

中图分类号: R979.1; R965 文献标志码: A 文章编号: 1006-4931(2026)07-0078-07
doi:10.3969/j.issn.1006-4931.2026.07.014



基于机器学习方法构建乳腺癌 TAC 化学治疗方案致骨髓抑制风险预测模型*

刘 阳^{1,2}, 李红健¹, 刘思明¹, 刘学涛², 吴建华¹, 焦 敏¹, 于鲁海^{1,2△}

(1. 新疆维吾尔自治区人民医院, 新疆 乌鲁木齐 830001; 2. 石河子大学药学院, 新疆 石河子 832000)

摘要:目的 采用机器学习方法构建并验证乳腺癌多柔比星+环磷酰胺+多西他赛(TAC)化学治疗(简称化疗)方案致骨髓抑制的风险预测模型。方法 回顾性收集医院2019年1月至2023年10月收治的接受TAC化疗方案的女性乳腺癌患者的病历资料。对数据进行预处理后,综合运用随机森林、极限梯度提升模型进行特征筛选,并结合单因素分析与多因素Logistic回归分析确定最佳特征变量。基于筛选出的特征,构建支持向量、逻辑回归、随机森林、极限梯度提升和类别型特征梯度提升5种预测模型,在测试集中进行性能验证与比较,筛选最优模型。采用Shapley加法解释法对最优模型进行可解释性分析。结果 共纳入240例患者,其中发生骨髓抑制170例(70.83%)。最佳特征变量为化疗前白细胞计数、淋巴细胞数、血清白蛋白水平、预防性使用集落刺激因子、化疗周期、体质量指数及基础疾病。热图分析结果显示,各特征间相关性较弱,均适宜纳入模型。随机森林模型为最优预测模型,受试者工作特征曲线下面积为0.965,预测性能与校准度良好。结论 基于机器学习方法构建的随机森林模型在预测乳腺癌TAC化疗方案致骨髓抑制方面具有较好的性能与临床实用性,可作为个体化风险评估的辅助工具。

关键词:机器学习;乳腺癌;TAC化学治疗方案;骨髓抑制;随机森林;预测模型

Construction of a Risk Prediction Model for Myelosuppression Induced by TAC Chemotherapy Regimen in Breast Cancer Based on the Machine Learning Approach

LIU Yang^{1,2}, LI Hongjian¹, LIU Siming¹, LIU Xuetao², WU Jianhua¹, JIAO Min¹, YU Luhai^{1,2△}

(1. Xinjiang Uygur Autonomous Region People's Hospital, Urumqi, Xinjiang 830001, China; 2. School of Pharmacy, Shihezi University, Shihezi, Xinjiang 832000, China)

Abstract: Objective To construct and validate a risk prediction model for myelosuppression induced by the doxorubicin +

*基金项目:新疆维吾尔自治区自然科学基金[2023D01C69]。

第一作者:刘阳,男,硕士,研究方向为临床药学,(电子信箱)2363901559@qq.com。

△通信作者:于鲁海,男,硕士,主任药师,研究方向为医院药学与药物分析,(电子信箱)1523264450@qq.com。

中药材的性质密切相关,且由于其一人一方的特点,各处方中的药材及性质不相同,仍有极少部分处方因药材的性质特殊导致制备的水丸不佳。后期将完善对特殊性质药材制备工艺的研究,以提升我院临方水丸的质量。

参考文献

- [1] 国家药典委员会. 中华人民共和国药典(四部)[M]. 北京:中国医药科技出版社,2020.
- [2] 张兆旺. 中药药剂学[M]. 北京:中国中医药出版社,2011:320.
- [3] 乔延江. 中药制剂学[M]. 北京:中国协和医科大学出版社,2017:99-102.
- [4] 汤丽芝,史 亚,年娟娟,等. 中药传统丸剂研究进展[J]. 陕西中医药大学学报,2016,39(3):107-109.
- [5] 洪燕龙,王俊杰,沈 岚,等. 中药临方制剂的应用现状、问题分析和对策[J]. 中国中药杂志,2021,46(15):3739-3745.
- [6] 胡志强,鲜洁晨,楚世慈,等. 中药临方制剂技术的发展现状及研究策略[J]. 中国中药杂志,2019,44(1):28-33.
- [7] 马 媛,高晓南,施秉华,等. 正交试验优化皮肤科个体化加工水丸制备工艺[J]. 中国处方药,2022,20(5):29-31.
- [8] 黄志峰,李得堂. 正交试验法优化丹七活血丸成型性工艺研究[J]. 中国药业,2020,29(13):42-43.
- [9] 陈新芳. 影响中药水丸溶散时限因素的探讨[J]. 中成药,2002,24(3):227.
- [10] 杜宏伟. 中药浓缩丸与水丸生产过程质量影响因素分析[J]. 中国药业,2011,20(22):54-55.
- [11] 张 臻,高天慧,傅超美,等. 中药丸剂剂型理论与应用现状关键问题分析[J]. 中国中药杂志,2017,42(12):2408-2412.
- [12] 潘淑明. 中药水丸溶散时限的影响因素[J]. 中国中医药现代远程教育,2010,8(11):88.
- [13] 陈 洁,丘振文,蔡庆群,等. 医院开展个体化浓缩丸工艺研究及质量回顾性评估[J]. 中国药业,2022,31(12):35.
- [14] 李 霞,张子龙. 医院中药丸剂加工的体会[J]. 中国医药科学,2016,6(22):212-215.
- [15] 杨晓东,刘兴文. 中药丸剂的染菌途径与防控措施分析[J]. 中国药业,2011,20(22):92-93.

