第 14 卷 第 3 期
2025 年 5 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 14 No. 3
May 2025

引文格式：
戴伟, 张浩轩, 陈方旭, 等. 基于元学习的小样本癌症亚型分类算法 [J]. 集成技术, 2025, 14(3): 87-101.
Dai W, Zhang HX, Chen FX, et al. A meta-learning-based algorithm for few-shot cancer subtype classification [J]. Journal of Integration Technology, 2025, 14(3): 87-101.

基于元学习的小样本癌症亚型分类算法

戴 伟[1,2] 张浩轩[1,2] 陈方旭[1,2] 彭 玮[1,2*]

[1]（昆明理工大学信息工程与自动化学院 昆明 650050）
[2]（昆明理工大学云南省计算机技术应用重点实验室 昆明 650050）

摘 要 癌症是一种与基因密切相关的疾病，具有多种亚型，各亚型在遗传、表型和治疗反应上存在显著差异。准确的癌症亚型分类对个性化治疗至关重要，有助于提高治疗效果。然而，基于患者基因表达数据的癌症亚型分类方法在样本不均衡的情况下，往往难以有效区分稀有亚型。为解决这一问题，本文提出一种基于元学习的癌症亚型分类方法 MFP-VAE (meta-learning few-shot prototype learning VAE)，专注于处理样本不均衡的数据集。该方法改进了样本抽取策略，以确保在元学习任务中不同亚型的样本得到平衡重视。该模型采用变分自编码器进行特征提取，并通过计算样本与癌症亚型对应原型之间的距离进行分类。实验结果表明，MFP-VAE 在两个公开癌症数据集上的表现优于现有方法，特别是在样本不平衡的情况下，显著提高了分类效果。此外，生存率分析显示，不同的癌症亚型在临床特性上具有显著差异。

关键词 癌症亚型分类；元学习；变分自编码器；小样本学习
中图分类号 TP399，R730.49 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20241012001
CSTR: 32239.14.j.issn.2095-3135.20241012001

A Meta-Learning-Based Algorithm for Few-Shot Cancer Subtype Classification

DAI Wei[1,2] ZHANG Haoxuan[1,2] CHEN Fangxu[1,2] PENG Wei[1,2*]

[1](*Faculty of Information Engineering and Automation*, *Kunming University of Science and Technology*, *Kunming* 650050, *China*)
[2](*Computer Technology Application Key Lab of Yunnan Province*, *Kunming University of Science and Technology*, *Kunming* 650050, *China*)
[*]*Corresponding Author*: *weipeng1980@126.com*

Abstract Cancer is a genetically related disease with multiple subtypes, each exhibiting significant differences in genetics, phenotype, and treatment response. Accurate classification of cancer subtypes is critical for personalized treatment, as it helps improve therapeutic outcomes. However, cancer subtype classification methods based on patient gene expression data often struggle to effectively distinguish rare

收稿日期：2024-10-12 **修回日期**：2025-01-15
基金项目：国家自然科学基金项目 (6192185, 62472202)
作者简介：戴伟，副教授，研究方向为计算机应用技术； 张浩轩，硕士研究生，研究方向为大数据、生物信息学； 陈方旭，硕士研究生，研究方向为生物信息学； 彭玮 (通讯作者)，教授，研究方向为生物信息学、机器学习，E-mail：weipeng1980@126.com。

subtypes in the presence of imbalanced samples. To address this issue, a cancer subtype classification method called MFP-VAE (meta-learning few-shot prototype learning VAE) is proposed, focusing on handling datasets with imbalanced samples. This method improves the sampling strategy to ensure balanced consideration of different subtypes in meta-learning tasks. The model employs a variational autoencoder for feature extraction and classifies samples by calculating the distance between the samples and their corresponding cancer subtype prototypes. Experimental results show that MFP-VAE outperforms existing methods on two public cancer datasets, significantly improving classification performance, especially under imbalanced sample conditions. Furthermore, survival analysis reveals that the distinguished cancer subtypes exhibit significant differences in clinical characteristics.

Keywords cancer subtype classification; meta-learning; variational autoencoder; few-shot learning

Funding This work is supported by National Natural Science Foundation of China (6192185, 62472202)

1 引 言

癌症的本质在于细胞失控异常增长，而失控的细胞可能侵入周围组织并导致转移[1]。因此，发现癌症的内在多样性和异质性是治疗的首要难题[2]。即使是相同类型的癌症，不同患者之间也可能因遗传差异而表现出不同的临床特征[3]。现代医学根据癌症的分子属性细分亚型，设计针对性的治疗方案[4-5]。例如，乳腺癌分为 *HER-2* 阴性 (Lum A)、*HER-2* 阳性 (Lum B)、三阴性乳腺癌 (Basal) 和 *HER-2* 受体阳性 (Her2) 4 种亚型[6]。胶质母细胞瘤分为经典型 (Classical)、间充质型 (Mesenchymal) 和前神经元型 (Proneural) 3 种亚型[7]。针对不同亚型的肿瘤形态和临床特点，有相应的靶向药物和治疗方案可供选择[8-9]。各亚型发病机理不同，导致样本数量差异显著[10-11]。精准识别癌症亚型对精准医疗和个性化治疗至关重要，不仅能揭示癌症的发展机制，还能为个体化治疗奠定坚实基础[12]。

现有的单组学癌症亚型分类大多将基因表达数据作为特征，临床上通过捕获基因表达值的变化诊断癌症的不同亚型。现有的癌症亚型分类方法主要有两阶段方法和端到端方法。

两阶段方法采用两个单独的步骤，第一步提取样本特征，第二步使用特征分类模型进行癌症亚型分类。第一步常用的方法包括主成分分析 (principal component analysis，PCA)[13]、非负矩阵分解 (non-negative matrix factorization，NMF)[14]、自动编码器 (autoencoder，AE)[15]、稀疏自编码器[16]、叠加自编码器 (stacked autoencoder，SAE)[17]和变分自编码器 (variational autoencoder，VAE)[18]等。第二步常使用的分类模型包括支持向量机 (support vector machine，SVM)[19]、k 近邻[20]、随机森林 (random forest，RF)[21]、神经网络[22]和卷积神经网络[23]等。例如，Maulik 等[24]通过特征选择和转导支持向量机进行癌症亚型分类。

端到端方法将特征提取与分类任务合并为一个整体。随着计算能力的提升和深度学习技术的发展，研究者开始应用深度神经网络进行癌症亚型分类[25]。例如，Chen 等[26]开发了 Deep Type 方法，旨在解决高维数据处理中的混淆和重叠问题。该方法将特征映射到特定隐空间，通过联合监督分类、无监督聚类和降维，学习具有聚类结构的癌症数据表示。Dai 等[27]提出了 ERGCN 方法，将基因表达谱和样本相似性网络输入图卷积

网络实现癌症亚型分类，效果显著。Zhao 等[28]提出了 Subtype-DCC 方法，将对比学习表示与聚类相结合，通过引入自监督学习范式，共同优化深度表示学习和聚类参数，从而学习合适的特征表示。尽管深度学习在高维数据上表现出色，但仍面临挑战，如数据稀缺时模型表现不佳和数据不均衡可能导致模型偏向多数类等。

为解决样本不平衡问题，研究者提出了一些应对策略。Park 等[18]通过深度自动编码器提取特征，并结合过采样技术，缓解样本不平衡现象。Wang 等[29] 提出了自适应学习的不平衡采样方法，能更有效地挑选高质量样本，提高分类准确率。这些方法虽然在一定程度上减少了样本不平衡的影响，但仍未完全消除该问题。

近年来，元学习逐渐成为应对样本不平衡的有效方法。通过少量样本学习机制，在数据稀缺场景下仍能保持优异性能，有效缓解因某亚型样本过度主导导致的分类偏差。该框架通过多任务训练，实现亚型间样本分布的均衡性，显著提升模型对稀有亚型的分类效果。

