文章编号:1671-4598(2025)11-0219-09 DOI:10.16526/j.cnki.11-4762/tp.2025.11.026 中图分类号:TP391 文献标识码:A

# 基于生成对抗网络的乳腺癌基因数据生成与挖掘

杨 锦,边太成,李晓晖,焦 强,朱习军
(青岛科技大学 信息科学技术学院,山东 青岛 266061)

**摘要**:针对组学数据挖掘中遇到的数据样本量少、数据高维度和特征泛化性差的问题,提出了结合残差网络与软阈值化方法的生成模型RS-CGAN;该模型通过一维卷积层和残差网络结构提升高维数据的特征学习能力,并引入残差软阈值化的生成器和残差注意力的判别器以降低噪声影响并防止过拟合;采用距离相似度惩罚项指导生成器学习,增强训练质量;提出基于结构因果模型的特征选择模块,通过构建因果结构图,实现群体平均因果治疗效应估计值的计算,识别具有泛化性和因果关系的生物标志物;实验结果表明,该方法在数据生成质量方面具有优势,且特征选择后的数据集在预测模型中的准确率提升了30.58%,最终识别10个乳腺癌生物标志物,其中4个已经过临床医学验证为风险位点,证明了该标志物选择方法的有效性。

**关键词**:生成对抗网络;数据增强;标志物挖掘;因果推断;乳腺癌基因数据

# Gene Data Generation and Mining of Breast Cancer Based on Generative Adversarial Networks

YANG Jin, BIAN Taicheng, LI Xiaohui, JIAO Qiang, ZHU Xijun
(School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

**Abstract**: There are the characteristics of small sample size of data, high-dimensional data, and feature generalization in the omics data mining. To address these issues, a generative model RS-CGAN combining the residual network with the soft thresholding method was proposed. The model improves the feature learning ability of high-dimensional data through a one-dimensional convolutional layer and residual network structure, and introduces a generator for the residual soft thresholding and a discriminator for the residual attention to reduce the influence of noise and prevent overfitting. The distance similarity penalty term is used to guide the learning of the generator to enhance the training quality. A feature selection module based on the structural causal model is proposed, and a causal structure diagram is constructed to calculate the estimated value of the average causal treatment effect of the population, and to identify the biomarkers with the generalization and causal relationship. Experimental results show that the method has advantages of data generation quality, the accuracy of the dataset after the feature selection in the prediction model is increased by 30.58%, and finally 10 breast cancer biomarkers are identified, of which 4 have been verified by clinical medicine as risks, which proves the effectiveness of the marker selection method.

**Keywords**: generative adversarial networks; data augmentation; marker mining; causal inference; breast cancer gene data

**收稿日期**:2024-09-19; **修回日期**:2024-11-14。
**基金项目**:山东省重点研发计划基金(2015GSF119016);青岛市科技惠民示范专(23-2-8-smjk-20-nsh);山东省产教融合研究生联合培养示范基地项目(2020-19)。
**作者简介**:杨 锦(2000-),女,硕士研究生。
**通讯作者**:朱习军(1964-),男,博士,教授,硕士生导师。
**引用格式**:杨 锦,边太成,李晓晖,等.基于生成对抗网络的乳腺癌基因数据生成与挖掘[J].计算机测量与控制,2025,33(11):219-227.

## 0 引言

根据国家癌症中心的报告，2020 年中国女性癌症死亡病例数达到 118 万，其中，乳腺癌死亡病例数占比 9.9%，到 2022 年，中国女性新发癌症病例数上升至 209 万，占总数的 46%，其中乳腺癌新发病例数占比高达 19.9%，已超越肺癌成为新发病例数最多的癌症[1]。乳腺癌由于早期症状不明显，很容易被忽视导致死亡率居高不下，被称为“全球女性的第一杀手”。世界卫生组织声明乳腺癌的早期发现能够有效降低死亡率，因此，早期防治成为降低乳腺癌发病率及死亡率的关键研究议题[2-5]。

乳腺癌的早期防治主要依赖于生物标志物的检测[6-10]，其中，单核苷酸多态性（SNP，single nucleotide polymorphisms）是基因组学研究中的一种重要生物标志物，通过对 SNP 的研究，能够深入理解肿瘤的遗传背景，为疾病的预防、诊断和治疗提供重要的分子靶标[11-13]。然而，在基于基因组学的相关研究中，由于检测成本高昂和伦理道德限制，可用患者样本数量通常较少，导致统计力不足、数据变异性和泛化性较低等问题，使得分析结果存在偏倚，成为组学生物标志物挖掘研究中的一大难题。

近年来，机器学习与深度学习技术因其在处理复杂数据、提取高级特征和识别模式方面的卓越能力，被广泛应用于组学数据增强方面，生成近似真实数据分布的组学数据，解决组学研究中数据量不足的问题[14]，2020 年，文献［15］提出 cscGAN 方法模拟生成并增强 RNA-seq 数据，该方法能够有效生成基因测序数据，但无法有效生成稀疏数据。文献［16］提出 GeneGAN 方法通过生成基因图像数据，来实现对原始数据的增强和扩充，但该方法无法通过标签直接生成对应基因数据，无法投入生物医学研究中。同时，这些方法缺乏对生成数据在生物标志物识别领域的后续研究与应用。

针对这些问题，本文提出一种结合组学数据增强与因果特征选择的标志物挖掘方法—残差软阈值化条件生成对抗网络（RS-CGAN，residual soft thresholding-conditional generative adversarial net），提高组学数据生成效果，并解决标志物的识别精度和可解释性问题，本研究主要贡献如下：

1）RS-CGAN 通过结合软阈值化和残差注意力，解决高维数据学习过程中容易引入噪声的问题，避免模型过拟合，提高数据生成质量。

2）RS-CGAN 通过设计组合损失函数，增强生成模型训练质量，解决模型拟合困难和无法有效生成稀疏数据的问题。

3）RS-CGAN 增加因果特征选择模块，构建乳腺癌与特征间的因果结构图，保证特征选择的普适性和泛化性，实现生物标志物的识别。

## 1 数据集与预处理

本研究下载英国生物银行（UKB，UK biobank）数据，分为结构化个体 SNP 数据（特征空间如表 1，特征维度）和个体表型数据（特征空间如表 2，特征维度），由于基因数据的复杂性和多样性，导致 SNP 数据中数据质量不一致、缺失率较高等问题，需要设计合理的数据预处理，主要包括基因数据清洗、样本特征清洗、样本映射、缺失率清洗、缺失数据插补。

