第 40 卷　第 6 期
2025 年 12 月

天津科技大学学报
Journal of Tianjin University of Science and Technology

Vol. 40　No. 6
Dec. 2025

DOI:10.13364/j.issn.1672-6510.20240105
网络首发日期：2025–07–28；网络首发地址：http://link.cnki.net/urlid/12.1355.N.20250728.1348.003

基于生成模型和基因表达数据的关键基因筛选

余　钱[1,2]，李雨蒙[2]，罗军伟[3]，董浩帆[1,2]，李　玉[1]，吴　信[2]

（1. 天津科技大学生物工程学院，天津 300457；2. 中国科学院天津工业生物技术研究所，天津 300308；
3. 河南理工大学软件学院，焦作 454003）

摘　要：基因表达数据可以在特定条件和时间下揭示疾病的病理机制，然而“维数灾难”，也就是小样本、高维度，限制了传统机器学习分类方法的效果，导致预测精度低、无法识别小样本和稳定性差等问题。本文结合数据增强和基因选择两个方面提出了一种新的方法，命名为 CVAE-CWGNA-DAE，尝试解决由“维数灾难”带来的问题。针对基因表达数据中存在的小样本问题，提出基于条件变分自编码器结合基于梯度惩罚的条件 Wasserstein 生成对抗网络的数据增强方法，通过与现有方法的比较，证明该方法在分类效果和稳定性上的优越性。为了解决基因表达中存在的高维度问题，同时为了验证生成数据的有效性，采用基于降噪自编码器和支持向量机递归特征消除（SVM-RFE）的基因选择方法。结果表明：利用数据增强后的数据集进行基因选择，所选出的基因在分类任务上的准确率在 5 种不同分类上均得到了提升。这些结果证明本文方法在缓解“维数灾难”方面的有效性，并在基因选择方面取得了显著的改进。

关键词：基因表达；维数灾难；数据增强；基因选择；自编码器；生成对抗网络

中图分类号：TP391；Q78　　**文献标志码**：A　　**文章编号**：1672-6510（2025）06-0001-08

Key Gene Screening Based on Generative Models and Gene Expression Data

YU Qian[1,2]，LI Yumeng[2]，LUO Junwei[3]，DONG Haofan[1,2]，LI Yu[1]，WU Xin[2]

（1. College of Biotechnology，Tianjin University of Science and Technology，Tianjin 300457，China；
2. Tianjin Institute of Industrial Biotechnology，Chinese Academy of Sciences，Tianjin 300308，China；
3. School of Software，Henan Polytechnic University，Jiaozuo 454003，China）

Abstract：Gene expression data can elucidate the pathological mechanisms of diseases under specific conditions and times. However，the “curse of dimensionality”phenomenon characterised by small samples and high dimensions，constrains the performance of traditional machine learning classification methods. This results in low prediction accuracy，an inability to recognise small samples，and poor stability. This article introduces a novel method，namely CVAE-CWGNA-DAE，which integrates data augmentation and gene selection in order to address the issues that arise from the “curse of dimensionality”. Firstly，in order to address the issue of the small sample size in gene expression data，a data augmentation method is proposed，which combines a conditional variational autoencoder with a gradient penalty-based conditional Wasserstein generative adversarial network. A comparison with existing methods demonstrates the superiority of this approach in terms of classification performance and stability. Secondly，to address the high dimensionality in gene expression data and verify the effectiveness of the generated data，this article employs a gene selection method based on a denoising autoencoder and SVM-RFE. The results reveal that the use of the augmented dataset for gene selection has resulted in an improvement in the accuracy of selected genes across five distinct classification tasks. Therefore，these results demonstrate the effectiveness of the proposed method in addressing the “curse of dimensionality” and achieving significant improvements in gene selection.

Keywords：gene expression；curse of dimensionality；data augmentation；gene selection；autoencoder；generative

收稿日期：2024–05–13；**修回日期**：2024–08–24
基金项目：国家自然科学基金资助项目（62372156）
作者简介：余　钱（1996—），女，湖南平江人，硕士研究生；**通信作者**：吴　信，研究员，wuxin@tib.cas.cn

adversarial network

引文格式：
余钱，李雨蒙，罗军伟，等. 基于生成模型和基因表达数据的关键基因筛选[J]. 天津科技大学学报，2025，40(6)：1-8.
YU Q，LI Y M，LUO J W，et al. Key gene screening based on generative models and gene expression data[J]. Journal of Tianjin university of science and technology，2025，40(6)：1-8.

基因表达是细胞对疾病状态和遗传变化的反应，通过调控特定基因的激活或抑制，实现特定指令的动态获取和转译[1-2]。RNA 转录组作为基因表达的重要产物，反映了细胞的当前状态，并有助于揭示疾病的病理机制。随着高通量测序技术的发展，大规模生物医学数据不断涌现，包括分子化合物结构、高通量测序数据、医学影像和电子健康记录等[3-8]。然而，由高通量测序产生的基因表达数据受限于伦理和实验挑战，导致数据的规模、种类和收集速度受限，从复杂的生物医学数据中提取有价值的信息仍是研究的关键[9]。基因表达数据包含数十万个基因，但样本量相对较小，存在“维数灾难”的问题[10]。在这数十万个基因中，真正和疾病发生过程有关的也只有少数基因[11]。足够的样本量可以确保基因表达数据分类结果的准确性和可靠性[12]。因此，如何有效增加样本量并筛选出与疾病发生过程相关的关键基因至关重要。