本研究聚焦于样本不平衡和数据稀缺问题，精心设计了一个多任务元学习框架，创新性地提出了基于元学习的分类方法 MFP-VAE(meta-learning few-shot prototype learning VAE)，在数据处理层面，通过均衡采样各亚型样本，有效缓解了样本数量差异导致的不平衡问题，确保模型在训练过程中能全面考量各类亚型的特征模式，避免对多数类样本的过度依赖。在模型分类机制上，将每类原型与其他样本的相似度作为分类依据，进一步提升了分类的准确性和可靠性。通过大量实验验证，该方法在多个癌症亚型分类任务中展现出卓越性能，在样本稀缺的亚型分类中表现格外突出，有力证明了其在应对小样本数据挑战方面的有效性。

本文的核心创新点主要体现在两个关键方面：

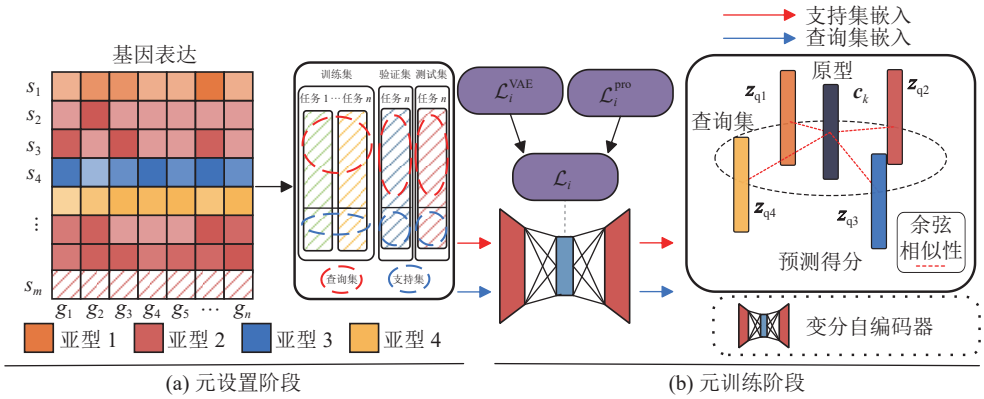
其一，构建了基于元学习的癌症亚型分类框架。该框架的独特之处在于能确保各亚型在元学习任务中拥有均衡的样本数量。这种均衡性不仅有效解决了样本不平衡问题，还为小样本亚型提供了充分的学习机会，使得模型能精准捕捉各类亚型的细微特征差异。VAE 提取的经过去噪和优化后的特征能使原型的构建更精准地反映各亚型的本质特征，从而在样本与原型进行相似度比较时，更准确地判断样本所属的亚型类别，显著提升了模型在多亚型分类任务中的综合性能。

其二，在多个具有代表性的癌症数据集上进行了全面验证。实验结果表明，本文方法在分类性能和模型泛化能力方面均取得了显著提升。特别是在稀缺亚型分类任务中，与传统方法相比，各项评价指标均有显著改善，充分彰显了该方法在处理小样本癌症亚型分类问题上的优势和潜力。

2　材料和方法

本研究的目标是解决多元分类问题，即通过基因表达数据提取特征，对样本进行癌症亚型划分。为此，本文提出了一个名为 MFP-VAE 的模型，包含元设置和元训练两个阶段，其整体框架如图 1 所示，其中图 1 (a) 表示元设置阶段，图 1 (b) 表示元训练阶段。

元设置阶段完成数据集划分和任务构建。首先将癌症样本数据集划分为训练集、验证集和测试集。训练集用于元学习，验证集用于元学习超参数评估和调整，测试集用于对模型进行评估。元学习的核心目标是学习一个基本模型 $f_\theta(\cdot)$(其中，f_θ 为参数 θ 控制的特征提取函数，输入为基因表达数据，输出为低维特征表示)，使其在癌症亚型预测任务上表现良好。元学习任务包括元训练任务、元验证任务和元测试任务，分别基于训练集、验证集和测试集抽样构建。每个元任务的数据进一步划分为支持集和查询集。支持集用于在当前任务上训练模型，而查询集则用于评估



图中 $\boldsymbol{c}_k$ 表示原型特征向量，$\boldsymbol{z}_{q1}$–$\boldsymbol{z}_{q4}$ 表示样本的潜在特征向量。

图 1　MFP-VAE 算法模型架构

Fig. 1　MFP-VAE algorithm model architecture

模型在当前任务上的泛化能力和性能。

元学习阶段，模型进行多次迭代，每次迭代为一个 epoch (过程)，每个 epoch 中包括多次 episode (回合) 的训练和验证。首先从元训练任务中抽取数据，使用基本模型 $f_\theta(\cdot)$(即特征提取器) 对支持集样本进行特征编码，并按类别计算各癌症亚型的原型特征向量，再计算查询集中每个样本与各类原型的距离，更新模型参数，上述过程为一个训练 episode。经过多次训练 episode 后，通过元验证任务对模型进行验证，比较各 epoch 的验证结果，选择最佳模型参数。整个过程重复多个 epoch，最终在测试集上评估模型。

测试过程中，选择多个元测试任务，利用支持集的数据预测查询集中的样本类别，最终性能通过多次元测试任务的平均结果确定。

2.1　元设置阶段

在元设置阶段，数据集划分和任务构建是模型学习的关键。本文采用严格的数据划分策略，结合五折交叉验证方法，将数据集划分为训练集、验证集和测试集，以减少数据划分的偶然性，提升模型的泛化能力。数据集均分为 5 个子集，每次选择 4 个子集用于训练，一个子集用于测试，重复 5 次。用于训练的子集进一步划分为训练集和验证集，同样采用五折交叉验证。

任务是元学习的基本单元，需构建训练、验证和测试任务集，任务集定义为 $\mathcal{D}=\{T_i\}_{i=1}^{I}$。每个任务 T_i 分为支持集和查询集，支持集用于训练，查询集用于评估模型的泛化能力。每个任务的目标是基于支持集的学习预测查询集的样本类别，最小化查询集上的损失。支持集采用小样本学习里的 N-way-K-shot 策略构建，对 N 类癌症亚型样本进行采样，采样个数为 K。考虑到癌症亚型样本数量的不均衡性特点，查询集也采用小样本学习里的 N-way-K-shot 策略构建，确保每类亚型样本都参与了模型的评估。查询集包含 N 个类别的 M 个样本。支持集和查询集互不相交。任务可以表示如下：

$$T_i=(S_i,Q_i),\left|S_i^k\right|=\frac{|S_i|}{N},\left|Q_i^k\right|=\frac{|Q_i|}{N},S_i\cap Q_i=\varnothing, \tag{1}$$

其中，S_i 表示任务 T_i 的支持集；Q_i 表示任务 T_i 的查询集；S_i^k 表示支持集 S_i 中属于第 k 个癌症亚型的样本；Q_i^k 表示查询集 Q_i 中属于第 k 个癌症亚型的样本；k 为样本的类别，$k\in\{1,2,\cdots,N\}$。以乳腺浸润性癌 (Breast Invasive Carcinoma，BRCA) 数据集为例，其包含 4 种癌症样本，数据划分和任务设置如图 2 所示。

