

骨肉瘤患者2年肺转移风险机器学习预测模型的构建

王杰¹, 陈仲毅², 訾贵洋¹, 岳斌^{1*}

¹青岛大学附属医院骨与软组织肿瘤科, 山东 青岛

²青岛大学青岛医学院, 山东 青岛

收稿日期: 2026年3月13日; 录用日期: 2026年4月6日; 发布日期: 2026年4月14日

摘要

目的: 构建基于机器学习技术的骨肉瘤患者2年肺转移风险预测模型, 并精准识别影响肺转移发生的关键预测指标, 为临床早期干预提供参考依据。方法: 回顾性收集青岛大学附属医院2016年1月至2025年12月期间收治的骨肉瘤患者临床资料。肺转移的定义为随访2年内通过影像学检查明确证实存在肺部转移病灶。采用Boruta算法(实施10折交叉验证, 筛选被10折中超过9折判定为重要的特征)、相关性分析及LASSO回归三种方法联合筛选核心预测变量, 构建8种不同的机器学习模型。结果: 经多步特征筛选后, 最终保留6个关键预测特征, 分别为白蛋白、软组织受累、血红蛋白、碱性磷酸酶、C反应蛋白及乳酸脱氢酶。在构建的8种机器学习模型中, 逻辑回归(LR)模型表现最优, 其训练集曲线下面积(AUC)为0.93, 测试集AUC为0.92, 敏感性达92%, 特异性为82%, 展现出良好的泛化能力和预测可靠性。结论: 基于多层感知机(MLP)构建的可解释性模型, 能够有效识别骨肉瘤肺转移的高危患者, 该模型预测效能稳定, 具备较高的临床推广应用价值。

关键词

骨肉瘤, 肺转移, 机器学习, 风险预测, 可解释人工智能

Construction of a Machine Learning Prediction Model for the 2-Year Lung Metastasis Risk in Patients with Osteosarcoma

Jie Wang¹, Zhongyi Chen², Guiyang Zi¹, Bin Yue^{1*}

¹Department of Bone and Soft Tissue Tumors, The Affiliated Hospital of Qingdao University, Qingdao Shandong

*通讯作者。

文章引用: 王杰, 陈仲毅, 訾贵洋, 岳斌. 骨肉瘤患者 2 年肺转移风险机器学习预测模型的构建[J]. 临床医学进展, 2026, 16(4): 2920-2929. DOI: 10.12677/acm.2026.1641549

²Qingdao Medical College, Qingdao University, Qingdao Shandong

Received: March 13, 2026; accepted: April 6, 2026; published: April 14, 2026

Abstract

Objective: To develop a machine learning-based model for predicting the 2-year risk of pulmonary metastasis in patients with osteosarcoma, and to accurately identify key predictive indicators associated with pulmonary metastasis, thereby providing evidence for early clinical intervention. **Methods:** Clinical data of patients with osteosarcoma treated at the Affiliated Hospital of Qingdao University from January 2016 to December 2025 were retrospectively collected. Pulmonary metastasis was defined as the presence of lung metastatic lesions confirmed by imaging examinations within a 2-year follow-up period. Core predictive variables were jointly screened using the Boruta algorithm (with 10-fold cross-validation, selecting features identified as important in more than 9 of the 10 folds), correlation analysis, and least absolute shrinkage and selection operator (LASSO) regression. Subsequently, eight different machine learning models were constructed. **Results:** After multistep feature selection, six key predictive features were ultimately retained: albumin, soft tissue involvement, hemoglobin, alkaline phosphatase, C-reactive protein, and lactate dehydrogenase. Among the eight constructed machine learning models, the logistic regression (LR) model demonstrated the best performance, with an area under the curve (AUC) of 0.93 in the training set and 0.92 in the test set. The model achieved a sensitivity of 92% and a specificity of 82%, indicating good generalization ability and reliable predictive performance. **Conclusion:** An interpretable model based on a multi-layer perceptron (MLP) can effectively identify osteosarcoma patients at high risk of pulmonary metastasis. The model demonstrates stable predictive performance and has considerable potential for clinical application.

