Advances in Applied Mathematics 应用数学进展, 2024, 13(9), 4182-4190
Published Online September 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.139399

Hans汉斯

基于深度多视图对比学习方法的多组学数据整合及预后预测模型构建

高新凤

青岛大学数学与统计学院，山东 青岛

收稿日期：2024年8月10日；录用日期：2024年9月2日；发布日期：2024年9月12日

摘　要

在癌症研究中，精准识别癌症亚型和评估患者预后对制定优化治疗方案至关重要。高通量测序技术生成的大量多组学数据为癌症预后研究提供了宝贵资源。深度学习方法能够有效整合这些数据，精确识别更多癌症亚型。在本研究中，我们分析了12种癌症的多组学数据集，并将其作为模型的输入。我们提出了一种基于卷积自动编码器的深度多视图对比学习模型(dmCLCAE)，该模型旨在利用多组学数据预测与生存相关的癌症亚型。为了验证模型的效果，我们对比了多组学因子分析算法(MOFA+)和深度学习模型(ProgCAE)在不同癌症类型分类中的表现。结果显示，dmCLCAE在区分不同生存亚型方面表现出显著优势，同时在预测一致性上也有更优异的表现。

关键词

多组学数据，卷积自编码器，对比学习，深度学习

Integration of Multi-Omics Data and Prognostic Prediction Model Construction Based on Deep Multi-View Contrastive Learning Methods

Xinfeng Gao

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Aug. 10th, 2024; accepted: Sep. 2nd, 2024; published: Sep. 12th, 2024

Abstract

In cancer research, accurately identifying cancer subtypes and assessing patient prognosis are

文章引用: 高新凤. 基于深度多视图对比学习方法的多组学数据整合及预后预测模型构建[J]. 应用数学进展, 2024, 13(9): 4182-4190. DOI: 10.12677/aam.2024.139399

crucial for developing optimized treatment strategies. The vast amount of multi-omics data generated by high-throughput sequencing technologies provides valuable resources for cancer prognosis studies. Deep learning methods can effectively integrate these data to accurately identify more cancer subtypes. In this study, we analyzed multi-omics datasets from 12 types of cancer and used them as input for our model. We proposed a deep multi-view contrastive learning model based on a convolutional autoencoder (dmCLCAE), designed to predict survival-related cancer subtypes using multi-omics data. To validate the model's performance, we compared it with the Multi-Omics Factor Analysis v2 (MOFA+) and prognostic model based on a convolutional autoencoder (ProgCAE) in classifying various cancer types. The results showed that dmCLCAE demonstrated a significant advantage in distinguishing different survival subtypes and exhibited superior consistency in predictions.

Keywords

Multi-Omics Data, Convolutional Autoencoder, Contrastive Learning, Deep Learning

1. 引言

在癌症研究中，确定癌症亚型和估计患者预后至关重要[1] [2]。高通量测序技术生成的大量多组学数据，是进行癌症预后研究的关键资源。随着生物技术的快速发展，各种组学数据的获取变得越来越方便。一些大型协作项目如癌症基因组图谱(TCGA) [3]、基因表达数据库(GEO) [4]和国际癌症基因组联盟(ICGC) [5]等，已经收集了成千上万个样本的多种组学数据，这为通过计算方法分析癌症亚型提供了前所未有的机会。早期方法通常在单一组学数据上应用聚类算法来预测癌症亚型[6]。然而，由于每种组学数据仅在一定层次上表征分子特征，多组学数据的整合可以为描述癌症亚型提供更全面的视角，并进一步加深我们对跨多个层次的生物分子复杂相互作用的理解[7] [8]。因此，利用多组学数据预测癌症亚型的计算方法受到了极大的关注。然而，高维生物数据包含着成千上万个单核苷酸多态性、基因和蛋白质，因此分析多组学数据是一项极具挑战性的任务[9]。为此，近年来涌现出了许多整合多组学数据的方法，其中大部分依赖于降维技术[10]。传统的统计线性降维方法包括因子分析和主成分分析(PCA)。近年来，深度学习技术取得了显著进展，并在多组学数据的分析中得到广泛应用，其强大之处在于能够分析非线性关系[11]。卷积神经网络(CNN) [12]在图像识别[13]和医学诊断[14]中被广泛应用。卷积自编码器(CAE) [15]结合了CNN 和自编码器的优点，在癌症预后分析中用于降维。然而，传统的监督学习在需要大量标注数据时可能受限于数据的稀缺性和高昂的标注成本。为了克服这些挑战，对比学习作为一种自监督学习范式引起了研究者的关注。对比学习不仅在自然语言处理和计算机视觉中取得了成功，也开始在生物信息学领域展现出潜力[16] [17]。它通过设计前置任务来构造正负样本对，训练特征模型，使得相似实例在映射空间中距离较近，而相异实例则距离较远。这种方法不仅能在标注数据有限的情况下提取有效的特征表示，还能够通过特征空间的调整来增强模型的泛化能力，从而在多组学数据的分析和应用中发挥更大的作用[18]。

