基于欠采样的影像组学机器学习模型术前预测子宫肌瘤高强度聚焦超声消融效果

崔运能[1,2],冯敏清[3,4,5],姚亮凤[2],严杰文[2],李闻瀚[6],黄燕平[6]
暨南大学附属第一医院[1]放射科,[3]妇科,广东 广州 510630;佛山市妇幼保健院[2]放射科,[5]妇科,广东 佛山 528000;[4]广州市第一人民医院妇科,广东 广州 510180;[6]佛山大学物理与光电工程学院,广东 佛山 528231

摘要:目的 探讨不同欠采样方法在解决小样本数据类别不平衡问题中的应用,以提高机器学习模型术前预测子宫肌瘤高强度聚焦超声(HIFU)消融效果的准确性。**方法** 收集在佛山市妇幼保健院就诊的140例HIFU治疗子宫肌瘤患者临床及影像学数据,其中高消融率组104例,低消融率组36例,提取患者MRI-T2WI影像组学特征,构建HIFU治疗机器学习预测模型。应用7种欠采样方法,即随机欠采样(RUS)、重复编辑最近邻(RENN)、全K最近邻(AllKNN)、近邻缺失-3(NM)、凝聚最近邻(CNN)、邻域清理规则(NCR)和实例硬度阈值(IHT),使用4种机器学习模型,即K最近邻(KNN)、随机森林(RF)、支持向量机(SVM)和多层感知机(MLP)共计构建28种预测模型处理类别不平衡数据,并通过5折交叉验证方法、以受试者工作特征曲线下面积(AUC)、准确率、召回率和特异性等评估各模型性能。**结果** 欠采样方法与机器学习模型交叉组合的结果为:4种最佳组合AUC即CNN-RF为0.772(95%置信区间:0.566~0.942)、NM-SVM为0.797(95%置信区间:0.600~0.950)以及CNN-KNN和NM-MLP均为0.822(95%置信区间分别为0.635~0.964、0.632~0.960)。各机器学习模型的AUC在欠采样后均显著增高,其中以MLP模型改善最明显;各模型的召回率也显著增加,即CNN-RF召回率增加0.389、NM-SVM为0.836、CNN-KNN为0.532、NM-MLP为0.372。**结论** 欠采样方法可有效解决小样本类别不平衡问题,为构建子宫肌瘤HIFU消融效果的机器学习预测模型提供新思路。

关键词:子宫肌瘤;磁共振成像;高强度聚焦超声;机器学习;预测模型;类别不平衡;影像组学;欠采样

Enhancement of radiomics-based machine learning models for predicting efficacy of high-intensity focused ultrasound ablation of uterine fibroids using undersampling methods

CUI Yunneng[1,2], FENG Minqing[3,4,5], YAO Liangfeng[2], YAN Jiewen[2], LI Wenhan[6], HUANG Yanping[6]
[1]Department of Radiology, [3]Department of Gynecology, First Affiliated Hospital of Jinan University, Guangzhou 510630, China; [2]Department of Radiology, [5]Department of Gynecology, Foshan Women and Children Hospital, Foshan 528000, China; [4]Department of Gynecology, Guangzhou First People's Hospital, Guangzhou 510180; [6]School of Physics and Optoelectronic Engineering, Foshan University, Foshan 528000, China

Abstract: Objective To improve the accuracy of machine learning models for preoperative prediction of high-intensity focused ultrasound (HIFU) ablation efficacy for uterine fibroids by correcting class imbalance in small sample datasets using undersampling methods. **Methods** Clinical and imaging data were collected from 140 patients with uterine fibroids undergoing HIFU treatment at Foshan Women and Children Hospital, including 104 with high ablation rates and 36 with low ablation rates. Radiomic features were extracted from MRI T2-weighted images (T2WI) of the patients, and machine learning models were constructed to predict HIFU treatment outcomes. Four machine learning algorithms, including k-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP), were coupled with 7 undersampling methods, namely Random Undersampling (RUS), Repeated Edited Nearest Neighbors (RENN), All k-Nearest Neighbors (AllKNN), Neighborhood Cleaning Rule-3 (NM), Condensed Nearest Neighbor (CNN), Neighborhood Cleaning Rule (NCR), and Instance Hardness Threshold (IHT), for handling class imbalance in the datasets. The 28 prediction models were evaluated using 5-fold cross-validation for areas under the receiver operating characteristic curve (AUC), accuracy, recall, and specificity. **Results** The best combinations of undersampling methods and machine learning models CNN-RF, NM-SVM, CNN-KNN, and NM-MLP had AUCs of 0.772 (95% *CI*: 0.566-0.942), 0.797 (95% *CI*: 0.600-0.950), 0.822 (95% *CI*: 0.635-0.964), and 0.822 (95% *CI*: 0.632-0.960), respectively. The AUCs of the machine learning models significantly increased after coupling with undersampling methods, with the MLP model showing the most pronounced improvement. The recall rates of the 4 combined models also improved significantly (by 0.389 for CNN-RF, 0.836 for NM-SVM, 0.532 for CNN-KNN, and 0.372 for NM-MLP). **Conclusion** The use of undersampling methods can effectively correct class imbalance in small sample datasets to improve the accuracy of machine learning models for predicting the efficacy of HIFU ablation for uterine fibroids.

Keywords: uterine fibroid; magnetic resonance imaging; high-intensity focused ultrasound; machine learning; prediction; class imbalance; radiomics; undersampling

收稿日期:2025-06-20
基金项目:广东省医学科研基金(B2019161);佛山市医学影像精准诊断工程技术研究中心(FS0AA-KJ819-4901-0049)
作者简介:崔运能,在读博士研究生,E-mail: letitb@163.com
通信作者:黄燕平,副教授,E-mail: yale.huangyp@fosu.edu.cn

子宫肌瘤是生殖系统最常见的肿瘤之一,在女性生命历程中,约70%的白人女性和80%的非裔美国女性会受此疾病困扰[1]。子宫肌瘤予社会及经济带来沉重负担,降低女性生活质量,是子宫切除术的主要原因[2]。

