技术创新

基于支持向量机的肝外胆管癌术前淋巴结状态预测模型构建

林泰安[1]，罗亚梅[1]，黄志伟[1]，杨　录[2,3]，要小鹏[1]
1.西南医科大学 医学信息与工程学院（泸州 646000）；2. 西南医科大学附属医院 放射科（泸州 646000）；3.核医学与分子影像四川省重点实验室（泸州 646000）

【摘　要】目的 开发一种预测肝外胆管癌淋巴结状态的机器学习模型。 **方法** 纳入101例接受根治性手术切除的肝外胆管癌病例，采用MaZda软件对多序列MRI图像病灶进行勾画，并且提取影像特征。利用最大-最小算法对影像特征进行归一化，并通过合成少数过采样算法进行分类数据平衡，生成新样本。采用Spearman相关性分析与最大相关最小冗余特征选择法进一步筛选影像特征，最终得到20个最具代表的特征。从新样本数据中任意选择80%样本作为训练集，剩余20%作为测试集，建立支持向量机（support vector machine，SVM）预测模型，并利用受试者工作特征曲线（ROC）评价模型性能。 **结果** 预测模型训练集的AUC为0.98，准确率为89.2%，灵敏度为92.9%，特异性为89.4%。测试集的AUC为0.83，准确率为82.2%，灵敏度为82.1%，特异性为80.9%。 **结论** 基于MRI影像的SVM预测模型具有良好的预测性能，可为临床医生提供对肝外胆管癌患者个性化的术前预测，辅助评估手术价值并做出适当的临床决策。

【关键词】肝外胆管癌；淋巴结转移；机器学习；支持向量机；辅助诊断
【中图分类号】R735.8，，TP18　　**文献标志码**　A　　DOI：10.3969/j.issn.2096-3351.2023.04.006

The prediction model based on support vector machine for preoperative lymph nodes status in extrahepatic cholangiocarcinoma

LIN Taian[1]，LUO Yamei[1]，HUANG Zhiwei[1]，YANG Lu[2,3]，YAO Xiaopeng[1]
1.*School of Medical Information and Engineering, Southwest Medical University, Luzhou 646000, China*；2.*Department of Radiology, The Affiliated Hospital of Southwest Medical University, Luzhou 646000, China*；3. *Nuclear Medicine and Molecular Imaging Key Laboratory of Sichuan Province, Luzhou 646000, China*

【Abstract】Objective To develop a machine learning model for predicting lymph node status of extrahepatic cholangiocarcinoma. **Methods** The multi-sequence MRI images of 101 patients with extrahepatic cholangiocarcinoma were retrospectively studied, which had underwent radical resection. MaZda software was used to delineate the lesions and extract the image texture features. The radiomics features were normalised using a max-min algorithm, and a synthetic minority oversampling algorithm was balanced the classification samples to generate new samples. The Spearman correlation analysis and the maximum correlation and minimum redundancy feature selection method were used to select the image features, and 20 most representative features were finally obtained. We arbitrarily selected 80% of the samples from the new sample data as the training set and the remaining 20% as the test set. The SVM prediction model was built and the performance of this model was evaluated using the subject operating characteristic curve (ROC). **Results** The AUC of the training group was 0.98, the accuracy was 89.2%, the sensitivity was 92.9%, and the specificity was 89.4%. The AUC of the test group was 0.83, the accuracy was 82.2%, the sensitivity was 82.1%, and the specificity was 80.9%. **Conclusion** The SVM prediction model based on MRI image had good prediction performance, which provided a personalized preoperative auxiliary diagnosis strategy for clinicians.

【Key words】Extrahepatic cholangiocarcinoma; Lymph node metastasis; Machine learning; Support vector machine; Assisted diagnosis

胆管癌（cholangiocarcinoma）是一种较为罕见的肝胆恶性肿瘤，具有扩散转移快、预后差、生存周期短等特点[1-2]。根据解剖位置通常分为肝内胆管癌（intrahepatic cholangiocarcinoma, iCCA）和肝外胆管癌（extrahepatic cholangiocarcinoma，eCCA）。过去20年间，大多数国家的胆管癌发病率呈逐年上升趋势，尤其是东南亚地区，其中eCCA占所有原发性肝癌的5%～30%[3-4]。目前，外科手术切除是eCCA患者获得长期生存的唯一