Keywords

Osteosarcoma, Pulmonary Metastasis, Machine Learning, Risk Prediction, Explainable Artificial Intelligence (XAI)

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

骨肉瘤(Osteosarcoma)作为临床最常见的原发性恶性骨肿瘤, 主要好发于青少年及年轻成人, 约占所有骨肿瘤发病总数的 35% [1]。肺转移是骨肉瘤患者最常见的远处转移类型, 据相关数据显示, 约 80% 的转移性骨肉瘤患者首发转移部位为肺部[2]。一旦发生肺转移, 患者的 5 年生存率将从初始的 60%~70% 急剧下降至 20% 以下[3]。因此, 早期精准识别肺转移高危患者, 对于制定个体化治疗方案、优化随访策略以及改善患者预后具有至关重要的临床意义。

目前, 临床上评估骨肉瘤患者预后时, 主要依赖肿瘤大小、Enneking 分期、组织学分级等传统临床指标, 但这些指标在预测肺转移发生风险方面的准确性存在明显局限[4]。近年来, 机器学习技术在医学预后预测领域呈现出巨大的应用潜力, 其能够自动捕捉各变量之间复杂的非线性关联, 有效提升预后预测的精准度[5]。

基于上述研究背景,本研究以青岛大学附属医院单中心临床数据为基础,联合采用 Boruta 算法、相关性分析及 LASSO 回归方法进行特征筛选,构建并对比分析 8 种机器学习模型的预测性能,旨在开发一款具备临床实用价值的骨肉瘤肺转移风险预测工具。

2. 资料与方法

2.1. 研究设计

本研究为单中心回顾性队列研究,研究数据来源于青岛大学附属医院骨科 2016 年 1 月至 2025 年 12 月期间收治的骨肉瘤患者[4]。纳入标准如下:(1) 经病理学检查明确诊断为骨肉瘤;(2) 年龄在 10 岁及以上、80 岁及以下;(3) 初诊时无远处转移迹象;(4) 接受规范化治疗方案(手术治疗联合或不联合化疗);(5) 随访时间 ≥ 2 年,或随访期间发生肺转移。排除标准包括:(1) 合并其他类型恶性肿瘤;(2) 既往有肺部疾病病史;(3) 关键临床或实验室检查数据缺失率超过 20%;(4) 随访期间出现失访情况。肺转移(阳性事件)定义为随访 2 年内通过胸部 CT 或 PET-CT 检查证实存在肺部转移病灶[6]。

2.2. 数据收集

根据文献回顾和临床重要性最终确定了 33 个候选预测变量。其中包括:人口学特征(年龄、性别)、临床特征(肿瘤部位、肿瘤大小、Enneking 分期、软组织受累、病理性骨折)、实验室指标(血红蛋白、白细胞计数、血小板计数、碱性磷酸酶、乳酸脱氢酶、白蛋白、C 反应蛋白、血沉)、治疗相关因素(手术方式、新辅助化疗、辅助化疗周期数)等。

2.3. 统计分析

本研究采用 SPSS 25.0 统计软件开展数据统计分析工作。采用 Shapiro-Wilk 检验(W 检验)验证数据的正态性:符合正态分布的定量资料以均数 \pm 标准差($\bar{x} \pm s$)表示,组间差异比较采用独立样本 t 检验;不符合正态分布的定量资料以中位数(四分位数)[M(Q₁, Q₃)]表示,组间差异比较采用秩和检验。分类资料以例数(百分数)[n(%)]表示,组间差异比较采用 χ^2 检验。检验水准 α 设定为 0.05, $P < 0.05$ 表示差异具有统计学意义[7]。

3. 模型构建与验证

3.1. 特征选择

在特征选择阶段,首先采用实施 10 折交叉验证的 Boruta 算法,该算法是一种基于随机森林的全相关特征选择方法,能够有效识别与研究结局相关的所有特征,同时可高效处理高维数据及复杂的特征交互关系[8]。筛选标准为 10 次交叉验证中均被判定为 confirm 的变量,即被超过 9 折验证判定为重要的特征。随后进行相关性分析,变量间相关性采用 Spearman 相关系数(r)表示,将相关性较强的变量($|r| \geq 0.7$)判定为冗余变量,并从后续分析中剔除。最后采用 LASSO 回归进一步筛选变量,确定最优特征子集[8]。