近年来，深度学习在生物信息学领域的应用逐渐增多，其中生成对抗网络(GAN)和变分自编码器(VAE)在基因表达数据分析中显示出巨大潜力[9,13]。这些生成模型能够合成几乎无限量的人工基因表达数据，从而缓解“维数灾难”问题。然而，在生物医学领域，特别是转录组测序技术(RNA-Seq)，利用 GAN 和 VAE 进行数据生成和增强的研究报道仍然较少。Wang 等[12]证明 GAN 在生成大量 RNA-Seq 数据方面优于扩散模型。Lacan 等[14]也证明了在 RNA-Seq 数据条件下进行数据增强的可行性。Ahmed 等[15]提出一种整合多组学数据和 GAN 的模型，该模型增强了合成数据中癌症结果的预测信号。Viñas 等[16]利用条件 GAN 生成保留组织和癌症特异性的真实转录组学数据。Marouf 等[17]开发 cscGAN 模型，生成真实的单细胞 RNA-seq 数据，增强了标记基因检测和分类器的可靠性。Dincer 等[18]利用基于 VAE 的模型从公共未标记的基因表达数据中提取低维特征，并且证明了这些低维特征表示的有效性。Kim 等[19]引入了一种基于 VAE 的生存预测模型，可以提取用于患者生存预测的基因显著特征。Bica 等[20]提出一种基于 VAE 的方法，该方法可以利用复杂、高维的基因表达数据，构建低维、有意义的表示，用来模拟细胞分化。Yu 等[21]结合 VAE 和 GAN 建立 MichiGAN 模型，该模型利用 VAE 捕捉底层数据分布，GAN 生成高质量数据，有效缓解了数据维度高的问题，其在基因表达数据生成中具有独特优势。

高效的基因选择方法对于识别疾病中的关键候选基因至关重要。目前，这些研究主要通过设计有效的分类算法或改进降维算法解决基因表达数据中的维度问题。Almutiri 等[22]提出一种卡方检验与支持向量机递归特征消除(SVM-RFE)相结合的基因选择方法，该方法在 11 个高维微阵列数据集上进行测试，在人工神经网络(ANN)中达到了最高精度。Danaee 等[23]提出堆叠降噪自编码器(SDAE)模型，该模型可以从高维基因表达数据中有效提取功能基因。Liu 等[24]使用去噪自编码器扩展基因表达数据样本，采用无限特征选择(IF-FS)对扩展数据进行基因选择，使用堆叠自编码器模型可以对结肠癌、乳腺癌和白血病的高维基因表达数据进一步分类。使用基因编码器模型、主成分分析、相关性和谱基因选择方法选择基因的初始水平，应用基于自编码器聚类的遗传算法对染色体进行评估，通过支持向量机、k 近邻算法(KNN)和随机森林算法对得到的基因子集进行分类，可以评估 6 个基准基因表达数据集上的性能[25]。Famitha 等[26]使用 PCA 降低特征空间的维度，然后将结果作为压缩特征提取应用于传统或多层稀疏自动编码器，用来发现在分类过程中使用信息的有限插图。

利用生成式深度学习模型进行数据增强，以应对大量 RNA-Seq 基因表达数据的稀缺性问题，并进一步解决特征基因的降维问题。这种方法可以提高基因表达数据分析的效率和准确性，为生物信息学研究提供新的视角和解决方案。因此，本文提出结合数据增强和基因选择的方法，命名为 CVAE-CWGNA-DAE。首先，采用基于条件变分自编码器(CVAE)和基于梯度惩罚的条件 Wasserstein 生成对抗网络(CWGAN-GP)对基因表达数据进行扩增，增加数据的样本量，解决基因表达的小样本问题。其次，采用

基于降噪自编码器(DAE)和 SVM-RFE 的基因选择方法，对基因表达数据进行降维，筛选出与疾病过程相关的关键基因，解决基因表达维度高的问题。

1　数据与方法

1.1　总体设计

CVAE-CWGNA-DAE 的总体设计如图 1 所示。首先，利用 CVAE 和 CWGAN-GP 进行数据增强，对基因表达数据进行扩增，增加数据中的样本量，从而解决基因表达数据中样本小的问题。然后，通过 DAE 和 SVM-RFE 的基因选择方法对基因表达数据进行降维，筛选出与疾病过程相关的关键基因，从而解决基因表达数据维度高的问题。

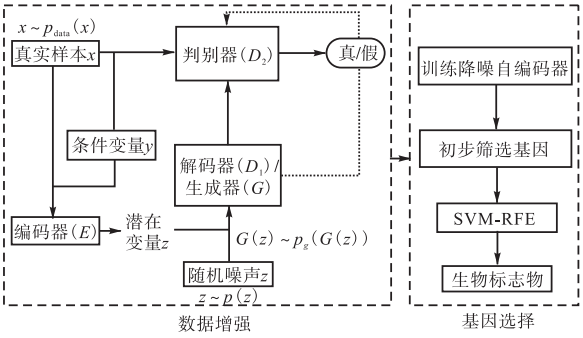


图 1　CVAE-CWGNA-DAE的总体设计

Fig. 1　Overall design of CVAE-CWGNA-DAE model

1.2　实验环境

Ubuntu/Window 11 操作系统，Python 3.7/R 4.3 开发语言，TensorFlow-GPU 框架，GPU 为 NVIDIA GeForce GTX 3080，开发环境为 Anaconda/PyCharm，内存为 DDR4-2666 ECC 16 GB × 12。

1.3　数据增强方法

为了将 CVAE 和 CWGAN-GP 结合起来，针对 CVAE 引入编码器(E)和解码器(D_1)，E 通过学习 $q(x;y)$ 将样本 x 映射到潜在变量 $z \sim N\left(\mu(x;y), \sum(x;y)\right)$，其中 y 是标签条件，使用重参数化技巧进行采样，$z_x = \mu(x;y) + \sum(x;y) \odot \varepsilon$。$D_1$ 通过输入条件 y 和从 E 中获取的 z_x 尝试重构 $\tilde{x} = p(z_x;y)$。针对 CWGAN-GP，其生成器(G)与 CVAE 的 D_1 共享相同参数，利用 CWGAN-GP 中的鉴别器(D_2)拟合学习原始样本和生成样本之间的差异，更好地了解真实样本分布。方法优化表示为

$$L_{\text{CVAE-CWGAN-GP}} = L_{\text{CVAE}} + L_{\text{CWGAN-GP}} \tag{1}$$

其中：$L_{\text{CVAE-CWGAN-GP}}$ 为本文方法的损失函数，L_{CVAE} 为 CVAE 的损失函数，$L_{\text{CWGAN-GP}}$ 为 CWGAN-GP 的损失函数。