2.2　元训练阶段

元训练阶段基于元设置阶段构建的任务集进

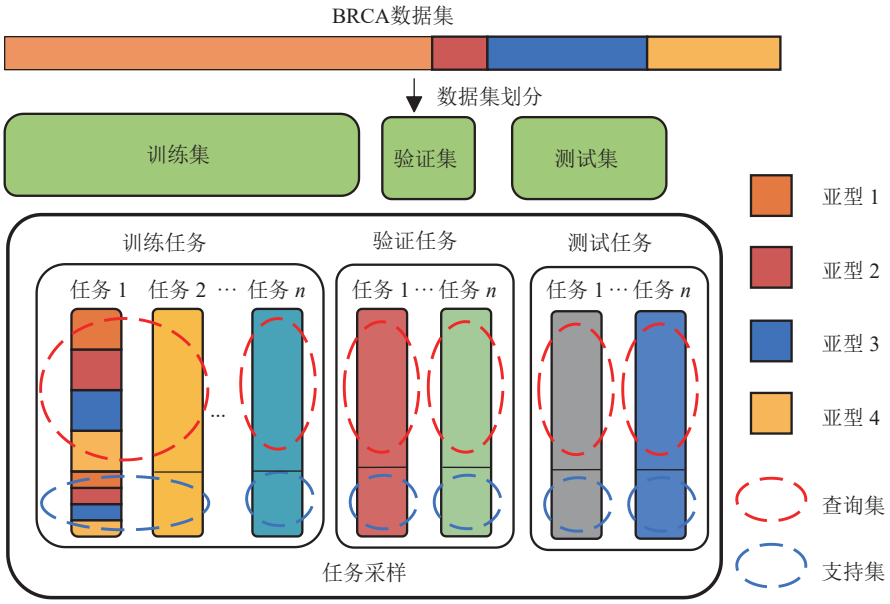


图 2　乳腺浸润性癌 (BRCA) 数据集的训练、验证和测试集划分

Fig. 2　Breast Invasive Carcinoma (BRCA) dataset training, validation, and test set partitioning

行模型训练。本文的基学习器由特征提取器 f_θ 和原型分类器构成，其中，特征提取器采用 VAE，输入维度为样本数据维度。VAE 能学习数据的潜在空间表示，通过编码器将输入数据映射到潜在空间，并逼近先验分布。VAE 作为特征提取器的优势如下：一是 VAE 能捕捉癌症亚型的关键特征，促进元学习中多任务间的共性特征提取；二是 VAE 能学习数据的连续表示，使模型对输入数据变化具有一定鲁棒性。VAE 的损失函数表示如下：

$$\mathcal{L}_i^{\mathrm{VAE}}(\theta,\phi,\boldsymbol{x})=-\mathbb{E}_{q_\theta(z|\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|z)]+\mathrm{KL}(q_\theta(z|\boldsymbol{x})\|p(z)) \tag{2}$$

其中，$q_\theta(z|\boldsymbol{x})$ 是由编码器参数 θ 定义的潜在变量 z 下数据 $\boldsymbol{x}$ 的条件概率分布；$p_\phi(\boldsymbol{x}|z)$ 为解码器参数 ϕ 定义的给定潜在变量 z 下数据 $\boldsymbol{x}$ 的条件分布；$p(z)$ 为潜在变量的先验分布。最小化真实后验分布 $p_\phi(\boldsymbol{x}|z)$ 与近似分布 $q_\theta(z|\boldsymbol{x})$ 之间的 KL 散度，并将其作为正则化项纳入损失函数中。该研究中的 BRCA 数据集输入层维度为 20 531，隐藏层为 256；多形性胶质母细胞瘤 (Glioblastoma Multiforme，GBM) 数据集的输入维度为 12 042，隐藏层为 256。

接下来将每个 episode 抽取到的任务 $\{T_i\}=\{S_i,Q_i\}$，输入编码器 f_θ，得到支持集 S_i 和查询集 Q_i 的样本在潜在空间的低维特征，再输入原型分类器进行分类。原型分类器首先通过计算支持集中的每类样本低维特征，得到该类亚型的嵌入中心作为原型 $\boldsymbol{c}_k$，再计算查询集中每个样本与各癌症亚型原型之间的距离。支持集亚型 k 的原型表示如下：

$$\hat{\boldsymbol{c}}_k=\frac{1}{|S_i^k|}\sum_{\boldsymbol{x}\in S_i^k}f_\theta(\boldsymbol{x}),\ \boldsymbol{c}_k=\frac{\hat{\boldsymbol{c}}_k}{\|\hat{\boldsymbol{c}}_k\|_2} \tag{3}$$

其中，$\hat{\boldsymbol{c}}_k$ 为未归一化的原型；$\boldsymbol{c}_k$ 为归一化后的原型；$f_\theta(\boldsymbol{x})$ 为编码器的输出；S_i^k 为支持集 S_i 中第 k 个癌症亚型样本的集合；$\boldsymbol{x}$ 属于该集合的样本；$|S_i^k|$ 为支持集 S_i 中第 k 个癌症亚型的样本数量。通过归一化 $\hat{\boldsymbol{c}}_k$ 得到原型 $\boldsymbol{c}_k$，并进行分类。

为确保亚型之间的区分度，查询集样本通过编码器得到的特征应与同类原型相似，并与其他亚型的原型有所区别，确保特征的区分性。原型对比损失 $\mathcal{L}_i^{\mathrm{pro}}$ 可表示如下：

$$\mathcal{L}_i^{\text{pro}} = -\log \frac{\exp(\text{sim}(f_\theta(\boldsymbol{x}_\text{q}), c_k))}{\sum_{k'=1}^{k} \exp(\text{sim}(f_\theta(\boldsymbol{x}_\text{q}), c_{k'}))} \tag{4}$$

其中，$\boldsymbol{x}_\text{q}$ 为查询集中的样本；$c_{k'}$ 为第 k' 个癌症亚型分类的原型，$k' \in \{1,2,\cdots,k\}$；$\text{sim}(\cdot)$ 为相似度函数。

最终，通过损失更新模型参数。总损失由 VAE 编码器损失和原型分类器损失构成，表示如下：

$$\mathcal{L}_i^{\text{total}} = \alpha \mathcal{L}_i^{\text{pro}} + (1-\alpha)\mathcal{L}_i^{\text{VAE}} \tag{5}$$

其中，α 为权重因子。

在元学习阶段，模型经历 N_E 次迭代，每个迭代称为一个 epoch。每个 epoch 中，模型经历 N_e 个训练 episode 和验证 episode。首先，每个训练 episode 从预先构建的元训练任务集中抽取任务。损失函数度量模型预测与真实标签间的差异，并通过反向传播更新参数。模型在完成多个训练 episode 后，通过元验证任务进行验证，并基于验证集的平均性能更新参数。完成多个 epoch 的训练后，对测试任务集进行测试，通过多次测试 episode 的平均结果评估模型性能。算法 1 为 MFP-VAE 算法模型伪代码，如图 3 所示。

对于一次训练，设有 N_E 个 epoch，N_e 个 episode，每个任务中有 $\boldsymbol{x}_\text{s}$ 个支持集，$\boldsymbol{x}_\text{q}$ 个查询集。对于每个任务，VAE 编码时间复杂度大致为 $O(D)$，其中 D 为输入数据的维度。整体算法的时间复杂度大致为 $O(N_\text{E} \times N_\text{e} \times (\boldsymbol{x}_\text{s} + \boldsymbol{x}_\text{q} + D))$。

算法 1 MFP-VAE 算法模型伪代码

输入： 基因表达矩阵 $\boldsymbol{X} \in R^{m \times n}$ 具有 m 个样本，其向量长度为 n，真实标签 $\boldsymbol{Y} \in R^s$，原型特征为 $\boldsymbol{c}_k$，支持集的样本为 $\boldsymbol{x}_\text{s}$，查询集的样本为 $\boldsymbol{x}_\text{q}$，S_i 和 Q_i 为任务 i 的支持和查询集，加权超参数为 α，学习率为 β，迭代次数为 N_E，每个任务的轮数为 N_e

输出： lable

随机初始化 θ

for $m = 1$ to N_E **do**

从训练集随机采样 N_e 个任务 $\{T_i\}_{i=1}^{N_\text{e}} \sim p(T)$

for $i = 1$ to N_e **do**

利用式(1)采样任务 T_i 的支持集 S_i 和查询集 Q_i

根据式(3)使用 $\boldsymbol{x}_\text{s}$ 计算 $\boldsymbol{c}_k$，$\boldsymbol{z}_\text{q} \leftarrow \text{VAE}(\boldsymbol{x}_\text{q})$，$\boldsymbol{x}_\text{q} \in Q_i$

