论著·智能医学与预测模型

EGWO-GAT:一种用于膀胱尿路上皮癌分期诊断的图注意力网络与增强型灰狼优化算法多组学整合模型

郭泓麟,秦茂洋,宋秋月,陈 欣,伍亚舟 陆军军医大学(第三军医大学)军事预防医学系军队卫生统计学教研室,重庆

［摘要］ **目的** 本研究提出了一种基于图注意力网络(graph attention network,GAT)与增强型灰狼优化算法(enhanced-grey wolf optimizer, EGWO)的多组学整合模型EGWO-GAT,实现对膀胱尿路上皮癌(bladder urothelial carcinoma, BLCA)肿瘤样本的分期预测。**方法** 基于在加州大学圣克鲁斯分校Xena功能基因组学探索器(University of California, Santa Cruz Xena,UCSC Xena)网站收集的404例癌症基因组图谱(The Cancer Genome Atlas, TCGA)BLCA样本,包含mRNA、DNA甲基化和微小RNA(microRNA,miRNA)数据,将3种组学数据分别进行预处理和差异分析之后得到其节点特征与边特征,以GAT为基础,引入EGWO进行超参数优化,采用多层感知机(multilayer perceptron,MLP)进行后续癌症分期预测。经5折交叉验证分析,将本研究创建的EGWO-GAT模型与多种经典机器学习分期模型的性能进行对比,并进行组学贡献分析和保留不同相似边条数的模型性能比较,采用准确率、精确率、召回率、F1分数及曲线下面积(area under the curve, AUC)作为性能评估的核心指标。**结果** 差异特征筛选结果显示,mRNA组学获得534个差异基因,DNA甲基化组学获得3 108个差异探针,miRNA组学获得114个差异miRNA。多模型对比结果表明,当整合所有组学数据类型且保留相似性排名前3的其他患者作为边时,EGWO-GAT模型性能最佳,其AUC值达到0.744,准确率达到0.711,精确率达到0.792,召回率达到0.782,F1分数达到0.785,其综合分类性能显著优于其余经典机器学习方法,且较GS-GAT模型在各项指标上均有明显提升。组学贡献分析显示,全组学(mRNA+DNA甲基化+ miRNA)整合的性能显著优于其他6种组学组合方式。相似性边数性能比较结果表明,保留前3条相似边时模型的AUC、准确率、精确率及F1分数均高于保留前5条或7条边的情况,综合性能最优。**结论** 本研究构建的EGWO-GAT多组学整合模型在BLCA分期中性能优异,可为精准分期提供技术支撑,解决因样本异质性引发的临床分期难题,对辅助个体化治疗及改善患者预后意义重大。

［关键词］ 膀胱尿路上皮癌;图注意力网络;增强型灰狼优化算法;多组学
［中图法分类号］ R319;R394.3;R737.14 ［文献标志码］ A

EGWO-GAT: A graph attention network and enhanced grey wolf optimizer-based multi-omics integration model for bladder urothelial carcinoma staging diagnosis

GUO Honglin, QIN Maoyang, SONG Qiuyue, CHEN Xin, WU Yazhou Department of Military Health Statistics, College of Military Preventive Medicine, Army Medical University (Third Military Medical University), Chongqing, China

［**Abstract**］ Objective This study proposes a multi-omics integration model, EGWO-GAT, based on graph attention network (GAT) and enhanced-grey wolf optimizer (EGWO), for staging prediction of bladder urothelial carcinoma (BLCA) samples. Methods A total of 404 BLCA samples from The Cancer Genome

［基金项目］ 国家自然科学基金面上项目(82173621,82574207)
［通信作者］ 伍亚舟,E-mail: asiawu@tmmu.edu.cn

2097-0927/Copyright © 2026, 陆军军医大学(第三军医大学)。这篇开放获取文章遵循 CC BY 许可协议 (https://creativecommons.org/licenses/by/4.0/)
引用格式:郭泓麟,秦茂洋,宋秋月,等. EGWO-GAT:一种用于膀胱尿路上皮癌分期诊断的图注意力网络与增强型灰狼优化算法多组学整合模型[J]. 陆军军医大学学报,2026,48(03):366-377. DOI:10.16016/j.2097-0927.202511119.

Citation:GUO H L, QIN M Y, SONG Q Y, *et al*. EGWO-GAT: A graph attention network and enhanced grey wolf optimizer-based multi-omics integration model for bladder urothelial carcinoma staging diagnosis[J]. J Army Med Univ,2026,48(03):366-377. DOI:10.16016/j.2097-0927.202511119.

Atlas (TCGA) were collected from the University of California, Santa Cruz Xena (UCSC Xena) functional genomics explorer, including mRNA, DNA methylation, and microRNA (miRNA) data. After separate preprocessing and differential analysis of the 3 omics datasets, node features and edge features were extracted. Using GAT as the backbone framework, EGWO was employed for hyperparameter optimization, and multilayer perceptron (MLP) was applied for cancer staging prediction. Through 5-fold cross-validation, the performance of EGWO-GAT was compared with classical machine learning models, along with omics contribution analysis and performance conparison of models retaining different numbers of similar edges. Accuracy, precision, recall, F1 score, and area under the curve (AUC) were used as core evaluation metrics. **Results** Differential feature screening identified 534 differentially expressed genes in mRNA, 3 108 differential probes in DNA methylation, and 114 differential miRNAs. Model comparisons showed that EGWO-GAT achieved optimal performance when integrating all omics data and retaining edges from the top 3 similar patients, with an AUC value of 0.744, an accuracy of 0.711, a precision of 0.792, a recall of 0.782, and an F1 score of 0.785. Its performance significantly surpassed other classical methods and outperformed the GS-GAT model across all metrics. Omics contribution analysis confirmed that full integration (mRNA+DNA methylation+miRNA) outperformed all other combinations. Result of the similarity edge number performance comparison demonstrated that retaining the top 3 similar edges yielded the highest AUC, accuracy, precision, and F1 score compared to the top 5 or 7 edges. **Conclusion** The EGWO-GAT model exhibits excellent performance in BLCA staging, providing reliable technical support for precise diagnosis. It addresses clinical challenges caused by sample heterogeneity and holds significant potential for guiding individualized treatment and improving patient prognosis.

[**Key words**] bladder urothelial carcinoma; graph attention network; enhanced-grey wolf optimizer; multi-omics

Supported by the General Program of the National Natural Science Foundation of China (82173621, 82574207). Corresponding author: WU Yazhou, E-mail: asiawu@tmmu.edu.cn

膀胱癌(bladder cancer,BCa)是全球公认的一种高发恶性肿瘤,作为泌尿系统常见的恶性肿瘤之一[1],据统计每年约有55万例新增确诊病例,且近年来其发病率呈现逐年上升的趋势。全球统计数据表明在男性群体新确诊的各类癌症中,BCa的发病例数排在第4位,病死率已上升至第8位,且患者年龄多集中于50~70岁[2]。膀胱尿路上皮癌(bladder urothelial carcinoma, BLCA)作为绝对主导类型,占BCa的90%以上;对于BLCA,其早晚期有显著的差异,美国癌症联合委员会(American Joint Committee on Cancer, AJCC)癌症分期手册作为全球肿瘤分期的金标准,明确将Stage 1—2定义为早期,Stage 3—4定义为晚期,这种差异涉及肿瘤的进展、治疗方案的选择和患者的预后[3-4]。因此,对于患者而言,明确BLCA分期是制定个性化治疗方案和评估预后的关键依据。随着高通量测序技术在近年的不断发展,癌症基因组图谱(The Cancer Genome Atlas, TCGA)已成为助力科研人员系统解析癌症发生机制、进展规律及耐药性分子基础的重要工具[5],其包含了许多癌症的多组学数据。多组学数据通过揭示肿瘤分子特征与生物学行为的关联,为癌症分期提供了新维度[6],可辅助病理学进行癌症分期,尤其在BLCA中展现出显著潜力[7]。

