文章编号:1674-2869(2024)02-0190-07

基于深度学习和多组学数据的肺腺癌分期预测研究

刘德真,李圆媛*
武汉工程大学光电信息与能源工程学院、数理学院,湖北 武汉 430205

摘 要:为解决癌症分期难以精准决策这一问题,对452例肺腺癌患者的信使核糖核酸(mRNA)转录数据、微核糖核酸(miRNA)转录数据和DNA甲基化3种组学数据进行集成融合,并采用随机森林算法进行分期预测。首先对从癌症基因组图谱(TCGA)数据库获取的3种组学数据进行预处理,将mRNA转录数据和DNA甲基化数据进行基因位点匹配,再使用4种不同的多组学集成策略对预处理后的组学数据进行集成,最后使用随机森林算法对集成后的数据进行分期预测并使用准确度、卡帕系数以及曲线下面积(AUC)作为预测效果的评价指标。研究结果显示,采用多组学集成策略在分期预测上具有更高的准确率,其中基于深度学习的集成策略的预测效果最好,评价指标分别为0.940、0.931和0.986,有希望应用于未来的肺腺癌分期预测中。

关键词:肺腺癌分期;深度学习;集成策略;随机森林算法
中图分类号:TP18;Q811.4 **文献标识码**:A **DOI**:10.19843/j.cnki.CN42-1779/TQ.202307022

Stage prediction of lung adenocarcinoma based on deep learning and multi-omics data

*LIU Dezhen, LI Yuanyuan**
School of Optical Information and Energy Engineering, School of Mathematics and Physics,
Wuhan Institute of Technology, Wuhan 430205, China

Abstract: To improve accuracy in decision-making in cancer staging, this study integrated three kinds of omics data, including messenger ribonucleic acid (mRNA) transcript data, micro ribonucleic acid (miRNA) transcript data and DNA methylation, from 452 lung adenocarcinoma patients, and used random forest algorithm to predict stages. First, three kinds of omics data obtained from the cancer genome altas (TCGA) database were preprocessed and the mRNA sequencing data were matched up with DNA methylation data at gene loci, then four different multi-omics integration strategies were adopted to integrate the preprocessed data, and finally a random forest algorithm was applied to the integrated data for the prediction of staging, and accuracy, Kappa coefficient and the area under the curve (AUC) were used to evaluate the performance of the prediction. The results show that adoption of the multi-omics integration strategies can achieve high accuracy. The integration strategy based on deep learning is considered as the most effective one, with accuracy, Kappa coefficient and AUC values of 0.940, 0.931 and 0.986, respectively, and it can offer relevant guidance for the lung adenocarcinoma staging prediction in the future.

Keywords: staging of lung adenocarcinoma; deep learning; integration strategy; random forest algorithm

收稿日期:2023-07-24
基金项目:国家自然科学基金(12001408)
作者简介:刘德真,硕士研究生。Email:1205125262@qq.com
***通信作者**:李圆媛,博士,教授。Email:yuanyuanli_wit@hotmail.com
引文格式:刘德真,李圆媛. 基于深度学习和多组学数据的肺腺癌分期预测研究[J]. 武汉工程大学学报,2024,46(2):190-196.

癌症是一种身体细胞生长、变化和繁殖失控的疾病，由于其高死亡率，科学家对其进行广泛的研究。据2018年的数据显示，全球有1 810万人被新确诊为癌症患者，而960万人在当年因癌症而丧生[1]。肺腺癌是癌症的一种，属于非小细胞肺癌的一种亚型，占所有肺癌的40%至55%。近几十年来，中国的肺腺癌发病率呈快速增长趋势[2]。此外，肺腺癌同样是一种复发性疾病：超过30%的患者即使接受手术治疗并进入晚期分期阶段后仍然面临着复发风险，并最终死于复发[3]。

传统的癌症分期诊断系统主要依据肿块大小，以及是否存在远程转移等宏观因素来确定临床分期。但这种方法会忽略微观因素，如肿块分子特征和免疫组织学特征等重要信息，这可能会导致早期癌变被忽视，从而错过最佳治愈时机。