(收稿日期:2025-05-28;修回日期:2025-10-21)

cyclophosphamide + docetaxel (TAC) chemotherapy regimen in breast cancer based on the machine learning approach. **Methods** Medical records of female patients with breast cancer admitted to the hospital from January 2019 to October 2023 and received TAC chemotherapy regimen were collected. After data preprocessing, the feature selection was performed by the Random Forest (RF) and Extreme Gradient Boosting (XGB) models, and the optimal feature variables were determined through the univariate analysis and multivariate Logistic regression analysis. Based on the selected features, five prediction models [support vector machine (SVM), Logistic Regression (LR), RF, XGB, and Categorical Boosting (CatBoost)] were constructed. Performance validation and comparison were conducted on the test set to select the optimal model. The interpretability of the optimal model was further analyzed using Shapley's additive interpretation (SHAP) method. **Results** A total of 240 patients were included, among whom 170 patients (70.83%) experienced myelosuppression. The optimal feature variables were pre-chemotherapy leukocyte and lymphocyte counts, serum albumin level, prophylactic use of colony-stimulating factors, chemotherapy cycles, body mass index, and underlying diseases. Heatmap analysis results showed weak correlations among the features, all of which were suitable for model construction. The RF model was identified as the optimal predictive model, with an area under the receiver operating characteristic curve (AUC) of 0.965, it had good predictive performance and calibration, which could serve as a reference for clinical decision-making. **Conclusion** The RF model constructed based on the machine learning approach has good performance and clinical utility in predicting myelosuppression induced by TAC chemotherapy regimens for breast cancer, and it can be used as a supplementary tool for personalized risk assessment.

Key words: machine learning; breast cancer; TAC chemotherapy regimen; myelosuppression; random forest; predictive model

乳腺癌为全球女性常见恶性肿瘤^[1],其治疗策略的优化一直是医学研究的焦点。尽管近年来的治疗手段不断进步,但化学治疗(简称化疗)仍是乳腺癌治疗不可或缺的一环^[2]。骨髓抑制是紫杉醇类、蒽环类、烷化剂类化疗药物常见的药品不良反应(ADR),严重影响患者的临床疗效和生存质量^[3]。目前,临床缺乏乳腺癌化疗过程中发生骨髓抑制的有效预测工具,难以早期识别风险。随着人工智能和机器学习技术的快速发展,其在医疗领域的应用逐渐成为研究热点^[4-5]。通过机器学习方法构建风险预测模型,可更精准地评估化疗风险,从而为临床治疗提供更个性化、更准确的决策支持。故本研究中借助机器学习方法构建了多个风险预测模型,经评估筛选出最佳预测模型,为临床提供强有力的预测工具,最大限度地减少治疗风险。现报道如下。

1 资料与方法

1.1 一般资料

病例选择:选取我院2019年1月至2023年10月收治的接受多柔比星 50 mg/m^2 +环磷酰胺 500 mg/m^2 +多西他赛 75 mg/m^2 (TAC)化疗方案治疗6个周期(以21 d为1个周期)的乳腺癌患者的临床资料。

纳入标准:女性;年龄 ≥ 18 岁;经病理证实为乳腺癌;每周化疗前后均接受血常规检查。本研究方案经我院临床科研伦理委员会批准(批件号:KY2024010901)。

排除标准:合并其他原发肿瘤;同步进行放射治疗;曾在外院接受阶段性化疗;合并脾功能亢进、骨髓或血液系统疾病;化疗前骨髓抑制已达Ⅱ度及以上。

1.2 数据收集

资料收集:收集患者的基线资料,包括年龄,民族,

体质指数(BMI),基础疾病,血常规指标[白细胞计数(WBC)、淋巴细胞数、中性粒细胞计数(Neut)、单核细胞计数(MON)、平均红细胞体积(MCV)、血小板体积分布宽度(PDW)],生化指标[血清白蛋白(Alb)、肌酐(Cr)、丙氨酸氨基转移酶(ALT)、尿酸(UA)、总胆红素(TBil)、血清总蛋白(TP)、乳酸脱氢酶(LDH)、天门冬氨酸氨基转移酶(AST)],凝血指标[活化部分凝血酶原时间(APTT)],乳腺癌分型,肿瘤TNM分期,化疗周期,预防性使用集落刺激因子等。

分组标准:本研究中以化疗过程中首次发生Ⅱ度及以上骨髓抑制为研究终点,参照世界卫生组织(WHO)抗肿瘤药物急性及亚急性毒性反应分度标准,化疗后符合以下任一标准即为骨髓抑制Ⅱ度及以上。1)血红蛋白(Hb)水平低于 95 g/L ;2)WBC低于 $3.0 \times 10^9/\text{L}$;3)中性粒细胞绝对计数(ANC)低于 $1.5 \times 10^9/\text{L}$;4)血小板计数(PLT)低于 $75 \times 10^9/\text{L}$ 。若发生两系及以上血细胞减少,以抑制程度最高者为准。将发生Ⅱ度及以上骨髓抑制的患者纳入病例组,将仅发生Ⅰ度和无骨髓抑制者纳入对照组。

