

文章编号 1004-924X(2026)18-2777-15

透射光谱数据增强策略的水体细菌识别

胡玉霞^{1,3*}, 黄书龙¹, 徐亚玲¹, 薛巧玲¹, 付金超¹, 胡敦^{2*}

(1. 安徽建筑大学 电子与信息工程学院, 安徽 合肥 230601;

2. 安徽第二医学院, 安徽 合肥 230601;

3. 安徽省古建筑智能感知与高维建模国际联合研究中心, 安徽 合肥 230601)

摘要:为实现水体细菌的准确识别,本研究采用透射光谱技术,结合生成对抗网络数据增强方法,对五种典型水体细菌进行分类识别。针对透射光谱样本获取成本高、易受仪器性能影响、小样本下模型泛化能力不足等问题,提出基于动态筛选生成对抗网络(DFGAN)的光谱数据增强策略,并构建卷积神经网络(CNN)分类模型。首先采集五种细菌在不同培养时段和梯度浓度下的 600 条透射光谱作为原始数据集;随后搭建深度卷积生成对抗网络(DCGAN)与 DFGAN 进行数据增强,从视觉形态、统计分布相似度和特征空间多样性三个维度评估生成光谱质量;最后对比支持向量机、K 近邻及 CNN 三类分类器的识别效果。结果表明:DFGAN 生成光谱的质量显著优于 DCGAN;DFGAN-CNN 方法所建立水体细菌识别模型性能最优,识别准确率达 95.00%,精确率 95.45%,召回率 95.00%,F1 分数 95.00%,具有最佳的准确性和鲁棒性。该方法为水体细菌检测提供了新途径,也为光谱分析领域的小样本问题解决提供了技术参考。

关键词:透射光谱;数据增强;水体细菌;识别;动态筛选生成对抗网络

中图分类号:O433 **文献标识码:**A

doi:10.37188/OPE.20263418.2777

CSTR:32169.14.OPE.20263418.2777

Expansion strategy of transmission spectral data for water bacteria identification

HU Yuxia^{1,3*}, HUANG Shulong¹, XU Yaling¹, XUE Qiaoling¹, FU Jinchao¹, HU Dun^{2*}

(1. School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei 230601, China;

2. Anhui Second Medical College, Hefei 230601, China;

3. International Joint Research Center for Intelligent Perception and High-Dimensional Modeling of Ancient Architecture of Anhui Province, Hefei 230601, China)

* Corresponding author, E-mail: yxhuhf@163.com; hudun@mail.ustc.edu.cn

Abstract: In order to achieve accurate identification of aquatic bacteria, this study employed transmission spectroscopy combined with a generative adversarial network (GAN)-based data augmentation approach to classify and identify five typical aquatic bacterial species. Addressing challenges such as the high cost of

收稿日期:2026-05-11; **修订日期:**2026-07-07.

基金项目:国家自然科学基金(No. 62105002);国家重点实验室开放课题(No. 6142804230105);安徽省高校省级自然科学研究重点项目(No. 2024AH050854);安徽第二医学院王建华科研创新团队项目(No. WJH2023GGWS002);安徽建筑大学科研项目(No. JZ202346);安徽省高校中青年教师培养行动项目(No. JNFX2025111)

acquiring transmission spectral samples, susceptibility to instrument performance variations, and insufficient model generalization under small-sample conditions, we proposed a spectral data augmentation strategy based on a dynamic filtering generative adversarial network (DFGAN) and constructed a convolutional neural network (CNN) classification model. Initially, 600 transmission spectra of the five bacterial species were collected across different cultivation periods and concentration gradients to form the original dataset. Subsequently, a deep convolutional generative adversarial network (DCGAN) and the proposed DFGAN were implemented for data augmentation, with the quality of generated spectra evaluated from three dimensions: visual morphology, statistical distribution similarity, and feature space diversity. Finally, comparative analyses were conducted on the recognition performance of three classifiers: support vector machine (SVM), K-nearest neighbor (KNN), and CNN. The results indicated that the quality of spectra generated by DFGAN was significantly superior to that of DCGAN. The aquatic bacteria identification model established by the DFGAN-CNN approach exhibited optimal performance, achieving an accuracy of 95.00%, precision of 95.45%, recall rate of 95.00%, and F1-score of 95.00%, thus demonstrating the highest accuracy and robustness. This method not only provides a novel approach for the rapid detection of aquatic bacteria but also offers technical references for addressing small-sample challenges in the field of spectral analysis.