为解决上述传统癌症分期系统的弊端，应从宏观因素转移，聚焦到能为癌症提供微观生物学特征的基因表达数据。研究表明，基因表达数据在分析癌症亚型分类、基因治疗和药物研发等领域有广泛应用[4]。通过对基因表达数据进行分析，可以发现与癌症发生、发展和转移相关的分子信号通路和生物标志物。这些生物标志物有助于评估癌症预后，并制定个体化治疗方案。Mallick等[5]通过设计一个5层深度神经网络分类器，对72名白血病患者的骨髓基因表达数据进行分析，对两类白血病进行分类，获得了98.2%的准确率；Xu等[6]对来自癌症基因组图谱（the cancer genome atlas，TCGA）与基因表达综合数据库的基因数据进行分析，建立基于基因表达数据的肺腺癌遗传标记，促进个体化治疗并为免疫治疗带来新分子标记的新方向；Naeem等[7]利用机器学习技术，通过比较转移性和局部前列腺癌的基因表达数据，识别与前列腺癌转移发展相关的差异表达基因和分子通路，从而确定潜在的生物标志物。

当前第三代测序技术的出现为治疗学和个性化医学开辟了新时代的道路[8]。对于多组学，可以通过表观基因组学加强基因表达数据中的模糊模式[9]，多组学间的互补信息可用于更好地解释分类结果[10]，提高预测性能[11]，并理解基因间复杂的分子路径。对多组学的研究通常包括不同类型、规模和分布的数据，并且只有很少样本，此外生物数据集是复杂且嘈杂的，因此正确地整合这些数据十分重要。如果集成不正确，即使添加更多组学信息，也无法显著提高模型性能，并且还会增加问题的复杂度。因此，挖掘相关信息并将不同组学数据正确地整合起来至关重要。

本文利用多组学集成策略来预测肺腺癌患者分期的方法，以3种不同的组学数据作为研究对象，应用随机森林模型，预测不同集成策略下的分期结果，并比较不同集成策略之间的优劣。同时，还将其与单一组学数据使用随机森林预测模型得到的结果进行对比。最终，找到了基于基因多组学数据最佳肺腺癌分期预测模型。

1　实验部分

本文选取452例癌症患者作为研究样本，使用信使核糖核酸（messenger ribonucleic acid，mRNA）、微核糖核酸（micro ribonucleic acid，miRNA）转录数据以及DNA甲基化数据作为组学数据进行实验。

首先，从TCGA数据库[12]中获取了相关的452例患者肺腺癌的3种组学数据，随后对这3种类型的数据进行预处理操作，并使用4种不同的集成策略对数据进行集成融合。将融合后的数据集过采样并进行最大最小值归一化处理，最终使用归一化数据构建随机森林模型来得到分期预测结果。

1.1　数据获取

TCGA数据库是由美国国家癌症研究所和国家人类基因组研究所合作建立的癌症研究项目，其涵盖了基因组、转录组、表观遗传、蛋白组等多个组学数据，提供了全方位、多维度的数据。本文使用TCGA数据库中的肺腺癌数据集，包含452例患者，不同分期的样本数量分布如表1所示。该数据集能够从官方提供的对应下载工具Genomic Data Commons Data Portal[13]网站（https://portal.gdc.cancer.gov/）中在线获取。

表1　癌症分期样例分布

Tab. 1　Distribution of cancer staging samples　例

样本	一期	二期	三期	四期
452	248	20	111	73

1.2　数据预处理

数据质量因素，如噪声、空值和数据类别不平衡，会直接影响算法的预测精度。因此需要对3种组学数据进行预处理。

首先，对每种组学数据进行去重处理，即对于重复的特征值取平均值，以确保每个特征具有唯一性。

由于基因转录组测序精度等因素的影响，某些基因可能会出现零表达或弱表达现象。这种现象可能会受到基因变异或者测序环境的影响。为保证数据健壮性，需要对低表达或弱表达的数据