本文提出了一种基于卷积自编码器的深度多视图对比学习模型(dmCLCAE)，该模型利用 CAE [19]整合多组学数据并通过将重构损失和对比损失纳入一个统一的框架进行预后预测，我们的模型能够同时将

样本区分信息编码到提取的特征表示中，并在嵌入空间中很好地保持样本的聚类结构。dmCLCAE 的输入数据包括拷贝数变异(CNV)、DNA 甲基化(甲基化)、RNA 测序(RNA-seq)和 miRNA 测序。我们将 dmCLCAE 应用于两组癌症数据集：其中一组包含 10 种含四个组学数据的癌症数据集，另一组包含 2 种含三个组学数据(不含 miRNA-seq)的癌症数据集。结果表明，我们的模型比任何其他已发表的模型具有更好的预测准确性。

2. 材料与方法

2.1. 生物多组学数据

近年来，生物组学数据库(如基因组、转录组、蛋白组、代谢组等)的数据越来越多，内容越来越全面。这些数据库收集和整理了大量的生物学信息。研究人员越来越重视将不同类型的组学数据结合起来进行分析。这种整合的趋势使得研究人员能够从多个层面对癌症进行更全面和深入的理解。通过整合不同的组学数据，科学家能够更准确地对不同类型的癌症进行分类。这有助于识别出癌症的不同亚型，从而为个性化治疗提供依据。当前，主要通过基因芯片和下一代测序技术来获取癌症相关的生物组学数据。基因芯片技术可以同时检测大量基因的表达水平，而下一代测序技术则能够对基因组进行高通量、精确的测序。为了有效利用这些海量的组学数据，科学界已经建立了许多公共数据库。这些数据库汇总了大量的多组学数据，使研究人员能够方便地获取和分析这些数据，从而推动癌症研究的发展。

本文的数据来源于癌症基因组计划——The Cancer Genome Atlas (TCGA)。为了深入研究癌症的分子机制及其生物标志物，本文从 TCGA 数据中心下载了 12 种癌症的多组学数据集。对于其中 10 种癌症，我们获取了四种主要组学类型的数据：RNA 测序(RNA-Seq)、DNA 甲基化、拷贝数变异(CNV)以及 miRNA 测序(miRNA-Seq)。而对于剩余的两种癌症类型——肺癌和胶质母细胞瘤，由于数据集中缺少 miRNA-Seq 信息，我们仅收集了 RNA-Seq、DNA 甲基化和 CNV 三种组学数据。此外，为了更全面地进行不同癌症类型的生存分析和预后评估，本文还下载了每种癌症患者的生存信息。

2.2. 基于统计降维方法的多组学整合算法 MOFA+

Argelaguet 等人开发的多组学因子分析算法(MOFA+)来整合和分析多组学数据。MOFA+通过将输入的数据矩阵分解为因子矩阵和权重矩阵，从而学习到输入数据的低维表示，捕捉不同组学数据中的共同和特定特征，便于后续的下游分析。具体而言，MOFA+处理 $N\times D_m$ 维的 M 个数据矩阵 $Y_1,\cdots,Y_M$，其中 N 表示样本数，D_m 表示第 m 个数据矩阵的特征数。MOFA+通过以下公式分解这些数据矩阵：

$$Y_M = ZW_m^{\mathrm{T}} + \varepsilon_m,\ m = 1,\cdots,M$$

其中，Z 为 $N\times K$ 的因子矩阵，K 为因子数；W_m 为第 m 个数据矩阵的权重矩阵，尺寸为 $D_m\times K$；ε_m 为噪声项，代表特定于每个矩阵的误差或噪声。MOFA+在概率贝叶斯框架中构建，所有未观测到的变量都被赋予先验分布。因子矩阵 Z 采用标准正态先验分布 $Z\sim N(0,I)$，权重矩阵 W_m 使用稀疏先验分布，以使得权重矩阵中的大多数元素为零，从而实现特征选择。噪声项 ε_m 也被赋予相应的先验分布，以控制其大小和影响。

