文章编号:1671-7554(2025)08-0079-07 DOI:10.6040/j.issn.1671-7554.0.2025.0118

基于多模态数据融合的多癌种风险预测模型

李千[1],杨帆[1,2,3],薛付忠[1,2,3]
(1.山东大学齐鲁医学院公共卫生学院医学数据学系,山东 济南 250012;
2.国家健康医疗大数据研究院,山东 济南 250003;3.山东大学齐鲁医院,山东 济南 250012)

摘要:目的 基于英国生物银行15种常见癌症的数据,通过多模态数据融合的方法构建多癌种风险预测模型,探讨基因组数据与临床数据在癌症预测中的应用,旨在提高癌症早期预测的准确性,并为个性化医疗提供数据支持。**方法** 首先对数据进行质量控制,此外,将高维的基因组数据转换为图像格式并应用卷积神经网络模型,将临床数据通过多层感知机进行建模;引入注意力机制,通过加权融合不同模态数据的特征,以优化预测效果。**结果** 通过融合基因组数据与临床数据,本研究构建的多模态数据融合模型在癌症预测的准确性上得到显著提升。经过卷积神经网络提取的图像特征和多层感知机提取的临床特征有效增强预测模型的能力,提升预测结果的准确性和鲁棒性。**结论** 提出一种基于基因数据与临床数据融合的多癌种风险预测方法,验证多模态深度学习方法在癌症早期预测中的效果。通过卷积神经网络、多层感知机及注意力机制等技术的结合,显著提高癌症预测的精度,为未来癌症的诊断和个性化治疗提供强有力的早期支持。

关键词:多癌种风险预测;多模态数据;基因组;临床;卷积神经网络;多层感知机
中图分类号:TP391 **文献标志码:**A

Multi-cancer risk prediction model based on multi-modal data fusion

LI Qian[1], YANG Fan[1,2,3], XUE Fuzhong[1,2,3]
(1. Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China;
2. National Institute of Health and Medical Big Data, Jinan 250003, Shandong, China;
3. Qilu Hospital of Shandong University, Jinan 250012, Shandong, China)

Abstract: Objective To develop a multi-cancer risk prediction model using data from 15 common cancers in the UK Biobank, employing a multi-modal data fusion approach, so as to explore the application of genomic and clinical data in cancer risk prediction, with the goal of enhancing early cancer detection accuracy and providing valuable insights for personalized medicine. **Methods** The rigorous quality control was performed to the data. High-dimensional genomic data were then transformed into image representations and processed using convolutional neural networks, while clinical data were modeled using multi-layer perceptron. An attention mechanism was incorporated to perform weighted fusion of features from both genomic and clinical modalities, aiming to optimize predictive performance. **Results** The integration of genomic and clinical data through a multi-modal fusion model resulted in a significant improvement in cancer prediction accuracy. Features extracted by convolutional neural networks from genomic data and by multi-layer perceptron from clinical data effectively augmented the predictive capability of the model, enhancing both the accuracy and robustness of the predictions. **Conclusion** This study introduces a novel multi-cancer risk prediction framework that integrates genomic and clinical data. The application of multi-modal deep learning techniques, including convolutional neural networks, multi-layer perceptrons, and attention mechanisms, significantly enhances early cancer prediction accuracy. The findings provide robust early support for cancer diagnosis and personalized treatment strategies, demonstrating the potential of multi-modal approaches in precision oncology.

收稿日期:2025-02-01
基金项目:国家自然科学基金(82273736)
通信作者:杨帆。E-mail:fanyang@sdu.edu.cn
薛付忠。E-mail:xuefzh@sdu.edu.cn

Key words: Multi-cancer risk prediction; Multi-modal data; Genomics; Clinical data; Convolutional neural networks; Multi-layer perceptron

癌症是一类由基因突变驱动的复杂疾病[1],其发生发展通常伴随着多种遗传与环境因素的交互作用。随着全球人口老龄化和环境风险的累积,癌症已成为全球性重大公共卫生问题[2]。据统计,2020年全球新增癌症病例约 1 930 万,死亡人数近千万[3]。癌症的高度异质性使得基于器官部位的传统分类和治疗策略面临挑战,早期精准预测与个体化干预亟需更有效的手段。