多组学数据的整合,对于理解癌症等复杂疾病有着重要的作用[8]。但其数据本身存在高异质性、高维度与高噪声等特点,传统的机器学习难以对其进行有效的整合。而图学习模型,作为一种新兴的人工智能(artificial intelligence,AI)方法,能够把数据的高维度特征压缩成低维度并且具有生物学意义的向量,用节点数据和边数据构建图结构网络,更加高效地利用多组学数据对癌症的发生发展进行分析与解释。目前主流的图学习模型大部分基于图卷积网络(graph convolutional networks, GCN),如LI等[9]利用GCN对28种癌症进行了泛癌分类;GUO等[10]提出基于GCN关系感知能力的注意力机制图卷积网络(attention-based graph convolutional network,AGCN)多组学整合方法,整合了基因表达、拷贝数变异与DNA甲基化数据这3种组学数据,对乳腺癌进行亚型分子分类;KESIMOGLU等[11]采用7种组学数据,构建SUPREME模型,该方法将GCN嵌入特征与节点特征相结合,并采用多层感知机(multilayer perceptron,MLP)作为分类

器,进行乳腺癌亚型的分类任务。WANG等[12]以基因表达、DNA甲基化和微小RNA(microRNA,miRNA)作为输入数据,采用GCN学习组学特异性嵌入特征并通过视图相关性发现网络融合这些嵌入特征,进而对乳腺癌、脑癌及泛肾癌的癌症亚型进行分类。

然而,这些基于GCN的框架,核心局限都在于对邻居节点的权重分配具有固定性和均一性,因此无法确定邻居样本在分类任务中的相对重要性。图注意力网络(graph attention network,GAT)[13]是一类面向图结构数据设计的深度学习模型,其核心创新在于将注意力机制嵌入模型架构,使模型能够动态学习图中目标节点与其邻居节点间关联的相对重要性,进而有效突破GCN依赖固定权重开展邻居信息聚合的固有局限,最终实现对图数据中蕴含的结构特征更精准地捕获,对于整合多组学进行分类任务有着较好的效果[14-15]。

需重点关注的是,GAT模型的性能表现不仅与模型本身构架有关,还与学习率、隐藏层维度、注意力头数等关键超参数的配置合理性相关,这些参数直接决定了生成的节点嵌入质量,进而影响下游分类任务的最终精度。对于大部分传统图学习模型,都采用网格搜索进行超参数的选择,但其存在计算成本高昂,容易陷入局部最优解等问题。针对这一问题,元启发式优化算法的出现为超参数选择提供了新的途径。灰狼优化算法(grey wolf optimizer, GWO)[16]作为元启发式算法的其中一类,通过模拟灰狼的社会等级结构与协作捕猎行为构建生物启发机制,既解决了传统优化方法中存在的盲目性问题,又降低了其容易陷入局部最优困境的风险,在机器学习模型的超参数优化任务中取得了显著的效果。

因此,本研究基于BLCA mRNA、DNA甲基化和miRNA的多组学数据,创建基于GAT框架的增强型灰狼优化算法(enhanced-grey wolf optimizer, EGWO)-GAT模型,创新性地采用EGWO对GAT嵌入生成过程中的超参数以及后续MLP分类器的核心超参数进行优化,最终对BLCA进行分期预测。

1 材料与方法

1.1 数据收集与预处理

本研究数据选取了加州大学圣克鲁斯分校Xena功能基因组学探索器(University of California, Santa Cruz Xena,UCSC Xena)网站的TCGA[17]组学数据,包含mRNA、DNA甲基化和miRNA 3种组学数据。mRNA数据作为基因转录的直接产物,是遗传信息从DNA向蛋白质传递的核心中间载体,其表达特征可直接调控细胞增殖、分化、凋亡及侵袭转移等肿瘤关键生物学过程[18]。DNA甲基化数据是表观遗传调控的核心方式,主要通过在基因启动子区域添加甲基基团(5-甲基胞嘧啶)抑制基因转录,无需改变DNA序列即可实现基因表达的长期稳定沉默[19]。miRNA数据是一类长度约22 nt的内源性非编码RNA,通过与靶向mRNA的3′非编码区(3′ untranslated region,3′UTR)互补结合,介导mRNA降解或抑制其翻译,实现转录后调控[20]。3种组学数据均只保留包含肿瘤组织的样本,排除重复样本与正常样本。根据临床信息,将TNM分期为Ⅰ/Ⅱ期的样本划分为早期样本,Ⅲ/Ⅳ期的样本划分为晚期样本[21]。最终将3种组学数据与临床信息的样本进行取交集,保留同时具有3种组学信息且有分期标签的样本共404例,其中早期样本132例,晚期样本272例。对于3种组学数据预处理流程图与各组学样本数据如图1所示。

对于mRNA数据,TCGA的原始数据包含60 660个基因的表达信息,从中分离出19 962个mRNA的表达数据。对其进行低表达基因过滤,仅保留在70%以上样本中表达量不为0的基因,剔除在多数样本中无表达的噪声基因,减少后续分析干扰,得到16 426个特征。使用DESeq2对数据进行早晚期差异分析,筛选阈值为调整后$P\leqslant 0.01$和$|\log_2^{(倍数变化)}|\geqslant 1$(确保生物学意义)。对于DNA甲基化数据,因其存在一定的缺失值,为避免缺失值对分析的干扰,先对数据进行过滤,仅保留缺失率≤20%的探针共403 434个,再使用KNN插补对剩余缺失值进行插补。后续使用limma对数据进行早晚期差异分析,筛选阈值为同时满足调整后$P\leqslant 0.01$(统计学意义显著)、$|\Delta\beta|\geqslant 0.1$(生物学意义显著)。

对于miRNA数据,过滤低表达miRNA,计算RPM≥1的样本数占总样本数的比例,仅保留比例≥30%的miRNA,保留414个miRNA特征。接下来使用limma-voom进行差异分析,仅保留调整后$P\leqslant 0.01$的miRNA。最终对3种组学数据的差异分析结果分别绘制火山图并进行Z-score标准化,为后续分析做准备。

1.2 EGWO-GAT模型构建

本研究所构建的EGWO-GAT框架图如图2所示,完整呈现EGWO-GAT模型从多组学数据输入到BLCA早晚期分期预测的全流程,核心围绕“建立多组学特征矩阵,构建患者相似性网络,生成GAT动态嵌入,实现下游分期任务”的逻辑。首先输入经预处理的3种组学标准化特征矩阵,每种组学数据均计算患者间分子特征相似性,并保留相似性排名前3的边构建患者相似