进行列剔除处理。具体的操作为：如果某一列特征中零表达或低表达数据占总数据20%以上，则需要将该列特征剔除。

基因表达数据通常是一个大矩阵，记录了不同实验条件下基因表达水平，并且经常有一些值缺失的情况。出现缺失值的原因可能是分辨率不足、图像损坏或者幻灯片上存在灰尘及划痕等。缺失值的存在会导致一些鲁棒性较差的基因表达分析算法失去有效性。mRNA和miRNA两种组学数据存在大量的缺失值，这些缺失值会影响后续的多组学集成分析。

因此，需要对2种组学数据进行缺失值的填补。将每个样本视为一次观测结果，将包含在该样本中的不同组学数据视为全部特征。可以利用包含对应特征不含空值的多个样本，针对含缺失值的样本进行缺失值填补。使用K近邻（k-nearestneighbor，KNN）缺失值填补算法[14]对mRNA和miRNA两种组学数据进行缺失值填补。具体来说，在处理含有缺失值的样本时，根据与其他样本的距离，选取距离最接近的K个观测样本，并将它们之间的距离进行加权平均，作为含有缺失值样本的填充数。在含有缺失值的情况下，通过忽略缺失值，并放大非缺失值的权重来计算欧几里得距离，即对于含缺失值样本$\boldsymbol{x}$与不含缺失值的样本$\boldsymbol{y}$，有：

$$\boldsymbol{x}=(x_1,x_2,\cdots,x_i,M,\cdots,x_n) \tag{1}$$

$$\boldsymbol{y}=(y_1,y_2,\cdots,y_i,\cdots,y_n) \tag{2}$$

其中，$\boldsymbol{x}$、$\boldsymbol{y}$均为1行n列的矩阵，n表示样本$\boldsymbol{x}$与$\boldsymbol{y}$的特征总数，M表示缺失值，两个样本之间的距离定义为：

$$d=\sqrt{\frac{n}{m}}\times\sqrt{\sum_{x_i,y_i\neq M}(x_i-y_i)^2} \tag{3}$$

其中x_i、y_i分别表示$\boldsymbol{x}$、$\boldsymbol{y}$中的第i个特征，m表示样本$\boldsymbol{x}$中非缺失值的特征总数，d表示样本$\boldsymbol{x}$与$\boldsymbol{y}$之间忽略缺失值并放大非缺失坐标的权重计算所得的欧几里得距离。

经过数据预处理操作后，组学数据集中的零表达和弱表达特征已被移除，同时miRNA和mRNA组学数据中的空值已完全填补。

1.3 mRNA、DNA甲基化数据的基因位点匹配

mRNA是通过DNA编码区域进行转录的，其测序数据可以反映基因表达情况，从而推断基因在生物学过程中的修饰情况。DNA甲基化是一种影响基因组表观修饰的方式，可控制基因表达。样本的DNA甲基化数据通过illumina 450k甲基化芯片上的探针获取，每个探针可采集CpG位点的甲基化β值。CpG位点与基因之间存在关联，图1展示了包含TSS1 500、TSS200、5′ UTR、1stexon、Gene body和3′ UTR等6个部分的基因示意图以及CpG位点分类。其中，对于基因部分，TSS代表转录起始位点，UTR代表非翻译区域，1stexon代表第1个外显子，Gene body代表编码区域，3′ UTR则表示3′非翻译区域。对于CpG位点部分，N Shelf与S Shelf都具有较高的甲基化水平，N Shore和S Shore具有的甲基化水平较低但同样产生甲基化，上述位点能与基因转录区域TSS1 500、TSS200和编码区域Gene body对应起来。与基因中非翻译区域及外显子所对应的CpG island区域通常无甲基化。

由图1可以确定CpG位点与基因转录区域存在对应关系。因此，在将mRNA转录数据与DNA甲基化数据进行匹配时，可以基于相同的基因位点深入分析2种组学数据之间的关联性，并降低数据维度。

