1 240(45.14)	257(46.73)	
Smoking status [n(%)]	271(8.22)	229(8.34)	42(7.64)	0.585 4
Alcohol consumption [n(%)]	73(2.21)	59(2.14)	14(2.55)	0.562 9
EC				
Age ($\bar{x}\pm s$, year)	59.98±11.53	59.37±12.55	60.75±10.63	0.776 9
Gender [n(%)]				0.384 9
Male	1 547(55.99)	1281(55.62)	266(57.83)	
Female	1 216(44.01)	1022(44.38)	194(42.17)	
Smoking status [n(%)]	223(8.07)	188(8.16)	35(7.61)	0.690 1
Alcohol consumption [n(%)]	69(2.50)	58(2.52)	11(2.39)	0.873 2

2.2　EC风险特征筛选　经数据清洗，在1 415项临床实验室数据中选取67个检测项目作为候选特征因子。使用RFE筛选特征得到共计47个因子，LASSO交叉验证方法共选择出27个因子。取二者重叠因子14个，LASSO回归分析在上述14个独立变量上进行，其中患EC作为因变量（图1）。LASSO可以压缩变量系数以防止过拟合，并解决严重的共线性问题[16]，结果显示：67个独立变量最终减少到10个（最小距离的标准误差的$\lambda=0.023$），癌胚抗原（carcinoembryonic antigen，CEA）、尿管型定量（urinary cast quantity，UCQ）、红细胞分布宽度标准差（red cell distribution width-standard deviation，RDW-SD）、血小板（platelet，PLT）计数、血小板分布宽度（platelet distribution width，PDW）、碱性磷酸酶（alkaline phosphatase，ALP）、尿酸（uric acid，UA）、肌酐（creatinine，Cr）、丙氨酸氨基转移酶（alanine aminotransferase，ALT）和总蛋白（total protein，TP）被确定为特征因子（$P<0.05$）。见图1和表2。

2.3　分类多模型的综合分析　采用XGBoost、LR、LightGBM、RF、AdaBoost和DT这6种ML算法以训练模型，并且重复10次，结果显示：在训练集中，XGBoost、LightGBM、RF和DT表现优异，而在测试集中，LightGBM具有最佳的诊断效能（表3）。因此，对上述模型进行DCA、校准曲线和PR曲线分析后，测试集DCA评估显示LightGBM模型具有出色的临床适用性（图2A）。校准曲线显示LightGBM模型具有相对较高的预测



A：Draw vertical lines at selected values using 10-fold cross-validation, where the optimal λ yields ten non-zero coefficients；B：In LASSO model, the coefficient profiles of 14 textural features are plotted against log(λ) sequence. Vertical dashed lines were drawn at the minimum mean squared error (λ=0.01) and the standard error of the minimum distance (λ=0.023).

图 1　LASSO 回归分析选择特征因子

Fig. 1　Feature factors selected by LASSO regression analysis

准确性（图 2B）。在训练集中，XGBoost、RF 和 DT 模型展现了最佳性能，AP 值最高（图 2C）。而在测试集中，LightGBM 模型则具有最高的 AP 值（图 2D）。综合分析表明：LightGBM 可被认为是 6 个 ML 模型中的最佳模型。

表 2　HC 组和 EC 组研究对象的临床指标

Tab. 2　Clinical indicators of subjects in HC group and EC group　　[$M(P25, P75)$]