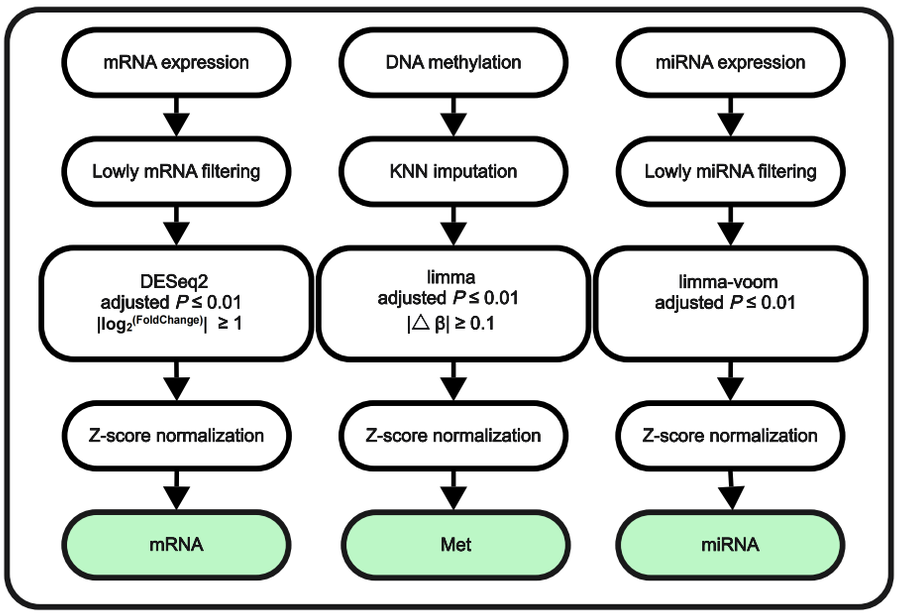


图1 数据预处理流程图

Figure 1 Flowchart of data preprocessing

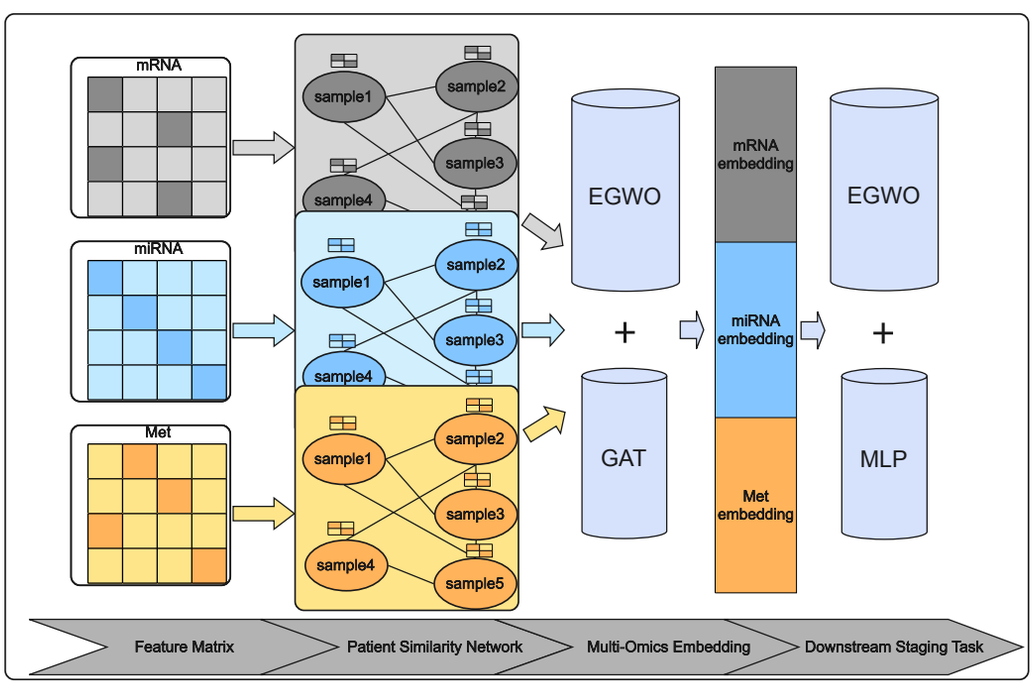


图2 EGWO-GAT框架图

Figure 2 EGWO-GAT framework diagram

性网络;随后各网络分别输入GAT以学习组学特异性节点嵌入,同时通过EGWO对GAT的学习率、隐藏层维度等关键超参数进行优化,保障嵌入对组学结构与特征关联的捕捉能力;再将3个组学的GAT嵌入横向拼接形成多组学融合特征矩阵,输入MLP执行下游分期任务,EGWO进一步优化MLP超参数并结合早停机制避免过拟合,最终输出患者早晚期分期结果,实现多组学信息的高效整合与智能优化驱动的精准分期预测。

1.2.1 增强型灰狼优化算法 本研究使用EGWO在GAT嵌入生成及下游早晚期分期任务2个阶段进行超参数寻优。对于模型的复杂性,基础的GWO在局部搜索精度与全局探索广度这2个方面还存在一定的局限性。因此,本研究构建EGWO算法,对于基准模型的局部搜索性能方面,在个体位置更新后,为部

分个体引入柯西噪声以打破局部最优陷阱;对于全局探索广度方面,增加探索狼机制,保留部分不依赖α/β/δ狼引导的独立个体,让它们基于历史最优解开展随机搜索或邻域搜索,以此避免群体过早收敛,达到拓宽全局搜索范围的目的。其具体步骤阐述如下。

首先初始化参数与狼群的定位,设定搜索空间维度、种群规模、最大迭代次数、探索狼比例、柯西噪声尺度、各维度搜索范围等参数,同时对狼群位置进行随机初始化生成,初始位置满足:

$$\text{lower}_d \leqslant X_{i,d}(0) \leqslant \text{upper}_d \quad (1)$$

其中lower_d是第d维搜索空间的下界,upper_d则是第d维的上界,$X_{i,d}(0)$代表第i只狼在第d维的初始位置,i是狼群中个体的编号,d对应具体超参数类型,0表示迭代初始状态(第0代);整个搜索过程需覆盖所有维度,即从d=1到总维度D的所有超参数要完成初始定位。

其次根据探索狼比例划分探索狼与开发狼的数量。开发狼基于当前最优解进行局部开发,探索狼则开展完全随机搜索,其位置更新服从均匀分布:

$$X_i^{\text{explore}}(t+1) \sim \text{Uniform}(\text{lower}_d, \text{upper}_d), \forall d \in [1, D] \quad (2)$$

其中$X_i^{\text{explore}}(t+1)$是第i只探索狼在第t+1代的位置;$\text{Uniform}(\text{lower}_d, \text{upper}_d)$表示在第d维的上下界之间进行均匀随机采样,确保搜索覆盖整个超参数取值范围;t是当前迭代次数,从0开始逐代递增;$\forall d \in [1, D]$意味着探索狼的位置更新需覆盖所有超参数维度。

然后更新开发狼位置,计算收敛因子:a从2衰减至0,平衡全局探索与局部精修,迭代初期搜索范围广,后期收窄。

$$a = 2 - t \times \frac{2}{T_{\max}} \quad (3)$$

开发狼位置更新:筛选出当前最优的α狼、次优的β狼和第三优的δ狼,开发狼位置通过3个公式向三者靠拢,公式中r_1到r_6是[0,1]区间的随机数,保证搜索多样性,$X_{i,d}(t)$为开发狼当前位置,最终得到3个引导后的候选位置。

$$X_{i,d}^{\alpha}(t+1) = \alpha_d(t) - (2ar_1 - a) \times |2r_2\alpha_d(t) - X_{i,d}(t)| \quad (4)$$

$$X_{i,d}^{\beta}(t+1) = \beta_d(t) - (2ar_3 - a) \times |2r_4\beta_d(t) - X_{i,d}(t)| \quad (5)$$

$$X_{i,d}^{\delta}(t+1) = \delta_d(t) - (2ar_5 - a) \times |2r_6\delta_d(t) - X_{i,d}(t)| \quad (6)$$