特征筛选:为构建简约且高效的预测模型,采用多步骤策略筛选最佳特征变量。首先分别应用随机森林和极限梯度提升模型进行初步特征选择,并结合Shapley加法解释评估特征贡献度,各自选取排名前20的特征,取二者的交集。随后对交集特征行单因素及多因素Logistic回归分析,最终选取具有统计学意义($P < 0.05$)的特征用于模型构建。

数据预处理:1)数据清洗与缺失值处理。通过评估各特征变量的缺失率,并以30%为阈值进行筛选。缺失

率超过 30% 的变量(包括碱性磷酸酶、血清钙),直接剔除;缺失率低于 30% 的变量,采用 K 近邻算法对 LDH、D-二聚体、凝血酶原时间(PT)、凝血酶原活动度(PTA)、国际标准化比值(INR)、纤维蛋白原(Fib)、APTT、癌胚抗原(CEA)、癌抗原(CA)-125、CA-15-3 等特征进行填补。最终共纳入 41 个特征变量进入后续流程。2)数据平衡处理。针对数据集中存在的正负样本不平衡问题,采用合成少数类过采样技术(SMOTE)生成新样本,以平衡各类别样本数量,避免模型训练产生偏倚^[6]。3)数据变换与标准化。为满足模型输入要求,对分类特征进行 One-Hot 编码处理;对连续型特征进行中心化和标准化处理,以消除不同量纲带来的影响,提升模型计算的收敛效率^[7-8]。

模型构建与优化:基于 Python 3.9.7 及 Anaconda 3 平台进行建模,将数据集按 7:3 随机划分为训练集与测试集。利用 Scikit-learn 构建 5 种机器学习模型,包括支持向量(SVM)模型、逻辑回归(LR)模型、随机森林(RF)模型、极限梯度提升(XGB)模型和类别型特征梯度提升(CatBoost)模型^[9-13]。采用 Bayes_opt 中的贝叶斯优化模型,结合五折交叉验证,以最大化受试者工作特征曲线(ROC)下面积(AUROC)为目标,对上述 5 种模型的超参数进行调优。随后使用最优超参数组合在训练集拟合模型,并在测试集评估其泛化能力。

模型评估:采用准确率、精确率、召回率及 F_1 值评估模型的预测性能,通过 AUROC 和精准召回曲线的曲线下面积评估模型的区分度,利用校准曲线和 Brier 分数评估模型的校准能力,绘制决策曲线,分析评估模型

的临床净收益。

1.3 统计学处理

采用 IBM SPSS Statistics 26 统计学软件分析。符合正态分布的计量资料以 $\bar{x} \pm s$ 表示,组间比较行独立样本 t 检验;不符合正态分布的计量资料以中位数(四分位数) $[M(P_{25}, P_{75})]$ 表示,组间比较行 Mann-Whitney U 检验。计数资料以率(%)表示,组间比较行 χ^2 检验。 $P < 0.05$ 为差异有统计学意义。

2 结果

2.1 病例分组

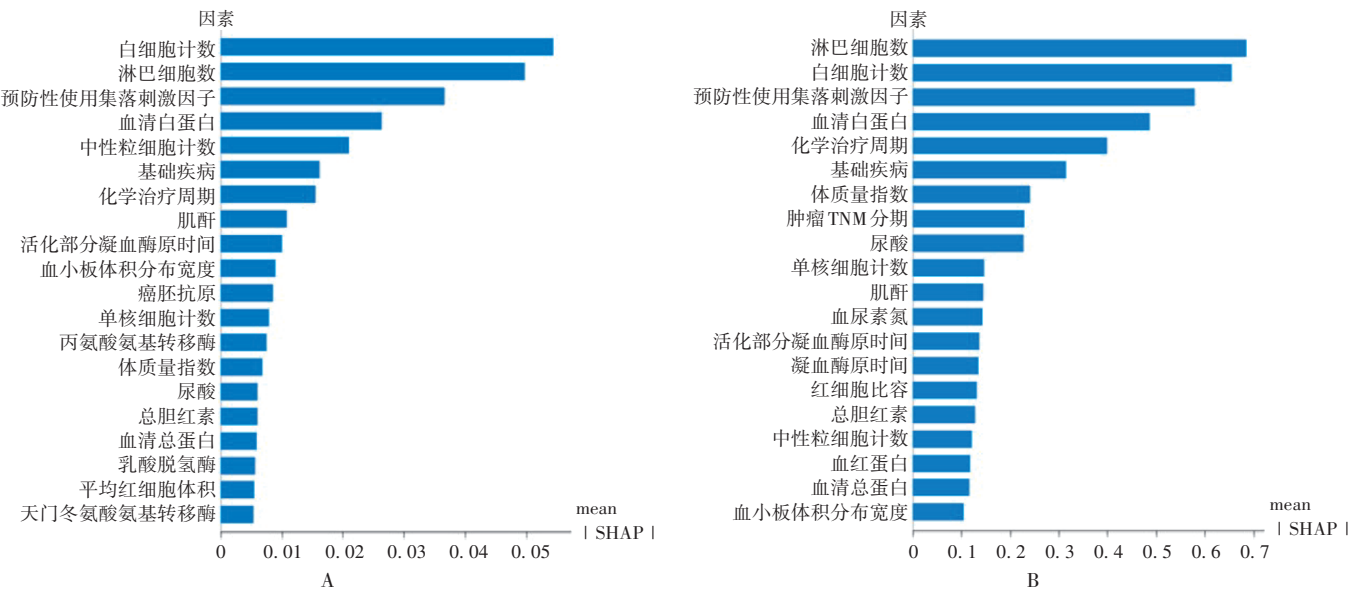
本研究中共纳入 240 例患者,其中对照组 70 例,病例组 170 例。训练集数据包括对照组 49 例、病例组 119 例,存在正负样本不平衡问题,采用 SMOTE 生成对照组 70 例,使训练集数据达到平衡后用于后续模型构建。