3.2. 模型训练、验证和评估

为构建骨肉瘤患者肺转移风险预测模型,本研究采用 8 种机器学习模型进行预测分析,包括支持向量机(SVM)、K 近邻(KNN)、随机森林(RandomForest)、LightGBM、AdaBoost、梯度提升(GradientBoosting)、逻辑回归(Logistic Regression)及多层感知机(MLP)。采用 10 折交叉验证法对数据集进行划分,分别用于模型训练与测试,通过贝叶斯优化方法确定各模型的最优超参数。采用准确率、受试者工作特征曲线下面积(AUC)、敏感性、特异性、阳性预测值、阴性预测值、精确率、召回率、F1 评分等多项指标综合评

价模型预测性能；采用校准曲线评估模型的校准度，运用决策曲线分析(DCA)评估模型的临床实用价值[9]。

4. 结果

4.1. 基线特征

本研究最终纳入 350 例骨肉瘤患者相关资料，经筛选后纳入分析。两组患者在大部分临床及实验室指标上均衡性良好，具体可见表 1。

Table 1. Comparison of baseline characteristics between training set and validation set

表 1. 训练集和验证集患者基线特征比较

特征名	ALL	测试集	训练集	P 值
年龄(岁)	35.051 ± 23.409	36.667 ± 23.768	34.359 ± 23.268	0.347
BMI (kg/m ²)	21.807 ± 3.821	21.966 ± 3.863	21.738 ± 3.809	0.726
肿瘤大小(cm)	8.087 ± 3.025	8.314 ± 2.927	7.990 ± 3.068	0.359
碱性磷酸酶(U/L)	251.141 ± 66.253	261.860 ± 65.390	246.547 ± 66.221	0.029
乳酸脱氢酶(U/L)	268.436 ± 49.351	265.037 ± 50.553	269.892 ± 48.860	0.472
血沉(mm/h)	47.922 ± 16.347	49.269 ± 17.653	47.345 ± 15.757	0.548
CRP (mg/L)	7.827 ± 3.052	8.077 ± 3.232	7.720 ± 2.972	0.309
血红蛋白(g/L)	112.710 ± 10.848	112.471 ± 11.316	112.813 ± 10.663	0.828
血小板计数(×10 ⁹ /L)	281.372 ± 40.960	282.540 ± 32.852	280.871 ± 44.033	0.249
白蛋白(g/L)	35.118 ± 2.997	35.050 ± 3.174	35.147 ± 2.923	0.648
化疗周期数	4.837 ± 2.583	5.238 ± 2.059	4.665 ± 2.763	0.241
症状持续时间(月)	3.485 ± 2.969	3.315 ± 2.412	3.558 ± 3.179	0.909
疼痛评分	6.251 ± 2.317	6.210 ± 2.277	6.269 ± 2.338	0.802
性别(n, %)				0.019
男	148 (42.286)	34 (32.381)	114 (46.531)	
女	202 (57.714)	71 (67.619)	131 (53.469)	
原发部位				0.829
0	118 (33.714)	38 (36.190)	80 (32.653)	
1	75 (21.429)	25 (23.810)	50 (20.408)	
2	42 (12.000)	11 (10.476)	31 (12.653)	
3	73 (20.857)	20 (19.048)	53 (21.633)	
4	42 (12.000)	11 (10.476)	31 (12.653)	
组织学亚型				0.937
0	177 (50.571)	56 (53.333)	121 (49.388)	
1	53 (15.143)	14 (13.333)	39 (15.918)	
2	41 (11.714)	13 (12.381)	28 (11.429)	
3	49 (14.000)	14 (13.333)	35 (14.286)	
4	30 (8.571)	8 (7.619)	22 (8.980)	