临时位置:开发狼的最终局部搜索位置需基于α/β/δ狼的引导位置取平均,形成临时位置。其中$X_{i,d}(t+1)$是第i只开发狼在第d维、第t+1代的临时位置,分子是受3只狼引导的3个位置之和,分母3是为了平衡3个引导方向的权重,避免单一最优解的偏见。再加入柯西噪声打破局部最优,既实现小幅精修,也保留跳出局部最优的可能。

$$\overline{X}_{i,d}(t+1) = \frac{X_{i,d}^{\alpha} + X_{i,d}^{\beta} + X_{i,d}^{\delta}}{3} \quad (7)$$

接着评估适应度,以模型的训练结果值作为适应度,按适应度对狼群进行排序,保留当前代中适应度最优的3个解,并更新其位置与适应度值,为下一轮迭代提供引导。

最后进行迭代终止与结果输出,当达到最大迭代次数时,迭代终止,返回最优解(即α狼的最终位置),该最优解对应模型的最佳超参数组合与最优模型性能。

1.2.2 患者相似性网络构建 GAT是图神经网络,需输入“节点特征矩阵”和“边结构信息”才能学习嵌入。对于经过数据预处理之后的节点特征矩阵,每1行代表1个节点,每1列代表1个组学特征。对于边结构信息,首先要构建患者相似性网络[22-23],作为GAT模型的核心输入基础,其目标是将多组学预处理后的患者样本转化为图结构,每个节点代表1例BLCA患者,如果2名患者在某一个组学数据上的相似性高于设定的阈值,则这2个节点存在边,边的权重大小由相似性计算结果决定。针对每种多组学数据,生成一个患者相似性网络。对于mRNA、DNA甲基化和miRNA这3个组学,其核心特征均为连续型数值,故采用Pearson相关系数衡量2个患者的同一组学特征向量间的线性相关程度,相关系数越接近1或-1,代表患者间该组学特征的相似性越高,因此适合此类数据的相似性计算。对于每1个组学,计算所有患者两两之间的相关性系数,都会生成1个404×404的相似性矩阵。为了避免网络过于复杂,且确保每条边的生物学意义,对于每1个节点,仅保留前3相似的其他患者作为边,最终每个网络的边数约为404×3=1 212。

1.2.3 GAT模型训练与嵌入生成 GAT嵌入的质量高度依赖模型超参数(如学习率、隐藏层维度),本研究采用EGWO实现超参数的自动化搜索。对于GAT超参数的选择以及寻优范围分别为学习率:[0.000 1, 0.01]、隐藏层维度:[32, 64]、训练轮次:[100, 300]以及注意力头数:[1, 2],其中注意力头数直接影响模型对节点间关联重要性的学习能力。通过25个狼个体的40次迭代搜索,以GAT验证损失的中位数为适应度函数,结合探索组(15%个体)的随机搜索与开发组(85%个体)的α/β/δ狼引导(含柯西噪声扰动),输出适配各组学网络数据特性的最优超参数组合。

随后构建2层GAT网络架构,核心是高效提取对癌症分期有用的关键信息,第一层通过指定数量的注意力头数,将输入特征映射至预设的隐藏层维度,其输出即为包含节点局部结构与特征关联的嵌入向量,

该嵌入会经过专门处理后再输入第二层网络,最终转化为“早期”或“晚期”的分类信号;训练时,模型会自动优化参数,还设置早停机制,保留表现最好的版本;最后提取第一层的核心信息,作为后续分期判断的关键依据。该流程通过2层网络的动态注意力机制(利用边权重自适应学习邻居节点的贡献度)与EGWO算法的超参数优化,突破了传统图模型依赖固定权重聚合邻居信息的局限,在保证嵌入生成效率的同时,提升了其对图结构与节点特征的融合能力。

1.2.4 基于EGWO优化的GAT嵌入下游分类方法

对于GAT模型生成的3个组学的节点嵌入,与对应组学的原始标准化特征进行横向维度拼接,形成最终输入特征矩阵。为避免过拟合误判与类别分布失衡,采用3层分层抽样划分策略,固定全局随机种子确保可复现,第一层将全量节点按8∶2比例划分为全量训练集与独立测试集,采用分层抽样保证2组数据的类别比例与全量数据一致。第二层则在全量训练集内采用5折无重复交叉验证,每折划分为子训练集与子验证集,进一步验证模型泛化能力。第三层将第二层得到的单折子训练集按4折交叉验证划分,生成4组超参优化训练集(用于超参数试训)和超参优化验证集(用于超参性能评估),为EGWO超参数搜索提供内部验证基准。

本研究采用MLP作为分类模型[24],对于多组学数据的分类任务,其能通过多层非线性变换捕捉特征间的复杂关系,契合多组学数据的复杂特性。对于早晚期分期任务,同样采用EGWO进行超参数的全局寻优,其核心机制与GAT嵌入生成阶段保持一致,但为了进一步保障搜索的深度与有效性,设置狼群数量为10、迭代次数为50。MLP超参数的的选择以及寻优范围分别为学习率:[0.000 1,0.1]、隐藏层维度:[16,32,64,128]、L2正则化系数:[0.001,0.1]。对于离散参数采用取整处理,并匹配有效值范围,而对于连续参数则直接进行优化。在EGWO超参数优化阶段,采用5折交叉验证框架,每一折的子训练集都会被进一步划分为4组超参优化训练集与1组超参优化验证集,以超参优化验证集的曲线下面积(area under the curve, AUC)值作为适应度函数,寻找最适合模型的最优超参数组合。在模型训练阶段,基于最优超参数组合再次初始化MLP模型,在当前折的子训练集上开展训练,通过早停机制避免过拟合现象的出现。

1.3 模型训练与测试

按8∶2分层抽样划分404例样本(维持早晚期比例),得到训练验证集(323例)与独立测试集(81例),训练验证集内实施5折交叉验证。

模型训练分2步:①GAT嵌入训练,基于3种组学数据构建患者相似性网络,输入2层GAT;用Adam优化器、BCEWithLogitsLoss计算损失,反向传播更新参数,结合早停机制避免过拟合。②MLP分类训练,拼接3组学GAT嵌入为融合特征,输入3层MLP;沿用Adam优化器与早停机制,确保模型学习多组学特征间的复杂关联。在模型优化阶段,采用EGWO对超参数进行智能搜索:针对GAT,搜索学习率、隐藏层维度、训练轮次、注意力头数4类超参数;针对MLP,搜索学习率、隐藏层维度、L2正则化系数3类超参数。在模型测试阶段,先通过5折交叉验证评估模型稳定性,再用独立测试集验证泛化能力。

本研究通过RStudio软件完成数据的预处理工作,在模型框架构建方面,采用Python软件中的PyTorch深度学习框架,使用Scikit-learn工具划分数据,计算各项性能指标。最终完成EGWO-GAT模型的构建与训练。