表 1 个体 SNP 特征空间

| 家族 | ID | 性别 | 表型类型 | 基因型 | 染色体号 | SNP 标识符 | 遗传距离 |
|---|---|---|---|---|---|---|---|
| 500612 | 500612 | 1 | 1 | 1 | 3 | rs1052133 | 0 |

表 2 个体表型特征空间

| ID | 性别 | 年龄 | 出生年 | 身高 | 腰围 | 臀围 | … | ICD-10 |
|---|---|---|---|---|---|---|---|---|
| 500612 | 1 | 45 | 1970 | 165 | 82 | 101 | … | C50 |

### 1.1 数据清洗

针对组学数据中存在大量无关特征，虚假关联的问题，设计如图 1 所示的数据清洗流程，实现 SNP 数据质控。

1）SNP 质控：最小等位基因频率与位点贡献信息成正相关，通过计算最小等位基因频率，结合贡献信息筛选位点。此外，通过哈温平衡过滤对 SNP 分布进行筛选，计算等位基因频率，如公式（1）：

$$p^2 + q^2 + pq = 1 \tag{1}$$

其中：$p$ 为等位基因 $A$ 频率，$q$ 为等位基因 $T$ 频率，根据公式（2）和（3）得到等位基因理想样本数，其中 $m$ 为总人数：

$$E_A = p^2 \times m \tag{2}$$

$$E_T = q^2 \times m \tag{3}$$

进一步根据公式（4）计算卡方检验：

$$x^2 = -\sum \frac{(O-E)^2}{E} \tag{4}$$

其中：$O$ 为观测值，$E$ 为公式（2）公式（3）所得理想值，取自由度为 1，$P$-value 阈值设置如表 3 所示。

2）表型特征清洗：采用基于组学专家知识对特征空间进行特征选择，删除了明显与乳腺癌患病无关特征（医疗器械、采血工具等），保留性别，年龄，体重，ICD-10 疾病分类等重要特征。

3）特征空间映射：以 ID 为主键，映射 SNP 特征空间与表型特征空间。

4）针对高缺失率的数据预处理：高缺失率数据无

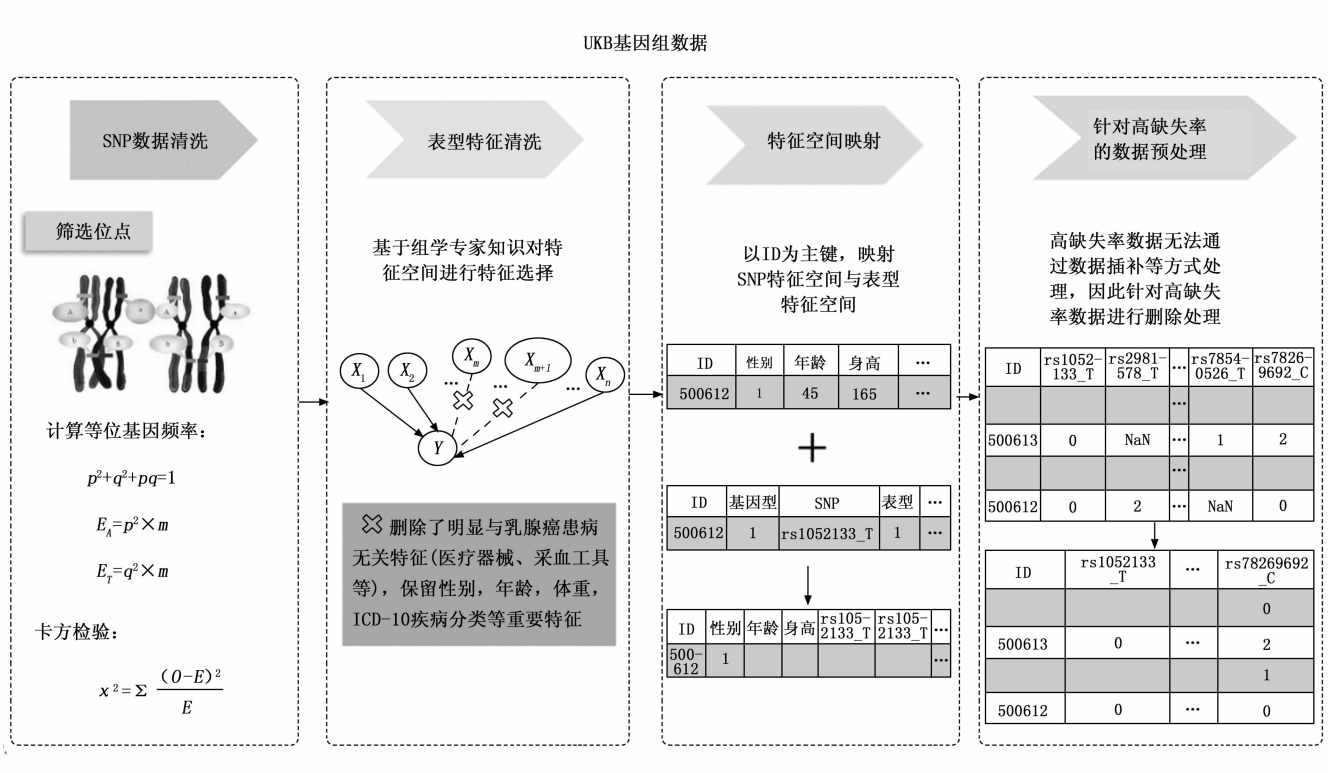


图 1　数据清洗流程

法通过数据插补等方式处理，因此针对高缺失率数据进行删除处理，数据质控参数设置如表 3 所示。

表 3　SNP 质控删除数据占比

| SNP 缺失质控方法 | 阈值设置 |
|---|---|
| SNP 缺失率 | 20% |
| 样本缺失率 | 20% |
| 最小等位基因频率 | 0.01 |
| 哈温平衡过滤 | 0.01 |