基金项目：四川省科技厅科技计划项目（2020YJ0151，2022YFS0616）；四川省科技厅区域创新合作项目（2021YFQ0002）
通信作者：要小鹏，E-mail:xp_yao@swmu.edu.cn
引信本文：林泰安，罗亚梅，黄志伟，等.基于支持向量机的肝外胆管癌术前淋巴结状态预测模型构建[J].西南医科大学学报，2023，46（4）：303-307.DOI：10.3969/j.issn.2096-3351.2023.04.006

治疗手段。根据文献报道，淋巴结转移（lymph node metastasis, LNM）被认为是胆管癌切除手术最重要的预后因素之一，因此在术前准确识别eCCA患者淋巴结转移状况对确定最佳治疗方案和预后效果评价极为重要[5-8]。

当前，临床上大多采用计算机断层扫描（CT），正电子发射断层扫描/计算机断层扫描（PET/CT），以及磁共振成像（MRI）进行胆管癌放射学诊断，尤其是MRI能更直观的显示病灶的形态学特征，在胆管癌病变评估和术前检查中普遍应用[9-11]。传统的影像诊断主要依靠医生的主观判断进行定性分析，易受主观因素影响而导致诊断结果出现偏差，进而产生误诊和漏诊。近年来，随着人工智能技术的不断发展，影像组学（radiomics）结合传统影像学和机器学习算法构建医学图像预测模型已成为研究热点，许多学者利用其对乳腺癌[12-13]、直肠癌[14]、肺癌等[15]患者的淋巴结转移进行了预测研究，取得了很好的预测效果。

上述研究主要是进行定性预测，而利用MRI影像数据开展定性和定量预测的研究较少。因此，本研究针对eCCA患者，提取MRI影像中DWI、T1WI、T2WI和ADC四种序列的图像特征，结合病理结果利用支持向量机（support vector machine，SVM）算法建立模型，用于预测eCCA患者术前淋巴结状态。

1 对象和方法

1.1 研究对象

本研究属于回顾性研究，获得西南医科大学附属医院伦理委员会批准，收集选取了2011年11月至2021年12月于西南医科大学附属医院接受根治性手术切除的101例eCCA患者病例资料（其中男性56例，女性45例，平均年龄57岁）。搜集数据包括患者临床症状、MRI图像和病理结果（包括淋巴结状态）。为保护患者隐私，所收集到的数据信息都进行了脱密处理。纳入标准：①患者术前未接受化疗等其他抗肿瘤治疗；②患者病变部位病理分析结果证实为肝外胆管癌。排除标准：①术前2周内无MRI影像；②MRI图像模糊；③存在不明或不明显的病变；④患者MRI资料不完整。

在最初收集的172例eCCA患者数据中，排除了71例，主要原因是检查信息不足，例如，没有术前完整的MRI图像（n = 38）、MRI图像模糊（n = 10）、存在不明或不明显的病变（n = 5）以及缺乏部分淋巴结病理活检数据（n = 18）等。最终有101名患者数据可用于构建eCCA的淋巴结预测模型，如图1所示。将eCCA患者分为两分类样本，其中淋巴结转移组为正样本，标记为1 [LNM（+）；n = 29；label = 1]；淋巴结未转移组为负样本，标记为0，[LNM（−）；n = 72；label = 0]。

