doi: 10. 3969 / j. issn. 1674 - 1242. 2025. 05. 006

利用特征选择模型提高癌症亚型识别精度的多组学数据处理方法

程锦元，黄棋文，傅浩翔
（景德镇陶瓷大学信息工程学院，江西景德镇 333403）

【摘要】癌症是一种具有高危害性、高度异质性和复杂性的疾病，精准识别癌症亚型对指导个性化治疗和改善患者预后具有重要意义。为此，该文提出了一种合理的多组学数据处理方法，以提高癌症亚型识别精度。该方法主要利用特征选择模型对具有高维、小样本特征的组学数据进行排序，并结合癌症亚型识别模型进行数据清洗，提高癌症亚型识别精度。经过三种癌症数据与三种癌症亚型识别模型的验证，该方法有效提高了多组学癌症亚型识别模型的识别精度。该文对该领域的工作提出了展望，为精准癌症亚型识别的研究与发展提供了新视野。

【关键词】特征选择；癌症亚型；多组学数据；精准识别

【中图分类号】R73，TP181　　【文献标志码】A　　文章编号：1674-1242（2025）05-0636-09

Multi-Omics Data Processing Method for Improving Cancer Subtype Identification Precision Using a Feature Selection Model

CHENG Jinyuan，HUANG Qiwen，FU Haoxiang
(School of Information and Engineering, Jingdezhen Ceramic University, Jingdezhen, Jiangxi 333403, China)

【Abstract】Cancer is a high-risk, highly heterogeneous, and complex disease, and precisely identifying cancer subtypes is crucial for guiding personalized treatment and improving patients'prognosis. To this end, a rational multi-omics data processing method is proposed to improve the precision of cancer subtype identification. This method primarily utilizes a feature selection model to reasonably rank omics data characterized by high dimensionality and small sample sizes, and integrates a cancer subtype identification model for data cleaning, aiming to enhance the precision of cancer subtype identification. Through the validation of three types of cancer data and three cancer subtype identification models, this processing method effectively enhances the identification precision of the multi-omics cancer subtype identification model. Finally, the prospect of this work is put forward, which provides a new perspective for the research and development of precise subtype identification.

【Key words】Feature Selection; Cancer Subtypes; Multi-Omics Data; Precise Identification

收稿日期：2024-11-17。

作者简介：程锦元（2000—），男，江西省景德镇市人，硕士研究生，从事生物信息数据统计分析研究。黄棋文（2003—），男，江西省赣州市人，本科学历，从事应用统计数据分析研究。傅浩翔（2005—），男，江西省吉安市人，本科学历，从事应用统计数据分析研究。

通信作者：程锦元，男，硕士研究生，电话：15083986013，邮箱：1216748202@qq.com。

0　引言

癌症是一种具有高危害性、高度异质性和复杂性的疾病。根据 2020 年相关统计数据，2020 年全球新发癌症病例约 1930 万例，癌症死亡病例为 1000 万例（不包括非黑色素瘤皮肤癌，NMSC），其中大多数新发病例（600 万例，占总数的 31.1%）和死亡病例（360 万例，占总数的 36.3%）发生在东亚地区 [1]。根据 2022 年相关统计数据，2022 年中国恶性肿瘤新发病例约 482.47 万例，癌症死亡病例约 257.42 万例 [2]。这是因为基因改变、异常甲基化、表观遗传变化和患者特异性特征都可能导致癌症的形成和增殖 [3-6]，进而导致癌症产生多种亚型，提高了癌症的异质性和复杂性。不同癌症亚型患者通常具有不同的基因组特征和临床特征，且其预后反应和治疗结果差异很大 [7,8]。因此，精准确定癌症亚型有助于改善癌症的诊断、治疗和预后，为精准医疗提供参考依据。

目前，可通过图像数据分析、多组学分析、结构分析等多种方法识别癌症亚型 [9-11]。随着高通量测序技术的快速发展，多组学分析得到了更快的发展 [12]。同时，癌症多组学数据集的不断积累 [13]，使整合多组学数据成为癌症亚型分析中一种经济且高效的策略 [14]。虽然多组学分析能为各种癌症亚型固有的不同分子特征提供全面的视角 [15,16]，但是，不同组学的异构特征、多组学数据中的噪声或冗余信息，以及组学数据的高维小样本特性，会大幅增加癌症亚型识别难度 [17,18]。如何更科学地处理多组学数据、提高癌症亚型识别精度，是本研究的重点。同时，精准的癌症亚型识别能挖掘出更具代表性的生物标志物，并为临床治疗策略提供支持。

如今，已有许多基于组学数据整合聚类的癌症亚型识别方法。早期这类研究可分为早期整合方法、后期整合方法与应用统计的建模方法 [19]，但这些方法如今已较少被使用。更多研究偏向于结合机器学习、降维技术、基于网络的方法、深度学习等方式 [20-23]。由于多组学数据具有高维、小样本的特性，大部分研究基于无监督方式进行，而且各类研究对同类数据的清洗方式存在差异，尚未形成相对标准的数据处理流程，这使对符合高维小样本特征的组学数据进行合理清洗显得尤为重要。