续表

肿瘤分级				0.19
0	41 (11.714)	16 (15.238)	25 (10.204)	
1	122 (34.857)	40 (38.095)	82 (33.469)	
2	187 (53.429)	49 (46.667)	138 (56.327)	
T 分期				0.279
0	104 (29.714)	28 (26.667)	76 (31.020)	
1	138 (39.429)	37 (35.238)	101 (41.224)	
2	106 (30.286)	39 (37.143)	67 (27.347)	
3	2 (0.571)	1 (0.952)	1 (0.408)	
N 分期				1.0
0	297 (84.857)	89 (84.762)	208 (84.898)	
1	53 (15.143)	16 (15.238)	37 (15.102)	
病理性骨折				0.908
0	314 (89.714)	95 (90.476)	219 (89.388)	
1	36 (10.286)	10 (9.524)	26 (10.612)	
软组织侵犯				1.0
0	224 (64.000)	67 (63.810)	157 (64.082)	
1	126 (36.000)	38 (36.190)	88 (35.918)	
钙化程度				0.95
0	130 (37.143)	38 (36.190)	92 (37.551)	
1	106 (30.286)	33 (31.429)	73 (29.796)	
2	114 (32.571)	34 (32.381)	80 (32.653)	
骨膜反应				0.096
0	44 (12.571)	20 (19.048)	24 (9.796)	
1	94 (26.857)	29 (27.619)	65 (26.531)	
2	119 (34.000)	31 (29.524)	88 (35.918)	
3	93 (26.571)	25 (23.810)	68 (27.755)	
MRI 强化程度				0.323
0	97 (27.714)	29 (27.619)	68 (27.755)	
1	115 (32.857)	40 (38.095)	75 (30.612)	
2	138 (39.429)	36 (34.286)	102 (41.633)	
新辅助化疗				0.003
0	52 (14.857)	6 (5.714)	46 (18.776)	
1	298 (85.143)	99 (94.286)	199 (81.224)	
化疗方案				0.006
-1	52 (14.857)	6 (5.714)	46 (18.776)	
0	217 (62.000)	74 (70.476)	143 (58.367)	
1	81 (23.143)	25 (23.810)	56 (22.857)	

续表

化疗反应				0.007
-1	52 (14.857)	6 (5.714)	46 (18.776)	
0	117 (33.429)	38 (36.190)	79 (32.245)	
1	181 (51.714)	61 (58.095)	120 (48.980)	
手术方式				0.002
0	214 (61.143)	72 (68.571)	142 (57.959)	
1	85 (24.286)	13 (12.381)	72 (29.388)	
2	51 (14.571)	20 (19.048)	31 (12.653)	
手术切缘				0.301
-1	51 (14.571)	20 (19.048)	31 (12.653)	
0	216 (61.714)	63 (60.000)	153 (62.449)	
1	60 (17.143)	14 (13.333)	46 (18.776)	
2	23 (6.571)	8 (7.619)	15 (6.122)	
辅助化疗				0.674
0	116 (33.143)	37 (35.238)	79 (32.245)	
1	234 (66.857)	68 (64.762)	166 (67.755)	
既往肿瘤史				0.105
0	341 (97.429)	105 (100.000)	236 (96.327)	
1	9 (2.571)	null	9 (3.673)	
Paget 病				0.127
0	344 (98.286)	101 (96.190)	243 (99.184)	
1	6 (1.714)	4 (3.810)	2 (0.816)	
放疗史				1.0
0	335 (95.714)	100 (95.238)	235 (95.918)	
1	15 (4.286)	5 (4.762)	10 (4.082)	

4.2. 特征选择

在训练集中应用 10 折交叉验证的 Boruta 算法, 选择 9 次都被判定为 confirm 的变量, 最终从 33 个候选变量中确认了 9 个重要变量(图 1(a))。相关性检验未发现强相关变量($|r| < 0.7$)。随后采用 LASSO 回归进一步筛选, 最终确定 6 个最优预测特征: 白蛋白、软组织受累、血红蛋白、碱性磷酸酶、C 反应蛋白及乳酸脱氢酶。各特征间共线性均 <5 , 之后纳入机器学习模型(图 1(b)、图 1(c))。

4.3. 模型建立与验证

将 8 种机器学习模型在训练队列与测试队列中的表现进行同步评估。随机森林存在严重的过拟合(图 2(a)、图 2(b))。相比之下, 逻辑回归(LR)展现了最优的泛化稳健性, 它不仅在两个队列中均维持了高水平的预测效能(训练集 AUC = 0.93, 测试集 AUC = 0.92), 更具备优异的灵敏度(0.92)和阴性预测值(0.82), 表明其在识别高危患者和排除低危人群方面均优于其他传统模型(图 3(a)、图 3(b))。决策曲线分析进一步证实, 在广泛的风险阈值范围内, LR 逻辑回归模型在临床应用中能提供良好的净获益(图 2(c)、图 2(d))。