1.4 分类性能比较

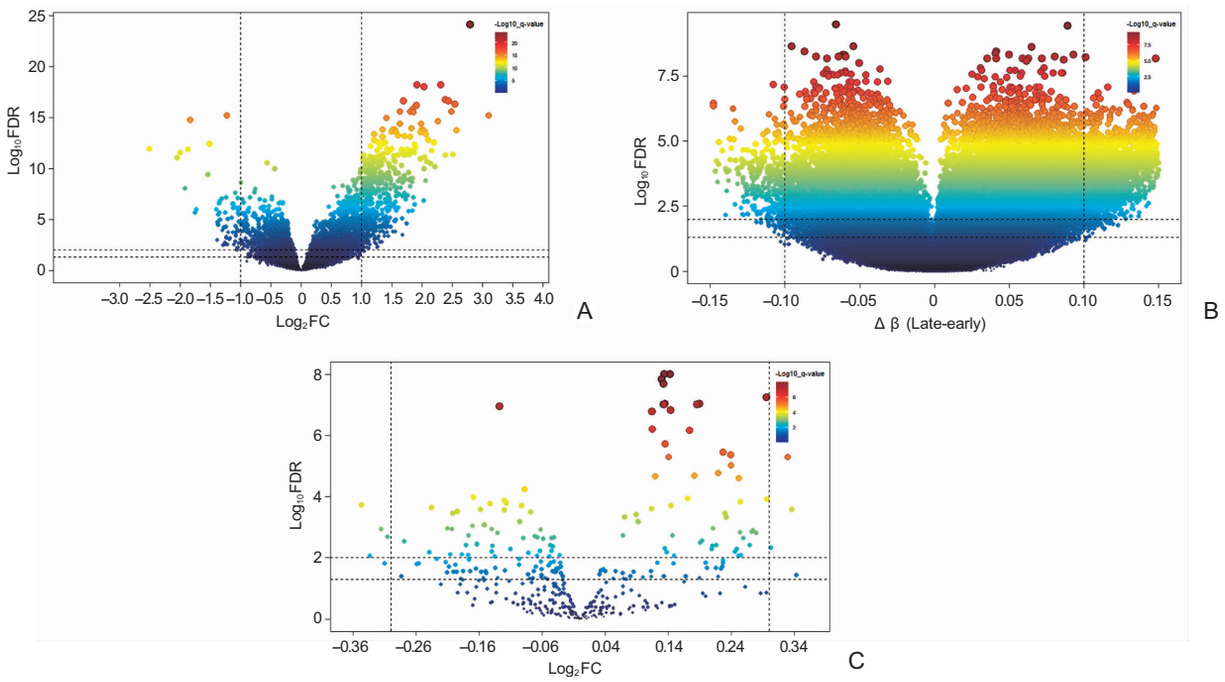
性能评估环节,针对二分类任务,本研究采用准确率、精确率、召回率、F1分数、AUC值5类核心指标评估性能,在独立测试集上进行性能评估,结果以$\bar{x}\pm s$形式统计,确保评估结果的无偏性,全面反映模型的稳健性与有效性。

为比较多种分类模型的性能,将本研究构建的EGWO-GAT模型,与Logistic-Regression、KNN、Naive-Bayes、AdaBoost、Decision-Tree(CART)等多种经典机器学习分类模型及采用网格搜索优化的GAT模型(GS-GAT)进行性能对比;此外,本研究还探讨了每种组学数据类型及其组合在BLCA分期中的贡献;在患者相似性网络构建阶段,对于每一个节点,保留相似边的条数也会决定模型最终的分期性能。因此,本研究还比较了分别保留3、5与7条边的情况对于最终模型性能的影响。

2 结果

2.1 差异特征筛选结果

对于mRNA、DNA甲基化与miRNA数据的特征分别进行差异分析并绘制火山图如图3所示。mRNA组学得到534个差异基因,包含上调基因471个,下调基因63个;DNA甲基化组学得到3 108个差异探针,包含上调探针2 780个,下调探针328个;miRNA组学得到114个差异miRNA,包含上调miRNA 61个,下调miRNA 53个。



A: mRNA; B: DNA methylation; C: miRNA

图3 多组学分析中差异表达的火山图

Figure 3 Volcano plots of differential expression in multi-omics analyses

2.2 多模型分类性能比较

如表1所示,EGWO-GAT模型在AUC、准确率和精确率3个指标上都取得了最高的性能,在召回率与F1分数2个指标上与最佳值0.811与0.790不相上下;对比单纯采用网格搜索的GS-GAT模型,EGWO-GAT模型在各个评估指标上都有较大优势。本研究还对7种模型绘制了混淆矩阵以及ROC曲线如图4所示。混淆矩阵内百分比为各类别样本内部占比(真实晚期样本中被误判为早期/正确预测为晚期的比例,或真实早期样本中被误判为晚期/正确预测为早期的比例)。

2.3 组学特异性预测贡献

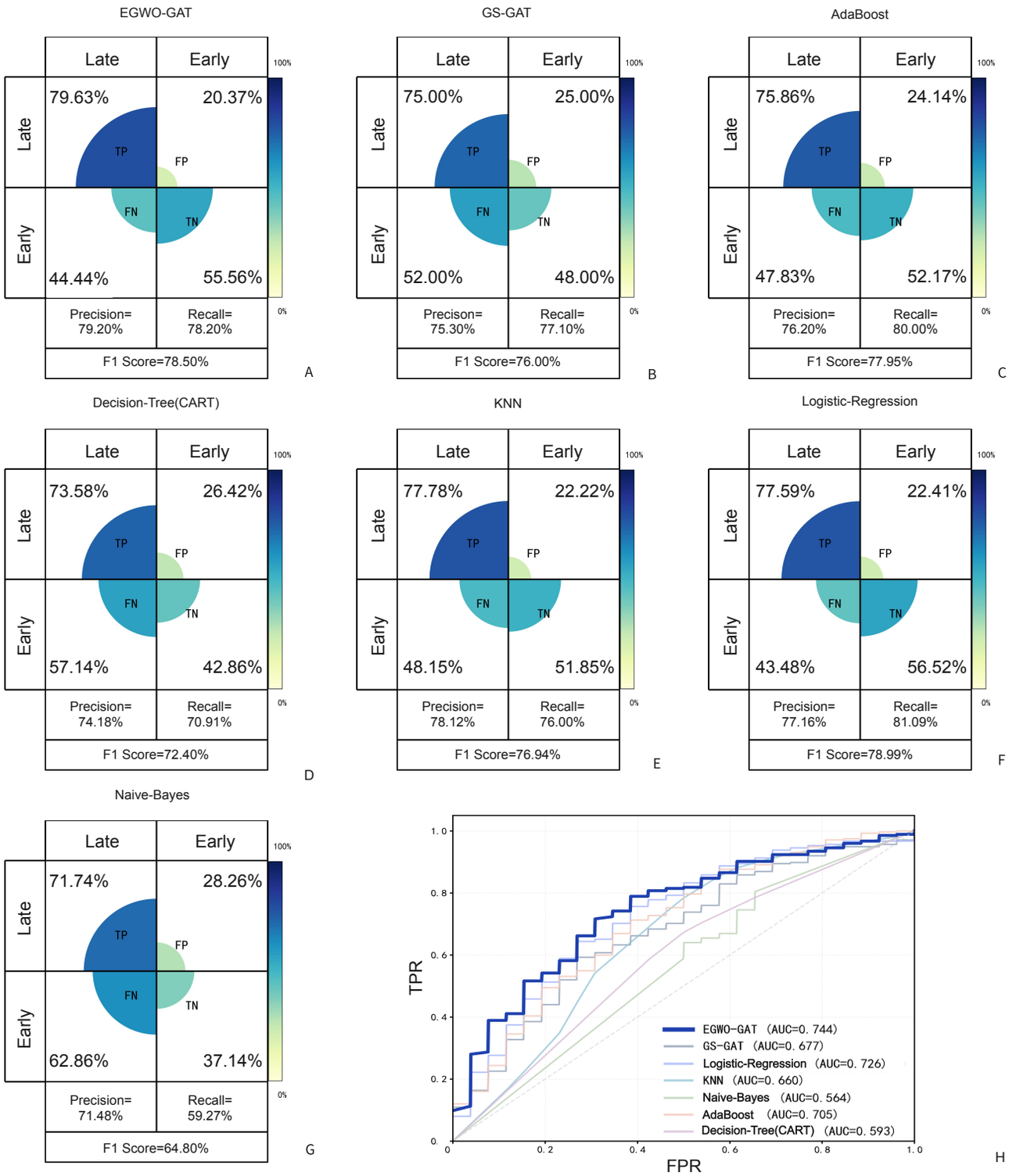
每种组学数据类型及其组合在BLCA分期中的贡献,其各项指标性能如表2所示,并绘制ROC曲线如图5所示。针对3种不同的组学数据类型,设计了7种组合:包括单组学、2种组学组合以及全部3种组学组合。单组学是所有组合的基础,其性能直接反映了该类组学数据本身携带的BLCA分期信息密度,在单组学分期模型中,miRNA单组学数据在AUC、准确率、精确率、F1分数4项核心指标上均为最优,是单组学中对分期贡献最核心的基础数据。在2种组学组合中,MET+miRNA的组合取得了较好的性能指标。而当EGWO-GAT使用所有组学数据类型时的性能,显著优