高通量测序技术的迅速发展推动了癌症的相关研究。当前,大量基因组及临床数据的积累,为肿瘤异质性解析与风险评估提供了前所未有的机遇。然而,多模态数据的高维性、异质性及潜在噪声,给传统分析方法带来极大挑战[4]。深度学习作为近年来迅猛发展的人工智能方法,在生物医学数据建模中表现出强大的特征学习与非线性建模能力,已被广泛应用于癌症分类、预后预测等任务[5]。

本研究旨在构建一种融合基因组与临床信息的多模态深度学习模型,实现多癌种风险预测。模型采用卷积神经网络(convolutional neural network, CNN)提取基因组数据中的潜在特征,并结合多层感知机(multi-layer perceptron, MLP)对临床变量进行建模;同时引入注意力机制,以加权整合不同模态的信息。该融合策略有助于提升模型对多源数据的表征能力和判别能力,从而提高风险预测的准确性与泛化性能,为癌症的早期诊断与个体化治疗提供可靠支持。

1 资料与方法

1.1 资料

1.1.1 数据来源

英国生物银行[6-7]是一项针对英国 50 多万人的大型队列研究数据库。该数据库招募 2006—2010 年间年龄在 40~69 岁之间的人,并进行追踪随访,收集的信息包括基因型数据、人体测量数据、生活习惯及个人与家族病史相关信息,项目号 98273。

本研究以英国生物银行数据库为来源,选取 15 种癌症类型的相关临床信息及单核苷酸多态性(single nucleotide polymorphism, SNP)[8]资料。涉及的癌症包括:乳腺癌、结肠癌、前列腺癌、子宫颈癌、直肠癌、黑色素瘤、卵巢癌、胃癌、膀胱癌、肺癌、肾癌、甲状腺癌、滤泡性淋巴瘤、多发性骨髓瘤及霍奇金淋巴瘤。利用 ICD-10[9]编码体系对这 15 种癌症进行详尽的标识与分类,见表 1。

表 1 15 种癌症 ICD-10 编码及数量

Table 1 ICD-10 codes and numbers of 15 cancers

癌症类型	ICD-10 编码	数量
胃癌	C16	916
结肠癌	C18	4 664
直肠癌	C20	2 293
肺癌	C34	3 533
黑色素瘤	C43	4 207
乳腺癌	C50	15 209
子宫颈癌	C53	580
卵巢癌	C56	1 588
前列腺癌	C61	10 167
肾癌	C64	1 665
膀胱癌	C67	2 987
甲状腺癌	C73	673
霍奇金淋巴瘤	C81	451
滤泡性淋巴瘤	C82	712
多发性骨髓瘤	C90	892

1.1.2 数据预处理

本研究提取英国生物银行中相关样本的 SNP 数据信息,利用 Plink 软件[10]对所有样本的 SNP 数据进行细致的筛选和质量控制,以确保所获取的基因型数据具有高质量和可靠性。设定缺失值过滤门槛,清除单一缺失数据的位点;在基因型质量过滤环节中,淘汰质量重复的位点;通过 Hardy-Weinberg[11]平衡筛选,剔除存在显著分布偏差的位点;考虑到连锁不平衡,筛选出 SNP 间的关联性;通过等位基因频率过滤,清除频率过低的位点。此外,研究采用 KNN 填充方法[12]填补缺失的 SNP 数据信息。在数据预处理后,SNP 特征总计 43 494 个,以 0、1、2 表示。共有样本数量 46 730 个(数据中存在 1 例患者罹患多种癌症的情况)。

本研究从英国生物银行数据中提取 15 个临床特征,以评估其在癌症风险预测中的潜在作用。临床特征包括:人口学特征(年龄、性别、种族)、生活方式相关特征(是否吸烟、是否饮酒)、个体的慢性疾病既往病史(高血压、糖尿病)、医学检查数据(血糖、血脂、血压、身高、体质量、BMI、心率、血氧饱和度),有助于深入理解生活方式、健康状况与癌症风险之间的复杂关系。其中连续变量经过标准化处理后转化为均值为 0、标准差为 1 的标准正态分布,离散型变量采用独热编码的方式转化为数值型变量。最终处理后的临床数据包含 46 730 个样本,每个样