Key words: transmission spectrum; data augmentation; waterborne bacteria; recognition; dynamic filtering generative adversarial network

1 引言

水体细菌污染已成为导致水环境恶化、威胁水生态安全及人类健康的重要因素。准确识别水体细菌种类对水环境保护和水质防控预警至关重要^[1-3],这些研究充分凸显了水环境细菌监测的紧迫性与重要意义。目前,水体细菌检测方法主要包括生物学方法和光谱学方法。其中,生物学方法主要包括:分离培养法^[4]、分子生物学方法^[5-6]和免疫学方法^[7]等技术,这些方法测量准确度高、有效,但存在样品预处理繁琐和耗时长等问题,不利于细菌现场实时检测。

多波长透射光谱技术^[8]是一种利用光与待测细菌微生物产生的透射光谱,来分析细菌微生物含量、结构和性质的技术。因该技术操作简便、无需复杂前处理和检测速度快等优点,逐渐成为水体细菌识别和检测的重要工具。甘婷婷等^[9]采用非线性最小二乘法结合多波长透射光谱,解析了水体常见致病菌的多波长透射光谱特征;冯春等^[10]利用多波长透射光谱结合支持向量机算法,完成了水体中大肠杆菌等5种常见致病菌的分类识别;Hu Y X等^[11]采用竞争性自适应重加权采样(Competitive Adaptive Reweighted Sampling,

CARS)与概率神经网络(Probabilistic Neural Network, PNN),实现了水体中大肠杆菌、金黄色葡萄球菌、肺炎克雷伯菌、鼠伤寒沙门氏菌4种典型致病菌的快速智能识别;Smith J M等^[12]基于紫外可见多波长透射光谱技术,结合多元统计分析构建物种识别模型,实现了大肠杆菌、肺炎克雷伯菌、铜绿假单胞菌、金黄色葡萄球菌4种临床致病菌的免试剂快速鉴别。这些研究表明,学者们将多波长透射光谱与机器学习、深度学习等人工智能算法相结合,在部分数据集的细菌识别和定量分析中取得了较好的结果。然而,基于人工智能算法的细菌透射光谱识别算法仍存在以下问题:1. 细菌透射光谱具有数据维度高、光谱相似性高等特点,其中与细菌细胞相关的有效特征易被大量无关特征覆盖,导致细菌细胞信息难以充分利用;2. 现有细菌透射光谱采集,需要较高的时间和人力成本,采集到的数据样本数量有限,可用于机器学习或深度学习训练的数据集偏少,最终造成模型泛化能力不足。

近年来,为解决数据量不足的问题,诸多学者引入深度学习方法,以生成对抗网络^[13](Generative Adversarial Network, GAN)为代表开展数据扩充研究,即通过训练神经网络模型来生

成仿真样本数据。Lu等^[14]综述了生成对抗网络在农业领域图像数据增强的应用进展;Huang等^[15]提出增强型平衡式生成对抗网络,有效解决了高度不平衡图像数据集中少数类样本生成困难、训练不稳定的关键问题;Lorencin等^[16]利用DCGAN生成高度逼真的医学图像样本,有效扩充了有限的医疗数据集,进而提高了膀胱癌图像分类和诊断的准确性。目前,已有学者将生成对抗网络应用于光谱研究领域;Li等^[17]将深度卷积生成对抗网络与近红外高光谱成像技术相结合,对不平衡的小麦数据集进行样本扩增,显著提升了小样本条件下不健康小麦籽粒的分类识别效果;李灵巧等^[18]通过深度卷积生成对抗网络提取拉曼光谱内部特征,并对抗生成新样本,实现了药品拉曼光谱数据集的高质量扩充,提升了分类精度;Lyu等^[19]采用一种带梯度惩罚的条件 Wasserstein 生成对抗网络(WGAN-GP)来合成高光谱反射率数据,克服了标签数据获取困难的瓶颈,增强了葡萄成熟度的无损检测评估模型。