虽然子宫肌瘤是良性疾病，但它引起患者不适症状，如月经过多、盆腔疼痛等，也可使子宫结构、功能发生变化，从而导致子宫内膜容受性降低、排卵障碍，甚至不孕。子宫肌瘤的治疗手段较多，包括药物治疗、腹腔镜肌瘤剔除术、腹腔镜热消融术和子宫动脉栓塞术等，但均具有其优点及不足之处[3]，如腹腔镜肌瘤剔除术为临床上最常用的手术方式，但它也存在局限性，属于侵入性术式，过于依赖主刀医生的手术技巧、可出现术中缝合困难等，也可能在后续妊娠期间存在手术出血和子宫破裂的风险[4]。因此，开发新型有效、安全的治疗方法仍有必要。

高强度聚焦超声（HIFU）是治疗子宫肌瘤的一种新型非侵入性方法，其有效性及安全性已经得到广泛认可[5]。它具备诸多优点，如无需手术切口、无辐射暴露、并发症率低、恢复时间短以及保留卵巢功能等[6]。HIFU通过将超声聚焦在目标组织，使其温度迅速升高至60°C以上，从而达到消融组织、破坏肿瘤，同时保留周围正常组织结构的目的[7]。然而，HIFU治疗效果在患者之间差异较显著，并非所有患者都能从治疗中获益。因此，筛选对HIFU治疗反应良好的患者至关重要[8, 9]，术前预测HIFU治疗子宫肌瘤的消融效果、可更好地指导临床治疗计划、制定个性化的治疗方案。

随着机器学习和医学成像技术的快速发展和广泛应用，影像组学整合了这两个领域的学科优势，以定量成像的方式引起较大的关注，成为近十年医学影像学最热门的研究领域[10]。影像组学通过获取医学图像中超出人眼的分辨能力的丰富信息，相较于人工分析具有明显优势，它可以从各种成像方法中定量化地提取更全面的、完整的包括信号强度/密度/回声、形状、大小、体积、纹理、肿瘤细胞结构和异质性等各种组织特征[11]。这些影像组学特征不仅为评估肿瘤表型提供了客观和定量的方法，而且在区分良性和恶性肿瘤以及预测治疗反应方面显示出巨大潜力[12]。最近的研究也已经将影像组学应用于预测妇科疾病的HIFU消融效果领域。Hocquelet等[13]通过单变量和多变量线性回归分析表明，MRI T2加权成像（MRI-T2WI）纹理特征与HIFU手术后子宫肌瘤的消融率显著相关。Li等[14]进一步构建一种影像组学预测模型，通过整合基于MRI-T2WI的影像组学特征和临床参数来预测MRI引导的HIFU子宫腺肌病治疗效果。Zheng等[15]研究了通过结合不同的特征选择方法和不同的机器学习模型来预测HIFU治疗子宫肌瘤的效果。这些研究表现，基于影像组学的机器学习模型作为术前预测的手段，具有较大的应用前景及开发潜力。

众所周知，临床数据主要由正常样本组成，异常或疾病样本只占一小部分，这是医学研究中普遍存在的类别不平衡问题。当机器学习模型在不平衡数据上进行训练时，预测结果往往会偏向多数类[16]。然而，在临床医学上，少数类通常更受到重视，但类别不平衡常会导致分类器忽视少数类，从而导致模型的预测性能不佳，这在小样本量时尤为明显[17]。目前，国内外尚鲜见针对预测HIFU治疗子宫肌瘤效果的样本不平衡问题的研究。因此，本研究旨在探索影像组学和机器学习在预测HIFU治疗子宫肌瘤效果中的应用，为解决该领域的样本不平衡问题提供新思路。

1 资料和方法

本研究主要流程包括4个关键步骤（图1）：数据准备、数据欠采样处理、分类器训练和模型选择，最终建立最佳预测模型。每个步骤的详细内容如下所述。

1.1 研究对象

本研究获得佛山市妇幼保健院伦理委员会批准（伦理编号：FSFY-MEC-2024-157）。

回顾性分析2018年9月~2019年12月在佛山妇幼保健院接受HIFU消融治疗子宫肌瘤患者资料。纳入标准：年满18岁或以上；患者有子宫肌瘤相关的不适症状；无MRI检查或钆对比剂注射禁忌症；肌瘤直径超过3.0 cm。排除标准：患者合并其他影响妇科器官或其他系统的严重疾病；肌瘤存在明显坏死或变性；≥2个肌瘤患者；在HIFU治疗期间严重不适、无法坚持完成治疗者；曾接受过HIFU消融治疗；术前MRI检查与HIFU治疗的时间间隔超过3个月。本研究共获得140例子宫肌瘤患者资料，年龄26~53（39.6±6.5）岁，治疗前MRI评估显示，子宫肌瘤体积为13.2~800.0 cm³，中位数为87.2（48.0~148.2）cm³。

参考既往类似研究、基于统计功效分析计算样本量，其公式为[5]：

$$n = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2 \times \sigma^2}{\left(\Delta AUC\right)^2}$$

其中$\alpha = 0.05$、$\beta = 0.2$、$\sigma = 0.1$、$\Delta AUC = 0.1$分别表示95%置信度、80%效能、AUC标准差及需要检测的典型AUC差异。经模拟分析验证，在样本量为140例时，设定条件下检验效能约为0.82，满足研究需求。

1.2 MRI检查

采用Brivo MR355型1.5T超导MRI系统（美国纽约GE医疗器械有限公司）、配套专用体部相控阵线圈实施术前术后全程评估。影像采集方案包含多平面多序列扫描：矢状面快速自旋回波T2加权成像（脂肪抑制技术），参数设定为TR 3270 ms/TE 70 ms，FOV 400×400 mm²，采集矩阵288×256；冠状面采用TR 2675 ms/TE 70 ms，FOV 300×300 mm²，相同矩阵配置；横断面T2WI参数调整为TR 2800 ms/TE 59 ms，T1WI序列参数为TR 680 ms/TE 12 ms，FOV统一设定为300×300 mm²，其中