Group	ALP [λ_B/(IU·L^{-1})]	UA [c_B/(μmol·L^{-1})]	Cr [c_B/(μmol·L^{-1})]	TP [ρ_B/(g·L^{-1})]	CEA [ρ_B/(μg·L^{-1})]
Control	75.000(65.000,87.000)	325.000(271.000,388.000)	73.000(59.000,91.000)	74.300(70.900,77.600)	1.790(1.180,2.640)
EC	85.000(71.000,100.000)	269.000(226.000,317.000)	85.000(69.000,97.000)	71.900(68.200,75.659)	2.430(1.560,3.680)
Z	−17.582	26.263	−15.869	16.571	−18.885
P	<0.001	<0.001	<0.001	<0.001	<0.001
Group	UCQ(V/μL)	RDW-SD(V/fL)	PLT(×10^9 L^{-1})	PDW(V/fL)	ALT[λ_B/(IU·L^{-1})]
Control	0.000(0.000,0.140)	42.900(41.500,44.700)	220.000(186.000,257.000)	15.500(12.800,16.200)	20.000(15.000,28.000)
EC	0.400(0.130,0.883)	44.000(41.900,47.000)	205.000(162.000,254.000)	12.500(10.800,14.600)	15.000(11.000,21.000)
Z	−34.378	−13.597	8.967	25.673	20.740
P	<0.001	<0.001	<0.001	<0.001	<0.001

2.4　最佳模型构建和评估　在训练集上进行了 LightGBM 分析和 5 折交叉验证，训练集平均分为 5 份，其中 4 份作为子训练集，1 份作为子验证集。结果显示：子训练集的平均 AUC 为 0.999（0.999～1.000），子验证集的平均 AUC 为 0.952（0.941～0.964），测试集的 AUC 为 0.946（0.932～0.960）。鉴于验证集 AUC 表现未超出测试集，或超出比例小于 10%，可认为模型拟合成功，学习曲线表明训练集和验证集具有强拟合性和高稳定性。见图 3。

2.5　EC 风险预测模型的 SHAP 解释　为直观阐释所选变量的作用机制，本研究采用 SHAP 方法展示其在模型中对 EC 风险的预测过程。图 4A 呈现了模型中 10 个最重要的特征，每个特征的重要性曲线中，所有患者的 SHAP 归因结果以不同颜色的点呈现：红色点代表高风险值，蓝色点代表低风险值。x 轴的 SHAP 值反映其对模型预测的重要性。CEA、RDW-SD 和 Cr 等指标的 SHAP 值随指标数值增加而增大，提示其为 EC 风险的正向贡献因子；而 TP 和 UA 等指标则呈现负向贡献，即指标

表3　6种机器学习模型区分HC组和EC组对象的算法性能

Tab. 3　Algorithm performance of six machine learning models in distinguishing subjects in HC group and EC group