为保证数据集质量，保留更多数据，避免样本不平衡。质控参数阈值设置如表 3 所示，缺失率阈值设定为不超过 20%，样本缺失率的上限定为 20%，最小等位基因频率考虑的下限为 0.01，哈温平衡测试的过滤阈值定为 0.01。

经数据预处理过程，得到包含 SNP 位点的乳腺癌样本数据集，特征空间如表 4 所示，共计 8 400 个特征，其中包括 8 395 种 SNP 位点特征［位点取值 0，1，2，(0：纯合基因，1：杂合基因，2：次等位纯合基因)］和 5 个表型特征，健康人群 43 361 名，患病人群 3 466 名，存在较大的样本不平衡性。

表 4　预处理 SNP 数据特征空间

| ID | 性别 | 年龄 | 体重 | 身高 | C50 | SNP 位点 |
|---|---|---|---|---|---|---|
| 500612 | 1 | 45 | 62 | 165 | 1 | 1 |

## 1.2　数据插补

SNP 数据缺失无法避免，面对不同缺失率，通常采用不同的预处理方法，对于高缺失率数据采用删除操作，但大量数据删除会丢失很多关键信息，造成宝贵数据的浪费，本研究强调数据质量在生成模型优化中关键作用的同时，提出减少数据浪费，因此需要针对低缺失率数据进行填补，传统统计学数据填补法主要有中值法、众数法等，这类方法简便快捷，但引入了较大混杂和噪声数据，鉴于此，本研究采用了生成对抗插补网络(GAIN，generative adversarial imputation nets)[17] 方法对缺失数据进行填补，学习数据分布，通过深度学习方式预测每个缺失值数据，确保数据质量，并能够根据其特征分布来有效地对缺失数据进行填补。

输入数据集 $X=(X_1, X_2, \cdots, X_d)$，由公式 (5) 得到由 0，1 构成的缺失数据掩码向量：

$$M=\begin{cases}1,\text{other}\\0,X_i=0\end{cases} \tag{5}$$

通过对 GAIN 的训练，由公式 (6) 得到缺失数据预测值：

$$\overline{X}=G(X,M,(1-M)\odot Z) \tag{6}$$

其中：$G$ 为模型生成器，$Z$ 为高斯噪声。$\overline{X}$ 为生成得到 SNP 数据缺失值。该方法相较于 Miss Forest[18] 等传统插补方法，以学习数据分布来对缺失值进行预测填补，减弱了个别特异性数据造成的混杂，提高了生成数

据的真实性和可靠性。但该方法仍存在缺陷，在面对高噪声的稀疏数据时，由于其网络结构简单，仅为两层全连接层，导致无法学习大部分稀疏特征，容易受到噪声影响，因此针对该问题，本研究为提升模型对离散 SNP 数据插补能力，提高组学数据质量，提出基于注意力的 GAIN 模型，通过在全连接层后增加注意力机制，添加 SoftMax 层，公式如下：

$$X = \text{softmax}(X) \tag{7}$$

提高模型在面对 SNP 稀疏数据和高缺失率数据时的学习能力，增强对关键特征的敏感度，根据不同缺失率在 spam、letter 和 MINST 数据集进行验证，计算原始 GAIN 模型与 Soft-GAIN 模型的均方误差（MSE，mean square error）进行对比，结果如表 5 所示，原始 GAIN 模型与 Soft-GAIN 模型在 spam、letter、MNIST 3 个公开数据集上进行对比，结果如下：

表 5 不同缺失率数据模型 MSE 对比结果

| 数据集 | 缺失率 | GAIN | Soft-GAIN |
|---|---|---|---|
| Spam | 0.5 | 0.054 6 | 0.053 2 |
| | 0.7 | 0.073 1 | 0.055 9 |
| | 0.8 | 0.118 7 | 0.057 1 |
| Letter | 0.5 | 0.142 7 | 0.141 |
| | 0.7 | 0.155 4 | 0.148 5 |
| | 0.8 | 0.211 8 | 0.156 4 |
| MNIST | 0.5 | 0.140 3 | 0.135 2 |
| | 0.7 | 0.224 | 0.216 3 |
| | 0.8 | 0.298 8 | 0.259 3 |

随着缺失率的增高，GAIN 模型损失增大，Soft-GAIN 损失较为稳定，可以看出，传统模型在针对高缺失率数据的插补工作中，无法有效生成缺失数据，模型拟合能力较差，Soft-GAIN 通过增加 SoftMax 层，提高了模型对关键特征分布的学习能力，对于高缺失率数据的缺失部分具有更好的预测效果，提高了模型数据插补能力，为后续模型效果提供了坚实的数据基础。

## 2 RS-CGAN

本文以生成对抗网络（GAN，generative adversarial networks）[19] 为大框架提出的 RS-CGAN 模型包括 3 个主要部分：结合残差软阈值化的生成器、结合残差注意力的判别器和基于结构因果模型（SCM，structural causal model）[20] 的特征选择模块。为针对性生成 SNP 数据，训练过程中每次输入一个 SNP 样本。生成器训练过程中，将服从高斯分布的噪声和标签变量作为生成器输入，为保证模型适配一维 SNP 输入数据，使用一维卷积层来有效捕捉数据局部特征，加入残差网络和软阈值化，保留输入信息的关键特征，最终通过 SoftMax 函数与全连接层，获得生成数据；将真实数据、生成数据和标签作为输入训练判别器，通过残差注意力提升模型对关键信息的提取能力，最终通过 SoftMax 函数与全连接层，输出对输入数据真假的预测值；成功训练生成模型后，能够生成近似真实 SNP 数据的高质量样本，将增强后的数据输入 SCM 特征选择模块，构建因果结构图并计算平均治疗效应值，根据效应值判断特征与标签的因果关系大小，获得因果乳腺癌标志物。

### 2.1 结合残差软阈值化的生成器设计

在 RS-CGAN 模型中，为了适应样本数据的特征学习，将生成器中的二维逆卷积层替换为一维逆卷积层，引入了具有收缩性质的软阈值化处理，能够有效地保留真实数据的关键信息，同时去除噪声。在模型中，输入的高斯噪声通过一维全局平均池化层和全连接层等操作，学习到了特征值的一组权重。这些权重通过矩阵乘