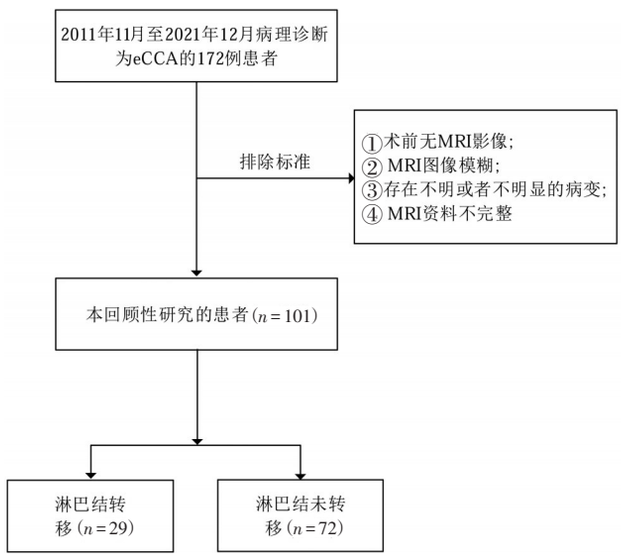


图1 病例纳入和排除流程图

Figure 1 Flowchart of case Inclusion and exclusion

1.2 感兴趣区域的勾画

本研究采用开源医学图像特征提取软件MaZda（version 4.6）对MRI图像的感兴趣区域（ROI）进行勾画。先将患者影像数据导入MaZda软件，根据医师判断手动沿ROI边缘分别对DWI、T1WI、T2WI以及ADC序列的图像进行勾画。为降低人为因素所导致的主观偏差，选取经验丰富的放射科医师在预先不知病理结果的情况下完成勾画，如图2所示。

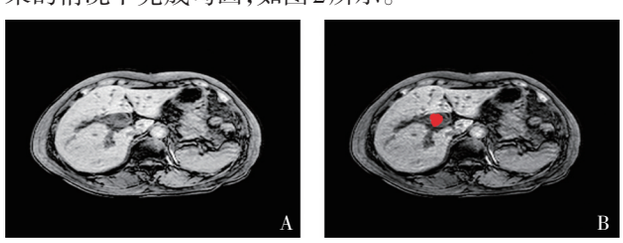


图2 MaZda软件勾画ROI示意图

Figure 2 ROI delineated by MaZda software

注：A：勾画前图像；B：勾画后图像。

1.3 特征提取与筛选

1.3.1 *数据预处理* 本研究采用MaZda软件对ROI区域进行放射学图像特征提取，从每个序列的MRI图像可以提取302个图像特征，共计获取1 208个放射学图像特征。为消除不同数据量纲影响，利用MATLAB软件采用最大-最小算法（Max-Min algorithm）对1 208个特征进行归一化。

1.3.2 *SMOTE算法* 根据机器学习的基本理论，如果数据样本存在严重的不平衡性，预测结果易出现偏异性，主要体现在预测结果会向样本量较多的一类倾斜。为有效解决样本数据不平衡问题，研究人员提出了合成少数类过采样算法（Synthetic Minority Oversampling Technique，SMOTE）来解决该类问题[16]。本研究中，淋巴结转移患者（n = 29）与未转移患者（n = 72）数据量很不平衡，直接影响预测模型的泛化性。从小样本数据

的角度看，适合采用SMOTE算法来实现样本数量平衡。通过计算出少数类样本的m个近邻值，从近邻值中随机选取n个样本进行随机线性插值来构建新的少数类样本，并将其与原数据合成，产生新的训练集，实现分类样本的数据平衡。

1.3.3 特征筛选　由于少样本高维度数据中存在大量的冗余特征，会直接影响预测模型性能。本研究采用Spearman相关性分析和最大相关最小冗余特征算法（Max-Relevance and Min-Redundancy, mRMR）对放射学图像特征进行初次筛选和二次筛选，尽可能降低冗余特征对预测模型的影响。首先，采用Spearman相关性分析遍历所有特征，计算两两相邻特征间的相关系数r。同时，将$|r| \geqslant 0.9$的特征从特征数据集中排除，完成特征初步筛选。其次，使用mRMR算法对剩余特征进行筛选，计算特征之间、特征与变量之间的互信息量分布，通过互信息量排序保留了前20个特征作为优选特征（表1），完成二次筛选。

1.4 预测模型的构建

支持向量机（Support Vector Machine，SVM）是VAPNIK在1995年所提出的机器学习算法，它的本质是利用核函数建立数据高维空间的非线性映射模型，对小样本数据分类有很好的预测效果[17]。本研究以放射学图像特征为自变量，术前淋巴结状态标签为因变量，采用SVM算法建立预测模型。在建模过程中采用Matlab计算软件进行数据预处理和特征选择，根据小样本数据划分训练集和测试集的准测，按照8:2的比例将病例样本分为训练集和测试集[18]。为提高预测模型精度和泛化性，在建模过程中通过网格搜索和交叉验证优化惩罚因子和核函数参数，并在测试集中利用受试者工作特征曲线（ROC）评价该模型性能。