Modol	SEN[a]	SPE[b]	ACC[c]	PPV[d]	NPV[e]	F1 score	Kappa	AUC[f]
Training cohort								
XGBoost	0.999 (0.999－1.000)	1.000 (1.000－1.000)	1.000 (0.999－1.000)	1.000 (0.999－1.000)	0.999 (0.999－1.000)	1.000 (0.999－1.000)	0.999 (0.998－1.000)	1.000 (1.000－1.000)
LR	0.808 (0.795－0.820)	0.778 (0.771－0.786)	0.792 (0.786－0.798)	0.753 (0.747－0.759)	0.828 (0.820－0.837)	0.779 (0.772－0.786)	0.583 (0.571－0.594)	0.866 (0.856－0.876)
LightGBM	0.977 (0.973－0.981)	0.981 (0.976－0.985)	0.979 (0.977－0.981)	0.977 (0.972－0.983)	0.981 (0.977－0.984)	0.977 (0.975－0.979)	0.958 (0.954－0.961)	0.998 (0.997－0.999)
RF	0.999 (0.998－0.999)	0.998 (0.996－0.999)	0.998 (0.997－0.999)	0.997 (0.995－0.999)	0.999 (0.998－0.999)	0.998 (0.997－0.999)	0.996 (0.994－0.998)	1.000 (1.000－1.000)
AdaBoost	0.851 (0.833－0.870)	0.882 (0.862－0.902)	0.868 (0.866－0.871)	0.859 (0.842－0.876)	0.877 (0.865－0.888)	0.855 (0.853－0.857)	0.734 (0.729－0.738)	0.942 (0.936－0.948)
DT	1.000 (1.000－1.000)	1.000 (1.000－1.000)	1.000 (1.000－1.000)	1.000 (1.000－1.000)	1.000 (1.000－1.000)	1.000 (1.000－1.000)	1.000 (1.000－1.000)	1.000 (1.000－1.000)
Test cohort								
XGBoost	0.865 (0.851－0.879)	0.887 (0.876－0.898)	0.877 (0.875－0.879)	0.865 (0.855－0.875)	0.887 (0.878－0.896)	0.865 (0.862－0.868)	0.752 (0.747－0.756)	0.948 (0.937－0.960)
LR	0.803 (0.791－0.814)	0.775 (0.757－0.794)	0.788 (0.777－0.798)	0.750 (0.735－0.765)	0.824 (0.816－0.833)	0.775 (0.766－0.785)	0.575 (0.555－0.595)	0.865 (0.844－0.885)
LightGBM	0.869 (0.852－0.886)	0.890 (0.879－0.901)	0.880 (0.873－0.888)	0.869 (0.858－0.880)	0.890 (0.878－0.903)	0.869 (0.860－0.877)	0.759 (0.744－0.774)	0.953 (0.942－0.964)
RF	0.851 (0.831－0.870)	0.883 (0.872－0.894)	0.868 (0.862－0.874)	0.859 (0.850－0.869)	0.876 (0.863－0.889)	0.855 (0.847－0.863)	0.734 (0.722－0.747)	0.941 (0.929－0.954)
AdaBoost	0.835 (0.812－0.858)	0.869 (0.835－0.903)	0.854 (0.843－0.864)	0.845 (0.813－0.876)	0.863 (0.851－0.876)	0.839 (0.831－0.847)	0.705 (0.685－0.725)	0.929 (0.915－0.943)
DT	0.796 (0.790－0.802)	0.827 (0.820－0.834)	0.813 (0.808－0.818)	0.794 (0.787－0.801)	0.829 (0.824－0.833)	0.795 (0.790－0.800)	0.623 (0.612－0.633)	0.811 (0.789－0.834)

a：Sensitivity；b：Specificity；c：Accuracy；d：Positive predictive value；e：Negative predictive value；f：Area under curves.

水平下降与EC风险升高相关。这种方向性分析与临床观察到的消耗性状态及炎症反应高度吻合。图4B通过平均绝对SHAP值对这10个风险因素进行排名。此外，本研究以1例非EC患者为典型示例（其SHAP预测分数较低，为0.04），进一步阐释模型的可解释性（图4C）。

