·论著·

基于复合核偏最小二乘模型的胶质瘤多模态MRI数据预测MGMT启动子甲基化合并IDH突变*

杨　琪[1]　闫俊羽[1,2,3]　刘　夏[1]　王　彤[1,2]　房瑞玲[1,4]　曾　平[5]　杨海涛[6]
罗艳虹[1,4]　余红梅[1,4]　张岩波[1,4]　杨国强[3△]　曹红艳[1,4△]

【提　要】 目的　探讨基于术前MRI影像特征，构建胶质瘤中MGMT启动子甲基化合并IDH突变的预测模型，为胶质瘤患者的预后评估与治疗策略提供理论依据。**方法**　收集来自TCGA/TCIA公共数据库的208例胶质瘤患者的术前MRI图像、临床及基因信息。根据MGMT启动子甲基化和IDH突变状态，将患者分为IDHmut&MGMTmet亚型与Other亚型，并比较两亚型间各因素的差异，分析分子亚型与生存结局的关系。从术前MRI图像中提取并量化CE-T1和T2-FLAIR序列的影像学特征，经标准化后，采用惩罚logistic回归筛选特征，使用基于遗传算法的融合核偏最小二乘算法（fused kernel partial least squares with the genetic algorithm，GA-fKPLS）、K近邻（K-nearest neighbor，KNN）、最小绝对收缩和选择算法（least absolute shrinkage and selection operator，LASSO）、逻辑回归（logistic regression，LR）和朴素贝叶斯（Naïve Bayes，NB）构建胶质瘤亚型预测模型，通过AUC、灵敏度、特异度、准确率、约登指数、G均值、马修斯相关系数和F值评估模型性能。**结果**　胶质瘤不同分子亚型在年龄、肿瘤病理分级、生存状况、MGMT启动子甲基化状态以及IDH突变状态均存在显著差异。Other亚型患者的死亡风险是IDHmut&MGMTmet亚型的2.675倍。通过特征筛选得到42个CE-T1特征和68个T2-FLAIR特征用于构建预测模型。GA-fKPLS预测模型在多项评价指标上均高于其他四种模型。**结论**　基于术前多模态MRI数据构建的GA-fKPLS模型在预测胶质瘤患者分子亚型方面表现最佳，可为胶质瘤患者的治疗决策、预后评估提供重要临床辅助价值。

【关键词】 胶质瘤　MRI影像组学　异柠檬酸脱氢酶　氧6-甲基鸟嘌呤-DNA甲基转移酶　GA-fKPLS模型
【中图分类号】 R445.2；R737.31　　**【文献标志码】** A　　**DOI** 10.11783/j.issn.1002-3674.2026.04.001

The Prediction of MGMT Promoter Methylation Combined with IDH Mutation in Glioma Using a Composite Kernel Partial Least Squares Model based on Multimodal MRI Data

Yang Qi, Yan Junyu, Liu Xia, et al (*Division of Health Statistics*, *Shanxi Provincial Key Laboratory of Major Diseases Risk Assessment*, *Shanxi Medical University*, *Taiyuan* 030001)

【Abstract】 Objective　To explore the construction of a predictive model for the identification of MGMT promoter methylation combined with IDH mutation in glioma based on preoperative MRI imaging features, providing theoretical reference for prognosis assessment and treatment strategy of glioma patients. **Methods**　Preoperative MRI images, clinical data, and genetic data of 208 glioma patients were collected from the TCGA/TCIA public databases. Based on IDH mutation and MGMT promoter methylation status, patients were classified into the IDHmut&MGMTmet subtype and the Other subtype. Differences in various factors between the two subtypes were compared, and the relationship between molecular subtypes and survival outcomes was analyzed. Radiomic features of CE-T1 and T2-FLAIR sequences were extracted and quantified from the preoperative MRI images. After feature standardization, penalized logistic regression was used for feature selection. Predictive models for glioma subtypes were built using fused kernel partial least squares with the genetic algorithm (GA-fKPLS), K-nearest neighbor (KNN), least absolute shrinkage and selection operator (LASSO), logistic regression (LR), and Naïve Bayes (NB). Model performance was evaluated based on AUC, sensitivity, specificity, accuracy, Youden's index, G-mean, Matthews correlation coefficient, and F-measure. **Results**　Significant differences in age, tumor pathological grade, survival status, MGMT promoter methylation, and IDH mutation status were found between the two molecular glioma subtypes. The death risk of the Other subtype was 2. 675 times higher than glioma patients of the IDHmut&MGMTmet subtype. 42 features from CE-T1 sequence and 68 features from T2-FLAIR sequence were selected for model construction. The GA-fKPLS model outperforms the other four models across multiple evaluation metrics. **Conclusion**　The GA-fKPLS model based on

* 基金项目：国家自然科学基金资助项目（82473739；82273742）；山西省基础研究计划（202303021211130）；山西省回国留学人员科研资助项目（2024-081）；山西省高等教育"百亿工程"科技引导专项（BYBLK003）；煤炭环境致病与防治教育部重点实验室（山西医科大学）开放课题基金（MEKLCEPP/SXMU-202415）

1. 山西医科大学卫生统计教研室，重大疾病风险评估山西省重点实验室（030001）
2. 山西医科大学医学科学院
3. 山西医科大学第一医院影像科
4. 煤炭环境致病与防治教育部重点实验室
5. 徐州医科大学公共卫生学院流行病与卫生统计学系
6. 河北医科大学流行病学与卫生统计学教研室
△通信作者：杨国强，E-mail：doctor_ygq@163.com；曹红艳，E-mail：caohy@sxmu.edu.cn

preoperative MRI features exhibits the best performance in predicting the molecular subtypes of glioma patients, and could provide significant clinical auxiliary value for treatment decision-making, and prognosis assessment in glioma patients.