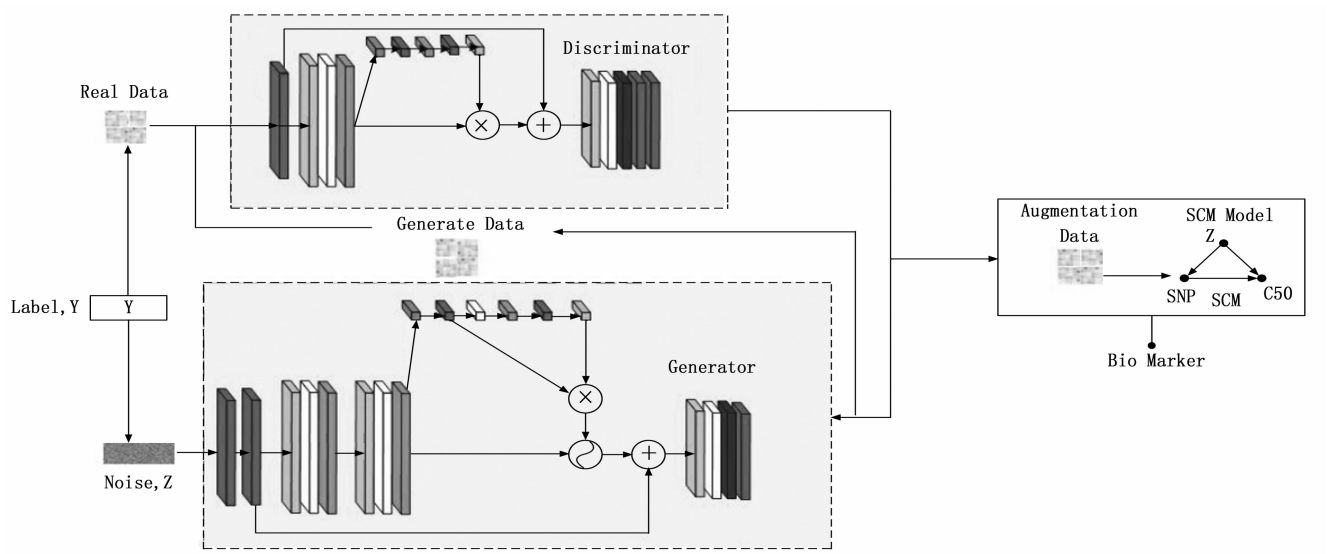


图 2 RS-CGAN 整体框架图

法计算得到一个阈值 $\lambda$，并被用于执行软阈值化去噪声操作，判断公式如下：

$$S_\lambda(x)=\begin{cases}x-\lambda, \text{if } x>\lambda\\ 0, \text{ if } |x|\leqslant\lambda\\ x+\lambda, \text{if } x<-\lambda\end{cases} \tag{8}$$

其中：$x$ 是输入数据，通过判断输入值是否大于阈值 $\lambda$，去除小幅度噪声分量，保证生成器在学习过程中过滤与生成分布差距较大的噪声，提高模型数据生成质量。

结合残差学习思想，使得模型能够通过学习输入与输出间的差异来加速训练，有助于解决深度网络中梯度消失和梯度爆炸问题，结合残差网络和软阈值化的方法能够使生成器有效学习数据的稀疏表示，从而解决 SNP 数据存在的高稀疏性问题。生成器框架如图 3 所示。

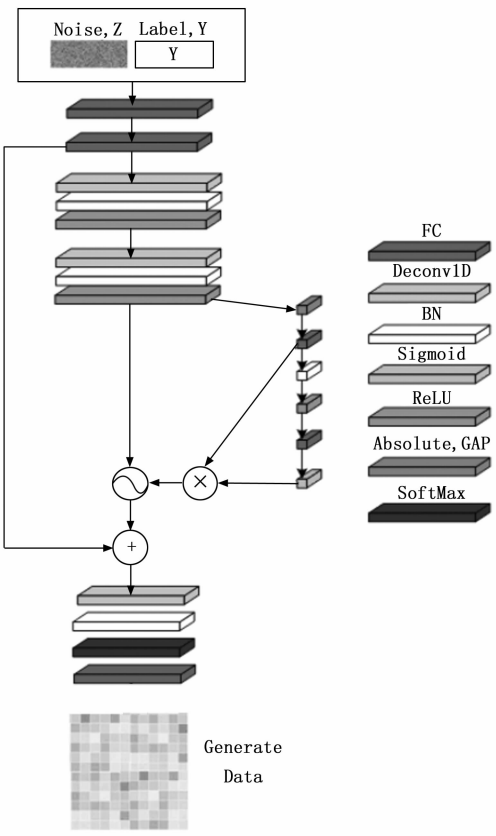


图 3　生成器框架图

## 2.2　结合残差注意力的判别器设计

本研究中，数据的稀疏性会导致判别器在学习关键特征方面的能力受限。因此设计了结合残差注意力模块的判别器，进一步提升对数据的特征表达能力。残差注意力判别器结构如图 4 所示。

判别器通过叠加残差注意力层，更加有效地学习和提取数据关键特征，使模型关注关键特征区域，准确区分真实和生成数据。残差结构帮助判别器在训练过程中保持稳定，降低过拟合风险，注意力机制结合有助于判别器捕捉数据长距离依赖关系，提高判别准确率和模型的泛化能力。

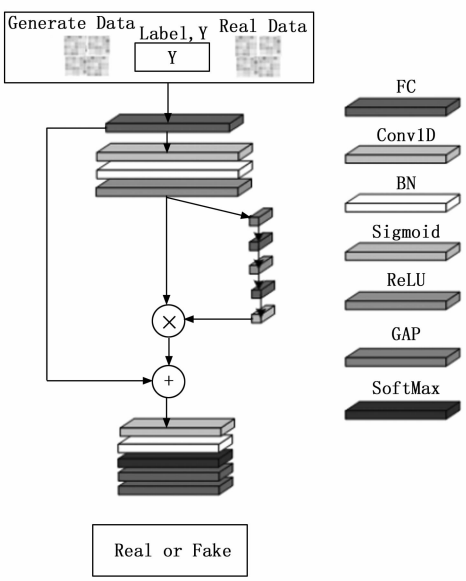


图 4　判别器框架图

## 2.3　组合损失函数设计

传统 GAN 模型基于 JS 散度构建对抗损失函数，来最小化真实数据与生成数据之间的距离，然而，仅使用对抗损失作为损失函数，存在训练不稳定和模型收敛速度慢的问题，针对此，为 RS-CGAN 设计了组合损失函数，包括对抗损失函数和相似度距离损失函数。