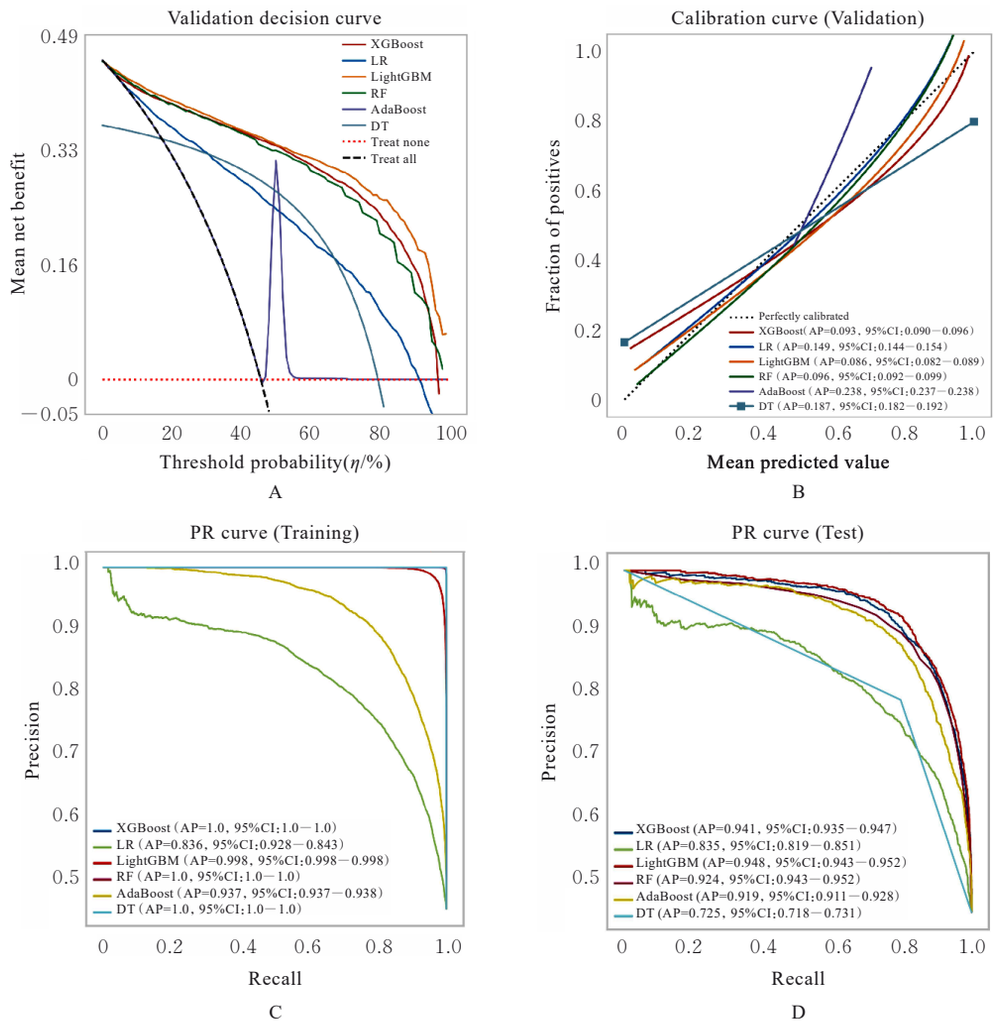
3　讨　论

EC是全球最常见的恶性肿瘤之一，2022年我国新发EC 22.4万例，死亡18.8万例，分别占全部恶性肿瘤的4.64%和7.28%[17]。尽管遗传易感性、生活方式和环境暴露等风险因素已得到广泛关注，但目前临床仍缺乏兼顾预测效能、检测可及性及临床实用性的EC风险预测模型。

本研究基于真实世界临床实验室常规检测数据，系统比较了多种ML模型在EC风险预测中的表现，结果显示：在多模型评估框架下，LightGBM模型在区分能力、校准度和临床净获益等方面均表现较为稳定，优于其余对照模型。

DT模型由于结构简单、对噪声敏感，在未进行严格剪枝或复杂度约束的情况下，易对训练数据产生过拟合。本研究中DT模型在训练集与测试集之间表现出显著性能差异，进一步印证了这一特性。因此，在模型选择过程中，本研究更加重视模型在独立测试集中的泛化能力，最终选择训练集和测试集性能均衡且稳定的LightGBM作为最优预测模型。

近年来，ML在医学领域的应用已从图像识别（如磁共振成像和内窥镜下肿瘤检出）及预后评估等方向逐步拓展，但多数研究依赖于影像学和遗传学等复杂检测技术，限制了其在基层医疗或资源有



A：DCA of test set；B：Calibration curves of test set（Closer to diagonal indicates better calibration）；C：PR curves and AP in training set；D：PR curves and AP in test set（Higher AP indicated better model performance）. Different colors represented different models；values were presented as mean and 95%CI.