$\mathcal{L}_{T_i}^{\text{total}}(\theta) = \alpha\ \mathcal{L}_{T_i}^{\text{pro}}(c_k, z_\text{q}) + (1-\alpha)\,\mathcal{L}_{T_i}^{\text{VAE}}$

更新 $\theta \leftarrow \theta - \beta \nabla_\theta\ \mathcal{L}_{T_i}^{\text{total}}(\theta)$

end for

从验证集随机采样 N_e 个任务 $\{T_i\}_{i=1}^{N_\text{e}} \sim p(T)$

for $i = 1$ to N_e **do**

利用式(1)采样任务 T_i 的支持集 S_i 和查询集 Q_i

根据式(3)使用 $\boldsymbol{x}_\text{s}$ 计算 $\boldsymbol{c}_k$，$\boldsymbol{z}_\text{q} \leftarrow \text{VAE}(\boldsymbol{x}_\text{q})$，$\boldsymbol{x}_\text{q} \in Q_i$

将 x_q 对比原型 $\boldsymbol{c}_k$ 并计算 ACC_i

end for

$$\text{ACC}_m = \frac{1}{N_\text{e}} \sum_{i=N_\text{e}} \text{ACC}_i$$

if $\text{ACC}_{\max} < \text{ACC}_m$ **then**

$\theta_{\text{meta}} = \theta$，$\text{ACC}_{\max} = \text{ACC}_m$

end if

end for

for $i = 1$ to N_e **do**

利用式(1)采样任务 T_i 的支持集 S_i 和查询集 Q_i

根据式(3)使用 $\boldsymbol{x}_\text{s}$ 计算 $\boldsymbol{c}_k$，$\boldsymbol{z}_\text{q} \leftarrow \text{VAE}(\boldsymbol{x}_\text{q})$，$\boldsymbol{x}_\text{q} \in Q_i$

将 $\boldsymbol{x}_\text{q}$ 对比原型 $\boldsymbol{c}_k$ 并计算 lable

end for

return lable

图 3 MFP-VAE 算法模型伪代码

Fig. 3 Pseudo-code of the MFP-VAE algorithm model

3 实 验

3.1 实验数据集

数据集来源于 TCGA 数据库 (https://cancergenome.nih.gov)，由美国国家癌症研究所和国家人类基因组研究所创建[30]。该数据库收录了超过 47 种癌症类型的基因数据和样本信息。

本文选取了两种癌症数据集：BRCA 和 GBM。BRCA 数据集包含 686 个样本，GBM 数据集包含 274 个样本。根据 Rappoport 等[31]的补充材料，该研究团队基于基因表达和生存分析数据进行数据处理。首先筛选含基因表达数据的样本，并整合标签。最终得到 646 名乳腺癌患者数据，分为 Lum A (356)、Lum B (46)、Basal (111) 和 Her2 (133) 4 种亚型；202 名 GBM 患者，分为 Classical (70)、Proneural (49)

和 Mesenchymal (83) 3 种亚型。各亚型的样本数量如表 1 所示。

表 1　相关癌症数据集

Table 1　Related cancer dataset

数据集	样本数量	数据维度
BRCA	646[356,46,111,133](686)	20 531
GBM	202[70,49,83](274)	12 042

3.2　评价指标

本文使用了两种评价指标：外部评价指标和内部评价指标。

外部评价指标包含精确率 (Precision)、召回率 (Recall)、F1 分数、准确率 (ACC)、马修斯相关系数 (MCC) 和调整兰德指数 (ARI)。它们通过比较算法输出与真实标签的吻合程度评估性能，表示如下：

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{6}$$

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{7}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

$$\mathrm{F1}=2\times\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{9}$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{10}$$

$$\mathrm{ARI}=\frac{\mathrm{RI}-E(\mathrm{RI})}{\max(\mathrm{RI})-E(\mathrm{RI})},\quad \mathrm{RI}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{11}$$

$$E(\mathrm{RI})=\frac{(\mathrm{TP}+\mathrm{TN})(\mathrm{TP}+\mathrm{FP})}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}},$$

$$\max(\mathrm{RI})=\frac{(\mathrm{TP}+\mathrm{TN})+(\mathrm{TP}+\mathrm{FP})}{2} \tag{12}$$

其中，TP 为真阳性样本数；TN 为真阴性样本数；FP 为假阳性样本数；FN 为假阴性样本数；RI 为原始兰德指数；E (RI) 为随机分配下的期望兰德指数。

内部评价包含轮廓系数 (Silhouette Index，SI) 和 Davies-Bouldin 指数 (DBI)。轮廓系数考虑了聚类的紧密度和分离度，用于评价一个聚类的质量。DBI 旨在通过量化聚类的紧密性和分离度衡量聚类质量，值越低，聚类效果越好。内部评价指标用于衡量聚类的质量和结构，反映样本之间的相似度，表示如下：

$$\mathrm{SI}=\frac{b(i)-a(i)}{\max(a(i),b(i))} \tag{13}$$

$$\mathrm{DBI}=\frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\left(\frac{S_i+S_j}{d(i,j)}\right) \tag{14}$$

其中，a (i) 为样本 i 的簇内平均距离；b (i) 为样本 i 到其他簇的最小平均距离；K 为聚类数；S_i 为簇 i 的标准差；d (i, j) 为簇 i 与 j 的距离。

3.3　实验设置

本研究采用重复 10 次的五折交叉验证方法评估模型性能。首先，将数据集随机划分为 5 个不重叠的子集。然后，选择其中 4 个子集用于训练和验证，另一个子集用于测试。每个用于训练和验证的子集均采用五折交叉验证，4 个子集用于元训练，一个子集用于元验证。在元训练与元验证之后，选择最优模型，最终在测试集中进行测试。上述过程重复 10 次，以减少评估结果对特定数据划分的依赖。

样本数据定义为基因特征表达值，经过筛选、归一化处理后，从数据集中抽取 48 个样本作为任务 T_i，并计算损失函数，目标是学习一个基本模型 $f_\theta(\cdot)$，使得该模型在癌症亚型预测任务上表现良好。

实验使用 BRCA 和 GBM 数据集，其中 BRCA 数据集包含 4 种癌症亚型，GBM 数据集包含 3 种癌症亚型。每个任务的支持集和查询集均采用 4-way-4-shot 构建，即从 4 个亚型中各抽取 4 个样本。优化器使用 Adam，学习率为 1×10^{-6}，训练轮次为 20，训练集 episode 为 50；支持集数量为 4 个，查询集为 8 个。

在其他领域的小样本学习中，支持集虽然可采用较小的 K-shot 值，但样本过少可能限制模型的泛化能力。在本研究中，由于某些亚型样本数

量较少，因此尽可能增加查询集数量。为测试参数影响，支持集 K 设为 1～5，查询集 K 设为 1～10，通过调整支持集与查询集的样本数量，系统比较了模型在不同 K 值组合下的性能表现，结果如图 4 和图 5 所示。

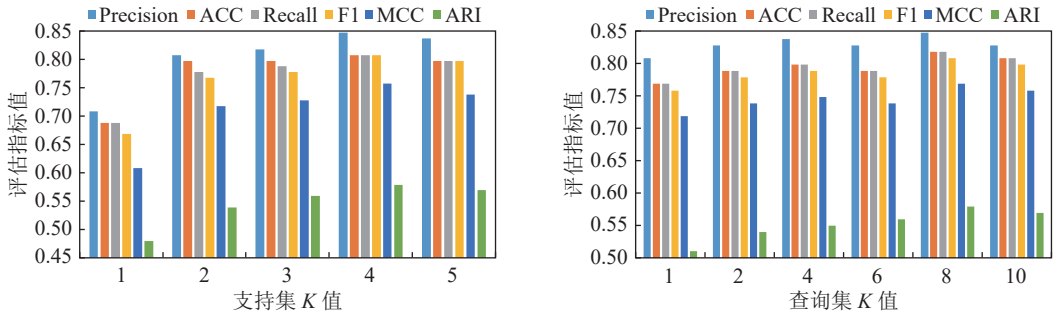
可观察到，随着支持集样本数量 K 的增加，模型训练效果逐渐提高。当设置 N=4、支持集 K 为 4、查询集 K 为 8 时，模型在两数据集上性能最佳。