2.2 特征选择工程

特征选择:采用 RF 和 XGB 模型进行特征选择,并通过 Shapley 加性解释(SHAP)法对特征重要性进行解释,根据贡献度对特征变量进行排序,分别选取排名前 20 的特征变量。详见图 1。

单因素分析:对 RF 和 XGB 模型筛选出的前 20 个特征变量取交集,并对交集特征进行单因素分析。结果显示,BMI、基础疾病、化疗周期、预防性使用集落刺激因子、WBC、淋巴细胞数、MON、Alb 8 个特征变量在两组间有显著差异($P < 0.05$)。详见表 1。

多因素 Logistic 回归分析:以是否发生骨髓抑制为因变量,将单因素分析中具有统计学意义的 8 个因素作为自变量进行赋值。详见表 2。多因素 Logistic 回归分



A. 随机森林特征选择 B. 极限梯度提升特征选择

图 1 SHAP 特征重要性矩阵图

A. Random forest feature selection B. Extreme gradient boosting feature selection

Fig. 1 SHAP feature importance matrix

表 1 TAC 化疗方案致骨髓抑制的单因素分析结果

Tab. 1 Univariate analysis results of myelosuppression induced by the TAC chemotherapy regimen					
因素		对照组(<i>n</i> = 70)	病例组(<i>n</i> = 170)	χ^2 / Z 值	<i>P</i> 值
体质量指数 [例(%)]	< 18.5 kg / m ²	1(1.43)	8(4.71)	9.354	0.009
	18.5 ~ 24.0 kg / m ²	22(31.43)	84(49.41)		
	≥ 24.0 kg / m ²	47(67.14)	78(45.88)		
基础疾病	无	47(67.14)	67(39.41)	16.066	< 0.001
	高血压	8(11.43)	25(14.71)		
	糖尿病	15(21.43)	78(45.88)		
化学治疗周期 [例(%)]	1 ~ 3 个	28(40.00)	112(65.88)	13.666	< 0.001
	4 ~ 6 个	42(60.00)	58(34.12)		
预防性使用集落刺激因子	否	5(7.14)	88(51.76)	41.596	< 0.001
	是	65(92.86)	82(48.24)		
白细胞计数 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), × 10 ⁹ / L]		5.41(4.46, 6.67)	3.85(3.58, 5.88)	5.590	< 0.001
中性粒细胞计数 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), × 10 ⁹ / L]		3.71(2.26, 4.77)	3.86(3.01, 4.74)	1.617	0.106
血小板体积分布宽度 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), %]		12.90(10.03, 15.80)	13.50(11.30, 15.80)	1.465	0.143
淋巴细胞数 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), × 10 ⁹ / L]		1.72(1.22, 2.13)	1.15(0.81, 1.39)	6.660	< 0.001
单核细胞计数 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), × 10 ⁹ / L]		0.53(0.41, 0.69)	0.47(0.39, 0.58)	2.447	0.014
总胆红素 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), μmol / mL]		8.75(6.52, 11.52)	9.40(7.30, 11.80)	1.121	0.262
血清总蛋白 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), g / L]		70.25(66.60, 73.80)	68.90(64.47, 72.80)	1.717	0.086
血清白蛋白 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), g / L]		42.44(40.74, 44.55)	39.92(36.58, 42.62)	4.742	< 0.001
肌酐 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), μmol / mL]		46.30(42.62, 51.59)	48.65(40.11, 55.02)	1.219	0.233
尿酸 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), μmol / mL]		266.21(221.75, 307.94)	244.22(198.02, 295.25)	1.270	0.204
活化部分凝血酶原时间 [<i>M</i> (<i>P</i> ₂₅ , <i>P</i> ₇₅), s]		28.10(26.45, 31.95)	27.45(25.90, 29.90)	1.411	0.158

表 2 多因素 Logistic 回归分析变量赋值
Tab. 2 Variable assignment of the multivariate Logistic regression analysis

变量	变量赋值
因变量 是否发生骨髓抑制	否 = 0; 是 = 1
自变量 体质量指数	< 18.5 kg / m ² = 0; 18.5 ~ 24.0 kg / m ² = 1; ≥ 24.0 kg / m ² = 2
基础疾病	无 = 0; 高血压 = 1; 糖尿病 = 2
化学治疗周期	1 ~ 3 个 = 0; 4 ~ 6 个 = 1
预防性使用集落刺激因子	否 = 0; 是 = 1
白细胞计数	-
淋巴细胞数	-
单核细胞计数	-
血清白蛋白	-

注：- 为未赋值。
Note：- refers to unassigned variables.