【Key words】 Glioma; MRI radiomics; Isocitrate dehydrogenase; O6-methylguanine-DNA methyltransferase; GA-fKPLS model

脑胶质瘤起源于神经胶质细胞，是成人颅内最常见的原发性恶性肿瘤之一，约占所有恶性脑肿瘤的80%，且死亡率高[1]。在世界卫生组织（world health organization，WHO）中枢神经系统肿瘤分类中，成人弥漫性脑胶质瘤划分为三类：星形胶质瘤（WHO 2~4级）、少突胶质瘤（WHO 2~3级）和胶质母细胞瘤（WHO 4级）[2]。WHO分类有助于胶质瘤患者的初步临床评估，但由于肿瘤具有高度异质性，即使形态学特征相似的胶质瘤患者也可能呈现不同的临床结果。通过分子标记物的检测可以进一步细分胶质瘤亚型，有助于精确地制定个体化治疗方案，提供可靠的预后评估依据[3-6]。

氧6-甲基鸟嘌呤-DNA甲基转移酶（O6-methylguanine-DNA methyltransferase，MGMT）和异柠檬酸脱氢酶（isocitrate dehydrogenase，IDH）是与胶质瘤患者预后密切相关的重要分子标志物[7]。MGMT启动子甲基化合并IDH突变的患者生存期最长，预后较好，准确预测MGMT启动子甲基化合并IDH突变对评估胶质瘤患者的预后有重要意义[8-9]。采用手术或穿刺活检标本进行基因序列分析能够准确获得IDH突变和MGMT启动子甲基化结果，但侵入性强、耗时较长，且存在因标本过小无法获取测序结果的风险[10]。

磁共振成像（magnetic resonance imaging，MRI）是胶质瘤患者术前的常规检查，具有无创性、及时性和价格合理等优点[11]。MRI检查可以无创地获取深层次组织或器官内肿瘤的静态结构和动态功能的多维图像，结合计算机图像处理能够提取肿瘤病灶的高通量定量特征[12-15]。利用MRI数据预测胶质瘤患者MGMT启动子甲基化和IDH突变具有可行性和诊断价值，对于胶质瘤患者的预后评估和临床治疗指导有重要意义。MRI影像包括多种序列，每种序列能提供不同的肿瘤信息。其中，对比增强后T1加权序列（post-contrast enhanced T1-weighted，CE-T1）使用钆等对比剂显示与血管生成相关的异常区域，有助于定位和评估肿瘤位置和大小。T2加权流体衰减反转恢复序列（T2-weighted fluid attenuation inversion recovery，T2-FLAIR）结合了T2加权成像与液体衰减反转恢复序列，能有效呈现水肿和周围组织中的异常信号，有助于评估肿瘤范围和潜在的手术边缘。单独使用CE-T1或T2-FLAIR只能从单个视角反映肿瘤信息，CE-T1和T2-FLAIR的结合能提供更全面的肿瘤信息。现有研究表明，单一影像学数据在脑肿瘤预测中的效能有限，而多模态影像学数据的结合能够显著提高预测准确性[16-17]。在基于MRI多模态数据预测胶质瘤患者MGMT启动子甲基化和IDH突变研究中，大多采用机器学习方法对影像学数据进行单模态建模，或采用串联影像学数据的方法将不同模态数据特征进行简单拼接，但都无法充分利用其数据信息，导致预测性能不佳[18-19]。基于遗传算法的融合核偏最小二乘算法（fused kernel partial least squares with the genetic algorithm，GA-fKPLS）[20]，通过将每种模态数据分别构建核矩阵，采用核融合策略整合不同模态数据，并利用遗传算法优化融合核的权重，能够有效地捕捉多模态数据中的潜在非线性关系，提升模型预测的准确性。

综上，本研究针对胶质瘤多模态MRI数据，即CE-T1和T2-FLAIR两种序列数据，采用GA-fKPLS进行多个数据集的整合预测，构建胶质瘤患者MGMT启动子甲基化合并IDH突变的预测模型，并与K近邻（K-nearest neighbor，KNN）、最小绝对收缩和选择算法（least absolute shrinkage and selection operator，LASSO）、逻辑回归（logistic regression，LR）和朴素贝叶斯（Naïve Bayes，NB）方法进行比较，评估不同模型预测性能，为开展胶质瘤患者针对性干预和预后评估提供参考依据。