表1 各模型分期性能展示($\bar{x}\pm s$)

Table 1 Comparative performance of different models in disease staging ($\bar{x}\pm s$)

Models	AUC	Accuracy	Precision	Recall	F1 score
Logistic-Regression	0.726±0.014	0.709±0.041	0.772±0.021	0.811±0.065	0.790±0.037
KNN	0.660±0.037	0.691±0.025	0.781±0.028	0.760±0.047	0.769±0.022
Naive-Bayes	0.564±0.004	0.563±0.011	0.715±0.005	0.593±0.016	0.648±0.012
AdaBoost	0.705±0.015	0.694±0.027	0.762±0.019	0.800±0.056	0.780±0.026
Decision-Tree(CART)	0.593±0.046	0.635±0.045	0.742±0.031	0.709±0.063	0.724±0.040
GS-GAT	0.677±0.022	0.672±0.041	0.753±0.028	0.771±0.077	0.760±0.038
EGWO-GAT	0.744±0.027	0.711±0.019	0.792±0.027	0.782±0.062	0.785±0.022

KNN: K-nearest neighbor; CART: Classification and regression tree; GS-GAT: Grid search-graph attention network; EGWO-GAT: Enhanced-grey wolf optimizer based-graph attention network



A~G: Confusion matrices of EGWO-GAT, GS-GAT, AdaBoost, Decision-Tree (CART), KNN, Logistic-Regression, and Naive-Bayes; H: Comparison of ROC curves between EGWO-GAT and other models

图 4 各模型的混淆矩阵以及 ROC 曲线

Figure 4 Confusion matrices and ROC curves of the models

表2 各组学组合分期性能展示($\bar{x}\pm s$)

Table 2 Staging performance of different omics combination ($\bar{x}\pm s$)

Omics combination	AUC	Accuracy	Precision	Recall	F1 score
mRNA	0.688±0.029	0.674±0.051	0.764±0.038	0.756±0.082	0.758±0.048
MET	0.681±0.078	0.679±0.042	0.776±0.055	0.756±0.113	0.760±0.043
miRNA	0.714±0.066	0.684±0.045	0.781±0.038	0.745±0.073	0.761±0.041
mRNA+MET	0.672±0.075	0.659±0.062	0.764±0.015	0.720±0.114	0.738±0.065
mRNA+miRNA	0.691±0.027	0.644±0.064	0.780±0.030	0.669±0.143	0.712±0.087
MET+miRNA	0.693±0.044	0.689±0.044	0.766±0.031	0.782±0.07	0.772±0.039
mRNA+MET+miRNA	0.744±0.027	0.711±0.019	0.792±0.027	0.782±0.062	0.785±0.022

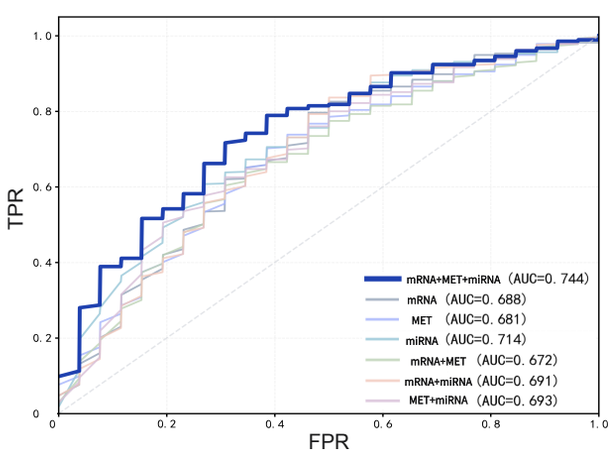


图5 不同组学数据组合用于BLCA分期的ROC曲线

Figure 5 ROC curves of BLCA staging performance among different omics data combinations

于其他6种组合方式,说明每一种组学数据对于模型整体的贡献都是必不可少的。

2.4 不同相似性边数性能比较

对于保留不同相似边对最终模型性能的影响并绘制极坐标条形图如图6所示。该极坐标条形图以极轴半径对应性能指标的数值大小,以径向分区区分3种边数设置(前3、前5、前7),每个边数分区内的5个条形分别对应5项评估指标,且通过颜色或标签明确区分。条形越向外扩,代表该指标的数值越高,可以看到保留前3条边数时的模型在AUC、准确率、精确率、F1数值等4项指标均为最高,说明其综合性能最优。