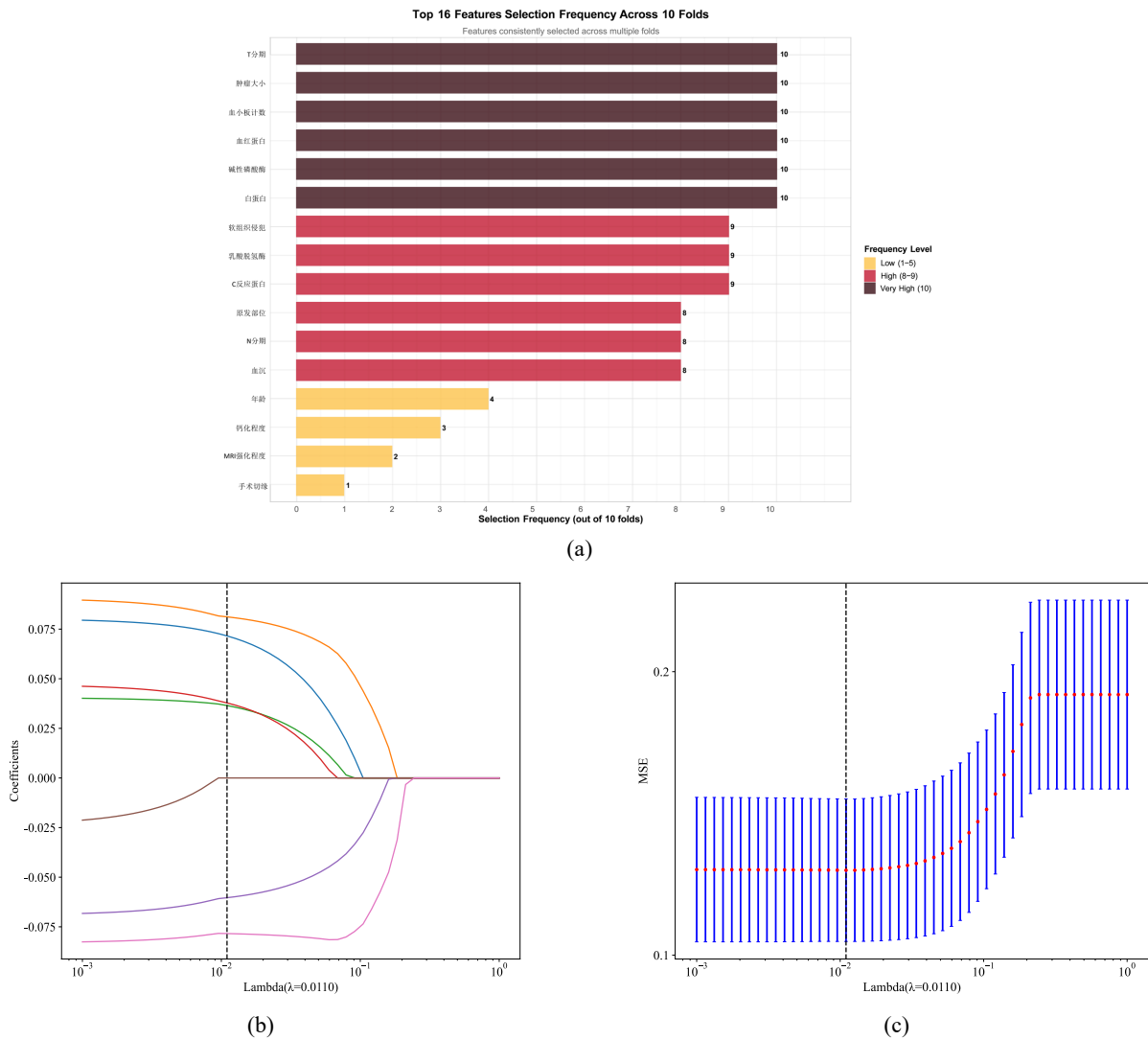
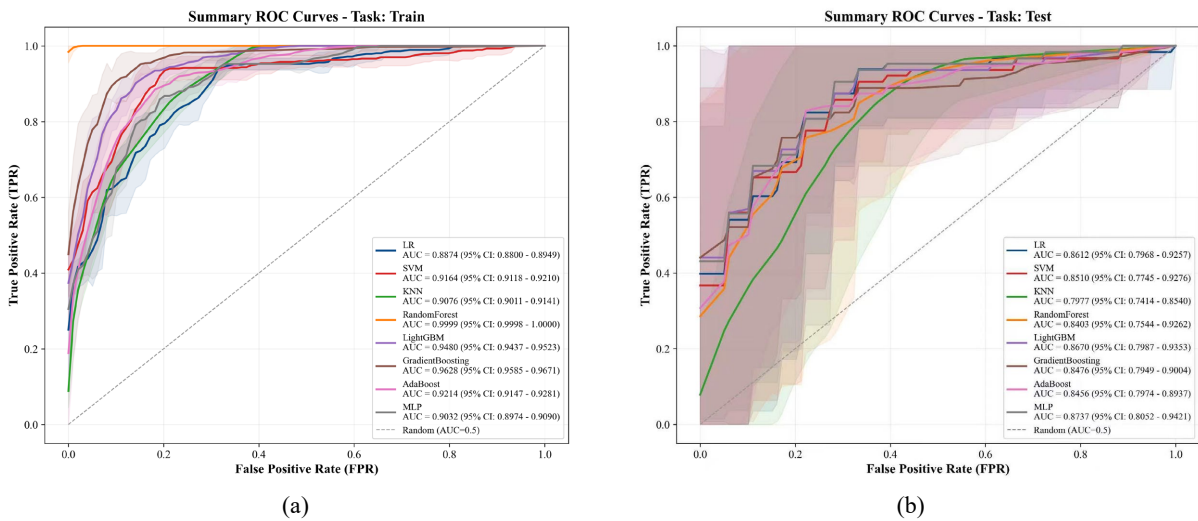