训练过程分为两个主要阶段：epochs 和 episodes。首先以 episode 为基本单位进行训练，在每个 episode 中，模型参数通过内循环训练更新，并在验证集上评估。重复多个 epochs，直至达到终止条件，记录验证集上最佳参数，用于独立测试集评估，检测模型的泛化能力。

如图 6 和图 7 所示，当增加内部训练循环 (episodes) 数量时，模型适应速度加快，可在较少的外部训练循环 (epochs) 内完成训练。但增加 epochs 数量会提高训练时间成本，过多 epochs 也会导致过拟合，所以需要找到平衡点。最终设定 episode 为 50 次，epoch 为 20 次。

3.4 实验结果与分析

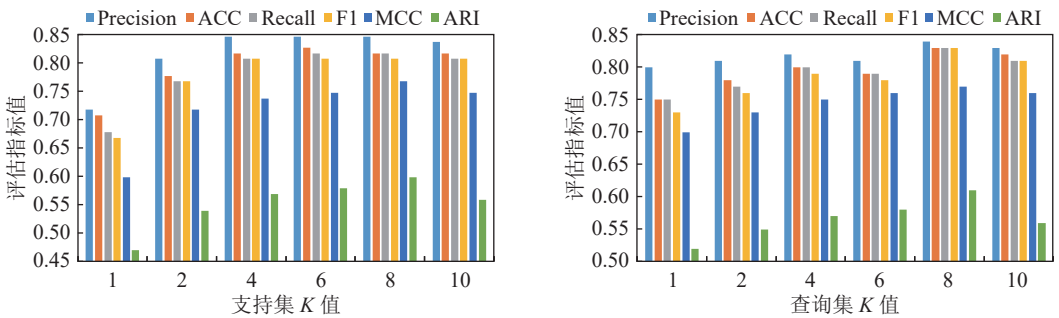
为评估模型性能，本研究进行了传统机器学习方法与 3 种新颖的深度学习方法的对比实验。传统机器学习方法采用 SAE[17]、VAE[18]、条件变分自编码器 (conditional variational autoencoder，CVAE)[32]、PCA[13]和 NMF[14]作为编码器，并结合 SVM[19]、RF[21]和多层感知机 (MLP)[33]作为分类器。深度学习方法包括 Deep Type[26]、ERGCN[27]和 Subtype-DCC[28]。Deep Type 基于监督分类损失、无监督聚类和降维技术；ERGCN 通过样本相似性构建网络，使用残差图卷积网络进行分



(a) 查询集 K=8 时, 支持集 K 值变化　(b) 支持集 K=4 时, 查询集 K 值变化

图 4　乳腺浸润性癌 (BRCA) 数据集超参数优化结果

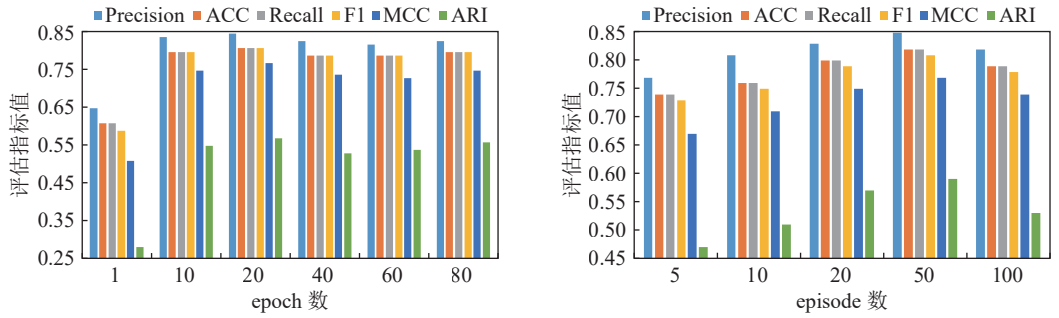
Fig. 4　Breast Invasive Carcinoma (BRCA) dataset hyperparameter optimization result



(a) 查询集 K=8 时, 支持集 K 值变化　(b) 支持集 K=4 时, 查询集 K 值变化

图 5　多形性胶质母细胞瘤 (GBM) 数据集超参数优化结果

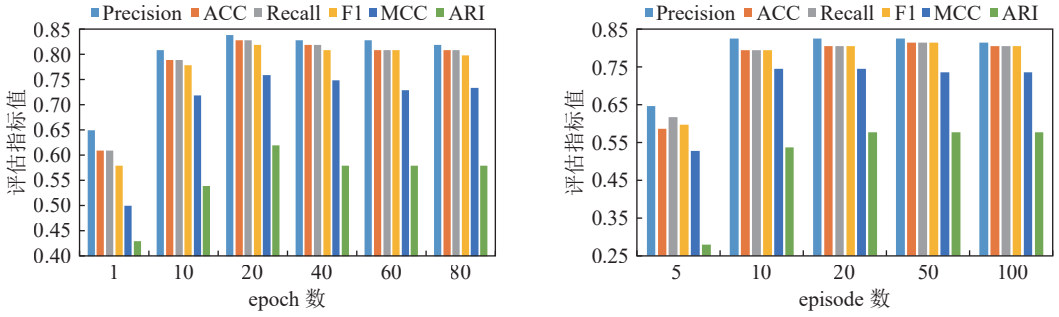
Fig. 5　Glioblastoma Multiforme (GBM) dataset hyperparameter optimization result



(a) episode＝50 时, epoch 数变化　　(b) epoch＝20 时, episode 数变化

图 6　乳腺浸润性癌 (BRCA) 数据集 epoch 和 episode 性能

Fig. 6　Breast Invasive Carcinoma (BRCA) dataset epoch and episode performance



(a) episode＝50 时, epoch 数变化　　(b) epoch＝20 时, episode 数变化

图 7　多形性胶质母细胞瘤 (GBM) 数据集 epoch 和 episode 性能

Fig. 7　Glioblastoma Multiforme (GBM) dataset epoch and episode performance

类；Subtype-DCC 是一种基于解耦对比学习的深度聚类方法，通过将对比学习引入深度聚类，同时在样本空间和聚类空间进行对比学习，并通过联合优化确保学习到的特征既有个体差异性，又适合聚类，属于自监督学习方法。本实验将其作为对比方法在内部指标上进行比较。

模型架构与参数设置：VAE 包括输入层和 3 个隐藏层 (1 024、512、256 个神经元) 的层次，SAE 包括输入层和 3 个隐藏层 (500、200、50 个神经元)，CVAE 的网络结构包括输入层和 3 个隐藏层 (512、512、128 个神经元)。训练轮次设为 300。PCA 和 NMF 采用默认参数，Deep Type[26]、ERGCN[27]和 Subtype-DCC[28]的参数设置均遵循对应文献的默认设置。

实验环境如下：

硬件方面：CPU 为 Intel Xeon Gold 5218，显卡为 GeForce RTX 3090，内存为 128 GB；

软件方面：操作系统为 Linux Ubuntu 20.04.2 LTS，Conda 4.12.0，Python 3.7.12，PyTorch 1.8，CUDA 12.2，scikit-learn 1.0.2。

3.4.1　外部评价指标对比

本文方法与其他方法在外部评价指标上的平均性能对比如表 2 和表 3 所示，最好结果由粗体标识，次优结果由下划线标识。经过 10 次五折交叉验证后，将平均测试集结果作为评价指标衡量模型性能。

在 BRCA 数据集上，与目前性能最佳的 ERGCN 模型相比，MFP-VAE 在精确率、F1 分数、召回率和马修斯相关系数方面均实现了显著提升，分别提升约 4%、5%、3% 和 5%；在 ACC 指标上，MFP-VAE 与 ERGCN 基本持平；而 ARI 指标略有下降。此外，与 Deep Type 相