析结果显示,除 MON 外,BMI、基础疾病、化疗周期、预防性使用集落刺激因子、WBC、淋巴细胞数、Alb 共 7 个特征变量具有统计学意义(*P* < 0.05)。详见表 3。

2.3 模型构建

相关性热图分析结果显示,7 个最佳特征变量间的

表 3 TAC 化疗方案致骨髓抑制的多因素 Logistic 回归分析结果
Tab. 3 Multivariate Logistic regression analysis results of myelosuppression induced by the TAC chemotherapy regimen

因素	变量赋值	β 值	标准误	OR(95%CI)	Wald 值	<i>P</i> 值
体质量指数	2	-	-	1	-	-
	0	2.492	1.242	12.086(1.059, 137.972)	4.024	0.045
	1	0.957	0.459	2.604(1.058, 6.406)	4.339	0.037
基础疾病	0	-	-	1	-	-
	1	1.120	0.706	3.055(0.768, 12.226)	2.516	0.113
	2	1.912	0.534	6.764(2.374, 19.274)	12.802	< 0.001
化学治疗周期	1	-	-	1	-	-
	0	1.117	0.465	3.055(1.227, 7.604)	5.762	0.016
预防性使用集落刺激因子	0	-	-	1	-	-
	1	2.596	0.600	13.407(4.136, 43.460)	18.715	< 0.001
白细胞计数	-	-0.327	0.158	0.721(0.529, 0.984)	4.262	0.039
淋巴细胞数	-	-2.457	0.506	0.086(0.032, 0.231)	23.539	< 0.001
单核细胞计数	-	-1.512	1.312	0.220(0.017, 2.884)	1.328	0.249
血清白蛋白	-	-0.191	0.054	0.826(0.743, 0.919)	12.441	< 0.001

注：- 为未计算。
Note：- refers to uncalculated factors.

相关性较弱,均可用于构建模型。详见图 2。分别采用支持向量机、多因素 Logistic 回归分析、RF、XGB、CatBoost

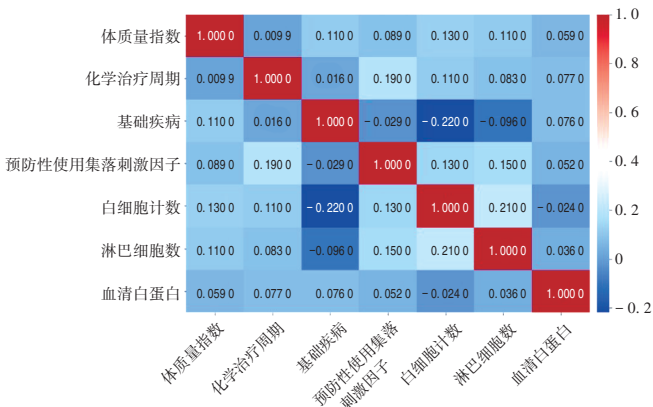


图 2 最优特征的相关性热图

Fig. 2 Correlation heatmap of optimal feature variables

5 种模型,基于训练集构建预测模型。采用五折交叉验证法控制模型过拟合,并应用贝叶斯优化模型反复训练迭代,获取各模型的最优超参数组合。

2.4 模型评价

在 5 种机器学习模型中,RF 模型表现最佳,其准确率为 0.903,精确率为 0.907,召回率为 0.961, F_1 值为 0.933,精度 - 召回率曲线下面积(AUPRC)为 0.985,AUROC 为 0.965。详见表 4 及图 3。在校准曲线对比中,RF 模型虽非最佳,但与其他模型差异较小。详见图 4。决策曲线分析结果显示,RF 模型在临床决策中具有较高的应用价值。详见图 5。综合比较各机器学习模型的预测性能,最终选择 RF 模型作为最终的预测模型。

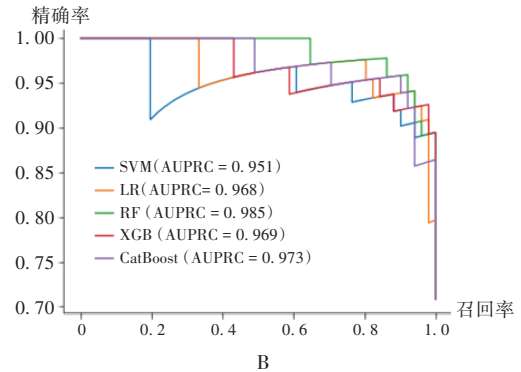
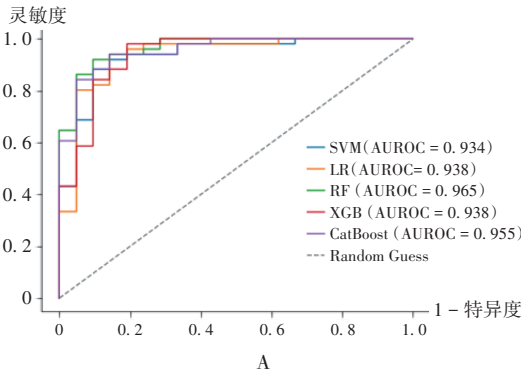
表 4 5 种机器学习模型预测性能比较