在 CVAE 中，新样本由 D_1 根据潜变量 z_x 的描述生成，它使用变分推理找到一个易于处理的条件概率分布 $q(z|x,y)$ 近似真实的后验概率 $p(z|x,y)$，用 KL 散度计算 $q(z|x,y)$ 和 $p(z|x,y)$ 之间的相似度，为

$$\begin{aligned} &D_{\text{KL}}\left[q(z|x,y)\|p(z|x,y)\right] = \\ &\quad E_{z\sim q}\left[\log q(z|x,y) - \log p(z|x,y)\right] \end{aligned} \tag{2}$$

CVAE 网络的目标就是最大化 $p(x)$，同时最小化 $D_{\text{KL}}\left[q(z|x,y)\|p(z|x,y)\right]$。因此，CVAE 的损失函数为

$$L_{\text{CVAE}} = L_{\text{rec}} + D_{\text{KL}}\left[q(z|x,y)\|p(z|x,y)\right] \tag{3}$$

其中：D_{KL} 为 KL 散度，L_{rec} 为重构数据 $\tilde{x}$ 与原始数据样本 x 之间的差异，本节中采用均方误差，如式(4)所示，L_{CVAE} 为 CVAE 的损失函数。

$$L_{\text{rec}} = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \tilde{x}_j\right)^2 \tag{4}$$

在 CWGAN-GP 中，以噪声样本 $z \sim p(z)$(从均匀分布或正态分布中采样)作为输入，生成网络 G 输出新数据 $G(z)$，其分布 p_z 应该接近于数据分布 p_x。同时，判别网络 D 用于区分真实数据样本 $x \sim p_x$ 和生成样本 $G(z) \sim p_g$。

$$L_{\text{CQWGAN-GP}} = L_G + L_{D_2} \tag{5}$$

$$L_G = L_{\text{rec}} \tag{6}$$

$$\begin{aligned} L_{D_2} = &-E_{z\sim p(z)}\left[D(x|y)\right] + E_{z\sim p_x}\left[D(x|y)\right] + \\ &\lambda E_{\hat{x}\sim p(\hat{x})}\left[\left(\left\|\nabla_{\hat{x}} D(\hat{x}|y)\right\|_2 - 1\right)^2\right] \end{aligned} \tag{7}$$

其中：$\tilde{x}$ 表示 CVAE 的重构数据；λ 约束惩罚项系数，仅在更新参数时使用；z 为随机噪声，$\hat{x}$ 表示从真实分布和生成分布之间的分布取出的样本。$\nabla_{\hat{x}} D(\hat{x})$ 表示样本 $\hat{x}$ 作为判别模型 D 输入时的梯度。目标函数为

$$\min_G \max_D L_{\text{CVAE}} + L_{\text{CWGAN-GP}} \tag{8}$$

综上所述，CVAE-CWGAN的算法描述见算法1。

算法 1：CVAE-CWGAN-GP

输入：原始真实数据 $X = (x_1, x_2, x_3, \cdots, x_n)$，条件变量 y；学习率 lr；鉴定器惩罚 λ；批量训练 bs；鉴定器模型训练次数 critic；训练迭代次数 Epochs。

输出：生成样本集 G。

1. 采用 Z-score 标准化和 one-hot 分别对输入的 X 和 y 进行预处理；
2. 初始迭代次数 epoch = 0；

3. 当 epoch ≤ Epochs：
4. 真实数据中选取 bs 个样本 $x=(x_i,y_i)$；
5. 训练 VAE，重构样本 $x \leqslant \hat{x}$；
6. 计算 $L_{\mathrm{CVAE}}=\frac{1}{n}\sum_{j=1}^{n}\left(x_j-\hat{x}_j\right)^2+D_{\mathrm{KL}}\left[q(z|x,y)\|p(z|x,y)\right]$，更新参数 θ_E 和 θ_{D_2}（θ_E 表示 VAE 中编码器（E）的参数，θ_{D_2} 表示 VAE 中解码器（D）的参数）；
7. 结束循环；
8. 设置 VAE 中编码器（E）的参数不变，将 VAE 中的解码器作为生成对抗网络的生成器；
9. 初始迭代次数 epoch = 0；
10. 当 epoch ≤ Epochs：
11. 从真实数据中选取 bs 个样本 $x=(x_j,y_j)$，生成 bs 个随机分布 $z=(z_1,z_2,z_3,\cdots,z_n)$；
12. 初始化鉴定器训练次数 $c=0$；
13. 训练鉴定器，当 $c \leqslant$ critic
14. 输入 z 使生成器（D_2）生成虚假数据 $\hat{x}$；
15. 计算 $L_D=-E_{z\sim p(z)}\left[D(x|y)\right]+E_{z\sim p_x}\left[D(x|y)\right]+\lambda E_{\hat{x}\sim p(\hat{x})}\cdot\left[\left(\left\|\nabla_{\hat{x}}D(\hat{x}|y)\right\|_2-1\right)^2\right]$，更新参数 θ_G（θ_G 表示生成对抗网络鉴定器（G）的参数）；
16. 训练生成器：
17. 输入 z 使生成器（D_2）生成虚假数据 $\tilde{x}$；
18. 计算 $L_G=\frac{1}{n}\sum_{j=1}^{n}\left(x_j-\tilde{x}_j\right)^2$，更新参数 θ_{D_2}（θ_{D_2} 表示生成对抗网络生成器（D_2）的参数）；
19. 结束循环。