表 2 乳腺浸润性癌 (BRCA) 数据集外部评价

Table 2 External evaluation of Breast Invasive Carcinoma (BRCA) dataset

方法	ACC	精确率	F1 分数	召回率	ARI	MCC
PCA＋SVM	0.704 40±0.02	0.694 75±0.04	0.662 03±0.02	0.661 45±0.06	0.503 17±0.02	0.680 84±0.04
NMF＋RF	0.737 33±0.02	0.701 59±0.03	0.779 97±0.03	0.704 75±0.02	0.458 00±0.06	0.696 78±0.02
SAE＋MLP	0.683 64±0.02	0.759 41±0.02	0.603 34±0.05	0.779 40±0.02	0.526 86±0.02	0.611 47±0.03
VAE＋SVM	0.800 00±0.03	0.775 11±0.02	0.648 64±0.02	0.730 92±0.03	0.464 81±0.04	0.614 06±0.05
CVAE＋MLP	0.687 17±0.01	0.728 20±0.02	0.675 31±0.02	0.706 55±0.05	0.430 86±0.03	0.623 56±0.02
Deep Type	0.689 24±0.01	0.701 22±0.04	0.741 17±0.02	0.731 94±0.02	0.421 19±0.02	0.625 64±0.02
ERGCN	**0.827 31±0.01**	0.809 38±0.01	0.770 55±0.01	0.793 12±0.01	**0.616 78±0.01**	0.734 99±0.01
MFP-VAE	0.826 01±0.01	**0.848 83±0.01**	**0.822 60±0.01**	**0.826 00±0.01**	0.600 14±0.08	**0.780 48±0.05**

表 3 多形性胶质母细胞瘤 (GBM) 数据集外部评价

Table 3 External evaluation of Glioblastoma Multiforme (GBM) dataset

方法	ACC	精确率	F1 分数	召回率	ARI	MCC
PCA＋SVM	0.768 27±0.03	0.773 25±0.02	0.789 53±0.02	0.766 12±0.02	0.514 38±0.05	0.692 43±0.06
NMF＋RF	0.738 24±0.02	0.752 43±0.04	0.701 07±0.03	0.701 96±0.03	0.447 15±0.05	0.623 91±0.05
SAE＋MLP	0.766 32±0.02	0.745 38±0.02	0.734 55±0.03	0.748 26±0.02	0.464 13±0.04	0.669 84±0.04
VAE＋SVM	0.742 07±0.02	0.733 01±0.03	0.714 36±0.02	0.724 85±0.02	0.433 26±0.09	0.646 59±0.02
CVAE＋MLP	0.808 18±0.02	0.794 67±0.01	0.772 83±0.02	0.781 54±0.01	0.573 24±0.05	0.726 87±0.02
Deep Type	0.683 45±0.01	0.662 81±0.02	0.629 94±0.04	0.651 72±0.05	0.378 93±0.07	0.577 02±0.05
ERGCN	0.860 89±0.01	0.845 05±0.02	0.839 81±0.02	0.845 26±0.02	0.626 16±0.03	**0.789 28±0.01**
MFP-VAE	**0.862 33±0.01**	**0.885 17±0.01**	**0.856 36±0.01**	**0.862 33±0.01**	**0.628 39±0.08**	0.787 57±0.05

比，MFP-VAE 在各项指标上均取得较优结果，平均提升率约 20%，并大幅领先传统机器学习的组合。对于性能的差异，本文做了细致的分析，具体如下：在癌症亚型分类任务中，当模型倾向于将所有样本预测为某一常见类别时，ACC 指标仍可能显示较好。在本实验中，鉴于数据分布的固有特点和模型处理样本的方式差异，ERGCN 在 ACC 指标上表现出一定优势。而本文模型由于采用平均采样策略，旨在均衡对待各类亚型，因此可能无法像 ERGCN 那样在 ACC 指标上达到较高水平。这是因为本文模型更注重对各类亚型的全面考量，而不是过度依赖数据集中占主导地位的亚型信息提高准确率指标。而 MCC 是一个更平衡的指标，尤其在样本不平衡的情况下，能更全面地评估模型性能。本文的 MFP-VAE 方法在 MCC 指标上超越了 ERGCN，提升达 4.5 个百分点，表明本文模型在处理样本不平衡问题上具有一定优势，能更准确地反映模型在不同类别样本上的分类能力。

为防止数据集的偏性，在 GBM 数据集上进行相同实验，取得了与 BRCA 数据集类似的结果。在该数据集上，ACC、精确率、F1 分数、召回率和 ARI 等指标均优于基线模型。而在马修斯相关系数指标上取得了次优结果，但与最优结果仅相差约 0.2 个百分点。总体而言，MFP-VAE 模型在两数据集上的表现充分展示了其在癌

症亚型分类任务中的有效性和独特优势，在处理样本不平衡问题和对不同类别样本分类能力的精准评估方面表现尤其突出。

由表 2 和表 3 可知，与其他基线模型相比，MFP-VAE 在处理复杂分类任务时表现出较强的优势。此外，ERGCN 作为次优方法，在各项指标上也取得了良好的结果。为体现 MFP-VAE 相较于 ERGCN 的优势，本文进一步比较了两种方法在各癌症亚型划分中的效果，如表 4 所示，其中最佳结果用粗体标识，稀缺亚型用下划线标识。

本研究改进了样本抽取策略，以确保不同亚型的样本在元学习任务中得到均衡重视。而在样本稀缺的情况下，常规方法无法有效分类稀缺亚型。在 BRCA 数据集中，*HER-2* 阳性亚型仅有 46 个样本，占数据集 7%，属于稀缺亚型。在 ERGCN 中，该亚型评估指标均远低于其余亚型，与最优亚型相比，召回率降幅达 22 个百分点，仅有 67.63%。而在 MFP-VAE 中，该亚型评估指标已达到平均水平，与最优亚型相比，召回率降幅仅有 2 个百分点，大大提升了模型在稀缺亚型的表现。与 ERGCN 相比，在稀缺亚型 *HER-2* 阳性中，MFP-VAE 的精确率、召回率和 F1 分数提升率分别约为 19%、21% 和 25%。GBM 数据集包含 3 种亚型，且分布相对均衡，分别包含 70、49 和 83 个样本。本文选择样本量最少的前神经元型亚型进行研究。在该亚型中，MFP-VAE 在各项指标中均取得了最优结果，与 ERGCN 相比，精确率、召回率和 F1 分数提升率分别约为 12%、15% 和 8%。此外，在 GBM 数据集上，3 种亚型在绝大多数指标中取得了最好结果。综上所述，与深度学习模型相比，本研究提出的 MFP-VAE 算法在稀缺亚型上的性能取得了明显提升，表明了样本抽取策略的有效性。

3.4.2 内部评价指标对比

BRCA 和 GBM 两个数据集在内部评价指标上的对比结果如表 5 所示，其中最优结果以粗体标识。

BRCA 数据集中，MFP-VAE 的 DBI 为 0.368 5，与 Deep Type 相比，降低了 0.054 8，与 ERGCN 相比，降低了 0.006 4，与 Subtype-DCC 相比，降低了 1.515 7，展现出更优的聚类性能。在 GBM 数据集中，MFP-VAE 的 DBI 为 0.326 6，与 Deep Type 相比，降低了 0.551 3，与 ERGCN 相比，降低了 0.010 8，与 Subtype-DCC 相比，降低了 0.774 0，其结果表现颇为优异。Subtype-DCC 作为先进的基线方法，即便在自监督学习情境下，依旧获取了优于传统机器学习方法的成效。上述结果表明，与所有基线方法相比，

表 4　乳腺浸润性癌 (BRCA) 数据集和多形性胶质母细胞瘤 (GBM) 数据集上的各亚型外部评价指标

Table 4　External evaluation indicators for each subtype of Breast Invasive Carcinoma (BRCA) dataset and Glioblastoma Multiforme (GBM) dataset