本研究提出了一种更精准的多组学数据处理方法，以提高癌症亚型识别精度。本研究通过 GRACES 特征选择模型 [24]，在实现患者与健康人有效区分的基础上，为所有特征赋予得分并对特征进行排序。随后，结合癌症亚型分类模型剔除得分较低的特征，并通过多种癌症数据验证该清洗方法的有效性。最后，在上述研究的基础上，通过引入多组学癌症亚型分类方法进一步验证该方法的适用性，具体流程如图 1 所示。该方法的重点在于对高维组学数据进行合理的清洗，且适用于各类多组学癌症亚型分类模型，从而达到提高模型亚型分类精度的效果。

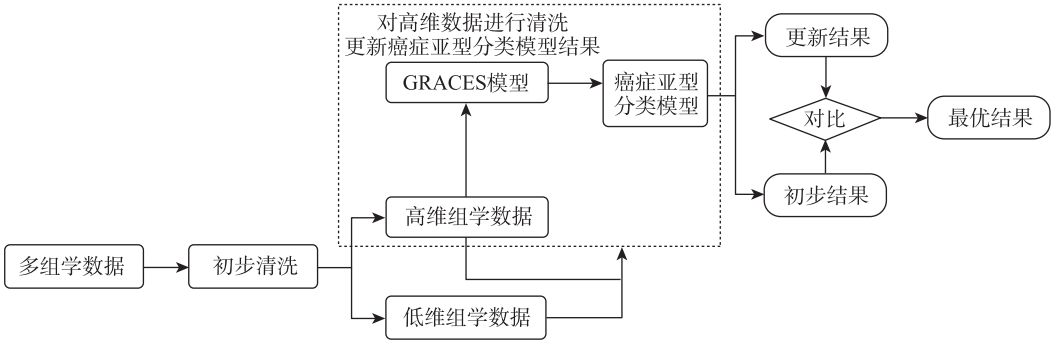


图 1　方法流程

Fig.1　Flow chart of the method

1　实验方法及材料

1.1　数据预处理

本研究从癌症基因组图谱（TCGA）网站下载了肾透明细胞癌（KIRC）、肝细胞癌（LIHC）与乳腺癌（BC）的数据，获取了 mRNA 表达数据、miRNA 表达数据、DNA 甲基化数据和临床数据。

其中，mRNA 表达数据与 DNA 甲基化数据具有样本量少、特征维度高的特点，将其归类为高维组学数据；miRNA 表达数据的样本量与特征维度差距不大，将其归类为低维组学数据。

首先，对患者的临床数据进行清洗：删除临床数据中被 AJCC 病理分期系统标注为 NA（缺失值）的样本；删除总生存期（OS）观察期开始后 30 天内死亡患者对应的样本 [25]。其次，将三种组学数据与清洗后的临床数据样本进行交集筛选，确定患者样本数量分别为 301 个、324 个、735 个。考虑到三种癌症患者样本的 miRNA 表达数据缺失率极高，分别为 71.96%、66.08%、68.14%，本研究保留 miRNA 样本缺失率小于 75% 的 KIRC 与 BC 患者样本，以及缺失率小于 70% 的 LIHC 患者样本，确保其他组学数据与保留后的 miRNA 患者样本一致，最终确定患者样本数量分别为 259 个、266 个、697 个。最后，删除这三种组学数据中缺失率大于 30% 的特征；进一步对 mRNA 表达数据与 miRNA 表达数据进行 $\log_2(X+1)$ 转换。其中，DNA 甲基化数据选择 Illumina Human Methylation 450 BeadChip 平台注释启动子区域的 CpG 位点，启动子区域被定义为转录起始位点 2 kbp 以内的区域 [26]；同时排除性染色体特征，过滤缺失率大于 30% 的特征，并用 0 值替代缺失值。具体内容如表 1 所示。

表 1 初步清洗的多组学数据数量

Tab.1 The number of multi-omics data initially cleaned

样本	样本数量 / 个	miRNA	mRNA	DNA 甲基化
KIRC	259	398	16725	50110
LIHC	266	484	15892	48895
BC	697	428	16752	50044

对于正常人样本数据，本研究首先对具有高维小样本特征的 mRNA 表达数据与 DNA 甲基化数据样本进行交集筛选，确定正常人样本；其次提取与患者样本一致的特征，将正常人样本与患者样本合并构建数据矩阵（维度为样本数 × 特征数），用 0 值替代缺失值后进行 $\log_2(X+1)$ 转换；最后对所有数据矩阵进行归一化处理，使数据取值范围为 0~1，以消除不同数据类型之间的尺度差异 [27]。包含正常人样本的数据将用于 GRACES 模型，为所有特征赋予得分，为后续的清洗工作提供依据。