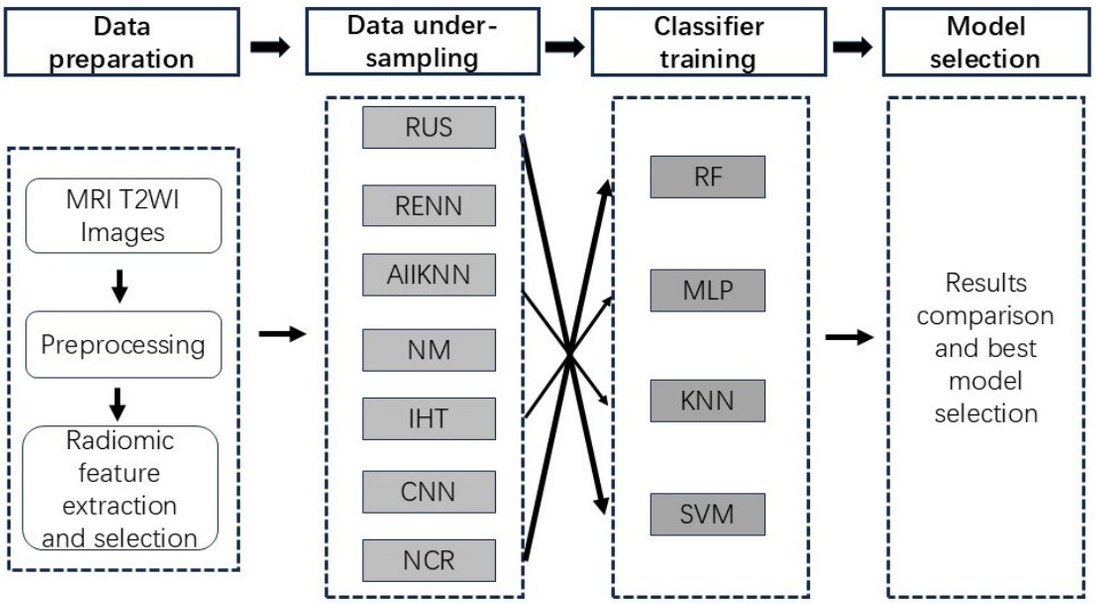

图1 主要数据处理步骤示意图。

Fig.1 Diagram of the main data processing procedures in this study. Seven undersampling methods are used in combination with 4 types of classifiers to construct predictive models for HIFU treatment of uterine fibroids.

T1WI矩阵设定为320×320。扩散加权成像(DWI)横断面采用b值800 s/mm²,TR 3700 ms/TE 80 ms,FOV设定为300×300 mm²,采用96×128矩阵采集。所有序列层厚为6~8 mm,层间距为1~2 mm。动态增强方案采用肘静脉高压注射器以3 mL/s流速推注钆喷酸葡胺对比剂0.2 mL/kg,延迟14 s启动肝脏三维容积快速扫描(LAVA)序列,关键参数为TR/TE=4/2 ms,翻转角15°,扫描后行横断面、矢状面、冠状面多平面重建(层厚6~8 mm,层间距为0)。为确保数据可比性,HIFU治疗前后严格保持相同扫描参数设置。所有影像数据经后处理工作站完成后,归档至医学影像存储与传输系统(PACS)进行后续分析。

在轴位T2WI图像上选取显示子宫肌瘤最大面积的层面,测量肿瘤左右径、前后径,并在矢状面上获取其上下径数值,按椭圆体体积公式"π/6×左右径×前后径×上下径"计算肿瘤组织体积[18, 19]。

1.3 HIFU治疗及评估

使用JC200聚焦超声肿瘤治疗系统(中国重庆海扶医疗科技有限公司)对子宫肌瘤进行消融治疗。患者俯卧在HIFU治疗台上,确保前盆壁与水囊紧密接触,以实现高效的声学耦合。通过尿道将导管置入膀胱内,避免膀胱体积过大影响超声引导的声学通道。治疗过程中通过对比增强超声成像在治疗前后实时观察血流动态,监测血管灌注和治疗效果即时评估的状态。进行消融操作时,每个子宫肌瘤靶组织以5 mm厚度进行逐层消融,确保全面覆盖肌瘤组织,且尽量减少能量对周围正常组织的影响,尤其注意识别子宫内膜组织,避免损伤。在治疗过程中,操作者保持与患者沟通、密切观察患者反应,以免出现严重并发症。

在完成HIFU治疗后短期内(<48 h)行MR增强扫描显示无灌注区体积(NPV)评估消融效果。使用T1WI增强扫描的静脉期图像进行NPV的测量,具体测量及计算方法与术前肌瘤体积类似。在本研究中,以NPV为原肌瘤体积的70%作为消融效果的判断标准[20],当NPV小于未治疗肌瘤体积的70%时,为低消融率,否则为高消融率,并将患者分为两组:即36例患者属低消融率组、104例患者属高消融率组。在本研究中,将低消融率组患者定义为阳性样本,高消融率患者为阴性样本,阳性与阴性样本的比例约为1:3。

1.4 肌瘤分割和图像预处理

通过手动分割MRI-T2WI图像获得肌瘤感兴趣区。在横断面T2WI图像上选择显示肌瘤最大的层面,使用3D Slicer软件(https://www.slicer.org/),在该图像上勾勒感兴趣区(ROI),确定肌瘤组织范围,分割肌瘤(图2)。

使用Python强度归一化包intensity-normalization(https://github.com/jcreinhold/intensity-normalization),以Nyul分段强度归一化方法来预处理MRI-T2WI图像[21, 22],使数据集具有一致的强度轮廓。使用PyRadiomics包(https://pyradiomics.readthedocs.io/en/latest/)从原始和小波变换MRI-T2WI图像中遵循图像生物标志物标准化倡议(IBSI)推荐的方法进行特征提取[23]。提取的特征包括一阶统计量、形状特征、灰度共生矩阵(GLCM)、灰度大小区域矩阵(GLSZM)、灰度游

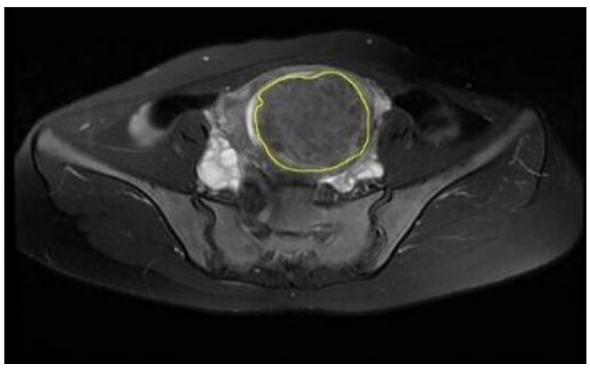

图2 勾勒ROI示意图
Fig.2 A schematic sketch of the region of interest (ROI).