1.4　基因选择方法

为了能够有效去除基因表达中的冗余基因，得到关键候选基因，采用结合 DAE 和 SVM-RFE 方法进行基因选择。

首先，建立 DAE 模型，为

$$y=f\left(\sum_{j=1}^{n}w_{ij}x_i+b\right) \tag{9}$$

在整个网络中采用 ReLU 作为激活函数，得到 $f(x)>0$。通过从数据集中按类型筛选基因表达数据，为每个类别训练单独的 DAE，重建给定的输入，采用均方误差衡量原始数据 x 和重构数据 $\tilde{x}$ 的差异，为

$$L_{\mathrm{DAE}}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\tilde{x}_i\right)^2 \tag{10}$$

通过特定类型的训练，得到每个类别的最佳训练模型，计算其权值影响（用 I_j 表示），为

$$I_j=\frac{w_j\cdot x_j}{\sum_{i=1}^{n}w_{ij}x_i+b}\times 100 \tag{11}$$

最终通过 DAE 得到初步筛选的基因子集，该基因子集进一步利用 SVM-RFE 进行筛选，得到最终的关键候选基因。

DAE-SVM-RFE 的算法见算法 2。

算法 2：DAE-SVM-RFE

输入：原始真实数据 $X=(x_1,x_2,x_3,\cdots,x_n)$，条件变量 y；KFold 划分 K。

输出：特征选择集 R。

1. 建立 DAE 的模型 $y=f\left(\sum_{i=1}^{n}w_{ij}x_i+b\right)$ 和激活函数 ReLU；
2. 当 $k \leqslant K$：
3. 训练各分类 DAE；
4. 保存预测最好的模型；
5. 读取模型；
6. 计算各个分类的影响因子：$I_j=\frac{w_j\cdot x_j}{\sum_{i=1}^{n}w_{ij}x_i+b}\times 100$，排序；
7. 选取影响前 1% 的特征基因；
8. 训练 SVM-RFE，得到特征基因集 R；
9. 返回特征选择集 R。

2　实验结果与分析

2.1　数据集与数据预处理

实验数据来源于癌症基因组图谱（TCGA）下载的基因表达数据。采用 Z-score 对基因表达数据进行标准化，如式（12）所示。为解决数据维度高的问题，除去基因表达中的线粒体基因，删除基因中基因表达量为 0 的样本占全部样本比值大于 1% 的基因。

$$x_i=\frac{x_i-\mu_i}{\sigma_i} \tag{12}$$

其中：x_i 表示样本 x 的第 i 维度，μ_i 和 σ_i 分别表示第 i 维度的均值和标准差。数据来源以及预处理见表 1，数据集的详细信息见表 2。

表 1　数据来源以及预处理

Tab. 1　Data sources and pre-processing

数据集	类别量	样本量	基因总数	过滤后的基因总数
Dataset	9	1 621	19 924	15 642
LIHC	2	424	19 924	16 482
PRAD	2	554	19 924	17 003
THCA	2	572	19 924	16 714

2.2　数据增强的参数设置

CVAE 参数设置：编码层为[1024，512，512，512，256]，解码层为[256，512，512，512，1024]，激活函数为 LeakyReLU，丢失率 dropout 为 0.3，迭代次数 Epochs 为 10000，学习率 lr 为 1×10^{-4}，优化器 optimizer 为 Adam，LeakyReLU 的 α 为 0.3，批量样

本 batch size 为 32。

表 2　数据集的详细信息

Tab. 2　Datasets details

癌症分类	样本总数	正样本	负样本
ACC（肾上腺皮质癌）	79	0	79
DLBC（弥漫性大 B 细胞淋巴瘤）	48	0	48
LAML（急性骨髓细胞样白血病）	151	0	151
LGG（低级别脑胶质瘤）	534	0	534
OV（卵巢浆液性囊腺癌）	429	0	429
TGCT（睾丸癌）	156	0	156
UCS（子宫肉瘤）	57	0	57
UVM（葡萄膜黑色素瘤）	80	0	80
MESO（间皮瘤）	87	0	87
LIHC（肝细胞肝癌）	424	50	374
PRAD（前列腺癌）	554	52	502
THCA（甲状腺癌）	572	59	513

CWGAN-GP 参数设置：判别器为[1024，512，512，512，256]，激活函数为 ReLU，学习率 lr 为 1×10^{-5}，优化器 optimizer 为 RMSprop，批量样本 batch size 为 32，约束惩罚项系数 λ 为 10，判别器训练次数 critic 为 2。

2.3　数据增强的结果分析

为了验证本文模型的优越性，实验采用模型 SMOTE[27]、Gene-CWGAN[28]与本文模型进行比较，并且使用 k 近邻算法（KNN）和决策树（DT）两种分类算法验证模型，通过五折交叉验证提高分类的可靠性。以准确率（用“平均值 ± 标准差”表示）、精确率、召回率和 F_1 作为样本分类稳定性和质量的指标，不同数据增强算法的准确性比较结果见表 3。