1）对抗损失函数：

在数据增强任务中，常用对抗损失函数包括传统 GAN 损失函数、条件生成对抗网络[21]的条件对抗损失和 Wasserstein GAN（WGAN）[22]的 EM 距离两种，其中，WGAN 损失需要施加约束，权重裁剪或梯度惩罚等额外计算会增加模型训练时间和成本，数据增强要求生成特定标签的数据，因此，对抗损失使用 CGAN 的条件对抗损失函数，计算过程如式（9）所示：

$$Los_{adv}=\min_G\max_D E_{x\sim \text{pdata}(x)}[\log D(x\mid y)]+E_{z\sim p_z(z)}[\log(1-D(G(z\mid y)))] \tag{9}$$

其中：$G$、$D$ 分别表示生成器与判别器，$x$ 表示真实数据，$z$ 表示高斯噪声，$y$ 表示标签变量。

2）相似度距离损失函数：

为了提高生成样本质量，加速模型收敛，对 RS-CGAN 生成方向进行约束，引入余弦相似度作为惩罚项。余弦相似度通过计算两样本间余弦夹角值衡量个体间差异，计算公式如下：

$$S(x,y)=\frac{x\cdot y}{\|x\|\ \|y\|} \tag{10}$$

根据式（10）得到惩罚项公式如下：

$$Loss_{\cos}\theta\cdot\left(1-\frac{G(z)\cdot x}{\|G(z)\|\ \|x\|}\right) \tag{11}$$

其中：$G(z)$ 表示生成样本，$x$ 表示真实样本，$z$ 表示高斯噪声，$\theta$ 表示正则项权重。根据以上设计，RS-CGAN 损失函数如下：

$$Loss = Loss_{\text{adv}} + Loss_{\cos} \tag{12}$$

## 2.4 基于 SCM 的特征选择

经典特征选择方法如过滤法、包装法和嵌入法，主要基于数理统计来评估特征间的相关性，但面对高维数据时计算复杂度较高。机器学习方法虽然缓解了计算问题，但往往无法保证所选特征的普适性，即难以判断通过模型学习得到的特征是否适用于整个社会群体的乳腺癌预测。

为了解决这一问题，本文提出基于 SCM 的特征选择模块，通过将输入位点数据构建成因果结构图，学习位点间与标签的关系，在 SCM 模型中，假设位点与是否患癌之间的治疗效应是线性关系，通过线性回归模型来估计平均治疗效应（ATE，average treatment effect），公式如下：

$$ATE = E[Y(1) - Y(0)] \tag{13}$$

其中：$Y(1)$ 和 $Y(0)$ 分别表示个体在接受治疗和不接受治疗情况下的潜在结果，$E[\cdot]$ 表示期望值。

$ATE$ 主要用于评估整个总体中接收治疗和未接受治疗之间的平均差异，在本研究中 $ATE$ 用于评估位点突变与未突变之前乳腺癌患病情况的平均差异。

基于结构因果模型的设计中，由于无法同时观察到同一个体的两种潜在结果，因此通常需要借助统计方法来估计 $ATE$，构建基于最小二乘法的线性回归模型，公式如下：

$$Y = \beta_0 + \beta_1 T + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon \tag{14}$$

其中：$Y$ 为是否患癌，$T$ 是目标位点变量，$X_2$，$X_3$，…，$X_k$ 是其他位点特征，$\beta_0$，$\beta_1$，$\beta_2$，$\beta_3$，…，$\beta_k$ 是回归系数，$\varepsilon$ 是误差项，计算 $T$ 的回归系数 $\beta_1$，该回归系数即为平均因果效应计算中的期望值。通过拟合线性回归模型，分别将 $T$ 取 1 和 2 时的估计值与 $T$ 取 0 时的估计值做差，并选择两者中最大值作为该位点对应患癌情况的因果效应估计值，并根据估计值筛选出有效生物标志物特征。

RS-CGAN 整体算法如下：

算法：RS-CGAN 模型算法

输入：乳腺癌 SNP 样本数据 $P_r(x) = (x_1, x_2, \cdots, x_n)$，迭代次数 Epoch，样本数 $M$；

输出：生成数据 $P_g$；

1）初始化生成器 $G$，判别器 $D$ 权重 $W_G$，$W_D$；

2）当 $t \leqslant$ Epoch 时：

3）输入 $M$ 个样本值 $x$ 和 $y$ 标签；

4）生成随机噪声 $z$，复制标签 $y$；

5）生成器训练：

6）残差收缩块计算阈值

7）软阈值化：$S_\lambda(x) = \begin{cases} x-\lambda, & \text{if} \quad x > \lambda \\ 0, & \text{if} \quad |x| \leqslant \lambda \\ x+\lambda, & \text{if} \quad x < -\lambda \end{cases}$

8）学习特征并生成数据：$\hat{x}_i = G(z_i, y_i)$

9）计算生成器 $G$ 损失函数：$G_{\text{loss}} = \frac{1}{M}\sum_{i=1}^{M}\log(1 - D(\hat{x_i} \mid y_i)) + \theta \cdot \left(1 - \frac{G(z) \cdot x}{\|G(z)\| \, \|x\|}\right)$

10）计算判别器 $D$ 损失函数：$G_{loss} = \frac{1}{M}\sum_{i=1}^{M}\log D(\hat{x_i} \mid y_i) + G_{\text{loss}}$