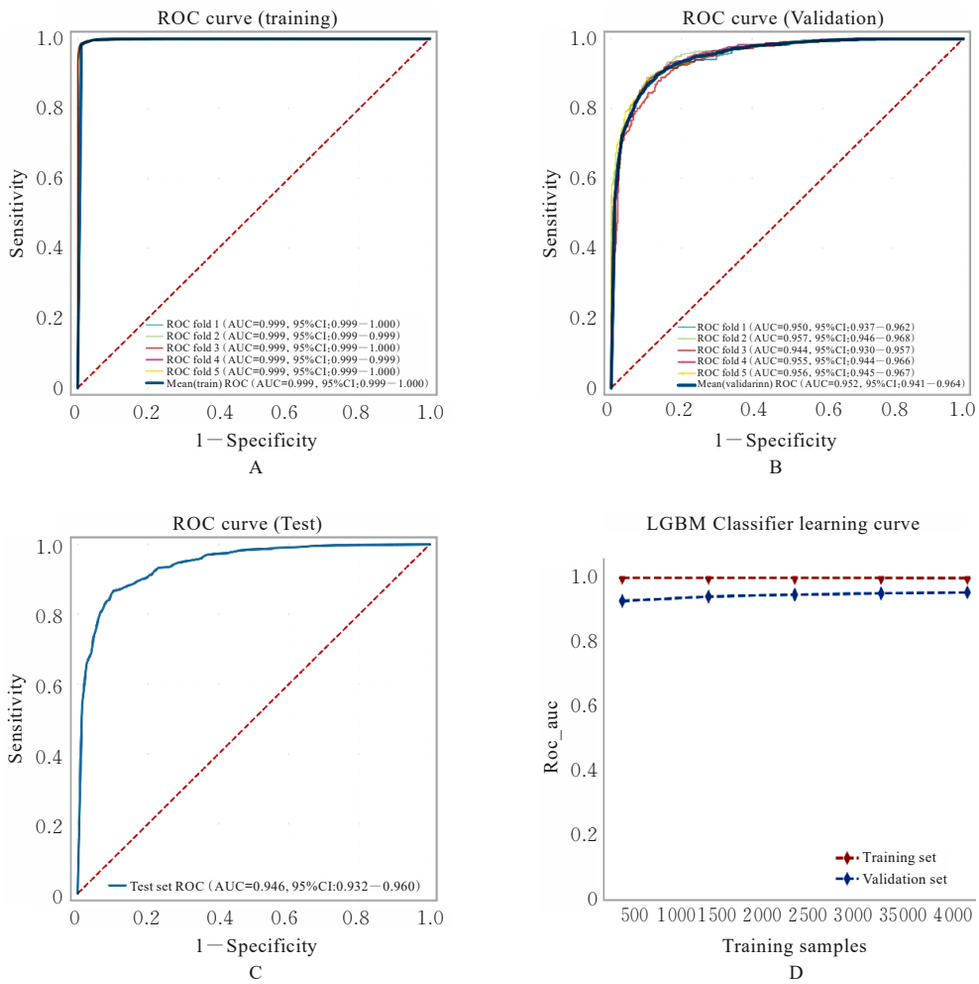
图 2　6 种 ML 模型性能综合分析

Fig. 2　Comprehensive performance analysis of six ML models

限地区的推广。本研究聚焦临床常规实验室检测数据（血常规、尿常规、生化指标及常见肿瘤标志物），核心优势体现在两方面：①常规实验室数据中蕴含的 EC 相关潜在关联模式，往往需多指标联合分析才能展现其临床意义，而 ML 算法可高效挖掘这种非线性、高维度的特征关联，这是传统统计方法难以实现的；②与影像学和基因组学检测比较，常规实验室检查具有获取便捷、成本低廉及普适性强的特点，尤其适用于欠发达地区或大规模人群筛查场景。本研究构建的风险预测模型并非旨在替代食管镜联合病理检查这一 EC 诊断“金标准”，而是作为初级保健或大规模筛查中的“风险分层工具”，通过识别高风险人群，引导其尽早接受进一步确诊检查，最终达到提升早期检出率、降低总体医疗负担的目标。