本包含 15 个临床特征。

1.2 方法

1.2.1 基于多模态数据融合的多癌种风险预测模型

本研究所提出的模型整体架构采用模块化设计,分别构建基因组风险预测模型和临床风险预测模型,并通过引入注意力机制实现跨模态特征的加权融合,模型框架图见图 1。

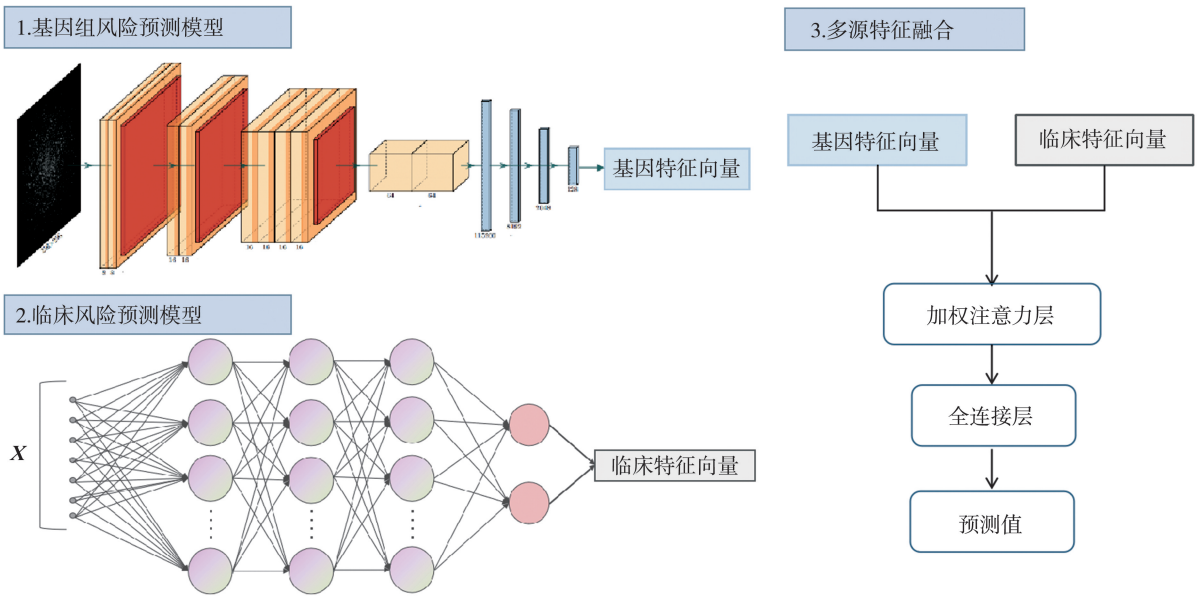

图 1 模型框架

Figure 1 Model framework

1.2.1.1 基因组风险预测模型

基因组数据具有高维性,采用传统的机器学习方法处理可能面临的维度灾难,但图像作为一种低维且密集的数据表示形式,能够通过空间关系捕捉高维数据中的潜在模式。尽管 SNP 数据本身并非图像数据,但它们之间的相关性可以通过图像中的空间分布来表示。因此,本研究采用DeepInsight[13]方法首先将高维的 SNP 数据转换为图像格式,为每一个样本生成唯一的 SNP 图谱。DeepInsight的核心思想在于将非图像样本转化为图像,以便利用 CNN 进行预测或分类,发挥其在图像处理方面的强大能力[14]。具体操作上,采用t-SNE算法对每个样本的 SNP 特征向量进行非线性降维,把高维特征数据映射到二维空间。然后,运用凸包算法确定包含所有特征点的最小矩形,并进行旋转处理。在特征矩阵中为各个特征定位后,便可根据特征值生成对应图像。若有多个特征在像素帧中位置重叠,则将这些特征值平均后置于同一位置。最后,对所有图像执行归一化处理,确保像素值在[0,255]范围内。特征在二维坐标系中的位置关系揭示了特征之间的相似程度[15]。