程矩阵(GLRLM)和灰度依赖矩阵(GLDM),共提取出140个病灶的467个影像组学特征,并使用z分数归一化方法进行归一化,保证数据集的一致性和可比性。

应用Mann-Whitney U检验识别在低消融率和高消融率肌瘤两组患者之间具有显著性差异(P<0.05)的特征。鉴于本研究样本量相对较小,使用Python的scikit-learn库的标准Relief算法(使用固定的随机种子0)进一步减少特征集[24],最终选取4个最关键特征即原图灰度最大值(original_firstorder_Maximum)、小波HL灰度共生矩阵相关性(wavelet-HL_glcm_Correlation)、小波HH灰度共生矩阵相关性(wavelet-HH_glcm_Correlation)、小波HL灰度共生矩阵长行程低灰度加强特征(wavelet-HL_glrlm_LongRunLowGray-LevelEmphasis)进行机器学习建模。

1.5 数据处理及模型构建

本研究的阳性、阴性样本的比例约为1:3,数据存在明显的不平衡,显著影响预测模型的训练,导致倾向多数类的偏差。因此,本研究采用欠采样方法来平衡数据集,引入7种不同方法,并将其与4种机器学习算法相结合,以建立稳健的HIFU治疗效果预测模型。本文所采用的7种欠采样方法及其主要处理步骤简要描述如下:

随机欠采样(RUS)[25]:从多数类中随机选择并移除样本,直到其大小与少数类相匹配;重复编辑最近邻(RENN)[26]:基于编辑最近邻(ENN)[27]方法,该方法使用K最近邻(KNN)从多数类中移除错误分类的样本,RENN迭代应用ENN,直到实现样本平衡;全K最近邻算法(AllKNN)[26]:源自ENN算法,AllKNN与RENN不同之处在于,它在每次迭代中逐步增加KNN中的K值,同时在每次迭代中移除所有误分类的多数类样本;近邻缺失-3(NM)[28]:这个两步算法首先从多数类中选择每个少数样本的M最近邻。然后,在这些邻居中,选择与N个最近少数样本的平均距离最大的多数样本进行欠采样;凝聚最近邻(CNN)[29]:所有少数类样本和一个随机选择的多数样本形成集合C,而剩余的多数样本形成集合S。根据C使用最近邻规则对S中的每个样本进行分类。将误分类的样本添加到C中,而正确分类的样本保持不变。此过程重复进行,直至在C中实现样本平衡;邻域清理规则(NCR)[30]:该方法使用K设置为3的KNN分类器对多数类进行欠采样。如果多数类样本被其3个最近邻误分类,则将其移除。相反,如果少数类样本被误分类,则将其相邻的多数类样本从数据集中移除;实例硬度阈值(IHT)[31]:该方法首先训练一个分类器,再使用它对所有实例进行分类,移除具有高分类误差的多数类样本。在本研究中,使用随机森林分类器进行IHT欠采样步骤。

使用4种机器学习模型-K最近邻(KNN)、高斯核的支持向量机(SVM)、多层感知机(MLP)和随机森林(RF)来预测HIFU治疗子宫肌瘤的效果。所有模型均使用Python 3.6开发环境中的Scikit-learn库(https://scikit-learn.org)进行训练。对于MLP,采用6隐藏层架构,神经元数量依次设置为64、256、128、32、16和8,并使用Adam优化器,其参数配置为α=0.001,β_1=0.9,β_2=0.999,初始学习率为0.001,随机状态种子为74。对于RF,将树的数量设置为125,随机状态种子为5。对于SVM,选择了径向基函数(RBF)核,优化的正则化参数C=68.6649,核系数γ=0.1,随机状态种子为528。对于KNN,将邻居数量设置为10,权重方案基于欧氏距离。

将7种数据欠采样方法与4种机器学习模型相结合,产生共计28种组合模型,用于筛选识别HIFU治疗效果的最佳预测模型。详细的模型训练和数据测试流程图如图3所示。使用分层K折方法将数据集划分为5个子集,确保每个子集保持两类样本的相似比例。利用五折交叉验证技术,其中4个子集用于训练模型,而剩余的子集作为测试数据。验证过程重复5次,每个子集轮流作为测试集,并计算所有交叉测试的平均性能,用以评估模型的泛化性能。在每一轮模型训练,只有训练样本进行了欠采样,而测试数据集保持不变,保留其原始分布,以反映真实条件下的泛化性能。为确定基准水平,本研究组同时进行没有欠采样的模型训练和测试实验,以便直接比较在平衡和不平衡数据上训练的模型。

使用受试者工作特征(ROC)曲线评估模型性能,并以曲线下面积(AUC)作为主要定量指标;通过1000次迭代的自助抽样法(bootstrap)计算平均AUC及其95%置信区间(CI),以确保统计可靠性,同时计算模型的准确率、召回率和特异性等指标,更全面地评估其预测性能。

2 结果

2.1 KNN相关模型性能

表1示KNN与各种欠采样方法相关的分类统计结果,其中CNN-KNN模型的平均AUC为0.822,优于其

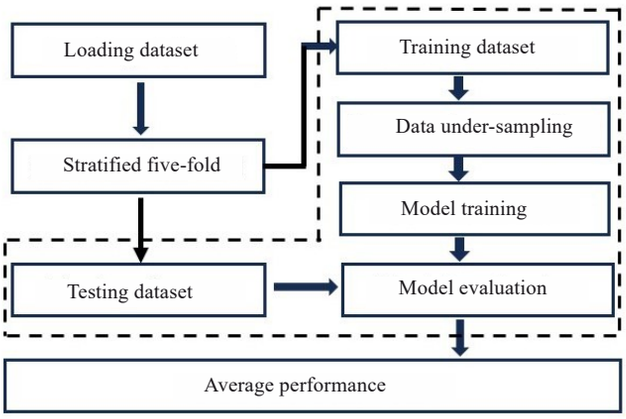

图3 模型建立和评估流程图

Fig.3 Flowchart of model establishment and evaluation in this study. The procedures boxed using dotted lines are repeated 5 times for the 5-fold cross-validation scheme.