2 结果

2.1 特征选择

利用Mazda软件对患者每个MRI序列影像分别提取302个放射学图像特征，总共获取1 208个放射学图像特征，然后进行归一化处理，消除量纲影响。由于上述特征中存在冗余信息和噪声信息，会对预测模型造成偏差，故通过筛选特征数据显示原始数据的本质结构特征。首先，使用SPSS软件进行Spearman相关性分析，筛选出具有高度相似性的特征，共计824个，删除后所得特征共计384个。然后，利用最大相关最小冗余特征选择法（mRMR）筛选剩余的384个放射学图像特征，根据计算互信息量结果，最终选取20个特征用于建立预测模型，如表1所示，其中序号1-10为T1WI序列图像特征，序号11-15为T2WI序列影像特征，序号16-20为ADC序列放射学图像特征，任选一组特征图形化描述，如图3所示。

表1 筛选后的图像特征表

Table1 Image features after screening

序号	图像特征名称
1	Perc.10%_T1WI
2	Perc.50%_T1WI
3	Perc.99%_T1WI
4	S(4,0)SumAverg_T1WI
5	135dr_RLNonUni_T1WI
6	WavEnLL_s-2_T1WI
7	WavEnLL_s-4_T1WI
8	WavEnHL_s-4_T1WI
9	WavEnLL_s-5_T1WI
10	WavEnHH_s-5_T1WI
11	Skewness_T2WI
12	Kurtosis_T2WI
13	AreaARM_T2WI
14	WavEnLH_s-4_T2WI
15	WavEnHH_s-5_T2WI
16	S(5,5)SumAverg_ADC
17	S(5,-5)Contrast_ADC
18	S(5,-5)DifEntrp_ADC
19	GrKurtosis_ADC
20	WavEnLL_s-3_ADC

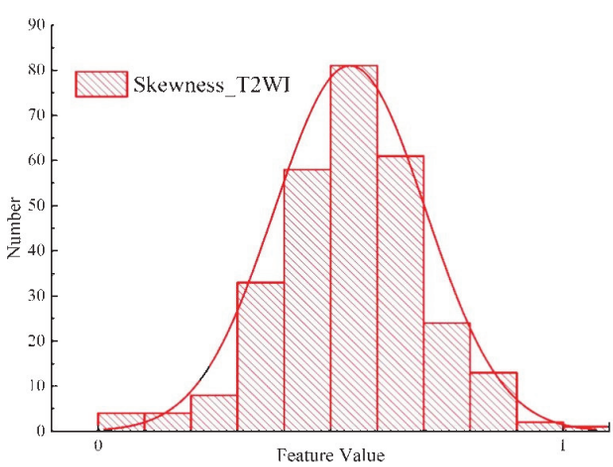


图3 Skewness_T2WI层特征直方图

Figure 3 Histogram of Skewness_T2WI layer feature

2.2 模型建立

在本研究中，eCCA患者淋巴结无转移病例为72份，淋巴结有转移的病例为29份，通过SMOTE算法设置合适的采样倍率N[LNM(-)，N = 2；LNM(+)，N = 5]后，总共构建了新样本289份[LNM(-)，n = 144；LNM(+)，n = 145]，从而基本实现了样本数据量平衡。从289份新样本数据中任意选择一定数量样本作为训练集，剩余的作为测试集，训练集与测试集之比为8:2。

本研究采用SVM算法，选用径向基（RBF）作为核函数，以筛选后的病灶放射学图像特征为自变量，淋巴结转移标签为因变量建立预测模型。将训练集中的两分类样本[LNM(+)；LNM(-)]输入SVM模型进行建模，

通过优化惩罚因子c和RBF核函数参数g，以获取预测模型最优的预测效果。

2.3 优化参数

为提高预测模型的性能，通过网格搜索和交叉验证优化惩罚因子c和RBF核函数参数g，以获取SVM模型最优的预测效果。通过程序计算可知不同的c、g值，对应不同的预测精度，如图3所示。从中搜索预测精度最好的一组c、g值，即为最优参数，则预测模型的最优参数为：c = 0.57435、g = 5.278，预测精度为86.6337%，如图4所示。