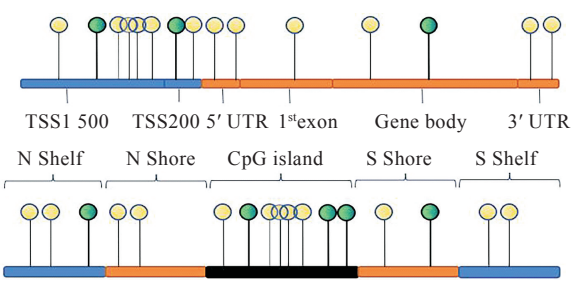


图1 CpG位点与基因的关系

Fig. 1 Relationship between CpG locus and gene

为实现数据匹配，从illumina官网上获取450k甲基化芯片的注释信息，并提取探针ID对应的基因位点信息，然后将其与mRNA转录数据中的特征（即基因）进行匹配，经1.2节的预处理操作及1.3节的基因位点匹配操作后，3种组学数据的特征数量如表2所示。在处理注释信息时需要遵循以下规则：（1）若探针点无对应的基因，则应删除该探针点。（2）若探针点对应多个基因，则选择最后一个基因作为对应的基因。

表2 组学数据特征数量

Tab. 2 Number of characteristics of omics data

组学数据	特征数量/个	
	原始数据	处理后
mRNA	60 660	14 019
miRNA	1 881	431
DNA甲基化	485 577	14 019

1.4　组学数据集成策略

利用miRNA、mRNA和DNA甲基化3种组学数据对肺腺癌患者进行分析。由于这3种数据具有相同的类型、规模和分布，因此直接预测肺腺癌分期，会难以挖掘不同组学信息之间的相关性。为此，采用4种集成策略（级联集成、基于深度学习的集成、基于核的集成和基于图的集成）挖掘组学信息间的相关性，并将经过归一化处理的3种组学数据进行融合，同时降低特征维度以挖掘不同组学之间的相关性，构建新样本训练数据。

1.4.1　级联集成　多组学的级联集成策略（基于连接）是将每个组学数据集拼接为1个大矩阵。该过程使所观察的样本量保持不变的同时增加变量的数目。这就导致矩阵更加复杂、噪声过多且维度过高，使得学习变得困难。并且不同组学数据之间的大小差异也可能导致模型学习不平衡，花费更多的时间去学习变量更多的组学，忽略了变量较少的组学数据。尽管如此，早期集成因其简单、易于实现，仍然被广泛使用。

在使用级联集成方法之前，本文对每个组学数据集分别进行主成分分析（principal component analysis，PCA）[15]、t-分布随机邻域嵌入（t-distributed stochastic neighbor embedding，t-SNE）[16]以及统一流形逼近与投影（uniform manifold approximation and projection，UMAP）[17]等降维处理。这样可以减少数据的维度，提高学习效果。对降维后的组学数据进行级联集成时，可以减少数据的噪声和复杂度，从而提高模型的准确性。

1.4.2　基于深度学习的集成　基于深度学习的集成方法是一种混合集成策略。该方法使用人工神经网络（artificial neural network，ANN），其由分层的许多神经元构成，是一种性能优越的机器学习模型[18]。ANN可以直接用于串联组学，也可以单独应用于每个组学。通过在单独的网络层中处理不同的组学数据来学习有意义的潜在特征表示。潜在特征表示可以看作是由模型的不同网络层所学到的新变量，这些变量能轻松连接到其他机器学习模型，以进行后续下游分析。

神经网络的隐藏层被视为连续的特征提取层。一些神经网络模型专门用于学习相关潜在特征表示，例如自动编码器（autoencoder，AE）[19]和受限玻尔兹曼机[20]。这两种无监督的神经网络架构能够从中心层或瓶颈层进行编码压缩，再现原始数据。中心层或瓶颈层的神经元数目越少，其潜在特征表示越紧凑，提取出的新特征维度越低。

本文使用AE作为基于深度学习的集成算法，AE是一种试图还原其原始输入的系统，其模型如图2所示。该模型将神经网络的隐层看作连续的特征提取层，并将中间的瓶颈层提取出来作为多组学数据融合的结果。