表 3　不同数据增强算法的准确性比较

Tab. 3　Comparison of accuracy among different data augmentation algorithms

数据集	模型	DT				KNN			
		精确率	召回率	F_1	准确率/%	精确率	召回率	F_1	准确率/%
LIHC	原始模型	0.91	0.94	0.93	97.41 ± 1.73	0.79	0.96	0.84	96.23 ± 2.28
	SMOTE	0.98	0.98	0.98	98.26 ± 1.24	0.89	0.88	0.87	91.71 ± 1.71
	Gene-CWGAN	0.98	0.98	0.98	98.65 ± 0.34	0.91	0.92	0.91	92.65 ± 0.83
	本文模型	0.99	0.99	0.99	99.45 ± 0.33	0.98	0.98	0.98	98.62 ± 0.47
PRAD	原始模型	0.65	0.61	0.62	92.42 ± 1.34	0.97	0.71	0.78	95.31 ± 0.38
	SMOTE	0.95	0.95	0.95	97.31 ± 0.92	0.91	0.91	0.90	90.64 ± 2.07
	Gene-CWGAN	0.96	0.96	0.96	97.34 ± 0.71	0.96	0.97	0.97	96.94 ± 1.00
	本文模型	0.98	0.98	0.98	98.63 ± 0.65	0.98	0.98	0.98	98.87 ± 0.64
THCA	原始模型	0.89	0.89	0.89	95.62 ± 1.84	0.85	0.94	0.89	95.98 ± 1.19
	SMOTE	0.99	0.99	0.99	97.47 ± 0.84	0.95	0.95	0.95	94.83 ± 1.09
	Gene-CWGAN	0.98	0.98	0.98	98.41 ± 0.45	0.99	0.99	0.99	99.27 ± 0.17
	本文模型	0.99	0.99	0.99	98.94 ± 0.22	0.99	0.99	0.99	98.88 ± 0.48
Dataset	原始模型	0.95	0.94	0.94	98.27 ± 0.54	0.98	0.94	0.95	98.27 ± 0.54
	SMOTE	0.99	0.99	0.99	99.29 ± 0.10	1.00	1.00	1.00	99.75 ± 0.12
	Gene-CWGAN	1.00	1.00	1.00	99.28 ± 0.19	1.00	1.00	1.00	99.76 ± 0.09
	本文模型	0.99	0.99	0.99	99.48 ± 0.09	1.00	1.00	1.00	99.84 ± 0.10

由表 3 可知，生成模型和 SMOTE 都能够解决小样本问题，进而提高分类准确性。本文模型在 DT 分类上，与原始模型相比，精确率、召回率和 F_1 指标均有所提升，在稳定性上也具有很好的优势。

在同样条件下，将 CVAE、CVAE-CGAN、CVAE-WGAN-GP（与本文模型在同样条件下添加了高斯噪声）与本文模型进行比对，分类算法采用多层感知机（MLP）和 KNN。不同数据增强算法的准确性比较结果见表 4。在 Dataset 数据集上，本文模型的准确率均为最高，与 CVAE-CGAN 相比，MLP 和 KNN 两种算法分别在准确率的均值上提升了 4.16% 和 5.71%。在 LIHC 上，CVAE 比本文模型在 MLP 算法上高 0.14%，但是在其他模型上本文模型的准确率均有所提升。

2.4　基因选择的参数设置与结果分析

DAE 参数设置：编码层为[256，32]，解码层为[32，256]，激活函数为 ReLU，迭代次数 Epochs 为 5 000，学习率 lr 为 0.01，优化器 optimizer 为 Adam，批量样本 batch size 为 128。

采用 CVAE-CWGAN-GP 算法生成基因表达数据，并通过随机筛选的方式，从生成数据的每个类别中随机选择 100 个样本，将其合并到已有的数据中，形成了新的数据集 Dataset1 作为训练数据集，数据集分布见表 5。所有癌症合集的基因选择分类准确率比较如图 2 所示，可以看出原始数据集 Dataset 和混合之后的数据集 Dataset1 都取得了很好的分类效果。

表 4　不同数据增强算法的准确率比较

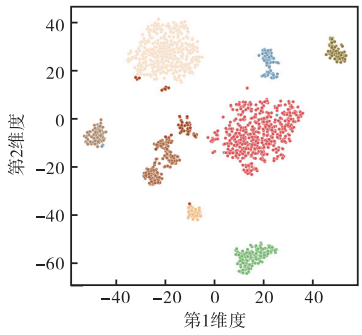
Tab. 4　Comparison of accuracy among different data augmentation algorithms

数据集	模型	MLP 准确率/%	KNN 准确率/%
LIHC	CVAE	99.59 ± 8.63	97.94 ± 3.18
	CVAE-CGAN	86.01 ± 28.63	96.70 ± 0.55
	CVAE-WGAN-GP	97.90 ± 5.76	98.97 ± 1.70
	本文模型	99.45 ± 0.11	98.62 ± 0.47
PRAD	CVAE	97.50 ± 5.20	96.45 ± 5.37
	CVAE-CGAN	97.49 ± 1.75	97.36 ± 1.09
	CVAE-WGAN-GP	97.81 ± 4.36	98.59 ± 2.16
	本文模型	99.07 ± 0.47	98.87 ± 0.64
THCA	CVAE	99.11 ± 1.40	98.25 ± 2.33
	CVAE-CGAN	97.27 ± 1.55	96.44 ± 0.99
	CVAE-WGAN-GP	97.79 ± 5.74	98.52 ± 1.92
	本文模型	99.50 ± 0.21	98.88 ± 0.48
Dataset	CVAE	99.46 ± 1.55	99.74 ± 0.56
	CVAE-CGAN	95.62 ± 8.82	94.13 ± 9.09
	CVAE-WGAN-GP	99.07 ± 2.70	99.78 ± 0.50
	本文模型	99.88 ± 0.12	99.84 ± 0.10

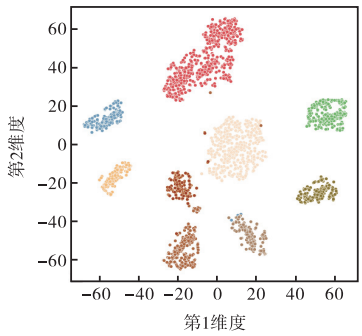
表 5　数据集分布

Tab. 5　Distribution of datasets

数据集名称	类别量	样本量	基因总数
Dataset	9	1 621	15 642
Dataset1	9	2 521	15 642



（a）Dataset 的基因选择分类准确率



（b）Dataset1 的基因选择分类准确率

注：蓝色表示 ACC，绿色表示 LAML，灰棕色表示 MESO，棕色表示 TGCT，深棕色表示 UCS，军绿色表示 UVM，橙色表示 DLBC，红色表示 LGG，粉白色表示 OV。