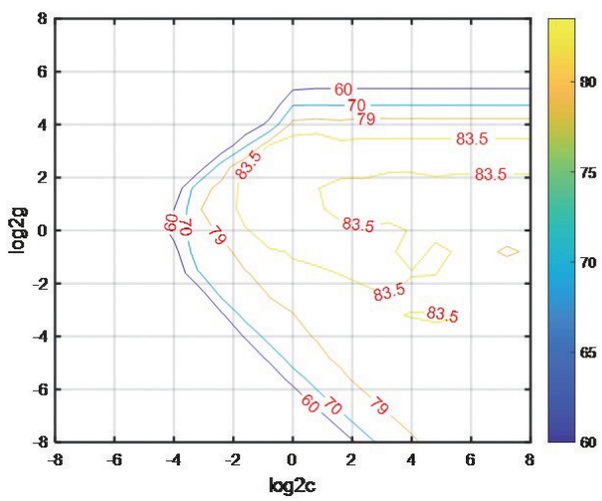


图3 参数c, g优化等高线图

Figure 3 Contour plots of optimized parameters c, g

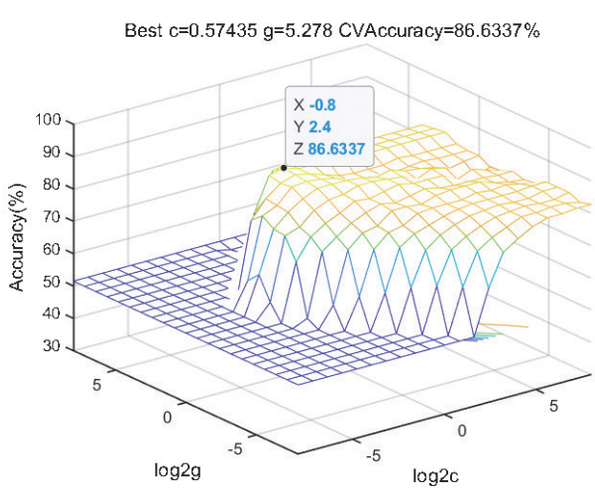


图4 SVM参数选择结果图

Figure 4 Results of SVM parameter selection

2.4 机器学习模型预测性能评价

为评价SVM预测模型性能，我们利用接收器操作特征曲线（ROC）来评估分类准确性，该模型的预测性能如图5所示。预测模型训练集的AUC为0.98，准确率为89.2%，灵敏度为92.9%，特异性为89.4%。预测模型测试集的AUC为0.83，准确率为82.2%，灵敏度为82.1%，特异性为80.9%。

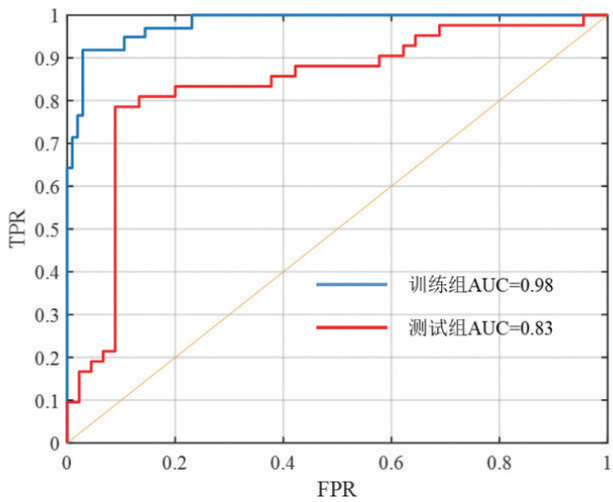


图5 训练集和测试集ROC曲线

Figure 5 ROC curves of training and testing groups

3 讨论

eCCA是一种高度侵袭性的恶性肿瘤，其病因可能与肝部胆管结石、原发性硬化性胆管炎等疾病有关。尽管eCCA患者的临床治疗策略不尽相同，但是根治性手术切除仍然是eCCA患者比较有效的治疗方法。一般情况下，eCCA患者存在术后预后差、5年生存率低等情况，早发现早治疗对根治eCCA有重要的临床意义。在临床诊断中，LNM是否转移被认为是预测eCCA患者术后生存率的重要因素之一[19]。