Tab. 4 Comparison of prediction performance of five machine learning models

模型	准确率	精确率	召回率	F_1 值	Brier 值	AUPRC	AUROC
SVM	0.888	0.905	0.941	0.923	0.084	0.951	0.934
LR	0.903	0.923	0.941	0.932	0.077	0.968	0.938
RF	0.903	0.907	0.961	0.933	0.084	0.985	0.965
XGB	0.861	0.918	0.882	0.900	0.082	0.969	0.938
CatBoost	0.903	0.940	0.921	0.931	0.080	0.973	0.955

2.5 模型可解释性分析

采用 SHAP 法对最佳预测模型(RF 模型)进行可解释性分析。通过 SHAP 值,评估各特征变量对模型预测结果的贡献程度,特征变量值越大代表该特征对模型预测的贡献越大。由 SHAP 特征重要性矩阵图(图 6)可知,在纳入的 7 个特征变量中,WBC 对模型的贡献最大。进一步通过 SHAP 图揭示各特征变量与结局变量间的效应方向。图中每个点代表一个样本,颜色反映数值的大小,颜色越红表明数值越大,颜色越蓝表明数值越小。分析结果显示,WBC、淋巴细胞数、预防性使用集落刺激因子、Alb、化疗周期、BMI 与结局变量呈负相关,而基础疾病与结局变量呈正相关。详见图 7。



A. ROC 曲线 B. P - R 曲线

图 3 5 种机器学习模型的 ROC 曲线与 P - R 曲线

A. ROC curves B. P - R curves

Fig. 3 ROC curves and P - R curves of five machine learning models

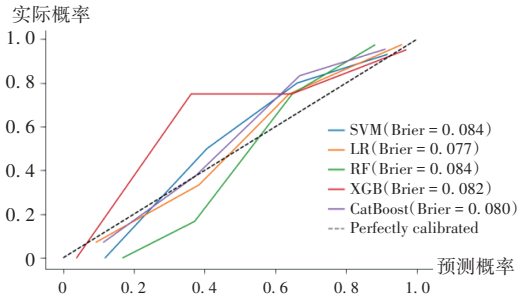


图 4 5 种机器学习模型的校准曲线

Fig. 4 Calibration curves of five machine learning models

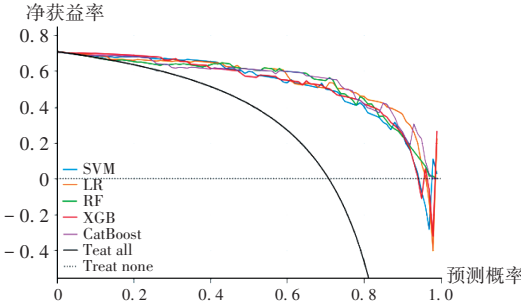


图 5 5 种机器学习模型的临床决策曲线

Fig. 5 Clinical decision curves of five machine learning models

3 讨论

近年来,机器学习方法的发展有效弥补了传统统计学方法与数学模型的局限性。机器学习方法不仅能识别变量间复杂的非线性关系,还能对高维数据进行

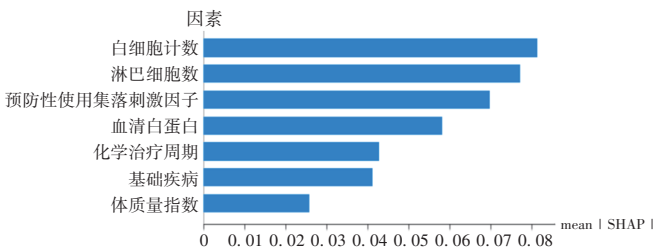


图6 SHAP特征重要性矩阵图

Fig. 6 SHAP feature importance matrix

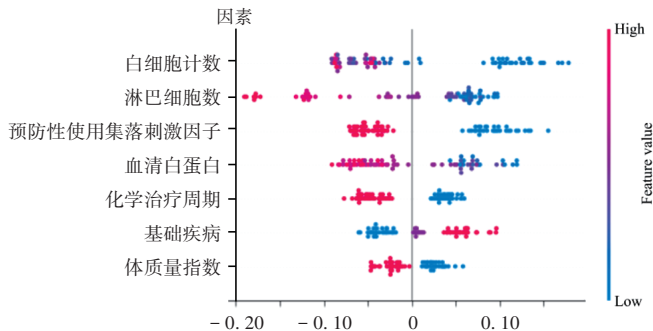


图7 SHAP图

Fig. 7 SHAP diagram

降维处理,降低模型过拟合风险,展现出优秀的预测能力^[14-15]。本研究中共收集了240例乳腺癌患者的临床资料,其中对照组70例,病例组170例,骨髓抑制发生率为70.83%。在训练集中,对照组样本量明显少于病例组,存在正负样本不均衡问题。为此,本研究采用SMOTE生成部分少数类样本,有效解决了样本不均衡问题,从而提高了模型的泛化能力,降低了过拟合风险。