1.2 统计方法

本节介绍的内容包括特征选择模型（GRACES）、多核学习亚型分类模型（hMKL）、相似性网络融合（SNF）、癌症亚型分类方法（SNFCC）、本方法介绍、癌症亚型差异表达基因分析。主要步骤如下：首先，本研究使用 GRACES 模型为符合高维小样本特征的多组学数据赋予得分；其次，结合 hMKL 模型通过消融实验确定特征的最佳清除率；最后，通过 SNF 与 SNFCC 两种癌症亚型分类方法，验证最佳清除率是否有效。

1.2.1 GRACES

GRACES 是 2023 年提出的一种基于深度学习的特征选择模型。该模型通过多重压差、高斯噪声和 f 校正，最大限度地降低优化损失，在完成分类任务的基础上迭代寻找一组最优特征。本研究使用该模型计算高维小样本数据中所有特征的得分。

1.2.2 hMKL

hMKL 是 2022 年提出的一种分层多核学习模型 [28]。首先，该模型优化单个组学数据类型的核参数，借鉴癌症整合多核学习（CIMLR）的思想，分别为每种组学数据构建一个复合核；其次，通过无监督多核学习方法，对单个组学数据的复合核进行加权线性组合，得到最终的融合核；最后，基于最终融合核，应用 k-means 聚类识别癌症亚型。本研究使用该模型对癌症患者进行亚型分类。

1.2.3 SNF

SNF 是 2014 年提出的一种基于非贝叶斯网络的癌症亚型分类方法 [29]。该方法主要分为两步：第一步，针对每种数据类型，在患者中构建一个相似性网络；第二步，在网络融合阶段，通过非线性组合迭代更新相似性网络，使它们的相似度逐渐提高，最终收敛到一个统一的融合网络。本研究使用该方法对癌症患者进行亚型分类。

1.2.4 SNFCC

SNFCC 是 2017 年 CancerSubtypes R 包 [30] 中提出的一种新的癌症亚型鉴定方法。该方法将 SNF 与一致性聚类（CC）相结合，既保留了 SNF 处理

多模态数据的融合能力与鲁棒性，又保留了 CC 对聚类结果的稳定性和可靠性。本研究使用该方法对癌症患者进行亚型分类。

1.2.5　本方法介绍

本方法的关键步骤是使用 GRACES 模型。原始 GRACES 模型无法获取所有特征的得分，只能得到一组最优特征。为此，首先，本研究对 GRACES 模型进行改进，使其能够保留所有可区分患者与健康人的特征得分。具体而言，本研究对原模型（https://github.com/canc1993/graces）的 real_test.py 文件进行修改，在代码的最后一步添加得分保留机制，确保在不改变原模型输出结果的基础上，保留所有特征的得分。其中，参数设定为 $\max_{features}=20$, $n_{iters}=5$, $n_{repeats}=1$），通过该参数设置可得到所有特征的五组得分，对这五组得分取均值以确定最终的特征得分大小。其次，根据最终特征得分从高到低进行排序。最后，结合癌症亚型分类模型剔除得分较低的特征，得到最优亚型分类结果。本方法的优势在于可对各类多组学癌症亚型分类模型或方法进行二次优化，提升癌症亚型分类精度。

1.2.6　评价癌症亚型的差异性

为探讨癌症亚型与患者生存结局之间的关系，本研究采用生存分析方法评价癌症亚型生存率的临床意义，通过 Kaplan-Meier 生存曲线图和 log-rank 检验评估癌症亚型与患者生存结局之间的关联 [31]。为进一步揭示癌症亚型的生物学意义，本研究分析了不同癌症亚型之间的 mRNA 差异表达情况：首先使用 DESeq2 R 软件包 [32] 筛选满足 $\log_2$ 倍变化 |FC|>1、Padj<0.001 的差异表达 mRNA（DEmRNAs）；然后使用 clusterProfiler 4.0 包对 DEmRNAs 进行基因本体（GO）和京都基因与基因组百科全书（KEGG）富集分析，阈值设为 Padj<0.05。