他测试模型。除了召回率表现较一般外,该模型达到最高准确率和特异性得分(表1)。

2.2 RF相关模型性能

表2示各重采样方法自RF模型得出的分类统计结果。其中,CNN-RF模型的平均AUC达0.772,为表现最佳测试模型。除召回率未达最高外,该模型获最高准确率和特异性得分(表2)。

2.3 SVM相关模型性能

表3示SVM模型上各重采样方法的分类统计结果。其中,NM-SVM模型平均AUC为0.797,优于其他测试模型,但RENN-SVM获最高召回率,而CNN-SVM模型在准确率方面表现出色,CNN-SVM模型在特异性方面表现最佳(表3)。

2.4 MLP相关模型性能

表4示MLP模型上各重采样方法的分类统计结果。其中,NM-MLP模型平均AUC为0.822,优于其他测试模型,且达最高准确率,提示其在正确分类大多数实例方面的效能,但RENN-MLP、NCR-MLP模型分别在召回率、特异性上为各模型最佳表现。

表1 与KNN相关的各种模型的性能

Tab.1 Performances of different models associated with KNN learning

Models	AUC (95% *CI*)	Accuracy	Recall	Specificity
RUS-KNN	0.792(0.586-0.958)	0.679	0.728	0.664
RENN-KNN	0.736(0.519-0.925)	0.643	0.678	0.637
AllKNN-KNN	0.708(0.465-0.909)	0.679	0.750	0.655
NM-KNN	0.769(0.558-0.939)	0.679	0.782	0.646
CNN-KNN	0.822(0.635-0.964)	0.714	0.675	0.733
NCR-KNN	0.734(0.507-0.927)	0.664	0.618	0.684
IHT-KNN	0.710(0.479-0.909)	0.664	0.728	0.646
KNN-baseline	0.784(0.571-0.955)	0.750	0.143	0.962

Bold values indicate the best among all the tested models.

表2 与RF相关的各种模型的性能

Tab.2 Performances of different models associated with RF learning

Models	AUC (95% *CI*)	Accuracy	Recall	Specificity
RUS-RF	0.768(0.556-0.932)	0.636	0.696	0.617
RENN-RF	0.722(0.504-0.907)	0.543	0.793	0.465
AllKNN-RF	0.692(0.476-0.883)	0.593	0.675	0.569
NM-RF	0.701(0.475-0.892)	0.579	0.586	0.580
CNN-RF	0.772(0.566-0.942)	0.700	0.725	0.694
NCR-RF	0.672(0.466-0.876)	0.614	0.504	0.656
IHT-RF	0.656(0.430-0.854)	0.550	0.750	0.482
RF-baseline	0.731(0.518-0.909)	0.750	0.336	0.895

Bold values indicate the best among all the tested models.

图4示两个最佳模型,即CNN-KNN与NM-MLP模型的ROC曲线,两者AUC较接近。

2.5 欠采样模型性能

与没有使用数据欠采样的模型相比,使用数据欠采

表3 与SVM相关各种模型的性能

Tab.3 Performances of different models associated with SVM learning

Models	AUC (95% *CI*)	Accuracy	Recall	Specificity
RUS-SVM	0.791(0.595-0.955)	0.728	0.807	0.702
RENN-SVM	0.702(0.363-0.828)	0.593	0.843	0.513
AllKNN-SVM	0.708(0.490-0.895)	0.629	0.778	0.579
NM-SVM	0.797(0.600-0.950)	0.664	0.836	0.607
CNN-SVM	0.782(0.577-0.950)	0.757	0.700	0.780
NCR-SVM	0.714(0.495-0.902)	0.671	0.643	0.684
IHT-SVM	0.734(0.511-0.912)	0.621	0.778	0.568
SVM-baseline	0.712(0.485-0.910)	0.743	0	1

Bold values indicate the best among all the tested models.

表4 与MLP相关的各种模型的性能

Tab.4 Performances of different models associated with MLP learning

Models	AUC (95% *CI*)	Accuracy	Recall	Specificity
RUS-MLP	0.791(0.593-0.949)	0.657	0.721	0.634
RENN-MLP	0.723(0.504-0.910)	0.607	0.818	0.541
AllKNN-MLP	0.729(0.520-0.911)	0.664	0.807	0.616
NM-MLP	0.822(0.632-0.960)	0.729	0.786	0.713
CNN-MLP	0.782(0.570-0.954)	0.679	0.728	0.666
NCR-MLP	0.730(0.499-0.913)	0.693	0.614	0.722
IHT-MLP	0.736(0.530-0.909)	0.629	0.811	0.568
MLP-baseline	0.710(0.467-0.911)	0.736	0.414	0.847

Bold values indicate the best among all the tested models.

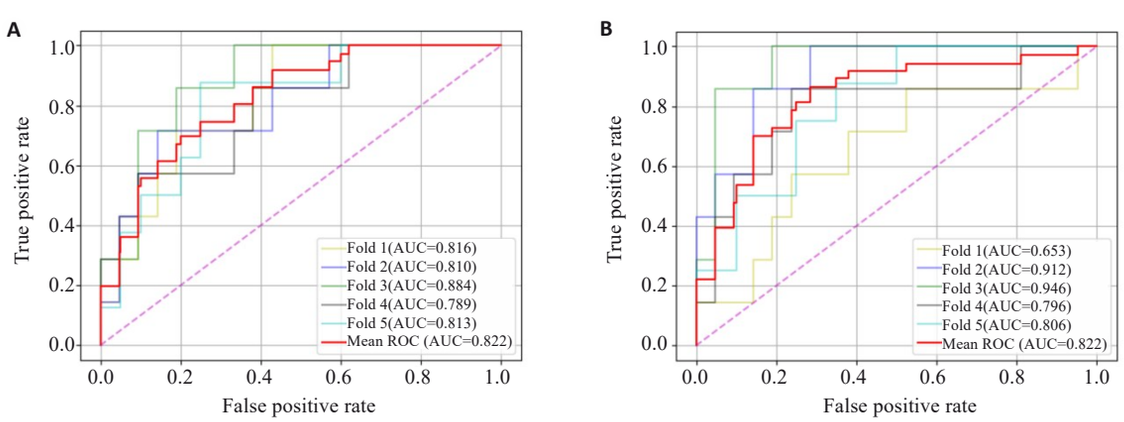

图4 五折交叉验证测试的两个最佳模型的ROC及AUC: (A)CNN-KNN和(B)NM-MLP

Fig.4 ROC and AUC for the 5-fold cross-validation tests of the two best models: CNN-KNN (**A**) andNM-MLP (**B**).