模型可解释性是 ML 方法在医学领域应用中的重要前提。为增强模型决策过程的透明度，本研究在模型构建的基础上引入 SHAP 方法，对关键特征在预测结果中的相对贡献进行了量化分析。该方法能够在不显著降低模型性能的前提下，直观呈现不同特征对预测结果的影响方向及重要程度，从而在一定程度上缓解 ML 模型“黑箱”特性所带来的理解障碍。

本研究筛选的核心指标具有一定的生物学合理性。既往研究已证实红细胞分布宽度（RDW）[19]、血小板相关参数（PLT 和 PDW）[20-21] 及 ALP 水



A: ROC curves and AUC of training set; B: ROC curves and AUC of validation set (Different colors represented five distinct results); C: ROC curve and AUC of test set; D: Learning curve (Red: Training set; blue: Validation set). Values were expressed as mean and 95%CI.

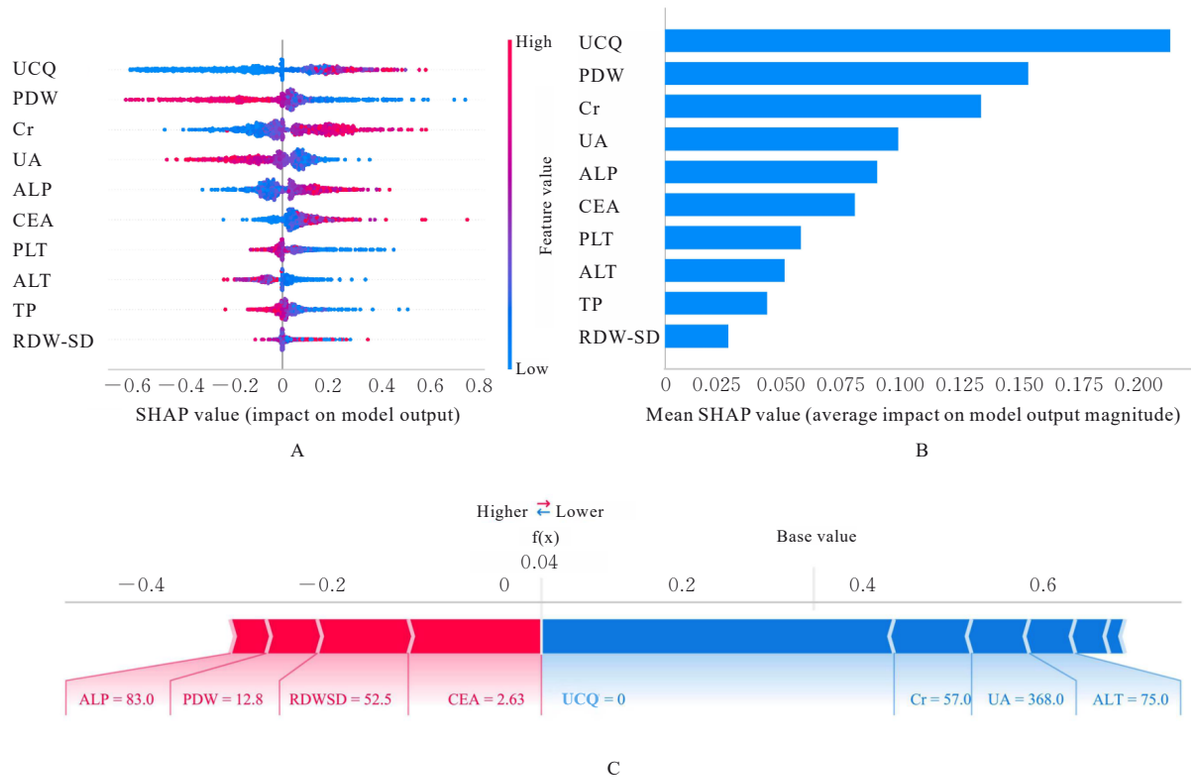
图3　LightGBM模型训练、验证和测试

Fig. 3　Training, validation, and testing of LightGBM model

平[22]与EC的发生发展或预后相关，本研究进一步将上述指标纳入风险预测模型，拓展了其在早期风险评估中的应用场景。此外，UA、Cr和TP等代谢与营养相关指标在模型中的重要性，提示全身代谢状态和营养状况可能在EC发生过程中发挥一定作用[23-24]。部分此前报道较少的指标（如UCQ和ALT）在本研究中显示出一定预测价值，其潜在生物学机制仍有待在后续研究中进一步探讨。

本研究亦存在一定局限性。首先，研究样本来源于单中心人群，地域和人群特征相对集中，模型的泛化能力尚需在多中心、不同地区及不同人群的外部独立数据集中进一步验证。其次，本研究未对EC患者的病理类型和临床分期进行分层分析，不同亚型及疾病阶段的风险特征可能存在差异。再次，尽管采用了RFE与LASSO相结合的特征筛选策略，并通过交叉验证及独立测试集评估模型性能，但较为复杂的特征筛选流程在一定程度上仍可能高估模型性能表现，相关影响需在外部验证中进一步评估。针对上述局限性，未来研究可从两个方面推进：其一，结合Kaplan-Meier生存分析与预后分层，评估模型对患者长期生存的预测价值，明确其在预后评估与治疗决策中的应用潜力；其二，开展卫生经济学评价与临床转化研究，将模型整合至临床决策支持系统，通过临床试点验证其在实际医疗场景中的可用性与成本效益，最终推动其在基层医疗筛查和高风险人群管理中的广泛应用。

综上所述，本研究基于临床实验室常规检测数据构建并验证了EC的ML风险预测模型，在多模型比较与可解释性分析框架下，系统评估了模型性能及关键特征的相对贡献。本研究结果为基于常规