在模型选取中，使用AE和去噪自编码器（denoising autoencoder，DAE）[21]2种模型。在处理数据输入小于0的部分时，使用LeakyReLU作为激活函数，并使用均方误差（mean squared error，MSE）作为损失函数L来衡量模型的重构误差，其中L为：

$$L = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2 \tag{4}$$

其中n是输入的总特征维数，x_i是输入特征，x_i'是重构特征。

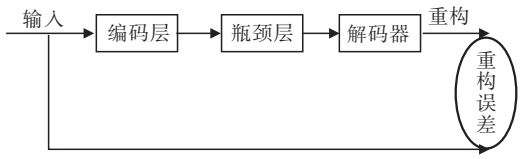


图2　自动编码器模型

Fig. 2　Automatic encoder model

1.4.3　基于核的集成　基于核的集成方法采用核函数对高维空间中的数据进行隐式操作，以捕捉观测样本之间的线性相关性。核模型采用相似度矩阵的形式来推断样本的相似性度量。使用不同的核函数（线性核、多项式核、径向核函数等）能够为相同的数据提供不同的相似性度量。故应根据组学的类型及其数据分布，通过交叉验证的方式选择正确的核函数来推断样本的相似性度量。

本文采用基于核的2种集成方法：核主成分分析（kernel principal component analysis，KPCA）[22]和核典型相关分析（kernel canonical correlation analysis，KCCA）[23]。在这些方法中，使用线性核、多项式核和径向核函数来处理数据。如表3所示，通过十折交叉验证实验，选择线性核函数对多组学数据进行基于核的集成融合，以达到更好的效果。在KPCA方法中，使用核函数将数据非线性映射到新的低维空间中；而在KCCA方法中，则利用线性核函数衡量不同组学数据之间的相关性，并进一步对其进行集成。这种基于核的集成方法可以有效降低高维度数据集，并提高模型泛化能力。

1.4.4　基于图的集成　基于图的混合策略是一种将不同组学数据进行整合集成以提高分析性能的方法。其基本思想是将每个组学数据视为单独的图，并通过将这些图融合在一起或建立多层（多通

表3 核函数选择

Tab. 3 Kernel function selection

KPCA	十折交叉验证准确率										平均准确率
线性核	0.909	0.902	0.927	0.889	0.901	0.903	0.929	0.910	0.926	0.893	0.909
径向核	0.826	0.856	0.833	0.799	0.837	0.850	0.819	0.878	0.836	0.830	0.836
多项式核	0.896	0.888	0.880	0.883	0.906	0.888	0.890	0.898	0.839	0.888	0.886
KCCA	十折交叉验证准确率										平均准确率
线性核	0.882	0.908	0.907	0.893	0.840	0.923	0.888	0.928	0.889	0.929	0.899
径向核	0.834	0.801	0.797	0.832	0.872	0.845	0.831	0.868	0.861	0.802	0.834
多项式核	0.895	0.846	0.867	0.873	0.889	0.901	0.864	0.873	0.904	0.879	0.879

路)连接的网络来集成这些信息。此外,可以利用每个组学数据的图,建立一个新的多组学信息变量,而无需依赖于融合网络方式。

本文使用拉普拉斯特征映(laplacian eigenmaps,LE)算法[24]对多组学数据进行基于图的集成。该方法首先将所有组学数据合并成1个大矩阵,然后基于欧氏距离构建无向图。接着,将该无向图转换为拉普拉斯矩阵,并计算其特征向量。最终,选取前10个特征向量作为输出矩阵。

1.5 过采样

类别数量不平衡问题会严重影响后续多分类模型的训练和误差计算。在新的训练数据中,4种分期类别的比例约为12∶5∶3∶1,因此需要对低比例类别的样本集进行数量上的扩充。

本文采用朴素随机上采样方法,对多组学习集成得到的肺腺癌样本数据中少数类样本集进行过采样,以避免数据不平衡问题。最终,得到1个肺腺癌四期分期数据集,其中各个类别比例均为1∶1∶1∶1。