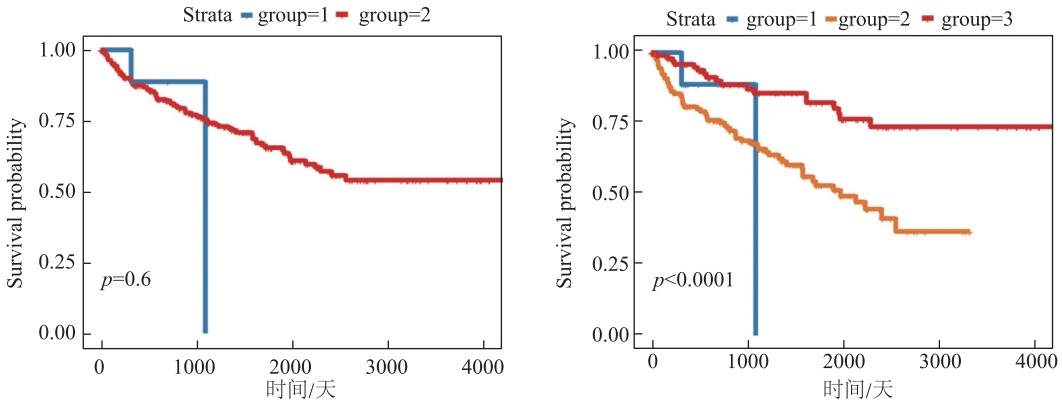
2　实验结果

2.1　癌症亚型分类结果

本研究以肾透明细胞癌（KIRC）为例，展示后续处理流程，这也是该方法的重点环节。

首先，本研究通过 hMKL 亚型分类模型对原始数据进行癌症亚型分类，并结合 Kaplan-Meier 生存曲线图得到初始癌症亚型分类结果。由图 2（a）可以看出，原始数据仅能识别出两种亚型，且这两种亚型间差异不显著。随后，本研究结合 GRACES 模型为高维组学数据赋予得分，按照得分从高到低对特征进行排序，以从最低分端按比例递增的方式剔除低得分特征。将剔除后的高维组学数据重新输入亚型分类模型，得到优化后的癌症亚型分类结果。

为确定最佳清除率，本研究开展了消融实验：按照清除比例与 GRACES 分值递增的方式（0 表示原始数据，0.8 表示剔除得分低于 0.8 的特征），通过 hMKL 亚型分类模型得到不同清除率下的亚型分类结果，结合 log-rank 检验的评估结果确定最佳 GRACES 清除率，具体结果如表 2 所示。



（a）原始数据得出的亚型生存曲线　（b）优化后得出的亚型生存曲线

图 2　KIRC 亚型的 Kaplan-Meier 生存曲线

Fig.2　Kaplan-Meier survival curves of KIRC subtypes

由表 2 可以看出，当剔除总特征中 15% 的低得分特征，或者剔除得分低于 0.8 的特征时，hMKL 模型的亚型分类效果最佳，能识别出与患者生存结局存在显著关联的癌症亚型，结果如图 2（b）所示。对比图 2（a）与图 2（b）可知，经过本方法处理后的高维组学数据，可提高癌症亚型分类模型的精准性，得到更优的癌症亚型分类结果。

表 2　GRACES 清除率确定表
Tab.2　The determination of GRACES clearance rate

清除率（比例）	mRNA	DNA 甲基化	χ^2 值	P 值
0	16725	50110	0.282	5.954×10^{-1}
5%	15889	47604	0.282	5.954×10^{-1}
10%	15052	45099	0.326	8.496×10^{-1}
15%	**14216**	**42594**	**19.242**	$\mathbf{6.629\times10^{-6}}$
20%	13380	40088	0.326	8.496×10^{-1}
25%	12544	37582	0.326	8.496×10^{-1}
30%	11708	35007	0.326	8.496×10^{-1}
清除率（以分数值为阈值）				
0.1	16667	49742	0.282	5.954×10^{-1}
0.2	16467	48684	0.282	5.954×10^{-1}
0.3	16159	47322	0.282	5.954×10^{-1}
0.4	15773	45937	0.282	5.954×10^{-1}
0.5	15379	44410	0.326	8.496×10^{-1}
0.6	14987	42862	0.326	8.496×10^{-1}
0.7	14566	41203	0.326	8.496×10^{-1}
0.8	**14131**	**39569**	**19.242**	$\mathbf{6.629\times10^{-6}}$
0.9	13418	37886	0.326	8.496×10^{-1}

为进一步确定优化后癌症亚型的可靠性，本研究基于组别、年龄和性别构建 Cox 回归模型，分析不同癌症亚型的预后差异。考虑到原始数据与优化后数据得到的组 1 生存率差异无统计学意义，且优化后新增的亚型组 2 与组 3 均源自原始数据的亚型组 2，因此将优化后的亚型组 2 与组 3 分别命名为 Group1 和 Group2，基于年龄和性别构建 Cox 回归模型，分析两者的预后差异，结果如表 3 所示（以 Group2 作为亚型间差异比较的参考，$P<0.05$ 表示差异具有统计学意义）。

从表 3 可知，Group1 的死亡风险是 Group2 的 2.99 倍（$P=5.6E\text{-}05$），两组患者 3 年死亡率分别为 29.1% 和 11.2%。这一结果证明，通过本方法优化高维组学数据，可对癌症亚型分类模型进行二次优化，得到更为精准的癌症亚型。

最后，为展示优化后癌症亚型之间的生物学差异，本研究根据优化得到的 Group1 与 Group2 癌症亚型，绘制基于两种亚型差异表达 mRNA 的热图，共筛选出 194 个差异表达 mRNA（DEmRNAs），其中 11 个上调、183 个下调。由图 3 可以看出，优化后得到的两种癌症亚型之间存在显著差异，基于此进一步开展 GO 和 KEGG 富集分析。