方法		BRCA				GBM		
		HER-2 阴性	*HER-2* 受体阳性	*HER-2* 阳性	三阴性乳腺癌	经典型	前神经元型	间充质型
ERGCN	精确率	**0.864 69±0.01**	0.764 12±0.01	0.723 45±0.01	0.857 97±0.01	0.839 05±0.02	0.740 72±0.02	0.866 13±0.02
	召回率	**0.897 74±0.01**	0.716 12±0.02	0.676 30±0.01	**0.883 73±0.01**	0.823 52±0.02	0.714 28±0.03	0.666 66±0.01
	F1 分数	**0.851 61±0.02**	0.706 12±0.01	0.656 37±0.01	**0.863 73±0.01**	0.813 84±0.02	0.764 71±0.02	**0.848 20±0.02**
MFP-VAE	精确率	0.845 80±0.02	**0.868 47±0.01**	**0.861 78±0.01**	**0.859 29±0.01**	**0.850 67±0.01**	**0.829 65±0.01**	**0.888 67±0.01**
	召回率	0.818 00±0.02	**0.843 00±0.01**	**0.821 00±0.01**	0.822 00±0.01	**0.848 00±0.01**	**0.819 00±0.01**	**0.805 37±0.01**
	F1 分数	0.811 05±0.01	**0.836 36±0.01**	**0.821 02±0.01**	0.821 94±0.02	**0.823 35±0.01**	**0.825 65±0.01**	0.847 63±0.01

表 5 乳腺浸润性癌 (BRCA) 数据集和多形性胶质母细胞瘤 (GBM) 数据集上的各亚型内部评价指标

Table 5 Internal evaluation indicators for each subtype of Breast Invasive Carcinoma (BRCA) dataset and Glioblastoma Multiforme (GBM) dataset

方法	BRCA		GBM	
	DBI	SI	DBI	SI
PCA＋SVM	4.264 5±0.16	−0.045 5±0.02	2.734 5±0.02	0.031 5±0.06
NMF＋RF	3.674 5±0.15	−0.056 4±0.02	3.243 1±0.02	−0.012 7±0.09
SAE＋MLP	3.142 1±0.17	0.012 4±0.03	3.162 5±0.05	−0.016 6±0.12
VAE＋SVM	3.534 5±0.14	−0.023 4±0.02	3.867 8±0.04	−0.056 4±0.10
CVAE＋MLP	1.031 3±0.21	0.513 4±0.01	0.903 1±0.05	0.334 9±0.05
Deep Type	0.423 3±0.05	0.596 8±0.02	0.877 9±0.04	0.337 7±0.04
ERGCN	0.374 9±0.02	0.782 2±0.00	0.337 4±0.01	0.782 8±0.01
Subtype-DCC	1.884 2±0.18	0.146 2±0.02	1.100 6±0.05	0.298 4±0.02
MFP-VAE	**0.368 5±0.04**	**0.787 1±0.01**	**0.326 6±0.01**	**0.798 5±0.01**

MFP-VAE 在聚类区分度上均具备显著优势。

3.4.3 消融实验

为深入剖析模型中不同编码器架构在癌症亚型分类任务中的贡献，并验证本研究采用的相关策略带来的改进有效性，本文精心设计了一系列消融实验。在本次实验中，本文聚焦于 VAE、AE 和多层感知机 (MLP) 等 3 种编码器。具体而言，本文通过将 MFP-VAE 模型中的编码器分别替换为 MLP 和 AE，构建了两个对比模型，并在 BRCA 和 GBM 数据集上进行实验，以探究不同编码器架构对整体模型性能的影响，同时保持其他参数恒定，以确保实验环境的一致性和结果的可比性，实验结果如表 6 所示，其中最佳结果用粗体标识。

在 BRCA 数据集上，不同模型的各项外部评价指标清晰地展示了各编码器架构的性能差异。基于 MFP-VAE 模型的表现尤其突出，在精确率、F1 分数和 MCC 等关键指标上具有显著优势。与采用 AE 作为编码器的次优模型相比，MFP-VAE 在精确率上实现了约 3.3% 的提升，这一提升表明 MFP-VAE 在准确识别正样本方面具有更强的能力，能更精准地判定样本所属的癌症亚型类别。在 F1 分数方面，MFP-VAE 的提升幅度达 8%。由于 F1 分数综合考量了精确率和召回率，因此其显著提高充分体现了 MFP-VAE 在平衡这两个重要指标方面的卓越表现。此外，

表 6 乳腺浸润性癌 (BRCA) 数据集和多形性胶质母细胞瘤 (GBM) 数据集消融实验

Table 6 Ablation experiments of Breast Invasive Carcinoma (BRCA) dataset and Glioblastoma Multiforme (GBM) dataset

数据集	方法	精确率	召回率	F1 分数	ACC	ARI	MCC
BRCA	MLP	0.805 89±0.02	0.787 73±0.03	0.804 85±0.01	0.807 37±0.01	0.521 82±0.13	0.706 81±0.06
	AE	0.815 89±0.02	0.807 73±0.01	0.764 85±0.02	0.816 37±0.01	0.596 82±0.09	0.756 81±0.05
	MFP-VAE	**0.848 83±0.01**	**0.826 00±0.01**	**0.822 60±0.01**	**0.826 01±0.01**	**0.600 14±0.08**	**0.780 48±0.05**
GBM	MLP	0.815 12±0.03	0.822 13±0.02	0.803 62±0.01	0.811 95±0.01	0.595 38±0.08	0.714 10±0.07
	AE	0.825 12±0.03	0.852 13±0.01	0.813 62±0.01	0.826 95±0.01	0.566 53±0.10	0.753 10±0.05
	MFP-VAE	**0.885 17±0.01**	**0.862 33±0.01**	**0.856 36±0.01**	**0.862 33±0.01**	**0.628 39±0.08**	**0.787 57±0.05**

在 MCC 指标上，MFP-VAE 比 AE 提升了两个百分点。MCC 作为一个对分类模型性能进行全面评估的重要指标，在处理样本不平衡问题时具有关键意义。在 GBM 数据集中，MFP-VAE 同样展现了优越的性能，取得了相似的结果。

在本研究的癌症亚型分类任务中，与传统的 AE 和 MLP 相比，MFP-VAE 作为编码器在模型性能提升方面发挥了关键作用。这一优势主要体现在多个方面，具体如下：MFP-VAE 能更有效地捕捉数据中的复杂特征表示，通过学习数据的潜在空间分布，将高维的基因表达数据映射到更具信息量的低维空间，从而更好地挖掘不同癌症亚型间的细微差异，为准确分类提供了坚实的基础；MFP-VAE 在处理样本不平衡问题上具有一定优势，其对数据分布的建模能力使得模型在面对不同数量的癌症亚型样本时，能更好地适应并学习具有代表性的特征。

3.4.4 生存分析

本节对 MFP-VAE 预测的癌症亚型进行生存分析，探讨不同亚型之间的生存关系。从 TCGA 数据库下载生存数据，构建生存分析数据集，包含样本编号、总体生存期、生存时间、生存结局和随访时间。基于亚型预测结果，绘制了 Kaplan-Meier 生存曲线。BRCA 和 GBM 数据集上的亚型生存情况如图 8 所示，图中标注了各亚型的中位生存时间，并通过 log-rank 检验计算 p 值 (p 值为假设检验中拒绝原假设的最小显著性水平)，以评估生存时间差异的显著性。