针对上述问题,本文采用基于深度卷积生成对抗网络(Deep Convolutional Generative Adversarial Network, DCGAN)和动态筛选生成对抗网络(Dynamic Filtering Generative Adversarial Network, DFGAN)透射光谱数据增强方法,生成细菌透射光谱,结合支持向量机(Support Vector Machine, SVM)、K近邻(K-Nearest Neighbors, KNN)、卷积神经网络(Convolutional Neural Network, CNN)分类算法,实现水体细菌微生物种类的准确识别。本研究旨在通过有效扩增

水体细菌透射光谱样本量,提高SVM, KNN和CNN学习模型的识别准确性与稳定性,以期透射光谱结合人工智能算法在水体细菌识别中的应用提供理论依据。

2 实 验

2.1 细菌菌悬液制备

粪链球菌(*E. faecalis*)和金黄色葡萄球菌(*S. aureus*),购自美国典型培养物保藏中心,肺炎克雷伯菌(*K. pneumoniae*)、枯草芽孢杆菌(*B. subtilis*)和鼠伤寒沙门氏菌(*S. typhimurium*)购自中国医学菌种保藏管理中心。

为了充分获取每种细菌光谱特征,需制备不同培养时间、不同梯度浓度的细菌悬浮液样本,具体流程如图1所示。首先,吸取1 mL活化后的细菌菌液,接种至装有牛肉膏蛋白胨液体培养基的锥形瓶中,每种细菌共设置20个平行样本,随后,将每个样本按培养时间(1~20 h)进行标记,并置于恒温培养箱中静置培养。培养过程中,每隔1小时取出相应时间点的样本,吸取5 mL培养液至离心管中,在12 000 r/min条件下离心10 min,弃去上清液,加入5 mL去离子水使其重新悬浮,再次离心。重复上述清洗步骤三次。将洗净的细菌加入5 mL去离子水,充分振荡使其均匀分散,得到20个培养时间点对应的均匀细菌悬浮液。最后,将每种细菌各培养时间点的菌液分为6个平行样本,按体积比(1:1, 1:3, 1:7, 1:15, 1:31, 1:63)与去离子水进行梯度稀释,得到6个梯度浓度的菌液。

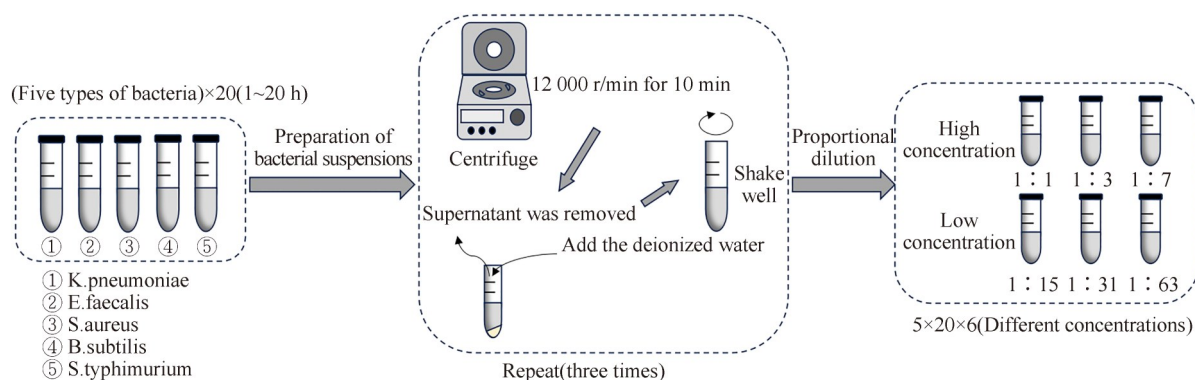


图1 五种细菌20个培养时间6个梯度浓度下菌液制备流程图

Fig. 1 Flowchart for preparing bacterial suspensions under 20 cultivation times and 6 concentration gradients for 5 types of bacteria