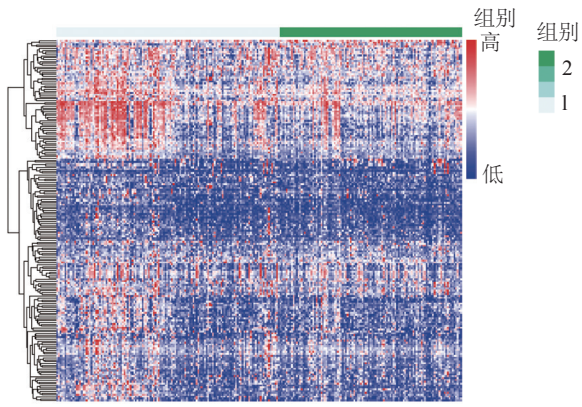
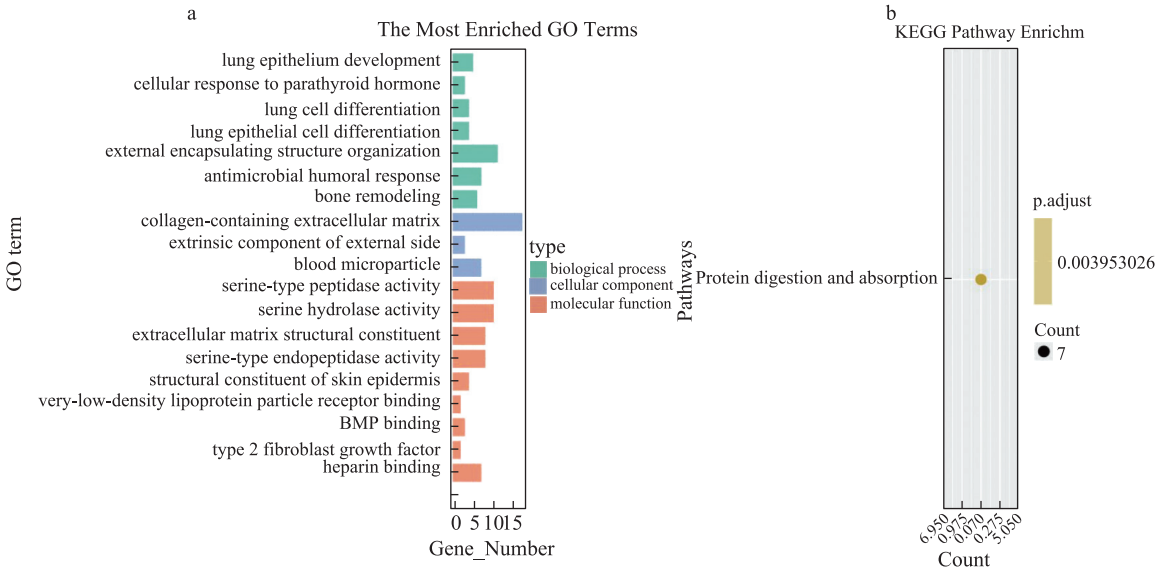


图 3　基于两种亚型差异表达 mRNA 的热图
Fig.3　Heatmap of the differentially expressed mRNAs between the two subtypes

表 3　247 例 KIRC 患者的 Cox 回归分析
Tab.3　Cox regression analysis of 247 KIRC patients

变量	系数和标准误差	Wald 统计量	P 值	风险比及 95% 置信区间
Group1	1.097159（0.272239）	4.030134	0.000056	2.995645（1.756953~5.107643）
年龄	0.021463（0.011676）	1.838317	0.066016	1.021695（0.998581~1.045345）
性别	0.139935（0.250426）	0.558786	0.576308	1.150198（0.704060~1.879039）

如图 4 所示，在 GO 功能分析中，两种亚型的差异基因显著富集于胶原蛋白相关的细胞外基质、外包囊结构组织、丝氨酸内肽酶活性和丝氨酸水解酶活性等多个生物学过程中，说明这两种亚型在细胞外基质重塑、蛋白质代谢及细胞微环境调控方面存在显著差异。此外，差异基因还涉及肺上皮细胞发育、骨重塑等功能，表明不同亚型间可能具有不同的组织特异性和功能状态。在 KEGG 通路分析中，虽仅显示蛋白质消化和吸收通路的富集，但这一结果进一步支持了两种亚型在蛋白质代谢和细胞外微环境



（a）GO 通路富集分析　　（b）KEGG 通路富集分析

图 4　差异表达基因功能注释

Fig.4　Functional annotation of differentially expressed genes

相关功能中的差异。

上述结果均表明，优化后得到的具有统计意义的 Group1 与 Group2 两种 KIRC 亚型分类合理，进一步证明本方法能有效提高多组学癌症亚型识别模型的亚型识别精度。

2.2　SNF 与 SNFCC 方法验证结果

本研究依据上述结论，使用确定的最佳清除率通过 SNF 与 SNFCC 方法进行验证。从表 4 可以看出，初步清洗的原始数据在这两种方法中表现出了较好的效果，但经过本方法处理后，癌症亚型分类精度进一步提升，log-rank 检验的 Chisq 值增大、*P* 值减小。同时，亚型数量发生变化。例如，在 SNF 模型中，亚型 1 数量从原本的 52 个减少至 50 个，亚型 2 数量从 131 个减少至 130 个，亚型 3 数量从 76 个增加至 79 个。