通过组合策略对乳腺癌患者的特征变量进行分析发现,化疗前WBC、淋巴细胞数、Alb、预防性使用集落刺激因子、化疗周期、BMI、基础疾病是预测骨髓抑制发生的主要特征。其中,化疗前WBC是最重要的预测指标,且该水平越低,发生骨髓抑制的风险越高。白细胞由骨髓中的造血干细胞不断分化而来,当WBC降低时,患者骨髓的造血功能可能已受到影响,此时化疗药物将进一步抑制骨髓的造血功能,从而增加骨髓抑制的发生风险。LYMAN等^[16]的研究证实,化疗前WBC低下是骨髓抑制发生的危险因素,与本研究结果一致。既往研究表明,未预防性使用集落刺激因子是成人化疗相关性中性粒细胞减少的危险因素^[16-17]。本研究结果显示,未预防性使用集落刺激因子会增加骨髓抑制的发生风险。

化疗前淋巴细胞数、Alb水平及BMI越低,骨髓抑制的发生风险越高。淋巴细胞作为免疫系统的关键组成部分,主要由骨髓中的造血干细胞分化而来,在机体免疫反应和抗感染过程中发挥重要作用。淋巴细胞数低下提示患者的造血功能及免疫系统可能已受损

或功能减弱,此时化疗可能进一步抑制骨髓的造血功能,从而增加骨髓抑制的发生风险^[18]。Alb是维持机体营养状态和渗透压平衡的关键蛋白质,能与多种药物及离子发生可逆性结合,发挥运输功能。Alb水平较低的患者,血浆中未结合的游离化疗药物浓度可能升高,增加的游离药物浓度与化疗过程中ADR的发生有关,从而增加骨髓抑制的风险^[19]。BMI广泛用于评估患者的营养状态,低BMI患者可能存在营养不良问题,进而影响体内药物代谢及骨髓造血功能^[20]。同时,合并基础疾病的患者发生骨髓抑制的风险高于无基础疾病者^[21]。此类患者因长期服用慢性病治疗药物,可能与化疗药物发生相互作用,影响化疗药物的代谢与清除,导致药物在体内蓄积,增强其对骨髓的毒性作用,从而增加骨髓抑制的发生风险。既往研究显示,化疗前淋巴细胞数低下、低Alb、低BMI及基础疾病是化疗致骨髓抑制的危险因素^[22]。与文献^[23]的研究结果不同,本研究中发现骨髓抑制的发生风险随着化疗周期的增加而降低。分析原因,3种药物(多柔比星、环磷酰胺联合多西他赛)常见ADR均为骨髓抑制,联合用药使骨髓抑制的总体发生风险升高。同时,本研究中关注的是化疗过程中首次出现的骨髓抑制事件,这可能是导致与既往研究结果存在差异的原因。

本研究中采用多种机器学习方法构建了预测模型,避免了单一模型可能出现的欠拟合或过拟合问题。通过模型间的对比,能识别在特定情境下预测能力最强的模型,从而提高模型的准确性、可靠性。在超参数优化方面,本研究中采用贝叶斯优化模型,相较于网格随机搜索等传统方法,贝叶斯优化模型能在每轮迭代中依据既往结果自适应调整超参数搜索方向,聚焦于更优的超参数空间区域,从而提升搜索效率。此外,RF模型在测试集的AUROC为0.965,具有良好的灵敏度和特异度。通过校准曲线、ROC及决策曲线分析可见,RF模型具有优异的预测能力,能准确估计事件发生概率,并为临床决策提供有价值的参考。在特征筛选环节,本研究中采用组合筛选策略,原因在于初始变量较多,传统统计显著性检验方法可能受到样本量与变量数量的影响。机器学习特征选择方法能捕捉数据中的复杂非线性关系,并识别特征间的交互效应,而传统统计检验方法则可评估特征与目标变量间的线性关系与差异。二者结合能更全面地评估特征的重要性,避免单一方法可能引入的偏倚或误差,同时降低模型复杂度,提升模型的泛化能力与预测性能。

综上所述,本研究中基于机器学习方法构建了乳腺癌TAC化疗方案致骨髓抑制的风险预测模型,其中

RF模型表现出良好的预测性能,可作为临床实践中的风险评估工具,为临床治疗方案的制订及个体化预防用药策略的实施提供参考。此外,本研究也存在以下局限性:1)本研究为回顾性研究,不可避免地存在选择偏倚;2)本研究基于单中心数据进行模型构建,未来需在多中心及前瞻性研究框架下进一步验证,以增强模型的稳健性与外推性。