为进一步深入分析数据特征并实现精确分类,本研究使用 CNN 模型。该模型采用多层次架构,包含多个卷积层、池化层及全连接层,逐层提炼特征。初始阶段,通过逐步提升输出通道的数量,卷积层负责捕获图像的局部特征,并揭示基因组数据的空间联系。在卷积层之后,通过最大池化层对特征图进行下采样,以提取高级特征表示。然后,利用一系列全连接层整合特征,提取数据中的深层语义。为增强训练成效并防止梯度消失,引入 ReLU 激活函数[16]。模型在初始化阶段采用 Kaiming 正态初始化方法对权重进行初始化,以加速收敛并提升训练稳定性[17]。为增强模型泛化能力,避免过拟合,本研究实施 Dropout 策略(丢弃率=0.5)[18]。本研究构建的 CNN 模型能够高效地从基因组数据中提取深层次特征。这些特征将在后续的特征融合阶段与其他模态的数据进行整合,从而为疾病分类或预测任务提供丰富的信息。

1.2.1.2 临床风险预测模型

在疾病预测任务中,临床数据通常包含患者的生理特征、疾病历史、治疗记录等重要信息。临床数据的处理和分析需要考虑这些特征之间的复杂关系和潜在的非线性交互。MLP 是一种经典的神经网络架构,能够通过其全连接层有效捕捉输入特征与输出之间的非线性关系,具有强大的表达能力和灵活的建模能力。

多层感知器架构由输入层、三层隐藏层及输出层构成。输入层接收经过标准化处理的临床数据特

征，随后经过多个隐藏层的非线性映射逐步提取其复杂模式。每一隐藏层中的神经元通过加权连接实现特征表达的逐级抽象，其前向传播过程可表示为：

$$\boldsymbol{h}^{(l)}=f(\boldsymbol{W}^{(l)}\boldsymbol{h}^{(l-1)}+\boldsymbol{b}^{(l)}), \quad (1)$$

其中，$\boldsymbol{h}^{(l)}$是第 l 层的输出向量，$\boldsymbol{h}^{(l-1)}$是 $l-1$ 层的输出向量，$\boldsymbol{W}^{(l)}$是第 l 层的权重矩阵，$\boldsymbol{b}^{(l)}$是第 l 层的偏置向量，f 是激活函数 ReLU。在隐藏层中使用激活函数，可以增强模型的非线性表达能力，并减轻梯度消失现象。MLP 模型能够自动从临床数据中提取特征，并逐渐学习最优的疾病预测模型。

1.2.1.3 多源特征融合

本研究将临床数据和基因数据作为两个独立的模态，采用注意力机制[19]来融合这两种模态的特征，以便模型能够自动学习每种模态对最终任务的重要性，并据此动态调整每个模态的权重。该阶段的核心目标是利用从不同模态中提取的特征，通过特征加权融合形成一个综合性的特征表示，这一表示能够充分捕捉临床数据与基因数据的交互信息。具体来说，本研究在特征融合阶段引入加权注意力层。首先，临床数据通过 MLP 得到一个临床特征表示 h_{clinical}，基因数据通过 CNN 得到一个基因特征表示 h_{genomic}。为每个模态的特征向量分配一个可学习的注意力权重，以决定该模态在最终预测中的贡献程度。该注意力权重是通过一个小型的全连接网络进行学习，将 MLP 和 CNN 输出的特征向量传递给注意力网络，得到每个模态的权重值 e_{clinical}、e_{genomic}。权重值经过 Softmax 归一化处理[20]，确保它们在 [0,1] 范围内，并且和为 1。最后，通过对每个模态的特征向量进行加权平均，得到融合后的特征向量，将其输入到全连接层进行处理。在全连接层的输出中，使用 Softmax 激活函数将结果转换为概率分布，使得每个类别的预测概率总和为 1。这一机制不仅提高了模型的分类性能，也使得结果更具可解释性，为后续的决策提供了依据。通过这种方式，模型能够输出每个类别的相对概率，进而为最终的疾病预测提供可信的决策依据。为了保证预测结果的精确性和稳健性，本研究采用了交叉熵损失函数衡量概率分布与真实标签分布之间的差异，公式如下：

$$L=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{c}y_{i,c}\log(\hat{y}_{i,c}), \quad (2)$$

其中，N 是样本数量，$y_{i,c}$是样本 i 的真实标签，$\hat{y}_{i,c}$是模型对样本的预测概率。交叉熵损失函数经常用于多分类问题，可衡量预测概率分布与实际标签分布之间的偏差。在训练阶段，为了适应动态学习率并加快收敛速度，采用 Adam 优化器[21]，模型采用 10 倍交叉验证训练。模型的具体参数设置见表 2。