2.3. 基于深度学习方法的多组学整合算法 ProgCAE

Liu 等人提出来了一种基于卷积自动编码器的新型深度学习模型(ProgCAE)。该方法能够高效进行表征学习和数据整合，有效捕捉不同组学数据之间的复杂关系。该模型首先通过卷积自动编码器整合多组学数据集，并将其转化为潜在表示。随后，利用潜在表示构建单变量 Cox 比例风险回归模型，以筛选出与患者生存显著相关的特征。通过这一过程，识别出对生存具有重要影响的特征。接着，利用这些特征

对患者进行聚类和分类。自编码器是一种神经网络模型，能够基于输入数据学习有效的编码方式。其结构如图 1 所示，通常由输入层、一个或多个隐藏层以及输出层组成。自编码器的目标是将输入数据压缩到隐藏层，并通过解码器从压缩表示中重建输入数据。自编码器包括两个主要部分：编码器(Encoder)和解码器(Decoder)。

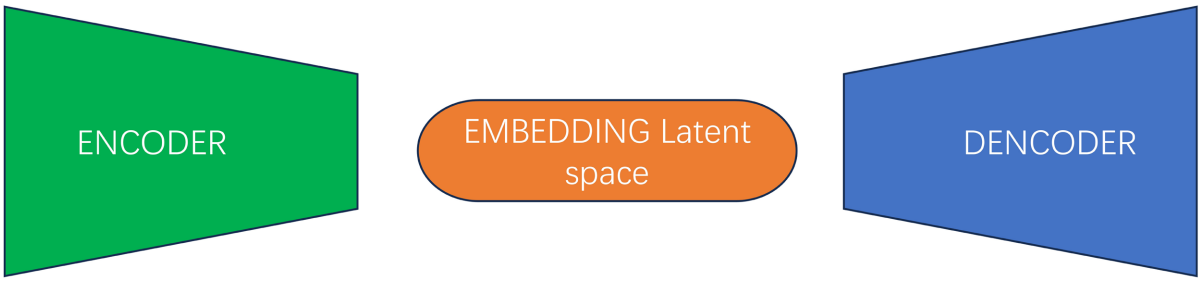

Figure 1. Autoencoder architecture
图 1. 自编码器结构

编码器通过一系列非线性变换和特征提取将输入数据映射到隐藏层，而解码器则通过类似的方式将隐藏层的表示重构回原始输入数据。对于输入 x，隐藏层值 z 为 x 的编码，即：

$$z = f\left(W^{(1)}x + b^{(1)}\right).$$

自编码器的输出值为

$$x' = f\left(W^{(2)}z + b^{(2)}\right),$$

其中 $f(\,)$ 为激活函数，$W^{(1)}$，$W^{(2)}$，$b^{(1)}$，$b^{(2)}$ 为网络参数，通过最小化重构损失训练得到。通过这种压缩和重建过程，自编码器能够学习到有效的数据表示，用于数据降维、特征提取和数据压缩等应用。卷积自编码器(Convolutional Autoencoder, CAE)是一种结合了卷积神经网络(Convolutional Neural Network, CNN)和自编码器(Autoencoder, AE)优势的模型。与传统神经网络相比，卷积神经网络能够利用卷积核从一组高度相关的组学特征中学习信息，能够有效捕捉复杂生物信息数据中的局部模式和特征，这有助于提高自编码器的降维效果。所提到的卷积自编码器框架如图 2 所示。