Figure 1. Boruta algorithm and LASSO screening of important variables
图 1. Boruta 算法和 LASSO 筛选重要变量



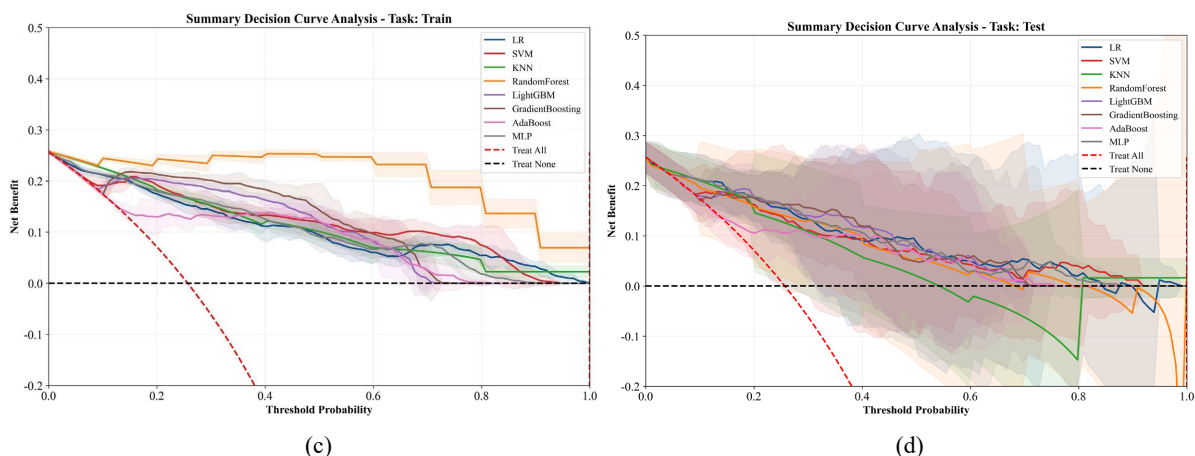
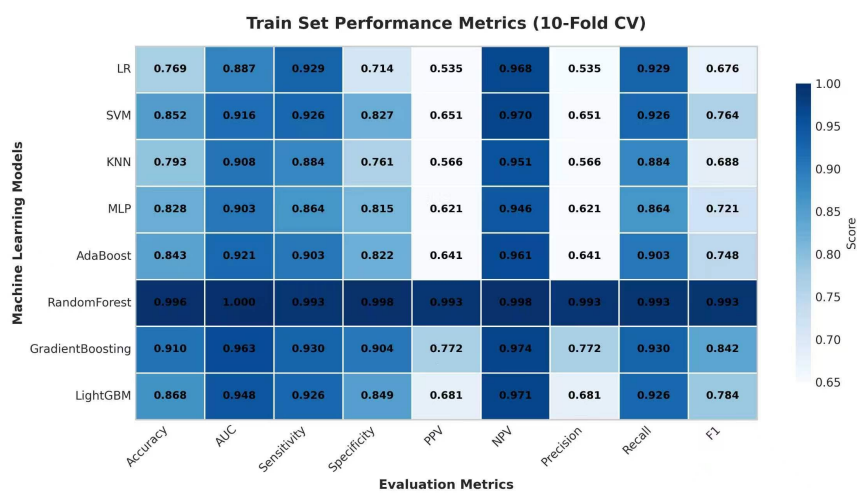
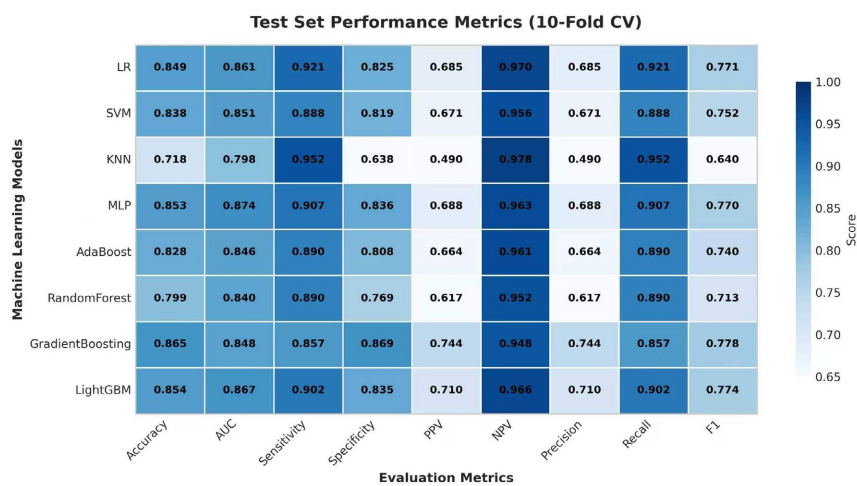