A：SHAP summary plot（Each line represented a feature；red dots indicated high feature values，blue dots indicated low feature values）；B：SHAP feature importance ranking；C：SHAP force plot for an individual non-EC case［Red indicated features increase EC risk，blue indicated features decrease risk；arrow length indicated effect size；bold number：predicted probability f(x)；base value：prediction without input］.

图 4 EC 风险预测模型的 SHAP

Fig. 4 SHAP of EC risk prediction model

检测数据开展 EC 风险建模提供了方法学参考，也为后续相关研究奠定了基础。

利益冲突声明：

所有作者声明不存在利益冲突。

作者贡献声明：

李蕊参与研究设计、数据分析和论文撰写，郝晓燕参与数据收集及整理，杨柳参与模型构建和验证，刘家云参与研究指导和论文审阅，何睦参与研究设计和论文审阅。

[参考文献]

[1] ZHU H C, MA X, YE T, et al. Esophageal cancer in China: Practice and research in the new era[J]. Int J Cancer, 2023, 152(9): 1741-1751.

[2] SHEIKH M, ROSHANDEL G, MCCORMACK V, et al. Current status and future prospects for esophageal cancer[J]. Cancers, 2023, 15(3): 765.

[3] ZHANG K Q, WU X W, WANG J H, et al. Circulating miRNA profile in esophageal adenocarcinoma[J]. Am J Cancer Res, 2016, 6(11): 2713-2721.

[4] CRAIG M P, RAJAKARUNA S, PALIY O, et al. Differential microRNA signatures in the pathogenesis of Barrett's esophagus[J]. Clin Transl Gastroenterol, 2020, 11(1): e00125.

[5] WANG L, JI F, LIU G, et al. Upregulation of circulating miR130a is correlated with development of Barrett's esophagus and esophageal adenocarcinoma[J]. Onco Targets Ther, 2018, 12: 1-7.