材料与方法

1. 研究对象

从癌症基因组图谱（the cancer genome atlas，TCGA）下载胶质瘤患者的临床信息、MGMT和IDH基因信息；从癌症影像档案库（the cancer imaging archive，TCIA）下载胶质瘤患者的影像数据。根据以下纳入排除标准筛选样本。纳入标准：①根据当前WHO标准，组织病理学确诊的WHO2~4级胶质瘤患者；②术前MRI数据可用，包括CE-T1和T2-FLAIR图像；③有明确的IDH突变状态和MGMT启动子甲基化状态；④有完整的生存结局和随访记录，生存结局包括生存时间（从术后病理诊断到死亡或最后随访，截尾生存时间为5年）和生存状态（存活或死亡）。排除标准：①病理分级不符合WHO2~4级，或分级不明确；②术前MRI数据不完善，缺少CE-T1或T2-FLAIR任一序列；③IDH突变或MGMT启动子甲基化状态不明确或缺失；④生存结局或随访时间记录不全，或失访者。最终得到208个样本。

2. 影像组学特征提取与量化

首先用FSL软件将CE-T1图像与相应T2-FLAIR

图像进行配准。再由两名具有脑部MRI诊断经验的放射科医生，在未知患者的病理分级、临床数据和其他影像学结果的情况下，将CE-T1图像导入ITK-SNAP（版本3.8.0）软件勾画肿瘤感兴趣区域（region of interest，ROI）。具体操作步骤如下，首先将图像从DICOM格式转换为NIFTI格式，然后进行图像分割，通过手动或半自动计算机辅助方式逐层勾画肿瘤ROI，从而获得肿瘤的三维感兴趣体积（volume of interest，VOI）。随后，将得到的VOI图像导入FAE 0.5.6软件，提取影像学特征。提取出的特征包括18个一阶直方图特征、14个形状特征以及75个纹理特征，其中纹理特征涵盖了24个灰度共生矩阵（gray level co-occurence matrix，GLCM）特征、14个灰度依赖矩阵（gray level dependence matrix，GLDM）特征、16个灰度运行长度矩阵（gray level run length matrix，GLRLM）特征、16个灰度大小区域矩阵（gray level size zone matrix，GLSZM）特征和5个邻域灰度依赖矩阵（neighbouring gray-tone dependency matrix，NGTDM）特征。同时，采用小波变换、平方、平方根、对数、指数和梯度6种图像滤波器对原始图像进行预处理。

最终，从患者CE-T1和T2-FLAIR序列中均提取了1688个影像学特征。在CE-T1序列中，由于两个特征（T1c_lbp-3D-m1_firstorder_Minimum和T1c_lbp-3D-m2_firstorder_Minimum）值均为零，因此将其排除，最终保留1686个特征。在T2-FLAIR中，两个特征（FLAIR_lbp-3D-m1_firstorder_Minimum和FLAIR_lbp-3D-m2_firstorder_Minimum）的值有98.15%为0，因此也被排除，剩余1686个特征。

3. 统计学分析

将同时发生MGMT启动子甲基化和IDH突变的胶质瘤患者归类为IDHmut&MGMTmet亚型（99，47.60%），其余患者归类为Other亚型（109，52.40%）。使用R 4.3.2进行统计分析。显著性水平$\alpha = 0.05$。

（1）基线资料和分子信息的统计描述

将患者年龄按照50岁（均值为50.88岁）划分为两类，使用卡方检验对胶质瘤亚型之间的临床及分子因素进行差异检验。为评估胶质瘤不同分子亚型的预后并确定高危预后亚型，控制年龄、性别和病理分期3个协变量，应用多因素Cox比例风险模型分析亚型与生存结局之间的关系。

（2）多模态MRI数据特征筛选

将CE-T1和T2-FLAIR序列进行标准化。两个序列均包含1686个特征，其中多数特征可能与预测胶质瘤分子亚型无关，因此需要进行特征选择以优化模型性能。本研究基于惩罚logistic回归分别对CE-T1和T2-FLAIR两个序列进行稳健特征选择[21]。首先，将样本按8∶2的比例随机分为训练集和测试集。以胶质瘤分子亚型为因变量，影像学特征为自变量，使用R包“glmnet”对训练集执行惩罚logistic回归以选择特征，通过5折交叉验证选择最佳λ值。记录100次样本随机划分的特征选择结果，选择出现频率≥10%的影像学特征作为最终筛选特征。

（3）胶质瘤分子亚型预测模型构建

以最终筛选特征为预测因子，基于GA-fKPLS构建胶质瘤分子亚型预测模型，并和KNN、LASSO、LR和NB所构建的预测模型进行比较。为提高预测结果的稳定性和可重复性，将样本随机划分过程在不同种子数下重复进行1000次。

①GA-fKPLS预测模型

GA-fKPLS[20]先利用高斯核函数$K(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)$，构造影像组学数据的核矩阵，基于核偏最小二乘算法（kernel partial least squares，KPLS）有监督学习框架构建融合核，将融合核函数（K_f）定义为多个核的非负线性组合，$K_f(x_i, x_j) = \sum_{l=1}^{L} w_l K_l(x_i, x_j)$，其中$K_l$为第$l$个模态数据的核矩阵，$w_l$为权重，$w_l > 0$，$\sum_{l=1}^{L} w_l = 1$。通过R包“GA”利用遗传算法对高斯核函数参数$\sigma$和权重$w_l$进行优化。

a. 基于fKPLS的核分类算法：

设X为$n \times p$的标准化预测变量矩阵，$Y = (Y_1, \cdots, Y_n)$为$n \times 1$的响应变量，其中$Y_i \in \{+1, -1\}$，其中+1类代表IDHmut&MGMTmet亚型，-1类代表Other亚型。