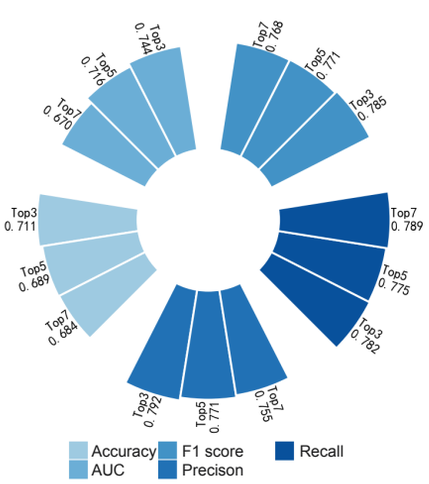


图6 性能极坐标条形图

Figure 6 Performance metrics polar bar chart

3 讨论

3.1 EGWO-GAT模型的核心创新与技术优势

BLCA是全球范围内高发的恶性肿瘤,其早晚期分期直接关系到治疗方案的制定与患者预后情况。而如何高效整合癌症多组学数据并进行下游任务是一个严峻的挑战,传统图学习模型在多组学数据整合及下游任务处理中表现出良好的适用性,但仍存在一定局限。本研究的核心创新在于GAT动态注意力嵌入+EGWO增强型智能优化的全流程协同设计,突破了传统多组学整合模型的两大瓶颈:针对主流的GCN不能充分关注样本间关联的相对重要性的缺陷,本研究采用GAT作为基础模型,能够有效捕捉样本间的相互关联,弥补GCN的不足。同时在此基础上,针对图模型普遍存在的超参数优化难题,创新性引入EGWO,在传统GWO狼群捕食机制的基础上,新增探索狼机制和柯西噪声扰动,找到了全局探索与局部开发能力之间的平衡点,这一改进经实验验证(对比GS-GAT模型),使AUC从0.677提升至0.744,F1分数从0.760提升至0.785,显著优于单纯使用网格搜索的GAT框架。EGWO并非孤立作用于GAT或MLP,而是深度融入模型构建的全流程,GAT生成的高维嵌入依赖最优超参数捕捉多组学之间的关联,MLP的分类性能依赖嵌入特征的质量,二者形成嵌入质量到分类精度的正向循环,这是本模型综合性能超越AdaBoost、Logistic-Regression等经典模型的关键。

3.2 多组学差异特征的生物学机制解读

从火山图结果可见,3种组学数据的差异特征呈现明显区别,且各自对应明确的肿瘤进展机制。对于DNA甲基化数据,上调位点(即晚期样本甲基化水平高于早期)占整体差异位点的90%,其核心生物学意义在于其导致抑癌基因沉默,这是驱动肿瘤从早期进展到晚期的关键表观遗传机制[19]。而对于mRNA数据,差异基因同样以上调为主,推测其病理意义是通过激活促癌生物学通路(如细胞周期、侵袭相关通路)推动肿瘤发展,也是早晚期差异的核心功能信号[18]。miRNA数据的上下调分布相对均衡(上调61个,下调53个),这种均衡性可能源于miRNA中的致癌miRNA与抑癌miRNA的协同作用,二者通过精准调控肿瘤从早期到晚期的进展速度,避免单一调控导致的肿瘤过度恶性化或调控失效[20]。

3.3 组学数据组合的分期贡献分析

组学贡献分析结果揭示了不同组学数据及组合对分期预测的作用差异,其性能差异的核心源于各组学数据独特的生理生化机制及信息互补性。miRNA单组学在3组单组学中的综合分期性能最优,这一优势源于其精准的转录后调控机制与BLCA分期的强关联性。miRNA作为转录后调控分子,能够调控肿瘤侵袭转移[25-26],其表达异常直接关联BLCA的上皮间质转化过程[27],是分期相关的关键分子标志物。mRNA单组学的贡献基于基因转录后的功能表型呈现。mRNA作为遗传信息从DNA到蛋白质的关键中间载体,其表达差异直接体现肿瘤细胞的功能状态。但mRNA易受转录层面的瞬时调控影响,单一维度难以全面覆盖肿瘤进展的分子机制,故性能略逊于miRNA。DNA甲基化单组学(AUC=0.681)的核心机制是表观遗传沉默驱动肿瘤进展。DNA甲基化的调控具有上游性,需通过影响mRNA表达间接作用于肿瘤表型,单一维度的信息传导存在局限,故单组学性能稍弱。两种组学组合中,MET+miRNA的组合性能相对突出,说明表观遗传调控与转录后调控具有互补性,二者协同增强了模型对分期相关特征的覆盖性。而全组学整合(mRNA+DNA甲基化+miRNA)性能在所有组学组合中表现最优,其生物学逻辑在于:mRNA反映基因表达状态、DNA甲基化体现代谢异常[28]、miRNA参与转录后调控,三者分别对应BLCA分期的基因、代谢、调控三个不同病理维度,全组学组合能同时整合这3个维度的信息,覆盖单组学或两组学组合无法捕捉的跨维度分期特征,实现最好的分期性能。

3.4 模型的临床转化价值与应用场景

从临床转化角度看,AJCC区分早晚期主要基于3大临床特征的显著差异:肿瘤浸润深度、治疗策略和预后结果。这样的差异主要是肿瘤从局限于膀胱壁内(可通过手术根治)到突破膀胱壁(需综合治疗)的转变。但在实际临床场景中,部分病例因病理切片质量、肿瘤的基因异质性等问题导致分期模糊,影响治疗方案的精准制定。而EGWO-GAT的预测结果可作为病理学分期的重要补充,这种分子特征+临床分期的联合策略,可为临床决策提供更可靠的依据。其临床转化价值可通过以下具体场景深度落地,对诊疗全程进行优化:①术前精准分期,优化新辅助治疗决策,通过多组学分子特征精准区分真正早期与隐匿进展期BLCA,避免过度手术或治疗不足,提升保膀胱率与治疗响应率。②基层医院标准化诊断,缩小诊疗差异:可将EGWO-GAT封装为轻量化工具,适配基层医疗资源有限场景,规避病理医师经验依赖,仅需输入核心组学数据即可输出标准化分期结果。③疑难病例双轨验证,提升会诊一致性,针对肿瘤异质性、切片质量不佳导致的分期争议,构建病理形态与分子表型的联合评估体系,为多学科会诊提供量化参考。

3.5 多组学整合的科学性论证

组学选择的合理性:本研究选取mRNA、DNA甲基化与miRNA 3种组学,是基于BLCA发生发展的分子机制,三者分别覆盖基因、代谢、调控核心维度,符合多组学互补整合原则,避免单一组学信息偏差。整合方式的优越性:采用GAT嵌入+横向拼接的整合策略,相对于仅采用GAT嵌入的核心优势,是同时保留样本关联特征与组学原始特征,实现结构关联+分子特性的双重信息互补,避免单一嵌入导致的信息损耗,最终提升分期预测的准确性与稳健性。相似性网络构建的严谨性:选择Pearson相关系数构建患者网络,是因为这3种组学数据均为连续型数值,该系数能有效衡量分子特征的线性关联[22];而保留前3相似边的设计,不仅基于实验验证(显示前3时综合指标最优),更符合GAT模型的核心需求,即仅纳入分子特征高度一致的邻居节点,减少低相似性样本带来的噪声干扰,使注意力机制能聚焦于分期相关的关键关联,避免GCN模型均一化权重导致的有效信息稀释。

本研究构建的EGWO-GAT模型也存在一定的局限性:①样本量较少,样本仅来源于TCGA数据库的404例BLCA样本,且为单中心数据;②疾病种类仅集中于BLCA,其他癌种泛化性不足;③EGWO的探索狼

比例与柯西噪声尺度等参数为经验设置,未进行多参数组合的敏感性分析。

综上所述,本研究创建的EGWO-GAT多组学融合模型,通过GAT动态注意力机制突破传统图模型的权重固定局限,结合EGWO实现超参数高效优化,证实多组学整合+智能优化策略在BLCA分期中的优势。未来通过数据拓展、模型深化与癌种延伸,EGWO-GAT有望成为癌症精准分期的重要辅助手段,助力癌症诊疗向个性化、分子化方向迈进。

利益冲突 本文所有作者声明不存在利益冲突

参考文献:

[1] 王海涛, 罗伟铭, 陈剑, 等. 中高危非肌层浸润性膀胱癌卡介苗灌注治疗后复发的临床预测模型构建与验证[J]. 陆军军医大学学报, 2025, 47(9): 959-968. DOI: 10.16016/j.2097-0927.202501061.
WANG H T, LUO W M, CHEN J, *et al*. Development and validation of clinical prediction model for post-treatment recurrence in high-risk non-muscle invasive bladder cancer after BCG intravesical instillation[J]. J Army Med Univ, 2025, 47(9): 959-968. DOI: 10.16016/j.2097-0927.202501061.