样的模型明显获得更高的预测性能，AUC增加0.069±0.036，以MLP模型的改善最明显（图5）。在准确率、召回率和特异性等具体评估指标中，以召回率的增加最明显，增加0.532±0.215，提示欠采样可在一定程度上减轻类别不平衡的影响、增强模型预测能力，显著提高了模型正确识别阳性病例的能力，使阳性类的召回率显著提高，突出了欠采样在解决类别不平衡问题和增强模型预测能力方面的有效性，凸显欠采样技术在提高模型对少数类的敏感性方面的重要性，从而提供更平衡、更准确的分类结果。

3 讨论

基于术前患者盆腔MRI影像学表现预测子宫肌瘤HIFU治疗效果，具有较大临床及社会意义，它可协助临床医生筛选患者、制定治疗计划，提高患者安全性，提供个性化医疗服务，进而达到良好的卫生经济效益、可惠

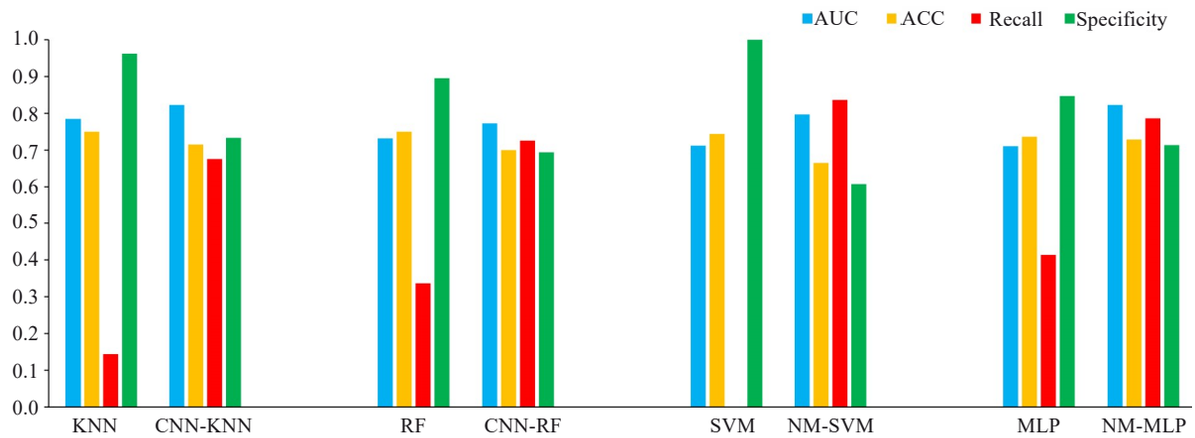

图5 使用、未使用欠采样的KNN、RF、SVM和MLP模型的预测性能比较(包含AUC、准确率、召回率和特异性)
Fig. 5 Comparison of prediction performances (AUC, accuracy, recall and specificity) among KNN, RF, SVM and MLP models without and with under-sampling. ACC: Accuracy.

及整个社会。在临床治疗领域,虽然人工评估已证实可在一定程度上发挥临床指导作用[32, 33],但其无法直接读取医学图像隐藏数字信息,也对无法定量衡量部分影像特征,使其效能大打折扣。鉴于此,影像组学技术应运而生,它以更全面、更深入的视角呈现影像学信息[10, 34]。机器学习恰是处理影像组学高维度信息的有用工具,在构建准确预测模型中发挥着关键作用。本研究主要关注数据不平衡问题,这是在小数据分析中建立机器学习预测模型时常见的问题。我们将7种数据欠采样方法与4种机器学习模型相结合,用于预测HIFU消融对子宫肌瘤的效果。研究结果显示,与未欠采样的基线模型相比,基于欠采样方法的模型显著提高了诊断性能。在测试的28种模型中,CNN-KNN和NM-MLP在AUC方面表现最佳,但NM-MLP在准确率特别是召回率方面表现更出色。高召回率能够更好地筛选出阳性即HIFU治疗效果不好的病人,可减少不必要的治疗,从而降低患者医疗负担。因此,推荐将NM-MLP作为优选的机器学习方法,用于基于MRI-T2WI影像组学特征开发治疗效果预测模型,优化子宫肌瘤个性化医疗和临床决策管理。

随着计算机技术和人工智能的快速发展,机器学习已成为复杂生物学建模的核心,用于肿瘤治疗预测、个性化用药方案设计、药物敏感性分析以及多组学数据整合,它通过挖掘海量医疗影像、基因组学和临床数据中的深层关联,显著提升了肿瘤疗效评估精准度与新型靶向疗法的开发效率。影像组学通过从医学图像中提取定量信息来增强疾病管理,在放射学领域获得了显著的关注[35]。既往研究已证实,将MRI影像组学特征纳入预测HIFU消融治疗子宫肌瘤效果中可进一步提供模型的性能[15, 36]。鉴于机器学习模型由数据驱动,故数据的质量和特性显著影响算法性能。其中,数据类别不平衡作为数据分析中常见问题,是许多医学研究工作的关键因素,可能会显著影响预测准确性和可靠性[37]。在本研究中,阳性病例数量相对较少,导致模型预测明显偏向多数类,表现为未经数据重采样的4个机器学习基准模型的平均召回率只有0.223±0.187。值得注意的是,SVM模型的召回率特别低,即为0,表明所有样本都被错误地分类为阴性,表明预测模型并没有发挥效用。为了解决这一问题,我们采用了欠采样技术,并测试了7种流行的欠采样方法,以评估其对性能的影响。选择特定的欠采样方法取决于多种因素,如任务的性质、数据分布、可解释性和计算成本,因此仅根据理论知识很难选择最佳方法,因为每种方法都有其自身的优势和局限性[38, 39]。因此,类似于本研究探索的HIFU治疗效果预测,经验性实验仍然是确定特定应用中首选重采样方法的常用方法。

本文从467个影像组学特征中选出最重要的4个进行预测模型的建模研究。选出的4个特征分别为:原图灰度最大值(original_firstorder_Maximum)、小波HL灰度共生矩阵相关性(wavelet-HL_glcm_Correlation)、小波HH灰度共生矩阵相关性(wavelet-HH_glcm_Correlation)、小波HL灰度共生矩阵长行程低灰度加强特征(wavelet-HL_glrlm_LongRunLowGrayLevelEmphasis)。暂时还没有看到的其它的专门探索这些特征的生物学或者临床意义的过往研究,但是根据相应的基础知识我们估计:原图灰度最大值可能反应肿瘤的高信号高密度区域,这部分可能形成热阻断,影响治疗的效果。小波图像的3个灰度共生矩阵特征可能反映组织纤维排列或者血管网排列的规则性及区域坏死等特征,因此也会影响HIFU治疗中的空化效应及热效应的传播扩散性能,影响治疗的效果。但是必须强调这些只是基于这些特征的计算方法、图像的基本知识和生理学常识做出的一些估计,有待后续进一步研究及验证,相关研究也正在成为影响影像组学大规模临床应用的一个重要课题[32]。