表 2 实验参数设置

Table 2 Experimental parameter setting

参数	设置
初始学习率	0.001
批量大小	32
优化器	Adam
训练轮数	200

1.2.2 实验指标

为全面评估所提出的多模态深度学习模型在疾病预测中的表现，本研究选择准确率（accuracy）、精确率（precision）、召回率（recall）和 F1 分数四个常用且有效的评估指标。计算公式如下：

$$\text{accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \quad (3)$$

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}, \quad (4)$$

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}, \quad (5)$$

$$\text{F1 分数}=2\times\frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}, \quad (6)$$

其中，TP 为真正例，TN 为真负例，FP 为假正例，FN 为假负例。这些指标能够从不同角度量化模型在分类任务中的效果，可以全面评估模型在多模态数据上的分类性能。

1.2.3 对比实验

为评估所提出的多模态深度学习模型的有效性，本研究选择以下五种对比方法：支持向量机（support vector machine，SVM）[22]、随机森林（random forest，RF）[23]、逻辑回归（logistic regression，LR）[23]、变分自编码器（variational autoencoder，VAE）[24]以及 DeepClinMed-PGM[25]。通过这些模型的对比，可以全面验证所提出的多模态特征融合方法是否能够有效提升疾病预测的性能。所有实验均在相同数据集上，确保结果的公平性和可比性。

1.2.4 消融实验

为了全面评估所提出的多模态特征融合模型中各关键模块对整体性能的影响，设置以下四种消融实验：①去除模型中的注意力机制，采用简单加权平均方法进行多模态特征融合，以评估注意力机制在提升模型性能中的作用；②剔除基因组数据，仅使用临床数据进行训练和预测，以分析基因组信息的贡献；③剔除临床数据，仅利用基因数据训练模型，进一步探讨临床数据在预测中的作用；④将多模态特

征融合方法替换为传统的拼接融合方法,通过直接拼接临床和基因特征并输入全连接网络,实现数据融合,以比较不同融合方式的效果。所有实验均在相同数据集上,确保结果的公平性和可比性。

1.2.5 外部实验

本研究原始模型针对多癌种风险预测任务进行构建与训练,为评估其在特定癌种内部异质性识别中的泛化能力与鲁棒性,本研究在 cBioPortal 平台中的 METABRIC 乳腺癌数据集[26](包含 1 157 个样本)和 TCGA 乳腺癌数据集[27](包含 721 个样本)上进行外部实验验证。外部实验聚焦于乳腺癌的亚型识别,具体包括 Luminal A、Luminal B、HER2-enriched和 Basal-like 四种分型。

2 结 果

2.1 基因组 SNP 图谱构建结果

本研究通过基因组 SNP 数据共构建出 46 730 个图像数据,图像尺寸统一,图谱质量稳定,无缺失样本,图 2 显示了部分样本的图谱。不同样本在图像的纹理结构、灰度分布及局部密度区域上存在差异。部分图谱呈现出高强度像素集中分布的区域,表明基因位点在空间映射中具有一定的聚集性。这些图谱差异为后续模型区分不同癌种提供了可识别的表征基础。

2.2 对比实验结果

多模态数据融合模型在各项评估指标上均优于所有对比方法,表现出更强的分类能力。其中,与传统机器学习模型相比,该模型在准确度和 F1 分数方面提升最为显著,说明其在整体判别性能和精确-召回平衡上具备更强优势。与 VAE 模型相比,多模态融合策略显著增强了模型对复杂特征的表达能力。尽管 DeepClinMed-PGM 在多模态处理方面也具备一定性能,但本研究模型在四项指标上均实现进一步提升,表明所提出的融合机制在不同模态间的信息整合效果更优,尤其在阳性样本识别的稳定性方面更具优势。见表 3。这些结果验证本研究提出方法在多模态疾病预测任务中的有效性与鲁棒性。