[6] KAZ A M, GRADY W M. Epigenetic biomarkers in esophageal cancer[J]. Cancer Lett, 2014, 342(2): 193-199.

[7] ERGUN P, KIPCAK S, BOR S. Epigenetic alterations from Barrett's esophagus to esophageal adenocarcinoma[J]. Int J Mol Sci, 2023, 24(9): 7817.

[8] YU Y, CHEN X X, CANG S D. Cancer-related long noncoding RNAs show aberrant expression profiles and competing endogenous RNA potential in esophageal adenocarcinoma[J]. Oncol Lett, 2019, 18(5): 4798-

4808.
[9] BHATT A, PARSI M A, STEVENS T, et al. Volatile organic compounds in plasma for the diagnosis of esophageal adenocarcinoma: a pilot study [J]. Gastrointest Endosc, 2016, 84(4): 597-603.
[10] DEO R C. Machine learning in medicine [J]. Circulation, 2015, 132(20): 1920-1930.
[11] DU Y H, MCNESTRY C, WEI L, et al. Machine learning-based clinical decision support systems for pregnancy care: a systematic review [J]. Int J Med Inform, 2023, 173: 105040.
[12] OBUCHOWSKI N A, BULLEN J A. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine [J]. Phys Med Biol, 2018, 63(7): 07TR01.
[13] FENLON C, O'GRADY L, DOHERTY M L, et al. A discussion of calibration techniques for evaluating binary and categorical predictive models [J]. Prev Vet Med, 2018, 149: 107-114.
[14] LI W K, GUO Q H. Plotting receiver operating characteristic and precision-recall curves from presence and background data[J]. Ecol Evol, 2021, 11(15): 10192-10206.
[15] BELKIN M, HSU D, MA S Y, et al. Reconciling modern machine-learning practice and the classical bias-variance trade-off[J]. Proc Natl Acad Sci U S A, 2019, 116(32): 15849-15854.
[16] SAUERBREI W, ROYSTON P, BINDER H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building [J]. Stat Med, 2007, 26(30): 5512-5528.
[17] 国家卫生健康委员会医疗应急司. 食管癌筛查与早诊早治方案(2024年版)[J]. 中华肿瘤杂志, 2024, 46(10): 913-914.
[18] MUSCHELLI J. ROC and AUC with a binary predictor: a potentially misleading metric [J]. J Classif, 2020, 37(3): 696-708.
[19] HAN F Y, LIU Y Q, CHENG S Q, et al. Diagnosis and survival values of neutrophil-lymphocyte ratio (NLR) and red blood cell distribution width (RDW) in esophageal cancer [J]. Clin Chim Acta, 2019, 488: 150-158.
[20] SU R X, ZHU J Y, WU S S, et al. Prognostic significance of platelet (PLT) and platelet to mean platelet volume (PLT/MPV) ratio during apatinib second-line or late-line treatment in advanced esophageal squamous cell carcinoma patients [J]. Technol Cancer Res Treat, 2022, 21: 15330338211072974.
[21] ZHOU X Y, CHEN H, ZHANG W F, et al. Predictive value of routine blood test in patients with early esophageal cancer: a matched case-control study [J]. J Cancer, 2021, 12(15): 4739-4744.
[22] WEI X L, ZHANG D S, HE M M, et al. The predictive value of alkaline phosphatase and lactate dehydrogenase for overall survival in patients with esophageal squamous cell carcinoma [J]. Tumour Biol, 2016, 37(2): 1879-1887.
[23] HUANG J J, MI N N, YANG J L, et al. U-shaped relationship between serum uric acid and gastric cancer risk: a large prospective cohort study [J]. Front Oncol, 2024, 14: 1482814.
[24] HADAVI S M S, OLIAEI S, SAIDI S, et al. Using data mining and association rules for early diagnosis of esophageal cancer[J]. Gulf J Oncolog, 2022, 1(40): 38-46.