本研究的实验结果显示,欠采样方法可整体提高预测性能,但对4种测试的机器学习模型的影响有明显的差异。在7种欠采样方法中,CNN对KNN和RF改进最明显,将AUC分别提升了0.038和0.041,而NM对SVM和MLP较有效,将AUC提升了0.085和0.112,RUS则为所有模型中次佳的欠采样方法。对于KNN和RF,只有2种欠采样方法(RUS和CNN)超过了未欠采样的基线模型。相比之下,与基线模型相比,5种方法(RUS、NM、CNN、NCR、IHT)提高了SVM的预测性能,所有欠采样方法均增强MLP的预测性能。对于MLP,7种欠采样方法的性能排名为:NM>RUS>CNN>IHT>NCR>AllKNN>RENN,所有方法都超过了基线模型。值得注意的是,尽管SVM和MLP在使用欠采样后均得到改善,其基线性能仍然不如RF和KNN,后者在4个基线模型中表现最佳,Li等[40]的研究也表明,使用NM-RF方法可提高不平衡数据模型的效能。因此本结果提示,为特定机器学习模型选择适当的欠采样技术,可最大化预测准确性。

除了AUC之外,模型的召回率等特定性能评估指标也显著提升。对于4种最佳模型——CNN-KNN、CNN-RF、NM-SVM和NM-MLP,召回率分别提高0.532、0.389、0.836和0.372。总体而言,所有欠采样方法均导致特异性有所降低,但召回率显著提高,这在本研究的临床工作中更具实际意义,因为准确预测消融无效的患者,对于筛选患者、避免接受无效的HIFU治疗、减少治疗管理中的不必要负担具有重要意义。尽管特异性较低可能导致假阳性率略有上升,从而给治疗计划带来一定复杂性,但这并不会显著影响患者预后,这些患者可以选择其他治疗方式。因此,通过欠采样技术解决数据不平衡问题,有望通过提高治疗效果病例的识别能力,支持更具个性化和有效的治疗策略,从而改善子宫肌瘤管理流程。后续可以通过加入更多有效特征参数、建立更加有效的预测模型等方法,在提高召回率的同时降低假阳率。

本研究存在局限性。首先,研究的样本量相对较小,这会增加过拟合风险、削弱模型泛化能力和统计功效,可能导致本文见到的如AUC的95%置信区间比较宽泛等现象,一定程度上也限制了模型对病人的普适性。未来争取纳入更多患者的研究,其多样化的HIFU治疗反应,以更好地验证和推广本研究结果。第二,本文使用的特征选择算法是在全局数据而不是在交叉验证每一折内部进行的,增加了一定的数据泄露风险,可能会过高估计预测模型的性能。第三,本研究仅使用了传统的机器学习方法,并未采用过去十年中发展迅速的深度学习技术。我们猜测,将来探索开发的深度学习预测模型,即使在患者数据有限的情况下,也会进一步提高预测模型的性能。第四,本研究仅使用MRI-T2WI图像中提取的影像组学特征来建立预测模型,但众所周知,其他疾病特征:如临床信息和肌瘤的形态学参数,及MRI其他序列图像也影响预测结果,故纳入更多的信息,可能会提高模型在实际临床环境中的预测准确性。第五,我们参考机器学习常用方法[14],直接对比模型的AUC数值,并没有进一步进行严格的统计学分析比较。最后,本研究焦点短期消融效果,然而临床上更关注的是患者的症状改善、肌瘤远期体积的缩小,即HIFU治疗后的长期效果,后续的研究应解决这一局限性,这将是未来研究中开发更全面和时间感知型预测模型的关键。

综上所述,本研究探讨基于MRI-T2WI图像的影像组学特征预测子宫肌瘤HIFU消融效果,聚焦于解决数据不平衡的问题。本研究结果提示,通过将7种欠采样方法与4种机器学习模型相结合,即使在小样本且不平衡的训练数据情况下,欠采样技术也可显著提高预测性能。值得注意的是,NM欠采样方法与MLP模型相结合,在所有28种可能的欠采样方法和机器学习模型组合中实现了最佳的预测结果。这些发现突出了在开发用于预测HIFU消融治疗子宫肌瘤成功的准确机器学习模型时,重采样技术在缓解小样本、不平衡数据集带来的挑战方面的实用性,有望改善子宫肌瘤管理中的临床决策和患者护理问题。

Declaration of interests: The authors declare no competing interests.

参考文献:

[1] Giuliani E, As-Sanie S, Marsh EE. Epidemiology and management of uterine fibroids[J]. Int J Gynaecol Obstet, 2020, 149(1): 3-9.

[2] Management of symptomatic uterine leiomyomas: ACOG practice bulletin, number 228[J]. Obstet Gynecol, 2021, 137(6): e100-15.

[3] Grube M, Neis F, Brucker SY, et al. Uterine fibroids - current trends and strategies[J]. Surg Technol Int, 2019, 34: 257-63.

[4] Haviv E, Schwarzman P, Bernstein EH, et al. Subsequent pregnancy outcomes after abdominal *vs.* laparoscopic myomectomy[J]. J Matern Fetal Neonatal Med, 2022, 35(25): 8219-25.

[5] Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics[J]. J Biomed Inform, 2014, 48: 193-204.

[6] Liu L, Wang T, Lei B. Ultrasound-guided microwave ablation in the management of symptomatic uterine myomas: a systematic review and meta-analysis[J]. J Minim Invasive Gynecol, 2021, 28(12): 1982-92.

[7] Jenne JW, Preusser T, Günther M. High-intensity focused ultrasound: principles, therapy guidance, simulations and applications[J]. Z Für Med Phys, 2012, 22(4): 311-22.

[8] Machtinger R, Inbar Y, Cohen-Eylon S, et al. MR-guided focus ultrasound (MRgFUS) for symptomatic uterine fibroids: predictors of treatment success[J]. Hum Reprod, 2012, 27(12): 3425-31.