最终,在 20 个培养时间、6 个梯度浓度下,每种细菌共制备了 120 组样本,5 种细菌共制备了 600 组细菌菌悬液样本。

2.2 细菌透射光谱获取

细菌透射光谱采用紫外-可见分光光度计(型号:UV2700i,日本岛津)进行获取,光谱测量波长范围设定为 220~800 nm,光谱仪分辨率为 0.1 nm,采样间隔设定为 1 nm,扫描速度为中速(约 6 nm/s)。以去离子水为参比,每个菌液样本测量三次,取平均值作为每个菌液样本的测量光谱。五种细菌在不同培养时间(1~20 h,间隔 1 h)及不同梯度浓度(6 个)的透射光谱如图 2(a)~2(e)所示,其归一化后的平均光谱如图 2(f)所示。

由图 2(a)~2(e)可知,肺炎克雷伯菌、粪链球菌、金黄色葡萄球菌、枯草芽孢杆菌和鼠伤寒沙门氏菌的透射光谱呈现出高度相似的谱型特征。240~300 nm 波长范围内,存在明显的吸收峰,其原因是细菌内生物分子(如核酸、蛋白质)对光的吸收;400~800 nm 波长范围无吸收峰,光

谱曲线整体趋于平缓。同种细菌在不同培养时段、不同梯度浓度下的光谱曲线整体趋势一致,但光密度值存在差异,表明同种细菌的光谱特征受细菌培养时间和梯度浓度的影响。

为比较不同种类细菌的光谱差异,首先对原始测量光谱进行 Min-max 归一化处理^[20],以消除梯度浓度差异带来的量纲差异,统一光谱的数值尺度。随后,为削弱不同培养时间引起的光谱波动,对归一化后的光谱取平均值,得到各类细菌对应的代表性平均光谱,如图 2(f)所示。五种细菌归一化后平均光谱光密度值随波长增大均呈下降趋势,但在部分波段存在明显差异。其中,枯草芽孢杆菌与金黄色葡萄球菌的光谱在全波段光密度强度差异显著,谱型辨识度较高;肺炎克雷伯菌、粪链球菌与鼠伤寒沙门氏菌的光谱曲线重合度较高,仅在 260~280 nm 特征吸收峰的峰高、峰宽,以及 300~340 nm 波段的基线斜率存在细微差异。这些光谱特征表明:为更有效表征不同细菌,需获取在多种培养时间和梯度浓度下的光谱数据,同时,考虑到部分细菌种间光谱差