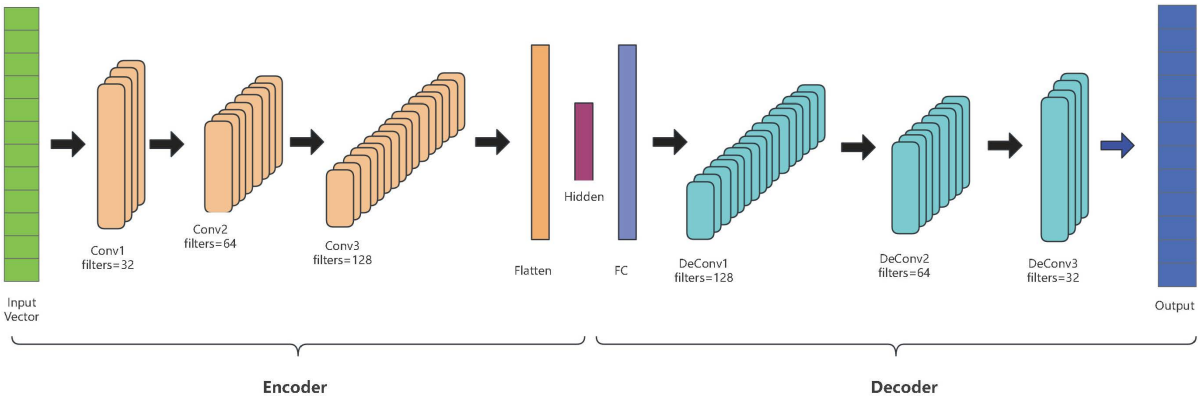

Figure 2. Convolutional autoencoder network architecture
图 2. 卷积自编码器网络结构

处理后的多组学数据通过卷积自编码器进行处理。数据经过多步骤的卷积操作后展平为高维长向量，接着通过全连接层提取潜在因子，并最终通过反卷积操作重构数据。通过重构误差来更新编码器

$z = F_w(x)$ 和解码器 $x' = G_{w'}(z)$ 的参数

$$L_{CAE} = L(x, x') = \frac{1}{n}\sum_{i=1}^{n}\left\| G_{w'}\left(F_w(x_i)\right) - x_i \right\|_2^2 .$$

其中 n 是数据集中的样本数，x_i 是第 i 个样本。

2.4. 基于深度学习方法的多组学整合算法 dmCLCAE

使用自编码器[20]提取的潜在表示虽然包含了样本的关键信息，但在区分样本方面表现不佳。这是因为自编码器主要关注恢复样本细节，而不是学习区分度强的特征。多组学数据中，不同组学之间的一致性信息没有被充分利用，导致潜在表示不能很好地反映多视图数据的综合特性。为了解决上述问题本文提出在模型中引入比学习方法。对比学习是一种自监督学习范式，通过对比正样本对和负样本对的相似性来学习区分度强的表示。正样本对是指同一数据的不同组学，负样本对则是指不同数据的组学。与传统的单组学对比学习依赖数据增强策略构建正样本对不同，本文利用多组学数据，直接将同一样本的不同组学作为正样本对，不同样本的组学作为负样本对。这样能够更好地利用多组学间的一致性信息。具体的，$z^v = \left(z_1^v, z_2^v, \cdots, z_n^v\right)$ 由第 v 个组学数据通过 CAE 得到的隐藏层值，则 $\left(z_i^{v_1}, z_i^{v_2}\right)$ 表示正样本对，$\left(z_i^{v_1}, z_k^{v}\right)$ $\left(i \neq k\right)$ 表示负样本对。首先给出任意一对隐藏层值的 cosin 相似性度量公式

$$S\left(z_i^{v_1}, z_k^{v_2}\right) = \frac{\left(z_i^{v_1}\right)\left(z_k^{v_2}\right)^{\mathrm{T}}}{\left\|z_i^{v_1}\right\|\left\|z_k^{v_2}\right\|},$$

充分利用多组学数据之间的联系，通过迭代将一个组学作为锚点，并从其他组学中枚举正样本对和负样本对。给出 $z_i^{v_1}$ 对比损失为

$$l_i^{v_1} = \sum_{v_1 \neq v_2}^{V} l_i^{v_1, v_2} = \sum_{v_1 \neq v_2}^{v} -\log \frac{\exp\left(s\left(z_i^{v_1}, z_i^{v_2}\right)/\tau\right)}{\sum_{v}^{V}\sum_{k=1}^{N}\exp\left(s\left(z_i^{v_1}, z_k^{v}\right)/\tau\right)},$$