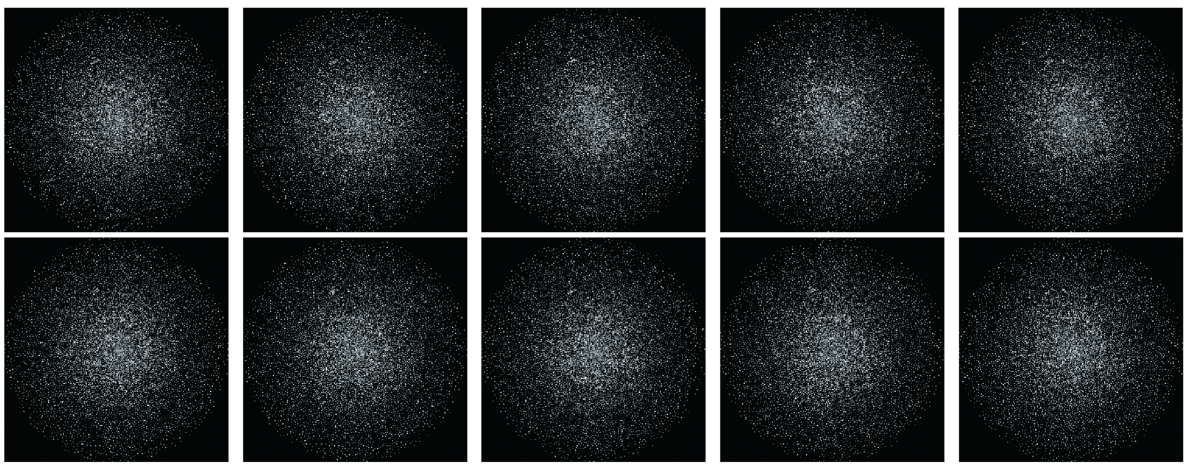

图 2 部分样本 SNP 图谱构建结果

Figure 2 Results of SNP map construction of partial samples

表 3 基准模型与多模态数据融合模型的性能对比

Table 3 Comparison of performance between baseline models and multi-modal data fusion model

对比方法	准确率	精确率	召回率	F1 分数
SVM	0.742	0.735	0.720	0.727
RF	0.760	0.745	0.730	0.737
LR	0.751	0.730	0.725	0.727
VAE	0.778	0.762	0.751	0.756
DeepClinMed-PGM	0.809	0.793	0.782	0.787
多模态数据融合模型	0.832	0.818	0.804	0.811

2.3 消融实验结果

去除注意力机制后,模型的准确率、精确率、召回率和 F1 分数均出现不同程度的下降。仅使用临床数据时,模型的整体性能较完整模型显著降低,表明基因数据在疾病预测中贡献显著。单独使用基因数据的模型表现更差,尤其是召回率下降明显,说明临床数据对于提升模型的敏感性具有重要作用。采用传统的拼接融合策略虽改善了模型性能,但其准确率、精确率、召回率和 F1 分数均低于采用加权注意力机制的融合模型。总体来看,消融实验充分体

现了各关键组件对提升模型性能的必要性和多模态融合策略的有效性。见表 4。

表 4 消融实验结果

Table 4 Ablation experiment results

对比方法	准确率	精确率	召回率	F1 分数
去除注意力机制	0.788	0.762	0.750	0.756
仅使用临床数据	0.761	0.748	0.735	0.741
仅使用基因数据	0.712	0.703	0.690	0.696
传统特征融合方法	0.804	0.792	0.780	0.786
多模态数据融合模型	0.832	0.818	0.804	0.811

2.4 外部实验结果

模型在 METABRIC 数据集上的准确率、精确率、召回率和 F1 分数均维持在较高水平,分别达到 0.846、0.823、0.817 和 0.820;在 TCGA 数据集上的对应指标略有下降,但仍保持稳定,分别为 0.829、0.815、0.801 和 0.808。两个数据集间模型性能指标的差异较小,表明该模型在不同来源和特征分布的数据中表现一致,体现了较强的泛化能力和鲁棒性。

3 讨 论

本研究提出一种基于多模态深度学习的疾病预测模型,通过有效融合临床数据与基因数据,能够充分挖掘不同数据源之间的潜在信息,从而提升预测的准确性。实验结果表明,该模型在分类性能、特征融合效果和泛化能力三个方面均表现出色:在英国生物银行数据集上,模型的准确率为 0.832,F1 分数为 0.811;通过消融分析证实,注意力机制贡献了 4.4%的准确率性能提升,临床数据与基因数据的有机结合显著提升了模型的整体性能表现;在 METABRIC 和 TCGA 乳腺癌数据集上,模型分别保持 0.846 和 0.829 的准确率,验证了其良好的泛化性。这些结果充分证明了多模态数据融合在癌症分类中的重要作用。