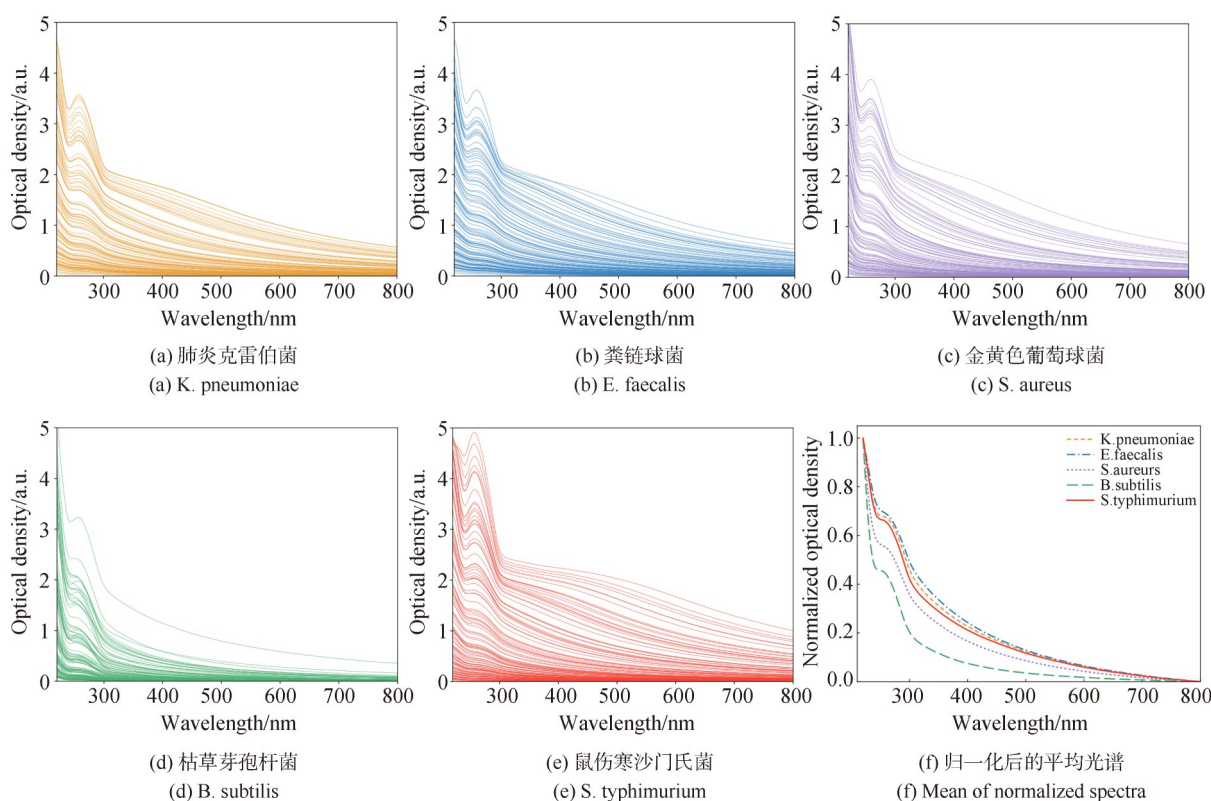


图 2 五种细菌的透射光谱

Fig. 2 Transmission spectra of five types of bacteria

异较小,还需进一步借助人工智能算法来进行辨识。

3 数据增强方法

3.1 深度卷积生成对抗网络(DCGAN)

为了扩充细菌透射光谱数据样本,并考虑到其作为一维连续序列数据——相邻波长点间具有紧密的物理关联与局部相关性,而常规方法(如:添加高斯噪声^[21]、平移光谱的波数点^[22]、将多条光谱以不同比率变换然后相加^[23])难以有效提取此类序列的局部特征,生成光谱与原光谱相关的特征会产生巨大变化,无法论证新光谱与原光谱的分类一致性。因此,本文构建了深度卷积生成对抗网络(Deep Convolutional Generative Adversarial Network, DCGAN)模型进行数据增强。DCGAN模型将卷积神经网络与对抗训练框架相结合,利用卷积层替代生成式对抗网络(Generative Adversarial Network, GAN)中的全连接层,以便有效地捕捉光谱数据的局部特征与全局趋势,从而提升模型训练稳定性与生成光谱的质量,该模型具体架构如图3所示(彩图见期刊电子版)。

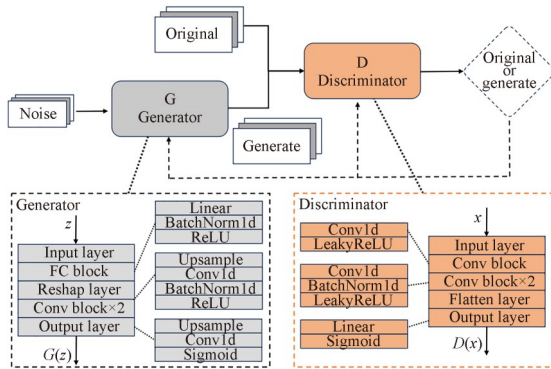


图3 深度卷积生成式对抗网络(DCGAN)的结构示意图

Fig. 3 Structure diagram of the deep convolutional generative adversarial network (DCGAN)

由图3可知(彩图见期刊电子版),DCGAN模型主要由随机噪声、生成器、判别器、全连接层与激活函数构成。生成器(见图3中灰色虚线框),以随机噪声向量 z 为输入,经过输入层后进入全连接块(FC Block),实现从低维噪声到高维特征的映射。随后,特征经重塑层(Reshape lay-

er)转换为一维特征图,依次通过两个结构相同的卷积块(Conv Block),最终经输出层(Output layer)生成与真实光谱维度一致的合成样本 $G(z)$ 。判别器(见图3橙色虚线框)采用一维卷积结构进行特征提取:输入光谱样本 x ,经输入层后依次通过两个卷积块提取特征,所得高维特征经展平层(Flatten layer)转换为一维向量,最终由输出层给出样本为真实数据的概率值 $D(x)$ 。