其中，log-rank 检验的 Chisq 值提高约 5，*P* 值减小幅度超过 94%，表明该方法能有效提高癌症亚

表 4　SNF 与 SNFCC 方法验证结果

Tab.4　Validation results of SNF and SNFCC methods

	清除率	亚型 1	亚型 2	亚型 3	Chisq 值	*P* 值
SNF	0	52	131	76	37.923	5.8235×10^{-9}
	15%	50	130	79	**43.567**	$\mathbf{3.4607\times10^{-10}}$
	0.8	50	130	79	**43.567**	$\mathbf{3.4607\times10^{-10}}$
SNFCC	0	54	129	76	37.964	5.7042×10^{-9}
	15%	51	131	77	**41.151**	$\mathbf{1.1592\times10^{-9}}$
	0.8	51	129	79	**43.409**	$\mathbf{3.7482\times10^{-10}}$

型分类精度。需要说明的是，不同癌症亚型分类模型或方法的算法与特性不同，其识别出的亚型在不同模型或方法间无可比性，需通过更多生物学分析验证亚型的实际意义。但经过本方法清洗后，各类癌症亚型分类方法的分类效果均有所提升，由此可得出结论：该清洗方法能提高癌症亚型识别精度。

2.3　LIHC 与 BRCA 癌症的实验结果

为展示该方法的适用性，本研究对肝细胞癌（LIHC）与乳腺癌（BC）数据进行实验，结果如下。表 5 展示了初步清洗的原始数据、经本方法确定最佳清除率处理后的数据，对应的亚型数量及 SNF 与 SNFCC 方法的验证结果。其中，在 hMKL 模型中，当 LIHC 数据以 0.5 为清除率时，mRNA 与 DNA 甲基化特征数量分别为 13534 和 39804（该数值为最优结果的高维多组学数据量，本研究未展示全部高维多组学数据量），共识别出 5 种癌症亚型；当 BC 数据以 0.4 为清除率时，mRNA 与 DNA 甲基化特征数量分别为 15601 和 45274，共识别出 3 种癌症亚型。

表 5　LIHC 与 BC 的全部实验结果
Tab.5　Full experimental results for LIHC and BC

癌症	分类方法	清除率	亚型数量					χ^2 值	P 值
LIHC	hMKL	0	10	54	56	93	53	23.503	1.01×10^{-4}
		0.5	10	51	55	97	53	**23.551**	$\mathbf{9.83\times10^{-5}}$
	SNF	0	27	51	60	78	50	17.853	1.32×10^{-3}
		0.5	46	39	60	72	49	**26.452**	$\mathbf{2.57\times10^{-5}}$
	SNFCC	0	23	49	59	82	53	17.643	1.45×10^{-3}
		0.5	20	49	58	81	58	**18.901**	$\mathbf{8.22\times10^{-4}}$
BC	hMKL	0	15		558		124	1.165	5.59×10^{-1}
		0.4	15		559		123	**1.195**	$\mathbf{5.50\times10^{-1}}$
	SNF	0	208		357		132	1.918	3.83×10^{-1}
		0.4	208		357		132	1.712	4.25×10^{-1}
	SNFCC	0	208		357		132	1.918	3.83×10^{-1}
		0.4	210		355		132	**2.023**	$\mathbf{3.64\times10^{-1}}$

由表 5 可知，LIHC 数据经本方法处理后，在 hMKL 模型、SNF 与 SNFCC 方法中均有所改善，具体表现为亚型数量发生变化，且 χ^2 值增大，P 值减小；BC 数据经本方法处理后，在 hMKL 模型与 SNFCC 方法中未获得显著改善，但亚型数量有一定的变化，在 SNF 方法中的应用效果不佳。经分析可知，KIRC 数据在 hMKL 模型中初始表现不佳，但在 SNF 与 SNFCC 方法中初始表现较好，因此本方法对 hMKL 模型的二次优化效果更显著；LIHC 数据在三种分类方法中初始表现均一般，但本方法仍能提高其亚型识别精度；BC 数据在三种分类方法中初始表现均不佳，导致本方法难以提升其亚型识别精度。由此可推断，本研究选用的癌症亚型分类方法不适用于结构复杂的 BC 亚型分类，进而影响本方法在 BC 数据中的应用效果。

3　总结与展望

癌症是一种复杂的异质性疾病，确定癌症亚型对发现潜在治疗靶点和实现精准治疗至关重要。本研究采用 GRACES 模型（一种适用于高维小样本数据的基于深度学习的特征选择模型）开展研究。首先，使用正常人与患者的组合数据进行二分类学习，在实现两者高精度区分的基础上，为所有特征赋予得分；其次，对所有特征进行排序，按照清除比例与分值递增的方式依次剔除低得分特征；最后，将剔除后的数据输入癌症亚型分类模型或方法中得到亚型，结合 log-rank 检验结果确定最佳清除率，以提高癌症亚型识别精度。