Figure 2. Subject work characteristics curves and clinical decision analysis curves of 8 machine learning algorithms
图 2. 8 种机器学习算法的受试者工作特征曲线和临床决策分析曲线



(a)



(b)

Figure 3. Model performance comparison of 8 machine learning algorithms
图 3. 8 种机器学习算法的模型性能比较

5. 讨论

骨肉瘤是临床最常见的原发性恶性骨肿瘤，肺转移现象的发生是影响患者预后效果、制约患者生存期改善的核心瓶颈。针对这一临床亟待解决的难题，本研究团队成功构建并完成了一种基于逻辑回归(LR)的肺转移风险预测模型的验证工作。经单中心数据验证显示，该模型具备良好的预测可靠性，其中测试集曲线下面积(AUC)达到 0.92，敏感性为 92%，其预测效果显著优于传统临床评估指标的预测能力。

在模型中，白蛋白被证实为最关键的保护性指标。白蛋白作为反映机体营养状况的核心指标，其水平降低通常预示着机体营养不良及免疫功能受损。已有相关研究证实，营养不良状态与肿瘤患者的不良预后密切相关[9]，对于骨肉瘤患者而言，低白蛋白水平可能因机体免疫功能下降，导致肿瘤细胞更易发生远处转移。此外，白蛋白还承担着抗氧化、抗炎以及维持机体胶体渗透压等多种重要生理功能，其功能的减退可能为肿瘤细胞通过血行播散、在肺部定植生长创造有利条件[10]。

软组织受累是骨肉瘤肺转移的核心风险因子，这一结论与骨肉瘤自身的生物学特性高度契合。软组织受累意味着肿瘤细胞已穿透骨皮质屏障，具备更强的侵袭能力和转移潜能。相关研究成果已明确证实，软组织受累是骨肉瘤肺转移发生的独立危险因素[11]，这与本研究的研究发现高度一致。同时，乳酸脱氢酶、碱性磷酸酶及 C 反应蛋白作为反映肿瘤负荷与机体全身炎症状态的重要指标，均与骨肉瘤肺转移风险呈现正相关关系。其中，乳酸脱氢酶作为糖酵解过程中的核心酶类，其水平升高提示肿瘤细胞代谢活动异常活跃[6]；碱性磷酸酶可反映成骨细胞的活性，与骨肉瘤的成骨特性密切相关；而 C 反应蛋白作为典型的急性时相反应蛋白，能够有效反映机体的炎症状态[12]。