在训练过程中,生成器的目标是生成足以“欺骗”判别器的逼真样本,而判别器则力图准确区分真实样本与生成样本,DCGAN的目标函数可表示为:

$$\min_G \max_D V(D, G) = E_{x \sim p_r} [\log D(x)] + E_{z \sim p_z} [\log (1 - D(G(z)))], \quad (1)$$

其中: x 是真实数据, z 是随机噪声, P_r 表示真实数据分布, P_z 表示随机噪声分布, $D(x)$ 表示 x 来自真实数据的概率。二者通过交替迭代、相互对抗优化,持续提升各自性能,当判别器无法准确分辨样本来源、输出概率稳定在0.5左右时,即标志着目标函数收敛至最小值,生成样本已达到理想效果,生成对抗网络取得成功。

3.2 动态筛选生成对抗网络(DFGAN)

考虑到DCGAN生成样本可能存在高度重合、多样性不足的问题,本文进一步构建了一种动态筛选生成对抗网络(Dynamic Filtering Generative Adversarial Network, DFGAN)进行数据增强。该DFGAN模型在DCGAN的基础上引入动态筛选机制(见图4)。

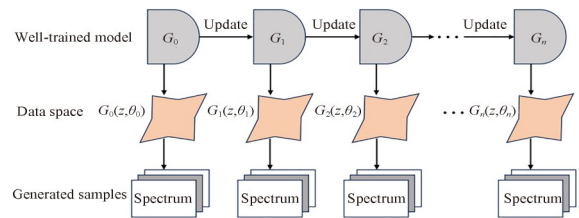


图4 动态筛选机制的原理

Fig. 4 Principle of the dynamic filtering mechanism

动态筛选机制核心原理:不提前终止对抗训练,在生成器与判别器初步收敛后,继续迭代更新网络参数。采用阶段性评估与逐类动态择优的筛选策略,每完成固定轮次训练即对当前候选

生成器 G_i 开展质量评估, 将其与对应类别已保存的历史最优模型进行对比, 保留综合评分更优的模型权重、淘汰表现较差的候选模型, 使每类细菌在整个训练过程中始终仅留存一组最优生成器参数。该机制以最大均值差异 (Maximum Mean Discrepancy, MMD) 和切片瓦瑟斯坦距离 (Sliced Wasserstein Distance, SWD) 这两项统计指标构建加权综合评分, 进而从整个模

$$MMD^2(X, Y) = \sum_{a \in \mathcal{A}} \left[\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_a(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m k_a(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k_a(x_i, y_j) \right], \quad (2)$$

其中, $k_a(x, y) = \exp\left(-\frac{|x - y|^2}{\alpha}\right)$ 。 X, Y 分别为

真实光谱数据集和生成光谱数据集, n, m 分别为两个数据集的样本数量, $\mathcal{A}$ 为多尺度超参数集合, k_a 是以 α 为尺度的径向基核函数。

SWD 用于度量真实光谱与生成光谱的局部分布相似度, 该指标侧重捕捉数据分布形态与局部结构差异, 对光谱细微形态变化更敏感; 其数值越小, 表明两种光谱的细微波形特征与指纹信息越相近, 其计算公式为:

$$SWD_p(X, Y) = \left(\frac{1}{L} \sum_{l=1}^L W_p^p(X\theta_l, Y\theta_l) \right)^{1/p},$$

$$W_p^p(X_\theta, Y_\theta) = \frac{1}{N} \sum_{i=1}^N |x_{(i), \theta} - y_{(i), \theta}|^p, \quad (3)$$

其中: L 是随机投影的数量, θ_l 是均匀分布在单位超球面上的第 1 个随机投影向量, p 为 Wasserstein 距离的阶数, $x_{(i), \theta}, y_{(i), \theta}$ 分别表示样本投影到 θ 方向后, 经升序排列后的第 i 个元素, N 为光谱样本的数量。