第一，将原始数据按8∶2比例随机分为建模数据（n_{mod}）和外部测试数据（n_{ext}）。建模数据进一步分为训练集和内部测试集，并采用5折交叉验证选择模型最优参数。

第二，在训练集中，对响应变量进行中心化处理，即$\tilde{Y}_{tr-k} = Y_{tr-k} - \bar{Y}_{tr-k}$，$\bar{Y}_{tr-k}$是去掉第$k$折观察值后的样本均值。

第三，为每个训练集的预测变量矩阵计算核矩阵K_{tr-k}，其中K_{tr-k}是$n_{tr} \times n_{tr}$的矩阵，$n_{tr} = 0.8n_{mod}$，可通过$\left(K_{tr-k(i,j)}\right)_{i,j=1}^{n_{tr}} = K_f(x_i, x_j)$来计算，$K_f(\cdot)$是经过GA优化的融合核函数。对$K_{tr-k}$进行中心化处理：

$$\tilde{K}_{tr-k} = \left(I - \frac{1}{n_{tr}} 1_{n_{tr}} 1_{n_{tr}}^T\right) K_{tr-k} \left(I - \frac{1}{n_{tr}} 1_{n_{tr}} 1_{n_{tr}}^T\right)$$

其中I是$n_{tr} \times n_{tr}$的单位矩阵，$1_{n_{tr}}$是长度为n_{tr}的全1向量。

第四，初始化Y的潜变量向量u_{tr-k}为$\tilde{Y}_{tr-k}$。

第五，计算X的潜变量向量$t_{tr-k} = \Phi\Phi^T u_{tr-k} = \tilde{K}_{tr-k} u_{tr-k}$，通过$t_{tr-k} \leftarrow t_{tr-k}/\|t_{tr-k}\|$归一化。$\Phi\Phi^T$可以通过使用原

始空间的数据计算得到的核矩阵K计算出来。

第六，计算权重$c_{tr-k}=\tilde{Y}_{tr-k}^{T}t_{tr-k}$。

第七，计算新的Y潜变量向量$u_{tr-k}=c_{tr-k}\tilde{Y}_{tr-k}$，通过$u_{tr-k}\leftarrow u_{tr-k}/\|u_{tr-k}\|$进行归一化。

第八，重复步骤第五到第七直到达到收敛标准[22]：$\left\|(t_{tr-k_{old}}-t_{tr-k_{new}})\right\|/\|t_{tr-k_{old}}\|<10^{-9}$。

第九，重新定义$\tilde{K}_{tr-k}$和$\tilde{Y}_{tr-k}$：$\tilde{K}_{tr-k}=(I-t_{tr-k}t_{tr-k}^{T})\tilde{K}_{tr-k}(I-t_{tr-k}t_{tr-k}^{T})$，$\tilde{Y}_{tr-k}=\tilde{Y}_{tr-k}-t_{tr-k}t_{tr-k}^{T}\tilde{Y}_{tr-k}$。重新定义的$\tilde{K}_{tr-k}$和$\tilde{Y}_{tr-k}$将用于提取下一组潜变量$t_{tr-k}$和$u_{tr-k}$。

第十，重复步骤第四到第九，直到达到最大约登指数，$YoudenIndex=sensitivity+specificity-1$。根据内部测试集的最大约登指数确定最优潜变量数量，记为m_{opt}。

第十一，使用以下公式对新数据进行预测：

$$f\left(X_{n_{new}\times p}\right)=sign\left(\tilde{K}_{new}U_{mod}(T_{mod}^{T}\tilde{K}_{mod}U_{mod})^{-1}T_{mod}^{T}Y_{mod}+\bar{Y}_{mod}\right)$$

其中，n_{new}为新数据的样本数，如果$f\left(X_{n_{new}\times p}\right)$为正，则将新数据分类为+1类；否则，将新数据分类为-1类[23]。$\tilde{K}_{new}$是包含新数据和建模数据的中心化核矩阵，其计算公式为：

$$\tilde{K}_{new}=\left(K_{new}-\frac{1}{n_{mod}}1_{n_{new}}1_{n_{mod}}^{T}K_{mod}\right)\left(I-\frac{1}{n_{mod}}1_{n_{mod}}1_{n_{mod}}^{T}\right)$$

其中，K_{new}是$n_{new}\times n_{mod}$的核矩阵，K_{new}元素为$(K_{mod})_{i,j=1}^{n_{mod}}=K_f(x_i,x_j)$，其中$\{x_i\}_{i=n_{mod}+1}^{n_{mod}+n_{new}}$，$\{x_j\}_{j=1}^{n_{mod}}$分别是新数据点和建模数据点。$K_{mod}$是基于建模数据的未经中心化的$n_{mod}\times n_{mod}$的核矩阵，$(K_{mod})_{i,j=1}^{n_{mod}}=K_f(x_i,x_j)$，$\tilde{K}_{mod}$是中心化核矩阵，$\tilde{K}_{mod}=\left(I-\frac{1}{n_{mod}}1_{n_{mod}}1_{n_{mod}}^{T}\right)K_{mod}\left(I-\frac{1}{n_{mod}}1_{n_{mod}}1_{n_{mod}}^{T}\right)$。$T_{mod}$和$U_{mod}$分别是提取的$m_{opt}$个潜变量的$n_{mod}\times m_{opt}$矩阵，其中$m_{opt}$是在第十步骤中确定的最优潜变量数量。