其中 τ 是温度参数用于控制控制模型对相似性和差异性的敏感程度。

定义对比学习损失为

$$L_{CL} = \frac{1}{V * n}\sum_{i=1}^{n}\sum_{v=1}^{V} l_i^v .$$

本文最终提出的模型损失函数表示为

$$L_{Total} = L_{CAE} + L_{CL} .$$

通过使用卷积自编码器，可以将各种组学数据降维到一个预定义的低维空间，从而获得简化且信息丰富的数据表示形式。这种低维表示不仅保留了原始数据中的关键信息，还降低了数据的复杂性，便于后续的特征选择。然后，本文构建单变量 Cox-PH 模型来评估单个因子对生存时间的影响。通过对每个因子分别建模并根据模型生成的 P 值进行筛选，可以识别出与患者生存显著相关的变量。

3. 结论

3.1. 模型分类效果

本研究收集了十二种癌症患者的多组学数据集，以及生存时间和存活状态数据，用以评估 dmCLCAE 的效果。由于这些患者样本缺乏标签，本文通过聚类方法将患者分组，以进行预后分型。为了综合评估模型的聚类准确性和生存预测效果，本文计算了轮廓系数和对数秩检验 P 值两个指标。首先，轮廓系数

是一种评估聚类效果的方法，取值介于[−1, 1]之间，值越大表示聚类效果越好。我们利用 Cox-PH 模型选择的特征进行 K-means 聚类，并在不同的 *K* 值下计算轮廓系数。其次，对数秩检验是一种比较多个生存曲线差异的常用方法，一般情况下，P 值小于 0.05 被认为具有统计显著性。首先，表 1 结合对数秩检验的 P 值和轮廓系数，确定了每种癌症的最佳亚型数目。显示了 dmCLCAE 在不同癌症聚类数目下的表现。结果表明，dmCLCAE 在不同 *K* 值下对每种癌症患者均实现了显著的分型(P 值 ＜0.05)，且轮廓系数均大于 0.1，说明聚类效果良好。最终，综合考虑 P 值和轮廓系数后，确定了每种癌症的最优聚类数目，其中 BLCA、BRCA、GBM、LUAD 和 LUNG 的最佳聚类数目为四个，SARC 最佳聚类数目为三个，其余癌症最佳聚类数目为两个。在此最佳聚类数目下，各种癌症的对数秩检验 P 值均低于 0.001，表明分型效果显著。

Table 1. Number of cancer subtypes identified under dmCLCAE
表 1. dmCLCAE 下确定的十二种癌症亚型数目

癌症	最优亚型数	2		3		4		5	
		P 值	轮廓系数	P 值	轮廓系数	P 值	轮廓系数	P 值	轮廓系数
ACC	2	1.00E−07	0.5207025	3.00E−06	0.39596155	2.05E−06	0.350057	7.11E−06	0.2946729
BLCA	4	0.002309084	0.30950302	0.000108597	0.23487175	8.56E−05	0.2228461	4.90E−05	0.20080127
BRCA	4	0.026639243	0.5462436	0.012729878	0.52440363	0.006938333	0.52509767	0.234111699	0.52573264
COAD	2	6.34E−05	0.30648416	0.001437175	0.25758713	0.000698961	0.25063124	0.000338717	0.22908437
GBM	4	0.001388466	0.38981345	0.000114685	0.3051706	4.66E−05	0.33175394	0.008657981	0.3039649
LUAD	4	0.000389779	0.2780115	2.98E−05	0.18462536	1.19E−06	0.17062436	0.000408619	0.13294406
LUNG	4	0.609864805	0.28362718	0.00958008	0.28988472	0.000263209	0.26026478	0.001169444	0.2337012
LUSC	2	1.90E−05	0.3579545	7.71E−05	0.23511612	0.001282313	0.22646917	0.002225513	0.19889073
PRAD	2	0.02430841	0.39426908	0.031824376	0.41576302	0.237154723	0.36427942	0.373457796	0.37346262
SARC	3	0.004447248	0.28042555	0.001316286	0.2588831	0.003227221	0.27802154	0.007532482	0.2612162
STAD	2	0.000308817	0.3048487	0.000386375	0.29102576	0.000215493	0.2482764	0.000832402	0.23739746
THCA	2	0.000368239	0.33038157	0.007984258	0.2612367	0.004741932	0.19961411	0.005716601	0.21560447