该动态筛选机制的具体执行流程:

(1) 初步收敛判据与筛选起点: 由图 5 生成器 G 和判别器 D 的训练损失曲线可知, 在训练迭代至 1 000 epochs (训练轮次) 之前, 生成器和判别器的损失波动明显; 训练迭代 1 000 epochs 后, 生成器和判别器的损失稳定在 1.0~1.4 区间, 表明模型对抗博弈达到动态平衡, 判别器已无法有效区分两类样本, 生成器已具备基本的光谱特征复现能力, 实现初步收敛, 因此将 1 000 epochs 设为动态筛选机制的起始轮次。

(2) 当前轮次的指标计算: 模型达到初步收敛后, 继续开展对抗迭代训练, 进一步优化生成效果; 同时设置固定筛选采样间隔, 每完成 500 个

型序列中筛选出每种细菌光谱对应的最优生成器。

其中, MMD 用于度量真实光谱分布与生成光谱分布的全局相似度, 不局限于单一均值差, 通过核函数映射到高维空间, 捕捉非线性、全局分布差异; 该指标值越小, 表明两组光谱数据的全局统计特征越接近, 其计算公式为:

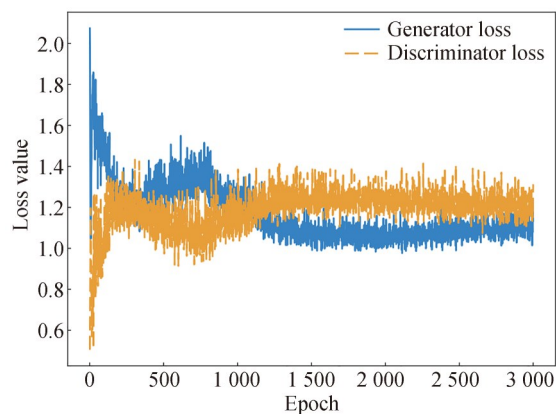


图 5 生成器和判别器的训练损失曲线

Fig. 5 Training loss curves for the generator and discriminator

训练轮次, 执行一次阶段性评估与权重筛选操作。在各筛选节点处, 先临时保存当前生成器完整网络权重作为候选模型, 再按细菌类别生成与原始训练集样本数量匹配的透射光谱样本。以原始训练集光谱为基准, 逐一计算各组生成样本对应的 MMD 与 SWD, 并根据加权综合评分公式求解各类别对应的综合评分 T :

$$T = MMD + \alpha \times SWD, \quad (4)$$

其中: 权重系数 α 取值为 10, 该参数设置依据为两项指标的实际数值量级: 本研究数据集内, MMD 数值分布在 0.02~0.06 区间, SWD 数值分布在 0.004~0.01 区间, 二者相差约一个数量级。在此将权重系数设为 10, 能够有效消除量纲差异产生的评价偏差, 使全局分布指标与局部细节指标在综合评分中拥有对等的决策权重, 防止数值量级更小的 SWD 指标被 MMD 掩盖。

(3) 逐类动态择优与淘汰: 完成当前轮次的

指标计算后,针对各类细菌,将当前候选模型的综合评分 T ,与该类别已存储的历史最优 T 值对比。若当前模型的 T 值更小,说明其样本生成效果更优,随即使用当前候选权重覆盖原有最优权重数据,更新该类别的历史最优记录;若当前模型的 T 值更大,则直接删除本次临时存储的候选权重数据,保留原有最优模型。根据该筛选机制,训练全程中每种细菌类别仅留存一组最优生成器权重,无需保存所有中间阶段模型数据,在确保筛选精度的前提下大幅度降低存储资源开销。