b. 遗传算法（genetic algorithm，GA）是受遗传学中的自然选择和遗传机制启发发展起来的一种优化算法，基本思想为模拟生物和人类进化的方法求解复杂的优化问题[24]。遗传算法通过借鉴进化生物学中的遗传、变异、选择和交叉等机制，能够高效地处理大规模搜索空间，找到全局最优解[25]。本研究中，将两种不同的MRI数据（CE-T1和T2-FLAIR）进行融合，则由4个参数（w_1、w_2、σ_1和σ_2）构成实验条件。实验条件就像一个包含4个基因（即4个参数的具体值）的染色体。通过预定义的适应度函数（约登指数）来评估每个染色体与相应实验条件相关的响应。随机初始化一个种群（大小为20），通过3次迭代进行进化，交叉率为80%，变异率为10%。每次进化后，选择适应度较高的个体进入下一代，直至达到最大迭代次数或满足收敛条件。

②其他预测模型

LASSO通过收缩或正则化回归系数来拟合预测模型[26]。通过惩罚，将不重要的特征系数趋向于零，从而实现特征选择，通过10折交叉验证选择最小误差对应的λ值作为最优参数。KNN通过计算样本与训练集中其他样本的距离，找到与之距离最近的K个训练样本，然后基于这K个邻居的标签来进行预测[27]。LR通过Sigmoid函数将线性模型的输出转换为概率值，计算样本属于某个类别的概率[28]。NB是通过贝叶斯定理计算后验概率，并选择具有最大后验概率的类别作为预测结果[29]。

（4）预测模型性能评估采用曲线下面积（area under the curve，AUC）、敏感性（sensitivity，Se）、特异性（specificity，Sp）、准确率（accuracy，ACC）、约登指数（Youden index）、G均值（G-means）、马修斯相关系数（Matthews correlation coefficient，MCC）和F值（F-measure）8种评价标准对5种模型的性能进行评价。其中AUC是较为全面的评价标准。并采用单因素方差分析和*Dunnett*多重比较检验，比较GA-fKPLS和其他四种模型的AUC。

（5）影像学特征与胶质瘤分子亚型的关联分析将胶质瘤的两种分子亚型作为因变量，最终筛选特征作为自变量，进行单因素logistic回归分析，筛选出$P<0.05$的影像学特征。进一步使用LASSO回归筛选特征，并采用10折交叉验证选择最小误差值对应的λ值。最后进行多因素logistic回归分析，$P<0.05$表示与分子亚型显著相关。通过上述分析，最终确定与胶质瘤分子亚型显著相关的影像学特征，并绘制森林图。

结　果

1. 患者临床资料与分子信息

胶质瘤两种分子亚型样本的临床资料和分子信息统计描述如表1。年龄、肿瘤病理分级、生存状态、MGMT启动子甲基化状态以及IDH突变情况在两亚型间差异均有统计学意义（$P<0.05$）。胶质瘤患者多因素Cox回归分析结果如表2。Other亚型患者死亡风险为IDHmut&MGMTmet亚型的2.675倍（$P<0.05$）。年龄、病理分级对生存结局有显著影响（$P<0.05$），年龄>50岁的患者死亡风险是年龄≤50岁的2.286倍，WHO 3级和WHO 4级患者的死亡风险分别是WHO 2级患者的3.262倍和6.970倍。

表 1　胶质瘤不同分子亚型临床资料和分子信息的比较

变量	IDHmut& MGMTmet(*n*)	Other(*n*)	χ^2	*P*
年龄(岁)			28. 887	<0. 001
≤ 50	65	30		
> 50	34	79		
性别			0. 021	0. 884
男	48	55		
女	51	54		
病理分级			100. 530	<0. 001
WHO 2级	49	7		
WHO 3级	44	23		
WHO 4级	6	79		
生存状态			71. 444	<0. 001
生存	79	22		
死亡	20	87		
MGMT启动子			77. 530	<0. 001
未甲基化	0	62		
甲基化	99	47		
IDH			178. 060	<0. 001
未突变	0	102		
突变	99	7		

注:分类变量以例数(n)表示。

2. 模型性能比较

通过惩罚 logistic 回归筛选得到 110 个影像学特征,其中 CE−T1 特征 42 个,T2−FLAIR 特征 68 个。用每种模型 1000 个结果的 AUC、Se、Sp、ACC、约登指数、F 值、MCC、G−means 的平均值比较所有模型的预测性能。如表 3 所示,GA−fKPLS 模型在 AUC、Se、约登指数、F 值、G−means 中表现最优,且在 ACC、MCC 表现次优。图 1 为 5 种模型的 AUC 值分布,GA−fKPLS 模型的 AUC 值显著高于其他模型(*P*<0. 05)。以上结果表明 GA−fKPLS 模型的预测性能优于其他模型。