图 2　所有癌症合集的基因选择分类准确率比较

Fig. 2　Comparison of gene selection classification accuracy for the entire cancer gene set collection

对数据增强前后的数据集进行基因选择，使用 5 种不同的分类算法对选出的基因进行分类。KNN、DT、逻辑回归（LR）、支持向量机（SVM）和随机森林（RF）为验证模型的分类算法，并通过十折交叉验证提高分类的可靠性。数据增强前后的基因选择分类准确率比较结果见表 6。

表 6　数据增强前后的基因选择分类准确率比较

Tab. 6　Comparison of gene selection classification accuracy before and after data augmentation

算法	Dataset 准确率/%	Dataset1 准确率/%
KNN	98.83 ± 0.93	99.44 ± 0.40
DT	94.69 ± 2.21	97.54 ± 0.83
LR	99.38 ± 0.55	99.76 ± 0.26
SVM	99.75 ± 0.41	99.76 ± 0.26
RF	99.57 ± 0.48	99.60 ± 0.31

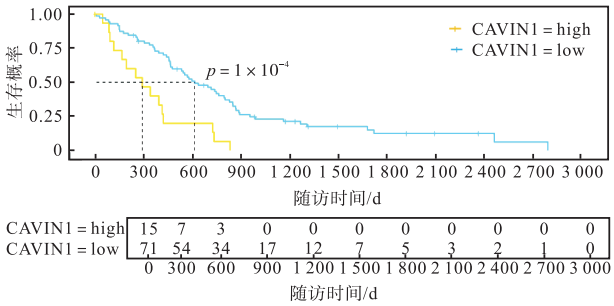
由表 6 可知，在 KNN 上数据增强后准确率的均值上升 0.61%，标准差的均值下降 0.53%；在 DT 上数据增强后准确率上升 2.85%，标准差的均值下降 1.38%；在 LR 上数据增强后准确率上升 0.38%，标准差的均值下降 0.29%。虽然在 SVM 和 RF 上数据增强后准确率的均值上升不是很明显，但是标准差的均值分别下降 0.15% 和 0.17%。

综上所述，本文成功利用 DAE 模型和 SVM-RFE 方法对基因进行筛选和分类验证，最终得到高准确率的分类结果和关键基因信息。这些结果表明本文方法在缓解“维数灾难”方面的有效性，并在基因选择方面取得了显著的改进。

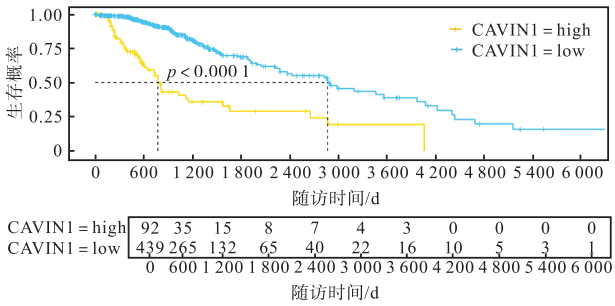
2.5　特征基因的生物意义

通过基因选择，最终猜测 CAVIN1、PDXP 和 CHCHD1 这 3 个候选基因的基因表达量对多种癌症存在影响。首先，这 3 个基因在原始数据集和混合之后的数据集中都重复出现；其次，PDXP 基因在肿瘤和癌症中具有潜在的作用[29-30]；CHCHD1 基因在肿瘤的发生和发展中有关键作用，对免疫微环境有影响，并且是线粒体呼吸链中的重要调节因子，对线粒体编码蛋白的表达有显著影响，而线粒体改变与癌症之间的联系已在多种癌症类型中被发现[31]，包括肝癌[32]、肺癌[33]、结肠癌[34]、乳腺癌[35]等。关于 CAVIN1 基因，Yi 等[36]证明其可以增强胶质母细胞瘤增殖，同时抑制肿瘤免疫反应。Bai 等[37]揭示了 miR-217 通过靶向 CAVIN1 抑制皮肤鳞状细胞癌的发展。Gould 等[38]揭示了 CAVIN1 水平升高可最大限度减少前列腺癌的发展，也有文献指出 CAVIN1 可能在一定程度上抑制肝癌细胞的发展[39]。

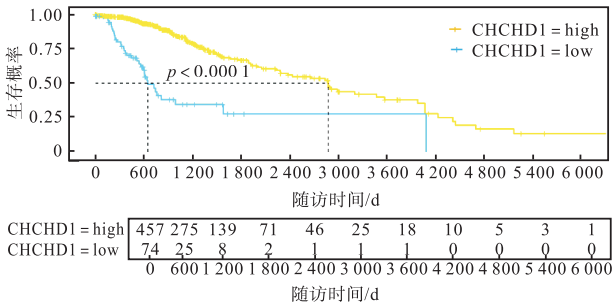
生存分析图如图 3 所示。由于 CAVIN1 两次在 MESO 癌症分类中重复出现，因此通过在 MESO 中的生存分析可以看出，在 MESO 中 CAVIN1 的表达和生存具有显著相关性（$P<0.001$），并且在 CAVIN1 中高表达的死亡风险要高于低表达。