11）梯度下降生成器损失，梯度上升判别器损失；

12）判别器无法判别真假或不符合循环条件时，退出循环；

13）基于生成数据构建结果因果模型；

14）根据生成数据计算平均因果效应：$ATE = E[Y(1) - Y(0)]$

15）基于因果效应值特征选择。

# 3 实验设计

本文使用经预处理的乳腺癌合成数据集（健康人群 43 361 名，患病人群 3 466），按 7∶1∶2 划分训练集、验证集和测试集。为验证 RS-CGAN 的数据增强能力和特征选择有效性，分别从数据增强对比实验、基于 SCM 效应估计值分析和基于特征选择的乳腺癌风险预测对比实验三部分进行比较分析。

## 3.1 评价指标

本文评价指标仍分为数据增强评价指标和特征选择评价指标。

1）数据增强评价指标：

距离相似度计算公式主要用于度量数据间的距离，如：汉明距离，欧式距离，余弦相似度和交叉熵等，面对不同问题，选择也不同。汉明距离通过计算两个向量间不同值的个数，通常用于比较长度相同的二进制字符串，但 SNP 数据是离散数值，因此并不适用。余弦相似度作为计算高维欧几里得距离的一种方法，主要通过计算两向量夹角的余弦来判断相似性，但该方法缺乏对生成数据大小的考虑。欧式距离是一种经典的距离度量方法，用于衡量两点在多维空间的直线距离，可以有效计算并评估生成 SNP 数据与真实 SNP 数据之间的相似度，计算公式如下：

$$d(g_i, r_j) = \sqrt{\sum_{k=1}^{n}(g_{ik} - r_{jk})^2} \tag{15}$$

其中：$g_i$ 是生成数据集中的样本，$r_j$ 是真实数据集中的样本，$n$ 是 SNP 位点数量，$g_{ik}$ 和 $r_{jk}$ 分别是样本 $g_i$ 和样本 $r_j$ 第 $k$ 个位点的值。通过公式（15）得到样本中每个位点差值的平方和，计算生成样本与真实数据整体的相似度，公式如下：

$$\bar{d}(G,R) = \frac{1}{m \times p}\sum_{i=1}^{m}\sum_{j=i}^{p} d(g_i, r_j) \quad (16)$$

$m$ 是生成数据集中样本的数量，$p$ 是真实数据中样本的数量。欧式距离越小，则表示模型生成的 SNP 数据与真实数据之间的相似度越高。

单一评价标准不能全面体现模型性能，除欧氏距离外，选择使用 *MSE* 作为数据增强工作的评价指标，相较于欧氏距离，*MSE* 更关注数据的数值精度，计算公式如下：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (17)$$

其中：$n$ 是 SNP 位点数量，$y_i$ 是真实数据，$\hat{y}_i$ 是生成数据。均方误差越小则表示生成数据与真实数据之间的差异越小，即生成质量越高。

2）特征选择评价指标：

对于特征选择有效性的验证，通过评估模型对原始数据集和特征选择数据集进行乳腺癌风险预测，选择准确率作为评价指标，准确率高则表示特征选择有效性。

准确率计算公式如下：

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (18)$$

其中：*TP* 表示正确的正例，*FP* 表示错误的正例，*FN* 表示错误的反例，*TN* 表示正确的反例。

### 3.2　数据增强对比实验

为证明方法在数据增强方面的有效性，本文与 4 种目前先进的生成模型方法进行对比，包括深度卷积生成对抗网络（DCGAN，deep convolution GAN）[23]、WGAN、WGAN-GP[24]、cscGAN。实验结果如表 6 所示。

表 6　数据增强模型实验结果对比

| 生成模型 | MSE | $\bar{d}$ |
|---|---|---|
| DCGAN | 0.237 | 0.316 5 |
| WGAN | 0.132 8 | 0.167 9 |
| WGAN-GP | 0.096 1 | 0.105 4 |
| cscGAN | 0.065 1 | 0.081 5 |
| RS-CGAN | **0.043 0** | **0.062 4** |

通过 MSE 与欧氏距离计算各对比模型生成数据与真实数据的差异，由表 6 可以看出，本文提出的 RS-CGAN 生成数据与真实数据最为相近，证明了其在一维 SNP 数据生成上表现出色，在生成质量上有显著提升。

### 3.3　因果效应估计值分析

通过构建因果图模型，计算各位点特征的回归预测值，根据 SNP 平均治疗效应估计值判断与乳腺癌患病之间的因果关系，并选择出效应值最大的 10 个位点作为乳腺癌因果生物标志物，各位点效应估计值如表 7 所示，其中，效应估计值的正负代表不同因果关系含义，正因果效应值表示 SNP 位点发生变异与是否患乳腺癌间呈正相关，负因果效应值表示 SNP 位点突变与患乳腺癌间呈负相关，效应值越趋向于 0 则表示位点变异与是否患癌无关。

表 7　平均治疗效应估计值

| 计算方式 | 位点 | 效应值 |
|---|---|---|
| ATE | rs78540526_T | 0.056 8 |
| | rs11571833_T | 0.038 2 |
| | rs34005590_A | −0.027 7 |
| | rs2981578_T | 0.022 1 |
| | rs4442975_A | −0.021 5 |
| | rs145342093_C | −0.019 4 |
| | rs56387622_C | −0.019 4 |
| | rs7297051_T | 0.019 3 |
| | rs78956371_C | 0.018 6 |
| | rs13002632_A | 0.015 2 |

根据因果效应估计值，确定表 7 中 10 个位点为乳腺癌生物标志物。

结构因果模型计算因果效应估计值结果如图 5，图 6 所示。绝对值越大，表明位点与患病的因果关系越大。由图 5 可以看出，共有 7 个位点对乳腺癌患病的因果估计值大于 0.01，其中包括 rs78540526 _ T、rs11571833 _ T 等位点，表示 rs78540526 和 rs11571833 位点的变异等位基因为 T，由其治疗效应值可以推断，随着等位基因变异为 T 的个数增长，该个体患乳腺癌的概率越大。并且经过文献［25］和文献［26］的临床医学实验和多基因风险评分验证，rs78540526 和 rs11571833 均为乳腺癌风险位点。