表 2　胶质瘤患者多因素 Cox 回归分析结果

变量	*b* (*S. E*)	*Z*	*P*	*HR* (95% *CI*)
组别	0. 984 (0. 352)	2. 793	0. 005	2. 675 (1. 341~5. 334)
性别	0. 267 (0. 201)	1. 330	0. 183	1. 306 (0. 881~1. 937)
病理分级				
WHO 3级	1. 182(0. 457)	2. 586	0. 010	3. 262(1. 331~7. 992)
WHO 4级	1. 942(0. 515)	3. 772	<0. 001	6. 970(2. 541~19. 118)
年龄	0. 827 (0. 241)	3. 434	<0. 001	2. 286 (1. 426~3. 666)

注:以 IDHmut&MGMTmet 亚型、女性、WHO 2 级、年龄≤ 50 岁为参考组。

表 3　5 种预测模型的评价指标

模型	AUC	Se	Sp	ACC	Youden	*F* 值	MCC	G−means
GA−fKPLS	0. 913	0. 857	0. 824	0. 840	0. 681	0. 836	0. 685	0. 838
LR	0. 747	0. 739	0. 754	0. 747	0. 493	0. 734	0. 498	0. 734
LASSO	0. 841	0. 838	0. 843	0. 841	0. 681	0. 833	0. 686	0. 838
KNN	0. 832	0. 771	0. 894	0. 835	0. 665	0. 815	0. 676	0. 828
NB	0. 807	0. 740	0. 874	0. 810	0. 614	0. 786	0. 626	0. 802

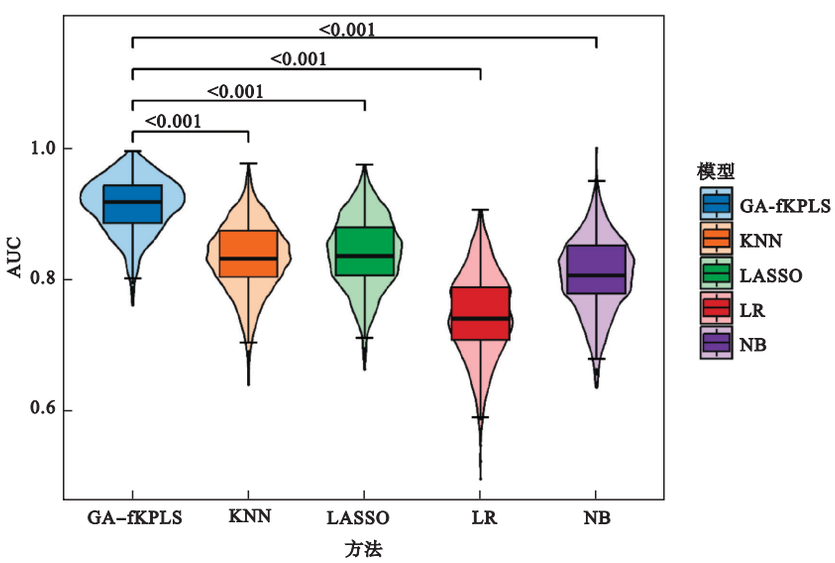


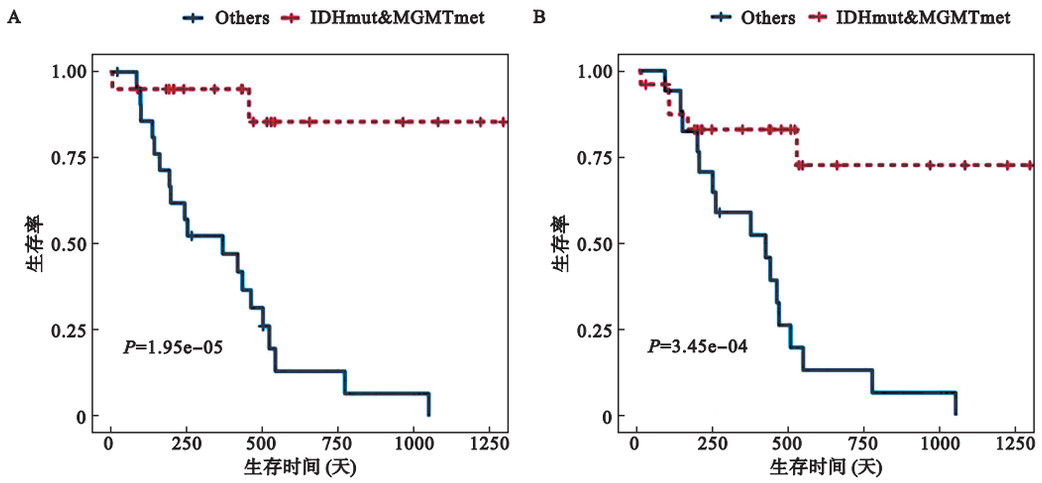
图 1　5 种模型 AUC 值的箱线图

3. GA−fKPLS 模型预测胶质瘤结果评价

为了说明 GA−fKPLS 模型识别胶质瘤分子亚型的临床意义,选择 166 个训练样本和 42 个测试样本随机分割的预测结果,MCC=0. 689(接近 GA−fKPLS 的 MCC 均值 0. 685)。分别基于测试集分子亚型的原始标签和预测标签,以及测试集的生存信息,使用 Kaplan−Meier 生存曲线和 log−rank 检验分析不同胶质瘤分子亚型患者之间的生存差异。如图 2 所示,图 2A 表明测试集胶质瘤两种分子亚型的生存概率显著不同(*P*<0. 05)。图 2B 显示该预测模型得到的生存曲线与原始数据接近,表示 GA−fKPLS 模型对于胶质瘤患者分子亚型预测效果较好。