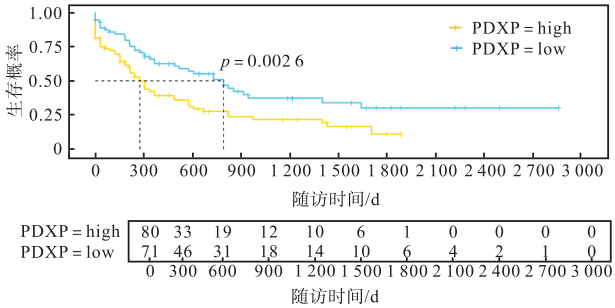
（a）CAVIN1 基因在 MESO 上的生存分析图



（b）CAVIN1 基因 LGG 上的生存分析图



（c）CHCHD1 基因在 LGG 上的生存分析图



（d）PDXP 基因在 LAML 上的生存分析图

图 3　生存分析图

Fig. 3　Survival analysis plot

3　结　语

本文从样本生成方面进行探索，利用现有数据进行数据增强，增加样本数量，用生成数据进行基因选择，获取最终的生物标志物。通过此方法能够克服样本数量不足的问题，为基因选择和生物标志物的发现提供更可靠的支持。这一研究方向的探索为解决基因表达数据分析中“维数灾难”问题提供了新的思路和方法，减少了下游分析的局限性，并通过生物信息学方法证明了这些方法的有效性和可行性，为基因表达数据的处理和分析提供参考。

参考文献：

[1] CHEN Y，LI Y，NARAYAN R，et al. Gene expression inference with deep learning[J]. Bioinformatics，2016，32(12)：1832–1839.

[2] FINOTELLO F，DI CAMILLO B. Measuring differential gene expression with RNA-seq：challenges and strategies for data analysis[J]. Briefings in functional genomics，2014，14(2)：130–142.

[3] IRWIN J J，SHOICHET B K. ZINC：a free database of commercially available compounds for virtual screening[J]. Journal of chemical information and modeling，2005，45(1)：177–182.

[4] HUNTER S，APWEILER R，ATTWOOD T K，et al. InterPro：the integrative protein signature database[J]. Nucleic acids research，2009，37(S1)：211–215.

[5] AL ABIR F，SHOVAN S M，HASAN M A M，et al. Biomarker identification by reversing the learning mechanism of an autoencoder and recursive feature elimination[J]. Molecular omics，2022，18(7)：652–661.

[6] POLDRACK R A，GORGOLEWSKI K J. OpenfMRI：open sharing of task fMRI data[J]. Neuroimage，2017，144：259–261.

[7] YAN K，WANG X，LU L，et al. DeepLesion：automated mining of large-scale lesion annotations and universal lesion detection with deep learning[J]. Journal of medical imaging，2018，5(3)：036501.

[8] WU P Y，CHENG C W，KADDI C D，et al. Omic and electronic health record big data analytics for precision medicine[J]. IEEE Transactions on biomedical engineering，2016，64(2)：263–273.

[9] WEI R，MAHMOOD A. Recent advances in variational autoencoders with representation learning for biomedical

informatics: a survey[J]. IEEE Access, 2020, 9: 4939–4956.

[10] ASYALI M H, COLAK D, DEMIRKAYA O, et al. Gene expression profile classification: a review[J]. Current bioinformatics, 2006, 1(1): 55–73.

[11] ZEEBAREE D Q, HASAN D A, ABDULAZEEZ A M, et al. Machine learning semi-supervised algorithms for gene selection: a review[C]//IEEE. 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET). Shah Alam: IEEE, 2021: 165–170.

[12] WANG Y, CHEN Q, SHAO H, et al. Generating bulk RNA-seq gene expression data based on generative deep learning models and utilizing it for data augmentation[J]. Computers in biology and medicine, 2024, 169: 107828.

[13] LEE M. Recent advances in generative adversarial networks for gene expression data: a comprehensive review[J]. Mathematics, 2023, 11(14): 3055.

[14] LACAN A, SEBAG M, HANCZAR B. GAN-based data augmentation for transcriptomics: survey and comparative assessment[J]. Bioinformatics, 2023, 39(S1): 111–120.

[15] AHMED K T, SUN J, CHENG S, et al. Multi-omics data integration by generative adversarial network[J]. Bioinformatics, 2021, 38(1): 179–186.

[16] VIÑAS R, ANDRÉS-TERRÉ H, LIÒ P, et al. Adversarial generation of gene expression data[J]. Bioinformatics, 2022, 38(3): 730–737.

[17] MAROUF M, MACHART P, BANSAL V, et al. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks[J]. Nature communications, 2020, 11(1): 166.

[18] DINCER A, CELIK S, HIRANUMA N, et al. DeepProfile: deep learning of patient molecular profiles for precision medicine in acute myeloid leukemia[EB/OL]. [2024–05–01]. http://doi.org/10.1101/278739.

[19] KIM S, KIM K, CHOE J, et al. Improved survival analysis by learning shared genomic information from pan-cancer data[J]. Bioinformatics, 2020, 36(S1): 389–398.

[20] BICA I, ANDRÉS-TERRÉ H, CVEJIC A, et al. Unsupervised generative and graph representation learning for modelling cell differentiation[J]. Scientific reports, 2020, 10(1): 9790.

[21] YU H, WELCH J D. MichiGAN: sampling from disentangled representations of single-cell data using generative adversarial networks[J]. Genome biology, 2021, 22(1): 158.

[22] ALMUTIRI T, SAEED F. Chi square and support vector machine with recursive feature elimination for gene expression data classification[C]//IEEE. 2019 First International Conference of Intelligent Computing and Engineering (ICOICE). Hadhramout: IEEE, 2019: 1–6.

[23] DANAEE P, GHAEINI R, HENDRIX D A. A deep learning approach for cancer detection and relevant gene identification[J]. Pacific symposium on biocomputing, 2017, 22: 219–229.

[24] LIU J, WANG X, CHENG Y, et al. Tumor gene expression data classification via sample expansion-based deep learning[J]. Oncotarget, 2017, 8(65): 109646–109660.

[25] UZMA, AL-OBEIDAT F, TUBAISHAT A, et al. Gene encoder: a feature selection technique through unsupervised deep learning-based clustering for large gene expression data[J]. Neural computing and applications, 2022, 34(11): 8309–8331.