本研究的创新性主要体现为方法学层面的 3 个突破:① 基因数据为疾病预测提供遗传学信息,临床数据则反映患者的病理、临床状况等方面的特征。将两者进行结合,不仅能够捕捉到疾病发展的生物学基础,还能考虑到临床表现等外部因素。②注意力机制对特征的融合起着至关重要的作用。该注意力机制通过对不同模式的数据赋予不同权重,可根据特征重要性动态调整数据,从而使模型能更高效地抽取与使用信息。③构建的通用性框架不仅为乳腺癌精准诊疗提供了可靠方案,其技术路线还可推广至其他疾病的预测研究。多模态特征融合策略具备较强的泛化能力,能够适应不同癌症背景下的异质性识别需求。

尽管本研究展示多模态深度学习模型在疾病预测中的优势,但也存在一些局限性。① 实验数据集的样本量较为有限,虽然取得显著的预测效果,但未来还需要基于更大规模的数据集进行验证,以进一步提升模型的泛化能力。②本研究中的基因数据主要来自癌症相关数据集,未来可以扩展到其他疾病的数据集,尤其是在非癌症领域,多模态数据的融合潜力仍然值得进一步探索[28]。③ 尽管本研究通过注意力机制有效地融合基因和临床数据,但在处理更复杂的多模态数据时,未来可以尝试引入更为先进的技术,如图神经网络[29]、强化学习[30]等方法。这些方法有可能进一步提高模型在特征融合和学习过程中的自适应能力,尤其在数据具有复杂结构和非线性关系时,能够更好地捕捉潜在模式。

综上所述,本研究提出的多模态深度学习模型显著提升疾病预测的精度。未来,该模型有望在实际临床应用中提供更精准的疾病预测和辅助决策支持,结合患者的遗传变异和临床指标,为个体化筛查方案的制定提供依据,减少漏诊和误诊。同时它还能辅助临床决策,优化治疗方案,降低不必要的治疗成本,为患者提供更精准的医疗服务。

参考文献:

[1] Bodmer WF. Cancer genetics[J]. Br Med Bull, 1994, 50(3): 517-526.

[2] Zhou MG, Wang HD, Zhu J, et al. Cause-specific mortality for 240 causes in China during 1990-2013: a systematic subnational analysis for the global burden of disease study 2013[J]. Lancet, 2016, 387(10015): 251-272.

[3] Ferlay J, Colombet M, Soerjomataram I, et al. Cancer statistics for the year 2020: an overview[J]. Int J Cancer, 2021. doi:10.1002/ijc.33588

[4] Song QX, Merajver SD, Li JZ. Cancer classification in the genomic era: five contemporary problems[J]. Hum Genomics, 2015, 9: 27. doi:10.1186/s40246-015-0049-8

[5] Ravì D, Wong C, Deligianni F, et al. Deep learning for health informatics[J]. IEEE J Biomed Health Inform, 2017, 21(1): 4-21.