4. 影像学特征与胶质瘤分子亚型的关联分析结果

影像学特征与胶质瘤分子亚型的关联分析得到 3 个与胶质瘤分子亚型显著相关的影像学特征,包括 2 个 CE−T1 特征和 1 个 T2−FLAIR 特征。如表 4 所示,



（A）：根据原始亚型标签绘制的测试集生存曲线；（B）：根据GA-fKPLS模型预测亚型标签绘制的测试集生存曲线。

图2 胶质瘤测试集不同分子亚型Kaplan-Meier生存曲线图

当T1c_wavelet. LHL_glcm_ClusterShade和T1c_original_shape_SurfaceVolumeRatio的表达量较高时，患者更可能属于IDHmut&MGMTmet亚型；而当FLAIR_wavelet. HLL_glszm_GrayLevelVariance的表达量较高时，患者则更可能属于Other亚型。

表4 显著影响胶质瘤分子亚型的影像学特征

Features	*P*	*OR*（95%*CI*）
T1c_wavelet. LHL_glcm_ClusterShade	0. 040	2. 100（1. 034~4. 266）
T1c_original_shape_SurfaceVolumeRatio	0. 040	3. 020（1. 052~8. 667）
FLAIR_wavelet. HLL_glszm_GrayLevelVariance	0. 042	0. 380（0. 150~0. 966）

讨 论

本研究基于无创的多模态术前MRI数据构建用于预测胶质瘤分子亚型（IDHmut&MGMTmet亚型和Other亚型）的GA-fKPLS模型，并与四种机器学习方法所构建的预测模型进行比较。GA-fKPLS模型的AUC显著高于四种机器学习模型，表明GA-fKPLS模型的预测性能更优，能较准确地预测胶质瘤患者的分子亚型，为胶质瘤患者预后评估以及个体化治疗方案的制定提供科学依据和实践指导。

IDH突变合并MGMT启动子甲基化、年龄小、病理分级低是胶质瘤患者的保护因素。IDH突变合并MGMT启动子甲基化的胶质瘤患者年龄小于其他分子状态的患者，且主要为低级别胶质瘤，其他分子状态的患者主要为胶质母细胞瘤，这与以往研究结果一致[30-31]。影像学特征T1c_wavelet. LHL_glcm_ClusterShade、T1c_original_shape_SurfaceVolumeRatio和FLAIR_wavelet. HLL_glszm_GrayLevelVariance与胶质瘤分子亚型显著相关，这些特征为胶质瘤分子亚型的精准识别提供了重要的影像学依据。

基于多模态MRI数据构建的GA-fKPLS预测模型，相较于仅将两个模态数据简单拼接构建的机器学习模型（如LR、LASSO、KNN、NB），具有更好的预测性能，说明GA-fKPLS可以有效地融合不同MRI模态数据，捕捉多模态数据中的潜在非线性关联，提升胶质瘤IDH突变合并MGMT启动子甲基化预测性能。

利用无创术前MRI数据获取胶质瘤IDH和MGMT分子标记，相较于病理检验，具有无创、高效且经济的优势。MRI是胶质瘤最常用的成像模式，因此基于多模态MRI数据构建的预测胶质瘤分子亚型（IDHmut&MGMTmet亚型和Other亚型）的GA-fKPLS模型具备临床应用潜力。随着磁共振波谱（magnetic resonance spectrum，MRS）和正电子发射断层扫描（positron emission tomography，PET）等新兴成像技术的普及，未来有望探索更多模态数据的融合，进一步提升对胶质瘤分子亚型的预测能力[32-33]。本研究缺乏独立的外部验证集，未来可纳入多中心数据来评估GA-fKPLS模型预测胶质瘤IDH突变合并MGMT启动子甲基化的普适性和稳定性。

综上所述，本研究基于术前胶质瘤多模态MRI数据构建的预测胶质瘤患者分子亚型的GA-fKPLS模型具有良好预测性能，能够有效鉴别MGMT启动子甲基化合并IDH突变，有助于在术前阶段进行胶质瘤患者的精准分型，为后续的治疗干预和预后评估提供重要参考，具有重要的临床应用价值。

参 考 文 献

[1] Ostrom QT, Bauchet L, Davis FG, et al. The epidemiology of glioma in adults: a "state of the science" review[J]. Neuro-Oncology, 2014,16(7):896-913.

[2] McNamara C, Mankad K, Thust S, et al. 2021 WHO classification

of tumours of the central nervous system: a review for the neuroradiologist[J]. Neuroradiology, 2022,64(10):1919-1950.

[3] Itakura H, Achrol AS, Mitchell LA, et al. Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities[J]. Science Translational Medicine, 2015,7(303):138-303.

[4] Yang G, Sha Y, Wang X, et al. Radiomics Profiling Identifies the Incremental Value of MRI Features beyond Key Molecular Biomarkers for the Risk Stratification of High-Grade Gliomas[J]. Contrast Media & Molecular Imaging, 2022(1):8952357.