[26] FAMITHA S, MOORTHI M. Deep learning approach for cancer detection through gene selection[C]//Proceedings of the Fourth Congress on Intelligent Systems, Singapore: Springer Nature, 2024: 333–345.

[27] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16(1): 321–357.

[28] HAN F, ZHU S, LING Q, et al. Gene-CWGAN: a data enhancement method for gene expression profile based on improved CWGAN-GP[J]. Neural computing and applications, 2022, 34(19): 16325–16339.

[29] KONG X, XU R, WANG W, et al. CircularLRRC7 is a potential tumor suppressor associated with miR-1281 and PDXP expression in glioblastoma[J]. Frontiers in molecular biosciences, 2021, 8: 743417.

[30] SCHULZE M, FEDORCHENKO O, ZINK T G, et al. Chronophin is a glial tumor modifier involved in the regulation of glioblastoma growth and invasiveness[J]. Oncogene, 2016, 35(24): 3163–3177.

[31] ZHAO L, TANG Y, YANG J, et al. Integrative analysis of circadian clock with prognostic and immunological biomarker identification in ovarian cancer[J]. Frontiers in molecular biosciences, 2023, 10: 1208132.

[32] HUANG Q, ZHAN L, CAO H, et al. Increased mitochondrial fission promotes autophagy and hepatocellular

（下转第 46 页）

Barley(1→3;1→4)-β-glucan and arabinoxylan content are related to kernel hardness and water uptake[J]. Journal of cereal science, 2008, 47(2): 365–371.

[21] ARDÖ Y, MCSWEENEY P, MAGBOUL A, et al. Biochemistry of cheese ripening: proteolysis[M]. Pittsburgh: Academic Press, 2017: 445–482.

[22] WANG W, JIA R, HUI Y, et al. Utilization of two plant polysaccharides to improve fresh goat milk cheese: texture, rheological properties, and microstructure characterization[J]. Journal of dairy science, 2023, 106(6): 3900–3917.

[23] CHEN T, WU Y, LIU F, et al. Unusual gelation behavior of low-acetyl gellan under microwave field: changes in rheological and hydration properties[J]. Carbohydrate polymers, 2022, 296: 119930.

[24] BANSAL V, KANAWJIA S K, KHETRA Y, et al. Steady and dynamic rheological properties of cheese dip: effect of milk proteins, fat and cheddar cheese[J]. Measurement: food, 2022, 8: 100066.

[25] QU R J, WANG Y, LI D, et al. Rheological behavior of nanocellulose gels at various calcium chloride concentrations[J]. Carbohydrate polymers, 2021, 274: 118660.

[26] PATEL G, MURAKONDA S, DWIVEDI M. Steady and dynamic shear rheology of Indian Jujube (*Ziziphus mauritiana* Lam.) fruit pulp with physiochemical, textural and thermal properties of the fruit[J]. Measurement: food, 2022, 5: 100023.

[27] ABDALLA A, ABU-JDAYIL B, ALSEREIDI H, et al. Low-moisture part-skim mozzarella cheese made from blends of camel and bovine milk: gross composition, proteolysis, functionality, microstructure, and rheological properties[J]. Journal of dairy science, 2022, 105(11): 8734–8749.

[28] FANG T, GUO M. Physicochemical, texture properties, and microstructure of yogurt using polymerized whey protein directly prepared from cheese whey as a thickening agent[J]. Journal of dairy science, 2019, 102(9): 7884–7894.

[29] MONSALVE-ATENCIO R, SANCHEZ-SOTO K, CHICA J, et al. Interaction between phospholipase and transglutaminase in the production of semi-soft fresh cheese and its effect on the yield, composition, microstructure and textural properties[J]. LWT-Food science & technology, 2022, 154: 112722.

[30] JIA Y, YAN X, HUANG Y, et al. Different interactions driving the binding of soy proteins(7S/11S) and flavonoids (quercetin/rutin): alterations in the conformational and functional properties of soy proteins[J]. Food chemistry, 2022, 396: 133685.

责任编辑: 郎婧

(上接第 8 页)

carcinoma cell survival through the ROS-modulated coordinated regulation of the NFKB and TP53 pathways[J]. Autophagy, 2016, 12(6): 999–1014.

[33] QIAN D C, KLEBER T, BRAMMER B, et al. Effect of immunotherapy time-of-day infusion on overall survival among patients with advanced melanoma in the USA (MEMOIR): a propensity score-matched analysis of a single-centre, longitudinal study[J]. The lancet oncology, 2021, 22(12): 1777–1786.

[34] TAILOR D, HAHM E R, KALE R K, et al. Sodium butyrate induces DRP1-mediated mitochondrial fusion and apoptosis in human colorectal cancer cells[J]. Mitochondrion, 2014, 16: 55–64.

[35] ZHAO J, ZHANG J, YU M, et al. Mitochondrial dynamics regulates migration and invasion of breast cancer cells[J]. Oncogene, 2013, 32(40): 4814–4824.

[36] YI K, ZHAN Q, WANG Q, et al. PTRF/cavin-1 remodels phospholipid metabolism to promote tumor proliferation and suppress immune responses in glioblastoma by stabilizing cPLA2[J]. Neuro oncology, 2021, 23(3): 387–399.

[37] BAI M, ZHANG M, LONG F, et al. miR-217 promotes cutaneous squamous cell carcinoma progression by targeting PTRF[J]. American journal of translational research, 2017, 9(2): 647–655.

[38] GOULD M L, WILLIAMS G, NICHOLSON H D. Changes in caveolae, caveolin, and polymerase 1 and transcript release factor(PTRF) expression in prostate cancer progression[J]. Prostate, 2010, 70(15): 1609–1621.

[39] HAO X, LI J, LIU B, et al. Cavin1 activates the Wnt/β-catenin pathway to influence the proliferation and migration of hepatocellular carcinoma[J]. Annals of hepatology, 2024, 29(1): 101160.

责任编辑: 郎婧