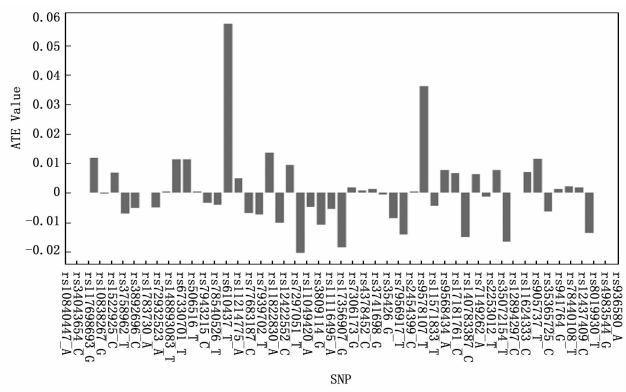


图 5　平均治疗效应估计值

由图 6 可以看出，同样有 7 个位点对患乳腺癌的因果估计值小于 −0.01，其中包括 rs34005590 _ A 和 rs4442975 _ A 位点，表示这些位点的变异等位基因为 A，且由治疗效应值推断，这些位点的突变为 A 的等位基因个数越多，个体越不容易患乳腺癌，表达为，该位点不发生突变的人群易患乳腺癌，这两个位点同样经过文献［27］的临床实验验证和多基因风险评分验证。

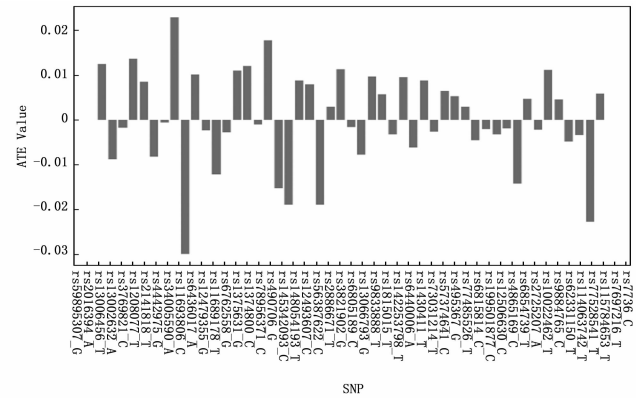


图 6 平均治疗效应估计值

### 3.4 特征选择对比实验

为证明本文方法的数据增强和特征选择能力，本实验采用原始多层感知机进行乳腺癌风险预测，分别对使用本文方法的增强数据集，对 WGAN 增强的数据集，SC-CGAN 增强的数据集，原始数据集以及特征选择后的原始数据集作对比，同样通过前文中构建的风险分类模型预测是否患癌，实验结果如表 8 所示，WGAN 增强的数据集风险预测准确率降低 4.74%，WGAN 并不适用于稀疏性矩阵生成，基于 RS-CGAN 增强数据，准确率提升 5.72%，证明 RC-CGAN 模型在稀疏数据增强方面具有显著提升，基于因果特征选择数据集乳腺癌风险预测准确率为 83.17%，对比原始数据增长了 30.58%。迭代训练 200 次后模型训练准确率拟合图如图 7 所示。

表 8 风险评估模型准确率比较

| 数据集 | 准确率/% |
|---|---|
| 原始数据集 | 52.59 |
| WGAN 增强数据集 | 47.85 |
| RS-CGAN 增强数据集 | 60.31 |
| 特征选择处理数据集 | 83.17 |

对比数据集在评估模型上的正确率训练结果如图 7 所示，可以看出，随着模型训练迭代次数增加，各数据集的预测准确率也逐渐增加，并于 150 轮后趋于稳定，经过特征选择后的数据在评估模型上预测正确率远高于其他数据集，证明了特征的有效性。

## 4 结束语

为了解决组学数据高维度问题，本研究提出了一种结合残差网络和软阈值化方法的生成模型-RS-CGAN。该模型通过一维卷积层和残差软阈值化技术，有效提高了在高维数据中的特征学习能力，避免了模型过拟合，同时设计组合损失函数，加速模型训练，提高生成数据质量。通过构建乳腺癌与 SNP 位点间的结构因果模型，计算群体平均因果治疗效应，解决了传统特征选择缺乏

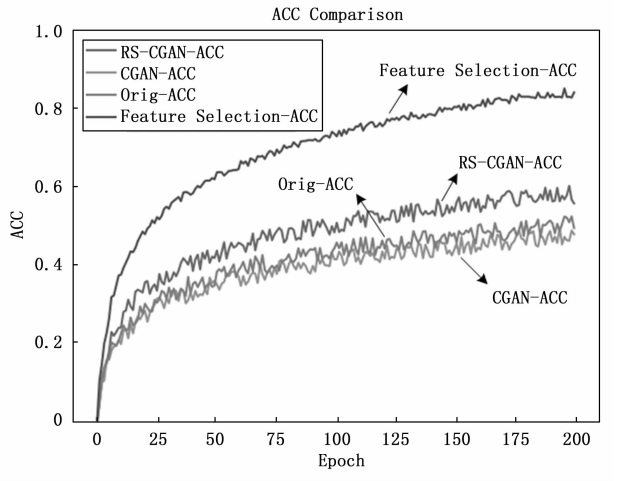


图 7 各数据集在风险预测模型上的正确率

泛化性和普适性的问题，实现生物标志物挖掘，其中 4 个位点已经过医学领域的临床实验验证，确定其作为生物标志物的有效性，证明了本文提出方法的有效性。

**参考文献：**

[1] HAN B F, ZHENG R S, ZENG H M. et al. Cancer incidence and mortality in China, 2022 [J]. Journal of the National Cancer Center, 2024, 4 (1): 1 - 10.

[2] 滕 熠，曹毛毛，陈万青. 中国癌症筛查的发展、现状与挑战 [J]. 中国肿瘤，2022，31 (7)：481 - 487.

[3] MANISHA B, JOANNA L, NAFTALI B, et al. ASO author reflections: breast cancer early detection: if you build It, she will come [J]. Annals of Surgical Oncology, 2024, 31 (3): 1653 - 1654.

[4] FOULADI H, EBRAHIMI A, DERAKHSHAN S. et al. Over-expression of mir-181a-3p in serum of breast cancer patients as diagnostic biomarker [J]. Molecular Biology Reports, 2024, 51 (1): 1 - 10.