[5] Smits M, van den Bent MJ. Imaging correlates of adult glioma genotypes[J]. Radiology, 2017,284(2):316-331.

[6] Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary[J]. Neuro-Oncology, 2021,23(8):1231-1251.

[7] Molenaar RJ, Verbaan D, Lamba S, et al. The combination of IDH1 mutations and MGMT methylation status predicts survival in glioblastoma better than either IDH1 or MGMT alone[J]. Neuro-Oncology, 2014,16(9):1263-1273.

[8] Tanaka K, Sasayama T, Mizukawa K, et al. Combined IDH1 mutation and MGMT methylation status on long-term survival of patients with cerebral low-grade glioma[J]. Clinical Neurology and Neurosurgery, 2015,138:37-44.

[9] Leu S, von Felten S, Frank S, et al. IDH/MGMT-driven molecular classification of low-grade glioma is a strong predictor for long-term survival[J]. Neuro-Oncology, 2013,15(4):469-479.

[10] 谢致岩. 扩散加权成像预测脑胶质瘤IDH突变与MGMT启动子甲基化共现的研究[D]. 青岛: 青岛大学, 2023.

[11] Ibrahim A, Primakov S, Beuque M, et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework[J]. Methods, 2021,188:20-29.

[12] Liu Z, Wang S, Dong D, et al. The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges[J]. Theranostics, 2019,9(5):1303-1322.

[13] Avanzo M, Wei L, Stancanello J, et al. Machine and deep learning methods for radiomics[J]. Medical physics, 2020,47(5):e185-e202.

[14] Song J, Yin Y, Wang H, et al. A review of original articles published in the emerging field of radiomics[J]. European Journal of Radiology, 2020,127:108991.

[15] 沙永建, 王效春, 谭艳, 等. MRI影像组学术前预测较低级别胶质瘤IDH突变合并MGMT启动子甲基化亚型的价值[J]. 磁共振成像, 2022,13(7):6-11.

[16] Tan Y, Zhang S, Wei J, et al. A radiomics nomogram may improve the prediction of IDH genotype for astrocytoma before surgery[J]. European Radiology, 2019,29(7):3325-3337.

[17] Kim M, Jung SY, Park JE, et al. Diffusion- and perfusion-weighted MRI radiomics model may predict isocitrate dehydrogenase (IDH) mutation and tumor aggressiveness in diffuse lower grade glioma [J]. European Radiology, 2020,30(4):2142-2151.

[18] Sun C, Fan L, Wang W, et al. Radiomics and qualitative features from multiparametric MRI predict molecular subtypes in patients with lower-grade glioma[J]. Frontiers in Oncology, 2022, 11: 756828.

[19] Korfiatis P, Kline TL, Coufalova L, et al. MRI texture features as biomarkers to predict MGMT methylation status in glioblastomas [J]. Medical physics, 2016,43(6Part1):2835-2844.

[20] Yang H, Cao H, He T, et al. Multilevel heterogeneous omics data integration with kernel fusion[J]. Briefings in Bioinformatics, 2020, 21(1):156-170.

[21] Meinshausen N, Bühlmann P. Stability selection[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2010,72 (4):417-473.

[22] Rännar S, Lindgren F, Geladi P, et al. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm[J]. Journal of Chemometrics, 1994,8(2):111-125.

[23] Huang X, Pan W. Linear regression and two-class classification with gene expression data[J]. Bioinformatics, 2003,19(16):2072-2078.

[24] Holland JH. Genetic algorithms[J]. Scientific American, 1992,267 (1):66-73.

[25] Zhuo L, Zheng J, Li X, et al. A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine[C]//Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images. Bellingham: SPIE, 2008.

[26] 夏志强,方羚,任凌云,等. 基于LASSO回归模型构建扩张型心肌病死亡风险的预测模型[J]. 中国卫生统计,2025,42(4):590-593.

[27] AlOmair O, M Zakaria O, Alabdullatif M, et al. Optimized KNN with domain-informed features and LIME explainability for improved breast cancer classification[J]. BMC Medical Informatics and Decision Making, 2026, 26(1): 138.

[28] Stoltzfus JC. Logistic Regression: A Brief Primer[J]. Academic Emergency Medicine, 2011,18(10):1099-1104.

[29] Efron B. Bayes´ Theorem in the 21st Century[J]. Science, 2013,340 (6137):1177-1178.

[30] Wirsching H, Weller M. Glioblastoma[J]. Malignant Brain Tumors: State-of-the-Art Treatment, 2017:265-288.

[31] Zhang S, Sun H, Su X, et al. Automated machine learning to predict the co-occurrence of isocitrate dehydrogenase mutations and O6 - methylguanine - DNA methyltransferase promoter methylation in patients with gliomas[J]. Journal of Magnetic Resonance Imaging, 2021,54(1):197-205.

[32] 范廷勇, 李建彬, 于金明. 功能影像在脑胶质瘤放射治疗靶区确定中的价值[J]. 中华肿瘤杂志, 2006,28(10):721-723.

[33] 李方明, 聂青, 赵文锐, 等. CHO PET功能显像:脑胶质瘤诊断和优化放化疗方案[J]. 中华神经外科疾病研究杂志, 2011,10(2): 154-157.

（责任编辑：邓　妍）