[2] HUANG X, DU G T, YANG Y, *et al*. Advancing bladder cancer management: development of a prognostic model and personalized therapy [J]. Front Immunol, 2024, 15: 1430792. DOI:10.3389/fimmu.2024.1430792.

[3] ABDEL-RAHMAN O. Validation of the eighth AJCC new substages for bladder cancer among different staging contexts [J]. Clin Genitourin Cancer, 2017, 15(6): e1095-e1106. DOI:10.1016/j.clgc.2017.07.025.

[4] AMIN M B, GREENE F L, EDGE S B, *et al*. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging[J]. CA Cancer J Clin, 2017, 67(2): 93-99. DOI:10.3322/caac.21388.

[5] LINEHAN W M, RICKETTS C J. The Cancer Genome Atlas of renal cell carcinoma: findings and clinical implications[J]. Nat Rev Urol, 2019, 16(9): 539-552. DOI:10.1038/s41585-019-0211-5.

[6] DONG Y Y, YANG W K, WANG J W, *et al*. MLW-gcForest: a multi-weighted gcForest model towards the staging of lung adenocarcinoma based on multi-modal genetic data[J]. BMC Bioinformatics, 2019, 20(1): 578. DOI: 10.1186/s12859-019-3172-z.

[7] DAI H J, YU Z J, ZHAO Y, *et al*. Integrating machine learning models with multi-omics analysis to decipher the prognostic significance of mitotic catastrophe heterogeneity in bladder cancer[J]. Biol Direct, 2025, 20(1): 56. DOI: 10.1186/s13062-025-00650-x.

[8] SUBRAMANIAN I, VERMA S, KUMAR S, *et al*. Multi-omics data integration, interpretation, and its application[J]. Bioinform Biol Insights, 2020, 14: 1177932219899051. DOI:10.1177/1177932219899051.

[9] LI B J, WANG T Y, NABAVI S. Cancer molecular subtype classification by graph convolutional networks on multi-omics data [C]//Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Gainesvillem, FL, USA: 2021: 1-9. DOI:10.1145/3459930.3469542.

[10] GUO H, LV X, LI Y Z, *et al*. Attention-based GCN integrates multi-omics data for breast cancer subtype classification and patient-specific gene marker identification [J]. Brief Funct Genomics, 2023, 22(5): 463-474. DOI: 10.1093/bfgp/elad013.

[11] KESIMOGLU Z N, BOZDAG S. SUPREME: multiomics data integration using graph convolutional networks [J]. NAR Genom Bioinform, 2023, 5(2): lqad063. DOI: 10.1093/nargab/lqad063.

[12] WANG T X, SHAO W, HUANG Z, *et al*. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification[J]. Nat Commun, 2021, 12(1): 3445. DOI: 10.1038/s41467-021-23774-w.

[13] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, *et al*. Graph attention networks [EB/OL]. 2017: arXiv: 1710.10903. https://arxiv.org/abs/1710.10903

[14] TANVIR R B, ISLAM M M, SOBHAN M, *et al*. MOGAT: a multi-omics integration framework using graph attention networks for cancer subtype prediction[J]. Int J Mol Sci, 2024, 25(5): 2788. DOI:10.3390/ijms25052788.

[15] ZHAO W Y, GU X, CHEN S Q, *et al*. MODIG integrating multi-omics and multi-dimensional gene network for cancer driver gene identification based on graph attention network model [J]. Bioinformatics, 2022, 38(21): 4901-4907. DOI:10.1093/bioinformatics/btac622.

[16] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimizer [J]. Adv Eng Softw, 2014, 69: 46-61. DOI: 10.1016/j.advengsoft.2013.12.007.

[17] TOMCZAK K, CZERWIŃSKA P, WIZNEROWICZ M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge[J]. Contemp Oncol, 2015, 19(1A): A68-A77. DOI:10.5114/wo.2014.47136.

[18] XU N, YAO Z M, SHANG G G, *et al*. Integrated proteogenomic characterization of urothelial carcinoma of the bladder[J]. J Hematol Oncol, 2022, 15: 76. DOI:10.1186/

s13045-022-01291-7.

[19] ZHANG C J, SHEN K J, ZHENG Y X, *et al*. Genome-wide screening of abberant methylated drivers combined with relative risk loci in bladder cancer[J]. Cancer Med, 2020, 9(2): 768-782. DOI:10.1002/cam4.2665.

[20] STEMPOR P A, AVNI D, LEIBOWITZ R, *et al*. Comprehensive analysis of correlations in the expression of miRNA genes and immune checkpoint genes in bladder cancer cells[J]. Int J Mol Sci, 2021, 22(5): 2553. DOI: 10.3390/ijms22052553.

[21] 李佳芮, 钱力, 沈俊杰, 等. 基于转录组学和变分自编码器的癌症分期诊断研究[J]. 陆军军医大学学报, 2025, 47(6): 613-622. DOI:10.16016/j.2097-0927.202412079.
LI J R, QIAN L, SHEN J J, *et al*. Cancer staging diagnosis based on transcriptomics and variational autoencoder[J]. J Army Med Univ, 2025, 47(6): 613-622. DOI:10.16016/j.2097-0927.202412079.

[22] 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. DOI: 10.1038/nmeth.2810.

[23] PAI S, BADER G D. Patient similarity networks for precision medicine[J]. J Mol Biol, 2018, 430(18): 2924-2938. DOI:10.1016/j.jmb.2018.05.037.

[24] POPESCU M, BALAS V E, PERESCU-POPESCU L, *et al*. Multilayer perceptron and neural networks[J]. WSEAS Trans Circuits Syst Arch, 2009, 8: 579-588. DOI:10.5555/1639537.1639542.

[25] ZHANG Z, WANG C, LIU T T, *et al*. MiRNA-182-5p promotes human bladder cancer proliferation and migration through the FOXF2/SHH axis[J]. Neoplasma, 2022, 69(2): 321-330. DOI:10.4149/neo_2021_210903N1266.

[26] ZOU Z H, CHENG Q, ZHANG S C, *et al*. MiR-205-3p suppresses bladder cancer progression *via* GLO1 mediated P38/ERK activation[J]. BMC Cancer, 2023, 23: 956. DOI: 10.1186/s12885-023-11175-9.

[27] NARWAL A, KUMARI K, KAUSHAL S, *et al*. The study of miRNA-200c expression and epithelial-to-mesenchymal transition-related transcription factors in the primary bladder urothelial carcinoma[J]. Urol Ann, 2023, 15(1): 35-42. DOI:10.4103/ua.ua_72_22.

[28] ZAGHLOOL S B, MOOK-KANAMORI D O, KADER S, *et al*. Deep molecular phenotypes link complex disorders and physiological insult to CpG methylation[J]. Hum Mol Genet, 2018, 27(6): 1106-1121. DOI: 10.1093/hmg/ddy006.

(**收稿**:2025-11-27;**修回**:2026-01-12)

(编辑 邓强庭)