[5] CALIFF R M. Biomarker definitions and their applications [J]. Experimental biology and medicine, 2018, 243 (3): 213 - 221.

[6] DALERBA P, SAHOO D, PAIK S, et al. CDX2 as a prognostic biomarker in stage II and stage III colon cancer [J]. New England Journal of Medicine, 2016, 374 (3): 211 - 222.

[7] CHENG B, LI J, LI X. et al. MiR-323b-5p acts as a novel diagnostic biomarker for critical limb ischemia in type 2 diabeticpatients [J]. Scientific Reports, 2018, 8 (1): 15079 - 15080.

[8] 李爱玲，宋 健. 生物标志物分类及其在临床医学中的应用 [J]. 中国药理学与毒理学杂志，2015，29 (1)：7 - 20.

[9] GAMBLE P, JAROENSRI R, WANG H. et al. Determining breast cancer biomarker status and associated morpho-

logical features using deep learning [J]. Communications Medicine, 2021, 14 (1): 1-10.
[10] KHADIJEH BARZAMAN, JAFAR KARAMI, ZEINAB ZAREI, et al. Breast cancer: biology, biomarkers, and treatments [Z]. 2020, 84 (6), 1-10.
[11] CHEN X, CHEN D, ZHAO Z, et al. Artificial image objects for classification of breast cancer biomarkers with transcriptome sequencing data and convolutional neural network algorithms [J]. Breast Cancer Research, 2021, 23 (1): 96-96.
[12] O'GRADY NICHOLAS, GIBBS DAVID L, ABDILLEH KAWTHER, et al. PRoBE the cloud toolkit: finding the best biomarkers of drug response within a breast cancer clinical trial [J]. JAMIA Open, 2021, 4 (2): 38-38.
[13] 陈诗慧，刘维湘，秦　璟，等. 基于深度学习和医学图像的癌症计算机辅助诊断研究进展 [J]. 生物医学工程学杂志，2017，34 (2)：314-319.
[14] BAFTI S M, ANG C S, MARCELLI G, et al. BioGAN: an unpaired GAN-based image to image translation model for microbiological images [J]. ArXiv Preprint ArXiv: 2306.06217, 2023, 2023 (6): 1-13.
[15] 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): 166-176.
[16] 曹一珉，蔡　磊，高敬阳. 基于生成对抗网络的基因数据生成方法 [J]. 计算机应用，2022，42 (3)：783-790.
[17] YOON J, JORDON J, SCHAAR M V D. GAIN: missing data imputation using generative adversarial nets [J]. International Conference on Machine Learning, 2018, 18 (6): 1-10.
[18] STEKHOVEN, DANIEL J, PETER B. MissForest-nonparametric missing value imputation formixed-type data [J]. Bioinformatics, 2011, 28 (1): 112-118.
[19] IAN GOODFELLOW, JEAN POUGET-ABADIE, MEHDI MIRZA, et al. Generative adversarial networks [J]. Commun. ACM, 2020, 63 (11): 139-144.
[20] JUDEA PEARL. Causality: models, reasoning, and inference [J]. Econometric Theory, 2003, 19 (4): 675-685.
[21] MIRZA M, OSINDERO S. Conditional generative adversarial nets [J]. ArXiv Preprint ArXiv: 1411.1784, 2014, 2014 (11): 1-10.
[22] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN [EB/OL]. (2017-12) [2024-03]. https://arxiv.org/abs/1701.07875.
[23] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL]. (2015-11) [2024-3]. https://arxiv.org/abs/1511.06434v1.
[24] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein GANs [EB/OL]. (2017-3) [2024-03]. https://arxiv.org/abs/1704.00028.
[25] FRENCH J D, GHOUSSAINI M, EDWARDS S L, et al. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers [J]. Am J Hum Genet, 2013, 92 (4): 489-503.
[26] THOMPSON E, GORRINGE K, ROWLEY S, et al. Reevaluation of the BRCA2 truncating allele c.9976A＞T (p. Lys3326Ter) in a familial breast cancer context [J]. Sci Rep 5, 2015, 5 (1): 1-10.
[27] WYSZYNSKI A, HONG C C. Kristin LamAn intergenic risk locus containing an enhancer deletion in 2q35 modulates breast cancer risk by deregulating IGFBP5 expression [J]. Human Molecular Genetics, 2016, 25 (17): 3863-3876.

(上接第 199 页)

[11] 朱晓敏，刘大千，费博雯，等. 局部通信条件下多无人机协同搜索方法 [J]. 系统工程与电子技术，2022，44 (12)：3783-3791.
[12] 郑伟铭，周贞文，徐　扬，等. 针对运动目标的多无人机协同鸽群优化搜索方法 [J]. 控制理论与应用，2023，40 (4)：624-632.
[13] 岳　伟，李超凡. 基于多蜂群的多无人机协同自适应搜索 [J]. 科学技术与工程，2022，22 (5)：2108-2115.
[14] 邓杨赟，李文光，葛佳昊，等. 多无人机协同搜索及干扰算法研究 [J]. 战术导弹技术，2023，(5)：10-18.
[15] 卢　卓，吴启晖，周福辉. 有人机/无人机智能协同目标搜索和轨迹规划算法 [J]. 通信学报，2024，45 (1)：31-40.
[16] 陈　黎，李芳芳，邹长虹. 基于动态贝叶斯网络和模板匹配的空中目标意图识别 [J]. 现代防御技术，2023，51 (2)：62-70.
[17] 董志鹏. 基于卷积神经网络的高分辨率遥感影像目标检测方法研究 [J]. 测绘学报，2023，52 (9)：1613.
[18] 石争浩，仵晨伟，李成建，等. 航空遥感图像深度学习目标检测技术研究进展 [J]. 中国图象图形学报，2023，28 (9)：2616-2643.
[19] 付　涵，范湘涛，严珍珍，等. 基于深度学习的遥感图像目标检测技术研究进展 [J]. 遥感技术与应用，2022，37 (2)：290-305.
[20] 余　翔，邓千锐，段思睿，等. 一种多无人机协同优先覆盖搜索算法 [J]. 系统仿真学报，2024，36 (4)：991-1000.