超声波、CT、PET/CT和MRI等医学影像技术的飞速发展，为肿瘤患者在诊断、治疗、疗效评估、复查检验等方面提供了更精准可靠的方法。然而，这些影像技术仍然受限于放射科医生的主观评价，诊断结果易受医生个人经验、临床水平等因素影响，缺乏定量评价。由于eCCA无特异性症状和肿瘤标志物，超声检查对识别胆管源性和转移灶有一定的缺陷，检查结果不明显[20]。尽管CT和MRI在eCCA术前检查方面有一定的优势，但对病灶体积小、强化方式特异性差的病例仍有较大的误诊漏诊率[21-22]。PET/CT价格昂贵，在良性病变如胆道感染或硬化性胆管炎中可能会受到假阳性结果的影响[23-24]。因此，利用机器学习算法对eCCA患者术前淋巴结状态进行预测是一种可行的定性异质分析方法。

影像组学的本质就是通过从影像数据中提取高维、定量影像特征，然后利用机器学习算法量化描述病灶异质性。HUANG等[25]通过随机森林算法预测iCCA的淋巴结转移状况，其敏感性为35.2%，特异性为91.8%，准确性为46.1%。本研究利用SVM预测对eCCA患者淋巴结是否转移进行了定量分析，计算结果显示AUC为0.92，具有优良的预测性能。尽管大量研究表明LNM与eCCA患者的预后密切相关，但淋巴结

清扫术(lymph node dissection,LND)的临床效果仍有一定的争议[26]。YANG等[27]的研究发现在147例患者中,54.4%(80例)接受LND,其中42.5%(34/80)被发现有淋巴结转移(LNM)。因此,对eCCA患者LNM进行术前精准评估代表了未来个性化治疗的重要方向。本研究仍有一定局限性,首先是数据样本较少,且患者数据来源于同一所医院,样本不可避免存在偏倚性;其次,病灶的ROI分割自动人工勾画,存在一定的主观性偏差;另外,本研究输入特征只包含了影像组学特征,而缺少了患者临床特征,会降低预测模型的泛化性。下一步,我们将开展多中心、多组学和多任务的研究作为工作重点,旨在建立性能更优异的预测模型,以提高本研究的临床应用价值。

4 结论

本研究中,我们利用放射学图像特征建立了一种预测肝外胆管癌术前淋巴结状态的机器学习模型,该模型在训练集和测试集中都显示出较高的准确性、敏感性和特异性,具有良好的预测性能。该模型可有助于对eCCA患者进行个性化预测,辅助临床医生评估手术价值并做出适当的临床决策。同时,也有助于对eCCA患者进行精准治疗,避免由于前期症状不明显而导致错过最佳手术时机,进而尽最大可能延长患者术后生存期。