尽管本研究在模型预测性能与临床解释性方面取得了阶段性进展，但仍存在一定的研究局限性。首先，本研究属于单中心回顾性研究，样本规模存在一定局限性，后续研究需在多中心、大样本的队列研究中进一步验证该模型的普适性，模型的预测结果能否直接转化为临床实践中患者预后的改善，仍需更多后续研究加以证实[9]。

综上，本研究成功研发并验证了一种基于 LR 算法的骨肉瘤患者 2 年肺转移风险预测模型。该模型通过整合白蛋白、软组织受累、血红蛋白、碱性磷酸酶、C 反应蛋白及乳酸脱氢酶等关键临床指标，不仅实现了优于传统临床指标的预测效果，还具备良好的泛化稳定性，为临床骨肉瘤肺转移的早期预警与干预提供了新的思路和工具。

声明

本研究获得青岛大学附属医院伦理委员会批准(审批号: QYFYWZLL42264)。

参考文献

- [1] 王凤伟, 王臻, 李靖, 等. 早期进展性骨肉瘤发生率与预后的单中心回顾性研究[J]. 中国骨与关节杂志, 2021, 10(8): 572-576.
- [2] 何涛, 唐冰, 熊敏, 等. 基于 SEER 数据库成人骨肉瘤肺转移病人的预后分析[J]. 骨科, 2024, 15(6): 534-540.
- [3] 李文乐, 胡朝晖, 王永辉, 等. 列线图在骨肿瘤疾病的应用与研究进展[J]. 中国骨与关节杂志, 2021, 10(8): 624-629.
- [4] 陈维, 潘德润, 陈卫国. 人工智能在骨肿瘤影像学中的应用研究进展[J]. 临床放射学杂志, 2025, 44(5): 953-956.
- [5] 刘璐, 郑新宇. 乳腺癌患者预后影响因素的 Cox 模型分析[J]. 中国卫生统计, 2021, 38(3): 458-461
- [6] 丁云. 炎症和营养相关指标与胃癌患者预后的相关性研究[D]: [硕士学位论文]. 恩施: 湖北民族大学, 2022.
- [7] Huang, G., Zhang, X., Xu, Y., Chen, S., Cao, Q., Liu, W., *et al.* (2024) Prognostic and Predictive Value of Super-Enhancer-Derived Signatures for Survival and Lung Metastasis in Osteosarcoma. *Journal of Translational Medicine*, 22, Article No. 88. <https://doi.org/10.1186/s12967-024-04902-8>
- [8] Pereira, H.M., Leite Duarte, M.E., Ribeiro Damasceno, I., de Oliveira Moura Santos, L.A. and Nogueira-Barbosa, M.H.

-
- (2021) Machine Learning-Based CT Radiomics Features for the Prediction of Pulmonary Metastasis in Osteosarcoma. *The British Journal of Radiology*, **94**, 20201391. <https://doi.org/10.1259/bjr.20201391>
- [9] Zheng, S., Chen, L., Wang, J., Wang, H., Hu, Z., Li, W., *et al.* (2023) A Clinical Prediction Model for Lung Metastasis Risk in Osteosarcoma: A Multicenter Retrospective Study. *Frontiers in Oncology*, **13**, Article 1001219. <https://doi.org/10.3389/fonc.2023.1001219>
- [10] Su, Z., Huang, F., Yin, C., Yu, Y. and Yu, C. (2023) Clinical Model of Pulmonary Metastasis in Patients with Osteosarcoma: A New Multiple Machine Learning-Based Risk Prediction. *Journal of Orthopaedic Surgery*, **31**, Article 10225536231177102. <https://doi.org/10.1177/10225536231177102>
- [11] Zumárraga, J.P., Baptista, A.M., Rosa, L.P.D.L., Caiero, M.T. and Camargo, O.P.D. (2016) Serum Values of Alkaline Phosphatase and Lactate Dehydrogenase in Osteosarcoma. *Acta Ortopédica Brasileira*, **24**, 142-146. <https://doi.org/10.1590/1413-785220162403157033>
- [12] Li, W., Liu, W., Hussain Memon, F., Wang, B., Xu, C., Dong, S., *et al.* (2022) An External-Validated Prediction Model to Predict Lung Metastasis among Osteosarcoma: A Multicenter Analysis Based on Machine Learning. *Computational Intelligence and Neuroscience*, **2022**, 2220527. <https://doi.org/10.1155/2022/2220527>