[6] Caleyachetty R, Littlejohns T, Lacey B, et al. United

Kingdom biobank (UK biobank): JACC focus seminar 6/8[J]. J Am Coll Cardiol, 2021, 78(1): 56-65.
[7] Roca-Fernandez A, Banerjee R, Thomaides-Brears H, et al. Liver disease is a significant risk factor for cardiovascular outcomes-A UK Biobank study[J]. J Hepatol, 2023, 79(5): 1085-1095.
[8] Louhelainen J. SNP arrays[J]. Microarrays, 2016, 5(4): 27. doi:10.3390/microarrays5040027
[9] Adler KG. ICD-10: our newest documentation dilemma[J]. Fam Pract Manag, 2015, 22(5): 7.
[10] Chang CC. Data management and summary statistics with PLINK[J]. Methods Mol Biol, 2020: 49-65. doi: 10.1007/978-1-0716-0199-0_3
[11] Gomes I, Collins A, Lonjou C, et al. Hardy-Weinberg quality control[J]. Ann Hum Genet, 1999, 63(6): 535-538.
[12] Petrazzini BO, Naya H, Lopez-Bello F, et al. Evaluation of different approaches for missing data imputation on features associated to genomic data[J]. BioData Min, 2021, 14(1): 44. doi: 10.1186/s13040-021-00274-7
[13] Sharma A, Vans E, Shigemizu D, et al. DeepInsight: a methodology to transform a non-image data to an image for convolution neural network architecture[J]. Sci Rep, 2019, 9(1): 11399. doi:10.1038/s41598-019-47765-6
[14] Son CS, Kang WS. Multivariate CNN model for human locomotion activity recognition with a wearable exoskeleton robot[J]. Bioengineering, 2023, 10(9): 1082. doi: 10.3390/bioengineering10091082
[15] 祝玉杰, 叶晟, 申利民. 基于t-SNE特征降维和K近邻的分类算法[J]. 电脑知识与技术, 2024, 20(34): 11-13.
[16] Liu YJ, Caglar T, Peterson C, et al. Integrating geometries of ReLU feedforward neural networks[J]. Front Big Data, 2023, 6: 1274831. doi:10.3389/fdata.2023.1274831
[17] de Pater I, Mitici M. A mathematical framework for improved weight initialization of neural networks using Lagrange multipliers[J]. Neural Netw, 2023, 166: 579-594. doi:10.1016/j.neunet.2023.07.035
[18] Hernández-Rodríguez JC, García-Muñoz C, Ortiz-Álvarez J, et al. Dropout rate in digital health interventions for the prevention of skin cancer: systematic review, meta-analysis, and metaregression[J]. J Med Internet Res, 2022, 24(12): e42397. doi:10.2196/42397
[19] 高宇, 李子昂, 魏正琦, 等. MR高分辨率血管壁成像影像组学联合注意力机制预测症状性颅内动脉粥样硬化狭窄患者卒中复发[J]. 中国医学影像技术, 2025, 41(2): 229-233.
GAO Yu, LI Ziang, WEI Zhengqi, et al. MR high-resolution vessel wall imaging radiomics combined with attention mechanism for predicting stroke recurrence in patients with symptomatic intracranial atherosclerosis stenosis[J]. Chinese Journal of Medical Imaging Technology, 2025, 41(2): 229-233.
[20] Shao H, Wang SF. Deep classification with linearity-enhanced logits to softmax function[J]. Entropy, 2023, 25(5): 727. doi:10.3390/e25050727
[21] 王建涛, 邵一川, 孙海静, 等. 改进的Adam优化算法在阿尔茨海默病医学图像分类中的应用[J/OL]. 计算机应用与软件. https://link.cnki.net/urlid/31.1260.tp.20241230.1745.005
WANG Jiantao, SHAO Yichuan, SUN Haijing, et al. Application of improved Adam optimization algorithm in medical image classification of Alzheimer's disease[J/OL]. China Industrial Economics. https://link.cnki.net/urlid/31.1260.tp.20241230.1745.005
[22] He SQ, Xiao B, Wei HJ, et al. SVM classifier of cervical histopathology images based on texture and morphological features[J]. Technol Health Care, 2023, 31(1): 69-80.
[23] 周晓燕, 魏申奥, 卢曼曼. 基于Lasso-logistic回归和随机森林模型的癌症患者抑郁影响因素分析[J]. 安徽医学, 2024, 45(9): 1177-1182.
[24] 付金露. 基于特征选择的乳腺癌患者预后模型研究[D]. 南昌: 江西财经大学, 2023.
[25] Wang ZH, Lin RC, Li YC, et al. Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction[J]. Precis Clin Med, 2024, 7(2): pbae012. doi:10.1093/pcmedi/pbae012
[26] Pereira B, Chin SF, Rueda OM, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes[J]. Nat Commun, 2016, 7: 11479. doi:10.1038/ncomms11479
[27] Network CGA. Comprehensive molecular portraits of human breast tumours[J]. Nature, 2012, 490(7418): 61-70.
[28] Mandal PK, Perry G. SWADESH: a comprehensive platform for multimodal data and analytics for advanced research in Alzheimer's disease and other brain disorders[J]. J Alzheimers Dis, 2022, 85(1): 1-5.
[29] Bessadok A, Mahjoub MA, Rekik I. Graph neural networks in network neuroscience[J]. IEEE Trans Pattern Anal Mach Intell, 2023, 45(5): 5833-5848.
[30] Lowet AS, Zheng Q, Matias S, et al. Distributional reinforcement learning in the brain[J]. Trends Neurosci, 2020, 43(12): 980-997.

(编辑:相峰)