[9] Mindjuk I, Trumm CG, Herzog P, et al. MRI predictors of clinical success in MR-guided focused ultrasound (MRgFUS) treatments of

uterine fibroids: results from a single centre[J]. Eur Radiol, 2015, 25 (5): 1317-28.

[10] Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis[J]. Eur J Cancer, 2012, 48(4): 441-6.

[11] Avanzo M, Wei L, Stancanello J, et al. Machine and deep learning methods for radiomics[J]. Med Phys, 2020, 47(5): e185-202.

[12] Yip SS, Aerts HJ. Applications and limitations of radiomics[J]. Phys Med Biol, 2016, 61(13): R150-66.

[13] Hocquelet A, Denis de Senneville B, Frulio N, et al. Magnetic resonance texture parameters are associated with ablation efficiency in MR-guided high-intensity focussed ultrasound treatment of uterine fibroids[J]. Int J Hyperthermia, 2017, 33(2): 142-9.

[14] Li ZC, Zhang J, Song Y, et al. Utilization of radiomics to predict long-term outcome of magnetic resonance-guided focused ultrasound ablation therapy in adenomyosis[J]. Eur Radiol, 2021, 31 (1): 392-402.

[15] Zheng Y, Chen L, Liu M, et al. Prediction of clinical outcome for high-intensity focused ultrasound ablation of uterine leiomyomas using multiparametric MRI radiomics-based machine leaning model [J]. Front Oncol, 2021, 11: 618604.

[16] Li DC, Liu CW, Hu SC. A learning method for the class imbalance problem with medical data sets[J]. Comput Biol Med, 2010, 40(5): 509-18.

[17] Galar M, Fernandez A, Barrenechea E, et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches[J]. IEEE Trans Syst, Man, Cybern C, 42(4): 463-84.

[18] Funaki K, Fukunishi H, Funaki T, et al. Magnetic resonance-guided focused ultrasound surgery for uterine fibroids: relationship between the therapeutic effects and signal intensity of preexisting T2-weighted magnetic resonance images[J]. Am J Obstet Gynecol, 2007, 196(2): 184.e1-6.

[19] Kim YS, Lim HK, Park MJ, et al. Screening magnetic resonance imaging-based prediction model for assessing immediate therapeutic response to magnetic resonance imaging-guided high-intensity focused ultrasound ablation of uterine fibroids[J]. Invest Radiol, 2016, 51(1): 15-24.

[20] Jiang Y, Qin S, Wang Y, et al. Intravoxel incoherent motion diffusion-weighted MRI for predicting the efficacy of high-intensity focused ultrasound ablation for uterine fibroids[J]. Front Oncol, 2023, 13: 1178649.

[21] Carré A, Klausner G, Edjlali M, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics[J]. Sci Rep, 2020, 10(1): 12340.

[22] Nyul LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization[J]. IEEE Trans Med Imaging, 19(2): 143-50.

[23] Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping[J]. Radiology, 2020, 295(2): 328-38.

[24] Kira K, Rendell LA. A practical approach to feature selection[C]// Proceedings of the 9th International Workshop on Machine Learning (ML 1992), Aberdeen, Scotland, UK, July 1-3, 1992.[S.l.]: Morgan Kaufmann Publishers Inc, 1992: 249-26.

[25] Prusa J, Khoshgoftaar TM, Dittman DJ, et al. Using random undersampling to alleviate class imbalance on tweet sentiment data [C]//2015 IEEE International Conference on Information Reuse and Integration. August 13-15, 2015. San Francisco, CA, USA. IEEE, 2015: 197-202.

[26] Tomek I. An experiment with the edited nearest-neighbor rule[J]. IEEE Trans Syst, Man, Cybern, SMC-6(6): 448-52.

[27] Wilson DL. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Trans Syst, Man, Cybern, SMC-2(3): 408-21.

[28] Zhang JP, Mani I: kNN approach to unbalanced data distributions: A case study involving information extraction. In: Proceeding of International Conference on Machine Learning (ICML 2003), Workshop on Learning from Imbalanced Data Sets: 2003; Washington D.C.: ICML; 2003: 1-7.

[29] Hart P. The condensed nearest neighbor rule (Corresp.)[J]. IEEE Trans Inform Theory, 14(3): 515-6.

[30] Laurikkala J. Improving identification of difficult small classes by balancing class distribution[M]//Artificial Intelligence in Medicine. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001: 63-6.

[31] Smith MR, Martinez T, Giraud-Carrier C. An instance level analysis of data complexity[J]. Mach Learn, 2014, 95(2): 225-56.

[32] Tomaszewski MR, Gillies RJ. The biological meaning of radiomic features[J]. Radiology, 2021, 298(3): 505-16.

[33] Zhao WP, Chen JY, Chen WZ. Dynamic contrast-enhanced MRI serves as a predictor of HIFU treatment outcome for uterine fibroids with hyperintensity in T2-weighted images[J]. Exp Ther Med, 2016, 11(1): 328-34.

[34] Fribbens C, O'Leary B, Kilburn L, et al. Plasma ESR1 mutations and the treatment of estrogen receptor-positive advanced breast cancer [J]. J Clin Oncol, 2016, 34(25): 2961-8.

[35] Rogers W, Thulasi Seetha S, Refaee TAG, et al. Radiomics: from qualitative to quantitative imaging[J]. Br J Radiol, 2020, 93(1108): 20190948.

[36] Zheng Y, Chen L, Liu M, et al. Nonenhanced MRI-based radiomics model for preoperative prediction of nonperfused volume ratio for high-intensity focused ultrasound ablation of uterine leiomyomas [J]. Int J Hyperthermia, 2021, 38(1): 1349-58.

[37] Walsh R, Tardy M. A comparison of techniques for class imbalance in deep learning classification of breast cancer[J]. Diagnostics: Basel, 2022, 13(1): 67.

[38] Guo HX, Li YJ, Shang J, et al. Learning from class-imbalanced data: review of methods and applications[J]. Expert Syst Appl, 2017, 73: 220-39.

[39] Kraiem MS, Sánchez-Hernández F, Moreno-García MN. Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. an approach based on association models[J]. Appl Sci, 2021, 11(18): 8546.

[40] Li M, Wu Z, Wang W, et al. Protein-protein interaction sites prediction based on an under-sampling strategy and random forest algorithm[J]. IEEE/ACM Trans Comput Biol Bioinform, 2022, 19 (6): 3646-54.

（编辑:余诗诗）