1.6 随机森林模型

随机森林是一种集成学习的方法,具有多个决策树组成的分类器。其通过构建多个决策树来降低过拟合和提高泛化能力,可以被用于特征选择和异常值检测。

随机森林中的每一棵决策树,都会从原始数据集中随机抽取一定数量的样本和特征,用这些样本和特征构建决策树。在决策树的构建过程中,采用随机特征选择和随机划分策略,即在每个节点上随机选择一部分特征进行划分。最后所有决策树的预测结果通过投票或者平均等方式进行集成,得出最终的预测结果。

为评估随机森林的性能,将数据集按照7∶3的比例划分为训练集和测试集。随着子决策树数量的增加,肺腺癌四期分期预测错误率呈现收敛趋势,并在子决策树数量达到800棵时达到收敛。

2 结果与讨论

本文对包含452例样本的数据进行预处理和归一化操作。针对3种单组学数据,采用t-SNE方法进行降维处理,得到了作为单组学数据对照组的降维结果。同时,采用级联集成方法(包含3种不同的降维方法)、基于深度学习的集成方法(2种不同的自编码器模型)、基于核的集成方法(2种不同的核融合方法)以及基于图的集成方法,将3种未降维的单组学数据进行融合,共计得到8份不同的融合数据。

针对3种单组学数据和8种融合后的集成数据,按照7∶3的比例将其划分为训练集和测试集。使用训练集对肺腺癌分期预测模型进行训练,并在测试集上对模型进行性能评估。最终,通过对不同的多组学数据集进行分析获得了相应的集成分类结果,并使用准确度(accuracy,ACC)、曲线下面积(area under the curve,AUC)[25]以及卡帕(Kappa)系数[26]等指标来评价基于集成方法的随机森林模型的性能。其中,ACC表示模型在测试样本中正确分类的比例。

在多分类任务中,使用多类接收者操作特征曲线下面积(multi-class receiver operating characteristic-area under the curve,mROC-AUC)。与二元分类任务不同,mROC曲线是一个多类别的分类曲线,可以同时评估多个类别的分类性能。如:在计算肺癌四期多分类AUC值时,首先将一期作为正例,其余三期作为反例,在测试集上计算随机森林模型的预测结果,并计算该类别的真阳性率(true positive rate,TPR)和假阳性率(false positive rate,FPR)。然后分别以二、三和四期作为正例来计算它们各自的TPR和FPR。最后针对每个类别绘制接收者操作特征(receiver operator characteristic,ROC)曲线并分别计算每个类别下面积的平均值,最终得到AUC值。

Kappa系数作为统计学中评估一致性的量化工具，被广泛应用于多分类模型的准确度评估。该系数通过比较观测一致性与随机一致性，提供对模型性能的综合评价。其取值范围在[−1，1]之间，取值越接近于1，模型预测结果与实际分类结果的一致性就越高。Kappa系数k的计算方法如下：

$$k = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

其中，p_o是每一类正确分类的样本数量之和除以总样本数，即总体分类精度。记总共有c个类别，再假设每一类的真实样本个数分别为$A_1, A_2, \cdots, A_c$，预测的每一类的样本个数分别为$B_1, B_2, \cdots, B_c$，则有：

$$p_e = \frac{A_1 \times B_1 + A_2 \times B_2 + \cdots + A_c \times B_c}{N \times N} \tag{6}$$

其中N为总样本个数。

由图3可知，对于单组学数据，单组学DNA甲基化数据在ACC、Kappa系数以及AUC值方面表现最佳，分别为0.836、0.785和0.968。将3种组学数据进行t-SNE级联集成后，其评价指标分别为0.846、0.794和0.971，明显优于仅使用单组学数据进行预测的效果。这表明通过使用多组学集成方法将不同的单组学数据整合在一起，可以提高肺腺癌分期预测的准确性。此外，与单组学相比，8种集成策略在不同程度上都使随机森林模型的预测准确性有所提升。其中，DAE的表现最为出色，其ACC、Kappa系数和AUC分别为0.940、0.931和0.986。