5 参考文献

[1] 贾明茹,曹飞,闫涛,等.肝门部胆管癌的诊断和治疗进展[J].国际外科学杂志,2022,49(8):563-566.
[2] 徐绽蕾.肝外胆管细胞癌的影像学诊断现状与展望[J].医学影像学杂志,2020,30(10):1936-1939.
[3] RUMGAY H, FERLAY J, DE MARTEL C, *et al*. Global, regional and national burden of primary liver cancer by subtype[J]. Eur J Cancer. 2022,161(1):108-118.
[4] ESNAOLA N F, MEYER J E, KARACHRISTOS A, *et al*. Evaluation and management of intrahepatic and extrahepatic cholangiocarcinoma[J]. Cancer, 2016, 122(9): 1349-1369.
[5] WELZEL T M, GRAUBARD B I, EL-SERAG H B, *et al*. Risk Factors for Intrahepatic and Extrahepatic Cholangiocarcinoma in the United States: A Population-Based Case-Control Study[J]. Clinical Gastroenterology and Hepatology, 2007, 5(10): 1221-1228.
[6] 黄皓川,余杰雄,伍隽华,等.肝内胆管癌手术治疗的术前诊断评估及临床治疗对预后的影响分析[J].齐齐哈尔医学院学报,2021,42(10):850-853.
[7] MCMILLAN R R, JAVLE M, KODALI S, *et al*. Survival following liver transplantation for locally advanced, unresectable intrahepatic cholangiocarcinoma [J]. Am J Transplant. 2022,22(3): 823-832.
[8] 杨春梅,舒健.胆管癌的影像学及相关人工智能研究进展[J].西南医科大学学报,2021,44(5):491-499.
[9] 张瑜,王继洲,陆朝阳,等.肝门部胆管癌淋巴结清扫的研究进展[J].中华肝胆外科杂志,2020,26(12):957-960.
[10] LI M D, LU X Z, LIU J F, *et al*. Preoperative Survival Prediction in Intrahepatic Cholangiocarcinoma Using an Ultrasound-Based Radiographic-Radiomics Signature [J]. J Ultrasound Med. 2022, 41(6):1483-1495.
[11] CHU H P, LIU Z L, LIANG W, *et al*. Radiomics using CT images for preoperative prediction of futile resection in intrahepatic cholangiocarcinoma [J]. Eur Radiol. 2021,31(4):2368-2376.
[12] CONTI A, DUGGENTO A, INDOVINA I, *et al*. Radiomics in breast cancer classification and prediction[J]. Semin Cancer Biol, 2021,72(7):238-250.
[13] PENG J B, PENG Y T, LIN P, *et al*. Differentiating infected focal liver lesions from malignant mimickers: value of ultrasound-based radiomics[J]. Clinical Radiology, 2022, 77(2): 104-113.
[14] BIBAULT J E, GIRAUD P, HOUSSET M, *et al*. Author Correction: Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer[J]. Scientific Reports, 2018, 8(1): 16914.
[15] CONG M D, FENG H, REN J L, *et al*. Development of a predictive radiomics model for lymph node metastases in pre-surgical CT-based stage IA non-small cell lung cancer[J]. Lung Cancer, 2020, 139: 73-79.
[16] ELREEDY D, ATIYA A F. Note on "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance"[J]. Information Sciences, 2019, 505: 32-64.
[17] BYVATOV E, SCHNEIDER G. Support vector machine applications in bioinformatics [J]. Appl Bioinformatics. 2003, 2 (2): 67-77.
[18] 周志华.机器学习[M].北京:清华大学出版社,2016.
[19] GEERS J, JAEKERS J, TOPAL H, *et al*. Predictors of survival after surgery with curative intent for perihilar cholangiocarcinoma. World J Surg Oncol. 2020 Nov 3;18(1):286.
[20] KONDO S, TAKAGI K, NISHIDA M, *et al*. Computer-Aided Diagnosis of Focal Liver Lesions Using Contrast-Enhanced Ultrasonography With Perflubutane Microbubbles[J]. IEEE Transactions on Medical Imaging, 2017, 36(7): 1427-1437.
[21] 肖潇,钟玲.扩散加权成像辅助磁共振胰胆管水成像对肝外胆管癌的诊断价值[J].肿瘤影像学,2019,28(5):325-330.
[22] WANG Y, LIU W, YU Y, *et al*. CT radiomics nomogram for the preoperative prediction of lymph node metastasis in gastric cancer[J]. European Radiology, 2020, 30(2): 976-986.
[23] NITTA N, OHGI K, SUGIURA T, *et al*. Prognostic Impact of Paraaortic Lymph Node Metastasis in Extrahepatic Cholangiocarcinoma[J]. World J Surg. 2021,45(2):581-589.
[24] FIZ F, MASCI C, COSTA G, *et al*. PET/CT-based radiomics of mass-forming intrahepatic cholangiocarcinoma improves prediction of pathology data and survival. [J]Eur J Nucl Med Mol Imaging. 2022 Aug;49(10):3387-3400.
[25] HUANG T, LIU H, LIN Z, *et al*. Preoperative prediction of intrahepatic cholangiocarcinoma lymph node metastasis by means of machine learning: a multicenter study in China. BMC Cancer. 2022 Aug 29;22(1):931.
[26] 王晓宇,尚华,周慧清,等.MRI在肝外胆管癌诊断中的应用[J].国际医学放射学杂志,2019,42(4):447-449.
[27] YANG F C, WU C K, BO Z Y, *et al*. The clinical value of regional lymphadenectomy for intrahepatic cholangiocarcinoma[J]. Asian J Surg, 2022,45(1):376-380.

(利益冲突:无)

(收稿日期:2022-10-29;修回日期:2023-02-05)