该方法经过多种癌症数据与癌症亚型分类模型或方法的验证，能有效提高癌症亚型识别精度。KIRC 多组学数据经该方法清洗后，在 hMKL 模型

中识别精度提升效果最显著，在 SNF 与 SNFCC 方法中也有一定的提升；LIHC 多组学数据经该方法清洗后，经三种方法验证均有提升；BC 多组学数据经该方法清洗后未获提升，推测是由于本研究选用的分类模型或方法对复杂的乳腺癌亚型分类效果不佳。但从整体结果来看，经该方法处理后的多组学数据，能提高不同多组学癌症亚型分类模型或方法的识别精度。

与其他研究类似，本研究也存在局限性：首先，本研究仅详细展示了 KIRC 数据的处理流程与验证结果，对 LIHC 与 BC 数据仅进行了简单的验证；其次，本方法虽能提高癌症亚型分类模型的精度，但对模型选择存在一定的依赖性。为此，未来研究需从 TCGA 数据库中获取更多癌症数据，结合各类癌症对应的有效多组学癌症亚型分类模型开展综合分析，以进一步证明该方法能实现更精准的癌症亚型识别。值得关注的是，将高精度识别癌症的高维特征与多组学癌症亚型分类模型结合，以识别更精准的亚型，是一种可行的新方式，但还需通过更多生物学分析，对比该方法处理前后癌症亚型的精准性，验证该方法在提高多组学数据癌症亚型识别精度方面的优势。

综上所述，本研究通过 GRACES 模型处理符合高维小样本特征的多组学数据，并结合多组学癌症亚型识别模型或方法，能有效提高癌症亚型识别精度，为精准癌症亚型识别提供了新策略和新思路。未来研究需正视本研究的局限性，通过更充分的实验验证该方法处理前后癌症亚型的精准性，并对识别出来的精确亚型进行深度挖掘，为精准医疗提供新见解。

参考文献

[1] FERLAY J, COLOMBET M, SOERJOMATARAM I, *et al*. Cancer statistics for the year 2020: an overview[J]. **Int J Cancer**, 2021.

[2] 郑荣寿，陈茹，韩冰峰，等 . 2022 年中国恶性肿瘤流行情况分析 [J]. **中华肿瘤杂志**，2024，46（3）：221-231.
ZHENG Rongshou, CHEN Ru, HAN Bingfeng, *et al*. Cancer incidence and mortality in china, 2022[J]. **Chinese Journal of Oncology**, 2024, 46(3):221-231.

[3] XU Z, ZHANG L, WANG M, *et al*. A novel subtype to predict prognosis and treatment response with DNA driver methylation-transcription in ovarian cancer[J]. **Epigenomics**, 2022, 14(18): 1073-1088.

[4] LI C, KE J, LIU J, *et al*. DNA methylation data-based molecular subtype classification related to the prognosis of patients with cervical cancer[J]. **J Cell Biochem,** 2020, 121(3): 2713-2724.

[5] VIKAL A, MAURYA R, PATEL B B, *et al*. Protacs in cancer therapy: mechanisms, design, clinical trials, and future directions[J]. **Drug Deliv Transl Res**, 2025, 15(6):1801-1827.

[6] MCMANUS H D, DORFF T, MORGANS A K, *et al*. Navigating therapeutic sequencing in the metastatic castration-resistant prostate cancer patient journey[J]. **Prostate Cancer Prostatic Dis**, 2025, 281(3):672-683.

[7] ZHAO J, ZHAO B, SONG X, *et al*. Subtype-DCC: decoupled contrastive clustering method for cancer subtype identification based on multi-omics data[J]. **Brief Bioinform**, 2023, 24(2):bbac1025.

[8] 杨瑞，武永丽，石项天，等 . 多模态超声及超声影像组学在预测乳腺癌分子亚型中的应用进展 [J]. **内蒙古医学杂志**，2024，56（10）：1232-1235.
YANG Rui, WU Yongli, SHI Xiangtian, *et al*. Progress in the application of multimodal ultrasound and ultrasound radiomics in predicting molecular subtypes of breast cancer[J]. **Inner Mongolia Medical Journal**, 2024, 56(10): 1232-1235.

[9] BAYKARA ULUSAN M, FERRARA F, MELTEM E, *et al*. MRI-only breast cancers are less aggressive than cancers identifiable on conventional imaging[J]. **Eur J Radiol, 2024**, 181: 111781.

[10] XIE Y, CHEN H, TIAN M, *et al*. Integrating multi-omics and machine learning survival frameworks to build a prognostic model based on immune function and cell death patterns in a lung adenocarcinoma cohort[J]. **Front Immunol**, 2024, 15: 1460547.