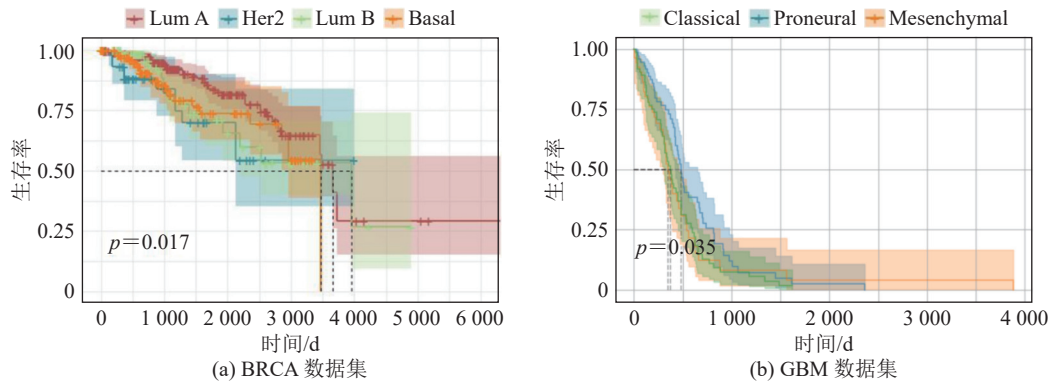


图 8　乳腺浸润性癌 (BRCA) 和多形性胶质母细胞瘤 (GBM) 数据集上的亚型生存情况

Fig. 8　Subtype survival outcomes in Breast Invasive Carcinoma (BRCA) dataset and Glioblastoma Multiforme (GBM) dataset

在 BRCA 数据集中，Lum A 亚型的中位生存时间为 3 669 d，Lum B 为 3 959 d，Basal 为 3 472 d，而 Her2 亚型因多数患者未达中位生存时间记为 NA (NA 指大多数患者无法活过中位生存时间)。GBM 数据集中，经典亚型中位生存时间为 372 d，前神经元亚型为 482 d，间充质亚型为 350 d。BRCA 数据集的 log-rank 检验 p 值为 0.017，GBM 数据集的 p 值为 0.035，均小于 0.05，表明模型能有效识别有显著生存差异的癌症亚型。

4　结　论

本研究提出一种新的基于元学习的癌症亚型分类方法，旨在解决癌症亚型分类中样本稀少和分布不均衡的问题。本文的主要贡献如下：构建多个元任务，均衡选择不同癌症亚型及样本数量，使模型在训练中接触到多样化样本组合。通过原型对比模块，计算查询集中每个样本与各亚型原型之间的距离，并分类。这一策略显著提高了模型对不同亚型间差异的理解，最终实现更准

确的分类结果。该方法在 BRCA 和 GBM 数据集上验证，结果表明，多项评价指标均优于现有先进模型，在稀缺亚型分类中的性能提升尤为显著，验证了少样本多任务学习的有效性。此方法不仅在提高分类准确性方面至关重要，还为解决生物数据分类中样本不均衡问题提供了新思路。此外，通过生存分析可知，所分类亚型在生存率方面存在显著差异，进一步验证了 MFP-VAE 模型的有效性。上述结果表明，本文方法为癌症亚型分类带来了新视角，具有重要的理论和实际意义。

参 考 文 献

[1] Dunn GP, Old LJ, Schreiber RD. The immunobiology of cancer immunosurveillance and immunoediting [J]. Immunity, 2004, 21(2): 137-148.

[2] Kent DM, Nelson J, Dahabreh IJ, et al. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials [J]. International Journal of Epidemiology, 2016, 45(6): 2075-2088.

[3] Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics [J]. British Journal of Cancer, 2013, 108(3): 479-485.

[4] Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2 000 breast tumours reveals novel subgroups [J]. Nature, 2012, 486(7403): 346-352.

[5] Bedard PL, Hansen AR, Ratain MJ, et al. Tumour heterogeneity in the clinic [J]. Nature, 2013, 501(7467): 355-364.

[6] Dai XF, Li T, Bai ZH, et al. Breast cancer intrinsic subtype classification, clinical use and future trends [J]. American Journal of Cancer Research, 2015, 5(10): 2929-2943.

[7] D'Aprile S, Denaro S, Lavoro A, et al. Glioblastoma mesenchymal subtype enhances antioxidant defence to reduce susceptibility to ferroptosis [J]. Scientific Reports, 2024, 14(1): 20770.

[8] Heiser LM, Sadanandam A, Kuo WL, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer [J]. Proceedings of the National Academy of Sciences, 2012, 109(8): 2724-2729.

[9] Prat A, Parker JS, Karginova O, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer [J]. Breast Cancer Research, 2010, 12: R68.

[10] Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications [J]. Proceedings of the National Academy of Sciences, 2001, 98(19): 10869-10874.

[11] Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes [J]. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 2009, 27(8): 1160-1167.

[12] Berger MF, Mardis ER. The emerging clinical relevance of genomics in cancer medicine [J]. Nature Reviews Clinical Oncology, 2018, 15(6): 353-365.

[13] Ding C, He XF. Cluster structure of *K*-means clustering via principal component analysis [C] // Proceedings of the Knowledge Discovery and Data Mining, 2004: 414-418.

[14] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401(6755): 788-791.

[15] Franco EF, Rana P, Cruz A, et al. Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data [J]. Cancers, 2021, 13(9): 2013.

[16] Fakoor R, Ladhak F, Nazi A, et al. Using deep learning to enhance cancer diagnosis and classification [C] // Proceedings of the 30th International Conference on Machine Learning, 2013: 1-7.

[17] Adem K, Kilicarslan S, Comert O. Classification and diagnosis of cervical cancer with softmax classification with stacked autoencoder [J]. Expert Systems with Applications, 2019, 115: 557-564.

[18] Park KH, Batbaatar E, Piao YJ, et al. Deep learn-

ing feature extraction approach for hematopoietic cancer subtype classification [J]. International Journal of Environmental Research and Public Health, 2021, 18(4): 2197.

[19] Li HL, Wang CH, Yuan BZ. An improved SVM: NN-SVM [J]. Chinese Journal of Computers, 2003, 26(8): 1015-1020.

[20] Yancey RE, Xin BC, Matloff N. Modernizing *k*-nearest neighbors [J]. Stat, 2021, 10(1): e335.

[21] 王爱平, 万国伟, 程志全, 等. 支持在线学习的增量式极端随机森林分类器 [J]. 软件学报, 2011, 22(9): 2059-2074.
Wang AP, Wan GW, Cheng ZQ, et al. Incremental extreme random forest classifier supporting online learning [J]. Journal of Software, 2011, 22(9): 2059-2074.

[22] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors [J]. Nature, 1986, 323(6088): 533-536.

[23] Ni PY, Su ZC. Deciphering epigenomic code for cell differentiation using deep learning [J]. BMC Genomics, 2019, 20(1): 709.

[24] Maulik U, Mukhopadhyay A, Chakraborty D. Gene-expression-based cancer subtypes prediction through feature selection and transductive SVM [J]. IEEE Transactions on Biomedical Engineering, 2013, 60(4): 1111-1117.

[25] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.

[26] Chen R, Yang L, Goodison S, et al. Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data [J]. Bioinformatics, 2020, 36(5): 1476-1483.

[27] Dai W, Yue WH, Peng W, et al. Identifying cancer subtypes using a residual graph convolution model on a sample similarity network [J]. Genes, 2021, 13(1): 65.

[28] Zhao J, Zhao BW, Song XT, et al. Subtype-DCC: decoupled contrastive clustering method for cancer subtype identification based on multi-omics data [J]. Briefings in Bioinformatics, 2023, 24(2): bbad025.

[29] Wang QY, Zhou Y, Zhang WM, et al. Adaptive sampling using self-paced learning for imbalanced cancer data pre-diagnosis [J]. Expert Systems with Applications, 2020, 152: 113334.

[30] Weinstein JN, Gollisson EA, Mills GB, et al. The cancer genome atlas pan-cancer analysis project [J]. Nature genetics, 2013, 45(10): 1113-1120.

[31] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark [J]. Nucleic Acids Research, 2018, 46(20): 10546-10562.

[32] Lim J, Ryu S, Kim JW, et al. Molecular generative model based on conditional variational autoencoder for *de novo* molecular design [J]. Journal of Cheminformatics, 2018, 10: 31.

[33] Yang X, Zheng Y, Xing X, et al. Immune subtype identification and multi-layer perceptron classifier construction for breast cancer [J]. Frontiers in Oncology, 2022, 12: 943874.