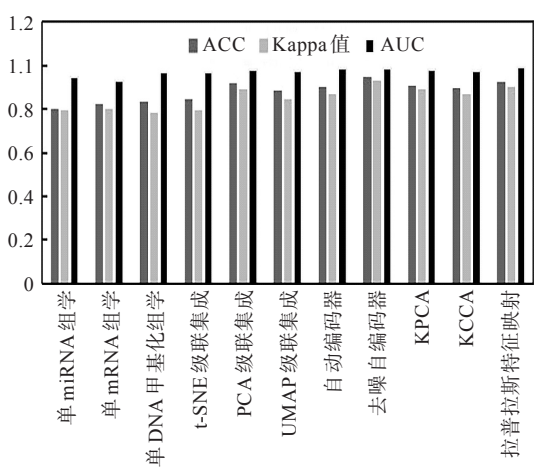


图3　单组学数据及多组学集成方法的评价指标对比（十折交叉验证）

Fig. 3　Comparison of evaluation indexes of one-group data and multi-group integration method(10 fold cross validation)

因此，将不同类型、规模和分布的基因组学数据进行集成，可以解决多组学数据集之间的一致性问题，并更好地捕捉不同组学数据之间的相互作用，从而使肺腺癌预测模型获得更高质量的数据集。综上所述，通过多组学数据集成和不同降维策略取得了较好的肺腺癌四期分期预测结果。DAE作为一种有效的集成方法展现出最佳预测能力，这为更深入地理解肺腺癌发展并进一步优化临床预测模型提供重要参考。

3　结　论

本研究首先对比了单组学和多组学数据在肺腺癌分期预测中的效果，结果显示多组学数据在预测分析上具有更好的表现。进一步地为了寻找最优的预测模型，探讨了级联集成、基于深度学习的集成、基于核方法的集成，以及基于图的集成在肺腺癌分期预测模型中的性能。对这些集成策略的综合评估表明，在肺腺癌分期预测中基于DAE的集成策略性能最优。DAE通过提取更具表现力的潜在特征表示，捕获了不同组学数据间的相互作用和变化模式，从而深入理解了肺腺癌分期的复杂性，并为预测模型提供了更可靠的输入。

未来的研究可以进一步扩展深度学习集成策略的应用范围，为复杂疾病的诊断、治疗提供更有力的支持。

参考文献

[1] LANCET T. Globocan 2018：counting the toll of cancer [J].Lancet(London，England)，2018，392(10152)：985

[2] ZHOU C C. Lung cancer molecular epidemiology in China：recent trends [J]. Translational Lung Cancer Research，2014，3(5)：270−279.

[3] GOODGAME B，VISWANATHAN A，MILLER C R，et al. A clinical model to estimate recurrence risk in resected stage I non-small cell lung cancer [J]. American Journal of Clinical Oncology，2008，31(1)：22−28.

[4] 余显学. 基于基因表达数据的癌症亚型发现双聚类方法研究[D]. 重庆：西南大学，2018.

[5] MALLICK P K，MOHAPATRA S K，CHAE G S，et al. Convergent learning-based model for leukemia classification from gene expression [J]. Personal and Ubiquitous Computing，2023，27(3)：1103−1110.

[6] XU Q，CHEN Y R. An aging-related gene signature-based model for risk stratification and prognosis prediction in lung adenocarcinoma [J]. Frontiers in Cell and Developmental Biology，2021，9：685379.

[7] NAEEM A，KHAN A H，AYUBI S U D，et al. Predicting the metastasis ability of prostate cancer using machine