为了更直观地评估聚类效果，本文引入了 Kaplan-Meier 生存曲线(简称 K-M 曲线)。K-M 曲线是一种在生存分析中广泛应用的非参数方法，用于估算在特定时间点上的生存概率。该方法适用于观察性数据，即研究对象在不同时间点进行观察，但不一定每个对象都被观测到整个研究期间。由于 K-M 方法不依赖于对生存时间分布的假设，因此适用于各种生存时间分布情况。曲线上的每个阶梯表示事件(如死亡或复发)发生的概率，常用于比较不同治疗组的生存情况、评估患者预后和研究预后因素的影响。它为理解患者群体的生存状况提供了一种直观的方式，并可以估计中位生存时间和特定时间点的生存率。图 3 展示了 dmCLCAE 聚类后进行生存分析的结果。

为了比较所提出模型的优越性，本文还应用了 MOFA+和 ProgCAE 模型对不同类型的癌症进行分类。以 dmCLCAE 选择的最佳聚类数目为基础，我们通过 MOFA+进行了聚类分析。在对 12 种不同类型的癌症进行分析后，发现 MOFA+在 ACC、GBM 和 LUAD 三种癌症类型中得到的生存亚型存在显著差异(P 值小于 0.05)。相比之下，ProgCAE 在不同聚类数目下的表现来看，其在 BRCA 癌症患者和 PRAD 癌症患者的 $K = 2$ 和 $K = 3$ 时差异不显著(P 值大于 0.05)。

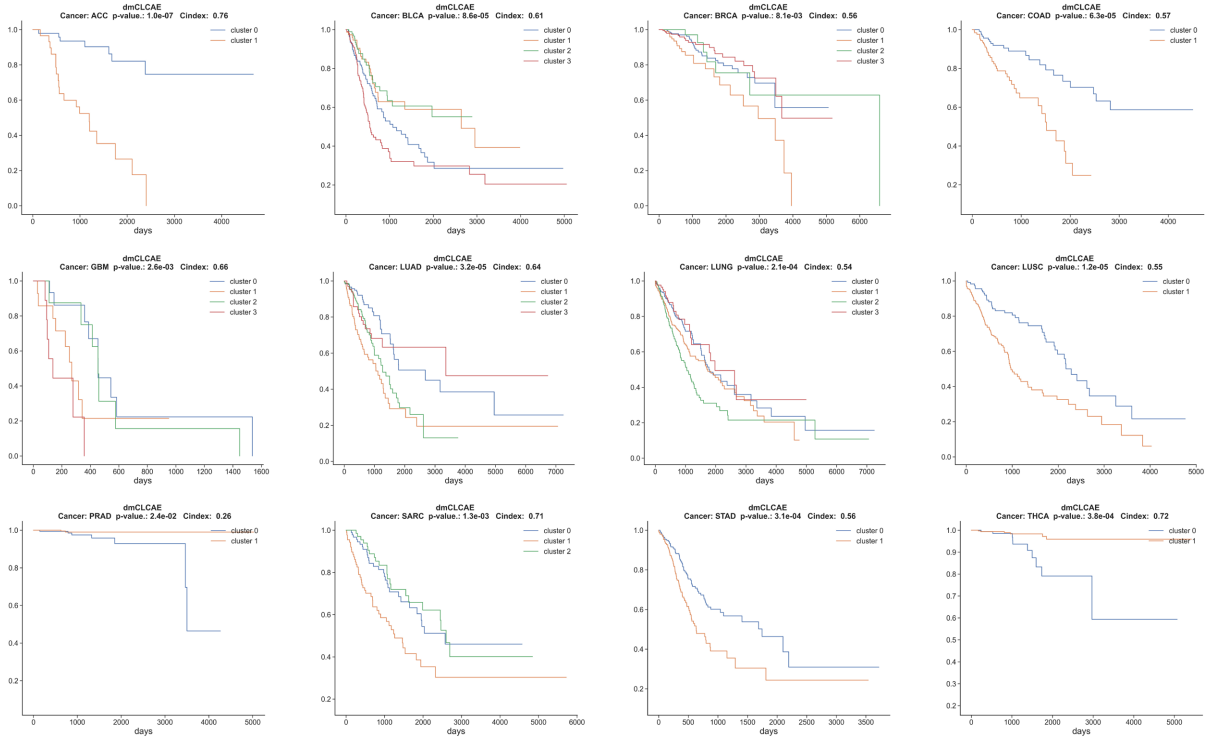

Figure 3. Kaplan-Meier curves for 12 types of cancer under dmCLCAE
图 3. dmCLCAE 的 12 种癌症的 Kalpan-Meier 曲线