参考文献

- [1] WANG LL, ZHANG SZ, WANG XC. The Metabolic Mechanisms of Breast Cancer Metastasis [J]. *Front Oncol*, 2020, 10: 602416.
- [2] ELSAYED B, ALKSAS A, SHEHATA M, et al. Exploring Neoadjuvant Chemotherapy, Predictive Models, Radiomic, and Pathological Markers in Breast Cancer: A Comprehensive Review [J]. *Cancers (Basel)*, 2023, 15(21): 5288.
- [3] 中国抗癌协会肿瘤临床化疗专业委员会, 中国抗癌协会肿瘤支持治疗专业委员会. 肿瘤化疗导致的中性粒细胞减少诊治中国专家共识(2023版) [J]. *中华肿瘤杂志*, 2023, 45(7): 575 – 583.
- [4] XIAO YW, WU J, LIN ZL, et al. A deep learning – based multi – model ensemble method for cancer prediction [J]. *Comput Methods Programs Biomed*, 2018, 153: 1 – 9.
- [5] ESTEVA A, KUPREL B, NOVOA RA, et al. Dermatologist – level classification of skin cancer with deep neural networks [J]. *Nature*, 2017, 542(7639): 115 – 118.
- [6] ALKHAWALDEH IM, ALBALKHI I, NASWHAN AJ. Challenges and limitations of synthetic minority oversampling techniques in machine learning [J]. *World J Methodol*, 2023, 13(5): 373 – 378.
- [7] AL'AREF SJ, SINGH G, VAN ROSENDAEL AR, et al. Determinants of In – Hospital Mortality After Percutaneous Coronary Intervention: A Machine Learning Approach [J]. *J Am Heart Assoc*, 2019, 8(5): e11160.
- [8] GARCIA – ORDAS MT, BENAVIDES C, BENITEZ – ANDRADES JA, et al. Diabetes detection using deep learning techniques with oversampling and feature augmentation [J]. *Comput Methods Programs Biomed*, 2021, 202: 105968.
- [9] HAO PY, CHIANG JH, CHEN YD. Possibilistic classification by support vector networks [J]. *Neural Netw*, 2022, 149: 40 – 56.
- [10] SCHOBEL P, VETTER TR. Logistic Regression in Medical Research [J]. *Anesth Analg*, 2021, 132(2): 365 – 366.
- [11] GELBARD RB, HENSMAN H, SCHOBEL S, et al. A random forest model using flow cytometry data identifies pulmonary infection after thoracic injury [J]. *J Trauma Acute Care Surg*, 2023, 95(1): 39 – 46.
- [12] ZOU Y, SHI Y, SUN F, et al. Extreme gradient boosting model to assess risk of central cervical lymph node metastasis in patients with papillary thyroid carcinoma: Individual prediction using SHapley Additive exPlanations [J]. *Comput Methods Programs Biomed*, 2022, 225: 107038.
- [13] LEE YW, CHOI JW, SHIN EH. Machine learning model for predicting malaria using clinical information [J]. *Comput Biol Med*, 2021, 129: 104151.
- [14] YAN P, DUAN SB, LUO XQ, et al. Development and validation of a deep neural network – based model to predict acute kidney injury following intravenous administration of iodinated contrast media in hospitalized patients with chronic kidney disease: a multicohort analysis [J]. *Nephrol Dial Transplant*, 2023, 38(2): 352 – 361.
- [15] 王昕阳, 魏静, 陈秋莎. 基于列线图模型评估 TG / HDL – C 和非 HDL – C / HDL – C 对腹膜透析患者预后的预测价值 [J]. *中国药业*, 2024, 33(5): 58 – 63.
- [16] LYMAN GH, KUDERER NM, CRAWFORD J, et al. Predicting individual risk of neutropenic complications in patients receiving cancer chemotherapy [J]. *Cancer*, 2011, 117(9): 1917 – 1927.
- [17] BA Y, SHI YK, JIANG WQ, et al. Current management of chemotherapy – induced neutropenia in adults: key points and new challenges: Committee of Neoplastic Supportive – Care (CONS), China Anti – Cancer Association Committee of Clinical Chemotherapy, China Anti – Cancer Association [J]. *Cancer Biol Med*, 2020, 17(4): 896 – 909.
- [18] CHEN WQ, PENG L, ZENG XL, et al. Predictors of Myelosuppression for Patients with Head and Neck Squamous Cell Carcinoma After Induction Chemotherapy [J]. *Clin Med Insights Oncol*, 2024, 18: 1373340457.
- [19] PEREZ – PITARCH A, GUGLIERI – LOPEZ B, NACHER A, et al. Impact of Undernutrition on the Pharmacokinetics and Pharmacodynamics of Anticancer Drugs: A Literature Review [J]. *Nutr Cancer*, 2017, 69(4): 555 – 563.
- [20] USISKIN I, LI F, IRWIN ML, et al. Association of relative dose intensity with BMI and pathologic complete response in patients treated with neoadjuvant chemotherapy for breast cancer [J]. *Breast Cancer Res Treat*, 2021, 186(1): 191 – 197.
- [21] CHAO C, RODRIGUEZ R, PAGE JH, et al. History of chronic comorbidity and risk of chemotherapy – induced febrile neutropenia in patients with non – Hodgkin lymphoma not receiving granulocyte colony – stimulating factor prophylaxis [J]. *Leuk Lymphoma*, 2015, 56(1): 72 – 79.
- [22] 黄家良, 夏坤健, 郭伟, 等. 三阴性乳腺癌化疗致重度骨髓抑制的危险因素分析及其预测模型的构建 [J]. *中山大学学报(医学科学版)*, 2023, 44(5): 886 – 892.
- [23] FERNANDES R, MAZZARELLO S, STOBEL C, et al. Primary Febrile Neutropenia Prophylaxis for Patients Who Receive FEC – D Chemotherapy for Breast Cancer: A Systematic Review [J]. *J Glob Oncol*, 2018, 4: 1 – 8.

(收稿日期: 2024 – 04 – 11; 修回日期: 2026 – 01 – 20)