learning classifiers [J]. Journal of Computing & Biomedical Informatics, 2023, 4(2): 1-7.
[8] AHMED Z. Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis[J]. Human Genomics, 2020, 14(1): 35.
[9] ZARAYENEH N, KO E, OH J H, et al. Integration of multi-omics data for integrative gene regulatory network inference [J]. International Journal of Data Mining and Bioinformatics, 2017, 18(3): 223-239.
[10] RAPPOPORT N, SAFRA R, SHAMIR R. MONET: multi-omic module discovery by omic selection [J]. PLOS Computational Biology, 2020, 16(9): 1008182.
[11] TINI G, MARCHETTI L, PRIAMI C, et al. Multi-omics integration—a comparison of unsupervised clustering methodologies [J]. Briefings in Bioinformatics, 2019, 20(4): 1269-1279.
[12] TOMCZAK K,CZERWIŃSKA P,WIZNEROWICZ M. The cancer genome atlas (TCGA): an immeasurable source of knowledge [J]. Contemporary Oncology, 2015, 19(1A): 68-77.
[13] JENSEN M A, FERRETTI V, GROSSMAN R L, et al. The NCI genomic data commons as an engine for precision medicine[J]. Blood, 2017, 130(4): 453-459.
[14] TROYANSKAYA O, CANTOR M, SHERLOCK G, et al. Missing value estimation methods for DNA microarrays[J]. Bioinformatics, 2001, 17(6): 520-525.
[15] WOLD S, ESBENSEN K, GELADI P. Principal component analysis[J]. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1/2/3): 37-52.
[16] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.
[17] BECHT E, MCINNES L, HEALY J, et al. Dimensionality reduction for visualizing single-cell data using UMAP[J]. Nature Biotechnology, 2019, 37(1): 38-44.
[18] PICARD M, SCOTT-BOYER M P, BODEIN A, et al. Integration strategies of multi-omics data for machine learning analysis [J]. Computational and Structural Biotechnology Journal, 2021, 19: 3735-3747.
[19] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.
[20] 张健, 丁世飞, 张楠, 等. 受限玻尔兹曼机研究综述[J]. 软件学报, 2019, 30(7): 2073-2090.
[21] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion [J]. Journal of Machine Learning Research, 2010, 11(12): 3371-3408.
[22] HUANG S Y, YEH Y R, EGUCHI S. Robust kernel principal component analysis [J]. Neural Computation, 2009, 21(11): 3179-3213.
[23] LAI P L, FYFE C. Kernel and nonlinear canonical correlation analysis [J]. International Journal of Neural Systems, 2000, 10(5): 365-377.
[24] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation [J]. Neural Computation, 2003, 15(6): 1373-1396.
[25] 班伟. 受试者工作特征曲线评估血清β2-MG、CEA、CA125、NSE、CYFRA21-1联合诊断早期肺癌价值[J]. 吉林医学, 2022, 43(5): 1384-1386.
[26] TANG W, HU J, ZHANG H, et al. Kappa coefficient: a popular measure of rater agreement [J]. Shanghai Archives of Psychiatry, 2015, 27(1): 62-67.

本文编辑:霍淑利

☆

(上接第189页)

[10] DONTSOV E V, PEIRCE A P. A multiscale implicit level set algorithm (ILSA) to model hydraulic fracture propagation incorporating combined viscous, toughness, and leak-off asymptotics [J]. Computer Methods in Applied Mechanics and Engineering, 2017, 313: 53-84.
[11] PEIRCE A, DETOURNAY E. Multiscale tip asymptotics for a deflating hydraulic fracture with leak-off [J]. Journal of Fluid Mechanics, 2022, 947: A17.
[12] WONG S W, GEILIKMAN M, XU G S. Interaction of multiple hydraulic fractures in horizontal wells [C]//SPE Middle East Unconventional Gas Conference and Exhibition. Richardson: SPE, 2013: 1-10.
[13] LONG G B, LIU S X, XU G S. et al. A perforation-erosion model for hydraulic-fracturing applications [J]. SPE Production & Operations, 2018, 33(4): 770-783.
[14] 龙恭博, 徐明皓, 吴浩宇, 等. 基于FrackOptima的PKN型水力裂缝数值模拟分析[J]. 武汉工程大学学报, 2023, 45(1): 101-105.
[15] 宋岩, 李卓, 姜振学, 等. 非常规油气地质研究进展与发展趋势[J]. 石油勘探与开发, 2017, 44(4): 638-648.
[16] 赵金洲, 陈曦宇, 李勇明, 等. 水平井分段多簇压裂模拟分析及射孔优化[J]. 石油勘探与开发, 2017, 44(1): 117-124.

本文编辑:霍淑利