3.2. 模型预测一致性

为了比较不同模型在预后价值上的表现，并评估 DmCLCAE 和 ProgCAE 在聚类效果上的差异，本文通过 Cox-PH 模型计算了一致性指数(C-index)。此外，为了提高结果的稳健性，我们采用了留出法和 5 折交叉验证。表 2 展示了 DmCLCAE 和 ProgCAE 的 C-index 对比结果。

Table 2. Comparison of C-index between DmCLCAE and ProgCAE
表 2. DmCLCAE 和 ProgCAE 下的 C-index 对比

癌症	dmCLCAE	ProgCAE
ACC	0.7573	0.7681
BLCA	0.6089	0.6333
BRCA	0.5601	0.5179
COAD	0.5707	0.5580
GBM	0.6633	0.7272
LUAD	0.6396	0.6193
LUNG	0.5402	0.5221
LUSC	0.5460	0.5023
PRAD	0.2644	0.4761
SARC	0.7097	0.6838
STAD	0.5567	0.5369
THCA	0.7176	0.6137

因此，综合上述结果 DmCLCAE 在区分不同生存亚型方面，结果显示出更为显著的差异，并且在预测一致性方面也表现更佳。如表 2 所示，在 12 种分析的癌症类型中，有 8 种显示出 DmCLCAE 具有更高的一致性指数。

4. 讨论

癌症是一种恶性疾病，在研究中准确地识别其亚型和预测患者的预后至关重要。伴随着大量组学数据的产生，研究人员获得了宝贵的资源。多组学数据的整合有助于减少来自不同平台的噪声，从而获得一致的生物信号，并揭示关键的生物学机制。然而，处理多组学数据中的高维生物数据仍是一个重大挑战，特别是当分析涉及数以千计的单核苷酸多态性、基因和蛋白质时，这一问题尤为突出。

为了应对多组学数据整合所遇到的挑战，本研究提出了一种基于卷积自编码器的深度多视图对比学习模型——dmCLCAE，该模型利用 CAE [19]来整合多组学数据，并通过将重构损失和对比损失统一在一个框架中进行预后预测。我们的模型不仅能够将样本的区分信息编码到提取的特征表示中，还能在嵌入空间中有效保持样本的聚类结构。与传统方法相比，这种方法结合了卷积神经网络和自编码器与对比学习的优势，能够处理高维数据并捕捉生物特征之间的相关性，从而获得更具表达力的潜在表示。此外，本研究利用患者的生存信息构建了单变量 Cox 比例风险回归模型，从潜在因子中筛选出具有统计意义的生存特征，并使用 K-means 算法将样本根据关键特征聚类成不同的生存亚组。

本研究在 12 个 TCGA 癌症组学数据集上应用了 dmCLCAE，发现它在预测癌症预后方面表现出显著优势。dmCLCAE 通过提取 RNA 表达、拷贝数变异、miRNA 表达和 DNA 甲基化数据中的隐藏特征，捕捉到了传统线性无监督方法难以检测的非线性关系，这些关系可能反映了重要的生物过程。通过生存分析筛选的特征能够有效地评估癌症预后。与其他方法相比，dmCLCAE 的预测结果更为一致和稳健，倾向于识别更多的生存特征。此外，利用 dmCLCAE 得到的聚类标签可以用于构建监督分类器，从而扩大其在生物学应用中的潜力。

参考文献

[1] Conesa, A. and Beck, S. (2019) Making Multi-Omics Data Accessible to Researchers. *Scientific Data*, **6**, Article No. 251. https://doi.org/10.1038/s41597-019-0258-4

[2] Hasin, Y., Seldin, M. and Lusis, A. (2017) Multi-Omics Approaches to Disease. *Genome Biology*, **18**, Article No. 83. https://doi.org/10.1186/s13059-017-1215-1