[11] PEROZ M, MANANET H, ROUSSOT N, *et al*. Clinical Interest in exome-based analysis of somatic mutational signatures for non-small cell lung cancer[J]. **Cancers (Basel)**, 2024, 16(17):3115.

[12] FAN Y, KAO C, YANG F, *et al*. Integrated multi-omics analysis model to identify biomarkers associated with prognosis of breast cancer[J]. **Front Oncol**, 2022, 12: 899900.

[13] 刘斯洋，林星辰，程丝，等 . 多组学大数据与医学发展 [J]. **科技导报**，2024，42（12）： 51-74.
LIU Siyang, LIN Xingchen, CHENG Si, *et al*. Multi-omics big data and medical development[J]. **Science & Technology Review**, 2024, 42(12): 51-74.

[14] XIE M, KUANG Y, SONG M, *et al*. Subtype-MGTP: a cancer subtype identification framework based on multi-omics translation[J]. **Bioinformatics**, 2024, 40(6).

[15] SHI T, YE X, HUANG D, *et al*. Cancer subtype identification by multi-omics clustering based on interpretable feature and latent subspace learning[J]. **Methods**, 2024, 231: 144-153.

[16] 章子怡，王棨临，张俊有，等 . 多组学数据驱动的机器学习模型在乳腺癌生存及治疗响应预测中的应用 [J]. **遗传**，2024，46（10）：820-832.
ZHANG Ziyi, WANG Xilin, ZHANG Junyou, *et al*. Machine learning applications in breast cancer survival and therapeutic outcome prediction based on multi-omic analysis[J]. **Hereditas**, 2024, 46(10): 820-832.

[17] GE S, LIU J, CHENG Y, *et al*. Multi-view spectral clustering with latent representation learning for applications on multi-omics cancer subtyping[J]. **Brief Bioinform**, 2023, 24(1): bbac500.

[18] 钟雅婷，林艳梅，陈定甲，等 . 多组学数据整合分析和应用研究综述 [J]. **计算机工程与应用**，2021，57（23）：1-17.
ZHONG Yating, LIN Yanmei, CHEN Dingjia, *et al*. Review on integration analysis and application of multi-omics data[J]. **Computer Engineering and Applications**, 2021, 57(23): 1-17.

[19] RAPPOPORT N, SHAMIR R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark[J]. **Nucleic Acids Res**, 2018, 46(20): 10546-10562.

[20] LI C, WU N, LIN X, *et al*. Integrated transcriptomic and immunological profiling reveals new diagnostic and prognostic models for cutaneous melanoma[J]. **Front Pharmacol**, 2024, 15: 1389550.

[21] RATHER A A, CHACHOO M A. Robust correlation estimation and UMAP assisted topological analysis of omics data for disease subtyping[J]. **Comput Biol Med**, 2023, 155: 106640.

[22] LIU C, FANG J, KANG W, *et al*. Identification of novel potential homologous repair deficiency-associated genes in pancreatic adenocarcinoma via WGCNA coexpression network analysis and machine learning[J]. **Cell Cycle**, 2023, 22(21-22): 2392-408.

[23] YADAV S, ZHOU S, HE B, *et al*. Deep learning and transfer learning identify breast cancer survival subtypes from single-cell imaging data[J]. **Commun Med (Lond)**, 2023, 3(1): 187.

[24] CHEN C, WEISS S T, LIU Y Y. Graph convolutional network-based feature selection for high-dimensional and low-sample size data[J]. **Bioinformatics**, 2023, 39(4):9730247.

[25] RAMAZZOTTI D, LAL A, WANG B, *et al*. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival[J]. **Nat Commun**, 2018, 9(1): 4453.

[26] GUSEV A, LEE S H, Trynka G, *et al*. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases[J]. **Am J Hum Genet**, 2014, 95(5): 535-552.

[27] 罗琴琴，雷锦志 . 单细胞转录组数据分析中的数学 [J/OL]. **生物信息学**，2024:1-37.
LUO Qinqin, LEI Jinzhi. Mathematics in single-cell RNA transcript ome analysis[J/OL]. **Bioinformatics**, 2024: 1-37.

[28] WEI Y, LI L, ZHAO X, *et al*. Cancer subtyping with heterogeneous multi-omics data via hierarchical multi-kernel learning[J]. **Brief Bioinform**, 2023, 24(1):bbac448.

[29] WANG B, MEZLINI A M, DEMIR F, *et al*. Similarity network fusion for aggregating data types on a genomic scale[J]. **Nat Methods**, 2014, 11(3): 333-337.

[30] XU T, LE T D, LIU L, *et al*. CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization[J]. **Bioinformatics**, 2017, 33(19): 3131-3133.

[31] SESSLER T, QUINN G P, WAPPETT M, *et al*. SurviveR: a flexible shiny application for patient survival analysis[J]. **Sci Rep, 2023**, 13(1): 22093.

[32] LOVE M I, HUBER W, ANDERS S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2[J]. **Genome Biol**, 2014, 15(12): 550.