[3] Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., *et al*. (2013) The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature Genetics*, **45**, 1113-1120. https://doi.org/10.1038/ng.2764

[4] Alameer, A. and Chicco, D. (2021) Geocancerprognosticdatasetsretriever: A Bioinformatics Tool to Easily Identify Cancer Prognostic Datasets on Gene Expression Omnibus (GEO). *Bioinformatics*, **38**, 1761-1763. https://doi.org/10.1093/bioinformatics/btab852

[5] Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., *et al*. (2011) International Cancer Genome Consortium Data Portal—A One-Stop Shop for Cancer Genomics Data. *Database*, **2011**, bar026. https://doi.org/10.1093/database/bar026

[6] Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., *et al*. (2003) Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets. *Proceedings of the National Academy of Sciences*, **100**, 8418-8423. https://doi.org/10.1073/pnas.0932692100

[7] Cabassi, A. and Kirk, P.D.W. (2020) Multiple Kernel Learning for Integrative Consensus Clustering of Omic Datasets. *Bioinformatics*, **36**, 4789-4796. https://doi.org/10.1093/bioinformatics/btaa593

[8] Nguyen, N.D. and Wang, D. (2020) Multiview Learning for Understanding Functional Multiomics. *PLOS Computational Biology*, **16**, e1007677. https://doi.org/10.1371/journal.pcbi.1007677

[9] Trunk, G.V. (1979) A Problem of Dimensionality: A Simple Example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 306-307. https://doi.org/10.1109/tpami.1979.4766926

[10] Rappoport, N. and Shamir, R. (2018) Multi-Omic and Multi-View Clustering Algorithms: Review and Cancer Benchmark. *Nucleic Acids Research*, **46**, 10546-10562. https://doi.org/10.1093/nar/gky889

[11] Reel, P.S., Reel, S., Pearson, E., Trucco, E. and Jefferson, E. (2021) Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnology Advances*, **49**, Article 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

[12] Springenberg, J.T., Dosovitskiy, A., Brox, T. and Riedmiller, M. (2014) Striving for Simplicity: The All Convolutional Net.

[13] Chauhan, R., Ghanshala, K.K. and Joshi, R.C. (2018). Convolutional Neural Network (CNN) for Image Detection and Recognition. 2018 *First International Conference on Secure Cyber Computing and Communication* (*ICSCCC*), Jalandhar, 15-17 December 2018, 278-282. https://doi.org/10.1109/icsccc.2018.8703316

[14] Sun, W., Zheng, B. and Qian, W. (2016). Computer Aided Lung Cancer Diagnosis with Deep Learning Algorithms. *SPIE Proceedings*, San Diego, California, 24 March 2016, 97850Z. https://doi.org/10.1117/12.2216307

[15] Masci, J., Meier, U., Cireşan, D. and Schmidhuber, J. (2011) Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In: Honkela, T., Duch, W., Girolami, M. and Kaski, S., Eds., *Artificial Neural Networks and Machine Learning—ICANN* 2011, Springer, 52-59. https://doi.org/10.1007/978-3-642-21735-7_7

[16] Tian, Y., Krishnan, D. and Isola, P. (2020) Contrastive Multiview Coding. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.-M., Eds., *Computer Vision—ECCV* 2020, Springer, 776-794. https://doi.org/10.1007/978-3-030-58621-8_45

[17] Oord, A.V.D., Li, Y. and Vinyals, O. (2018) Representation Learning with Contrastive Predictive Coding.

[18] 胡深, 钱宇华, 王婕婷, 李飞江, 吕维. 基于对比学习的超多类深度图像聚类模型[J]. 计算机科学, 2023, 50(9): 192-201.

[19] Poirion, O.B., Jing, Z., Chaudhary, K., Huang, S. and Garmire, L.X. (2021) Deepprog: An Ensemble of Deep-Learning and Machine-Learning Models for Prognosis Prediction Using Multi-Omics Data. *Genome Medicine*, **13**, Article No. 112. https://doi.org/10.1186/s13073-021-00930-x

[20] Liu, Q. and Song, K. (2023) Progcae: A Deep Learning-Based Method That Integrates Multi-Omics Data to Predict Cancer Subtypes. *Briefings in Bioinformatics*, **24**, bbad196. https://doi.org/10.1093/bib/bbad196