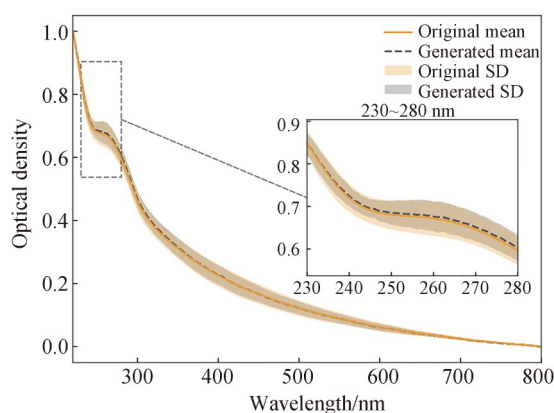
(4)筛选终止判定:当训练迭代达到预设最大轮次 10 000 轮时,停止迭代并终止筛选流程。此时各类细菌对应的留存生成器权重,为全训练周期内综合评分最低、生成光谱数据质量最优的模型参数,该机制规避了单一全局模型对部分细菌类别生成效果较差的缺陷,最终实现五类细菌生成光谱质量的整体最优。

4 结果与讨论

4.1 基于 DCGAN 和 DFGAN 的透射光谱增强分析

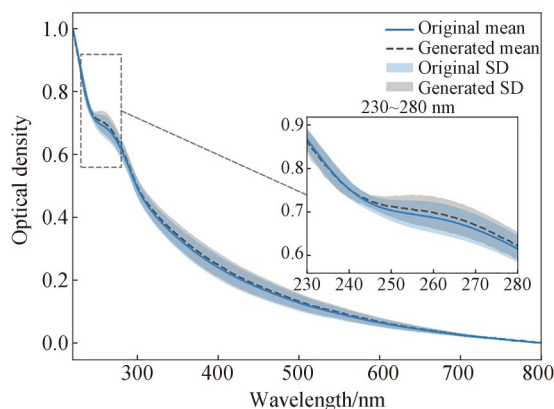
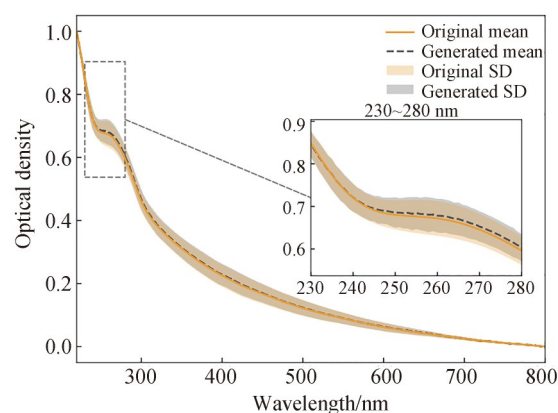
本文分别采用 DCGAN 与 DFGAN 网络对原始光谱数据增强,相关参数设置如下:输入噪声维度(z_dim)为 100,批次大小($batch_size$)为 16,迭代训练轮数($epochs$)为 10 000,学习率($learning\ rate$)为 0.000 2,针对 5 种不同培养时间不同梯度浓度的细菌样本,每种细菌样本选取 80 条原始光谱。

为验证 DCGAN 与 DFGAN 网络在细菌光谱数据增强中的有效性,计算每种细菌 80 条原始光谱的均值谱及 ± 1 个标准差(Standard Deviation, SD)谱带^[24](以均值谱为中心,向上叠加 1 倍标准差为上边界,向下减去 1 倍标准差为下边界),并将 DCGAN,DFGAN 生成光谱对应的均值谱与 ± 1 倍标准差谱带与原始光谱统计结果进行比较,结果如图 6 所示(左列为 DCGAN 生成样本,右列为 DFGAN 生成样本)。



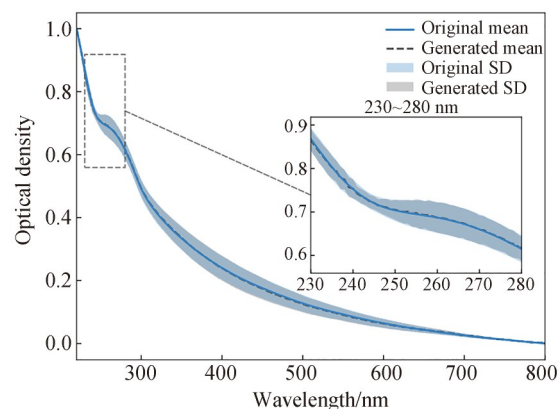
(a) 肺炎克雷伯菌

(a) *K. pneumoniae*



(b) 粪链球菌

(b) *E. faecalis*



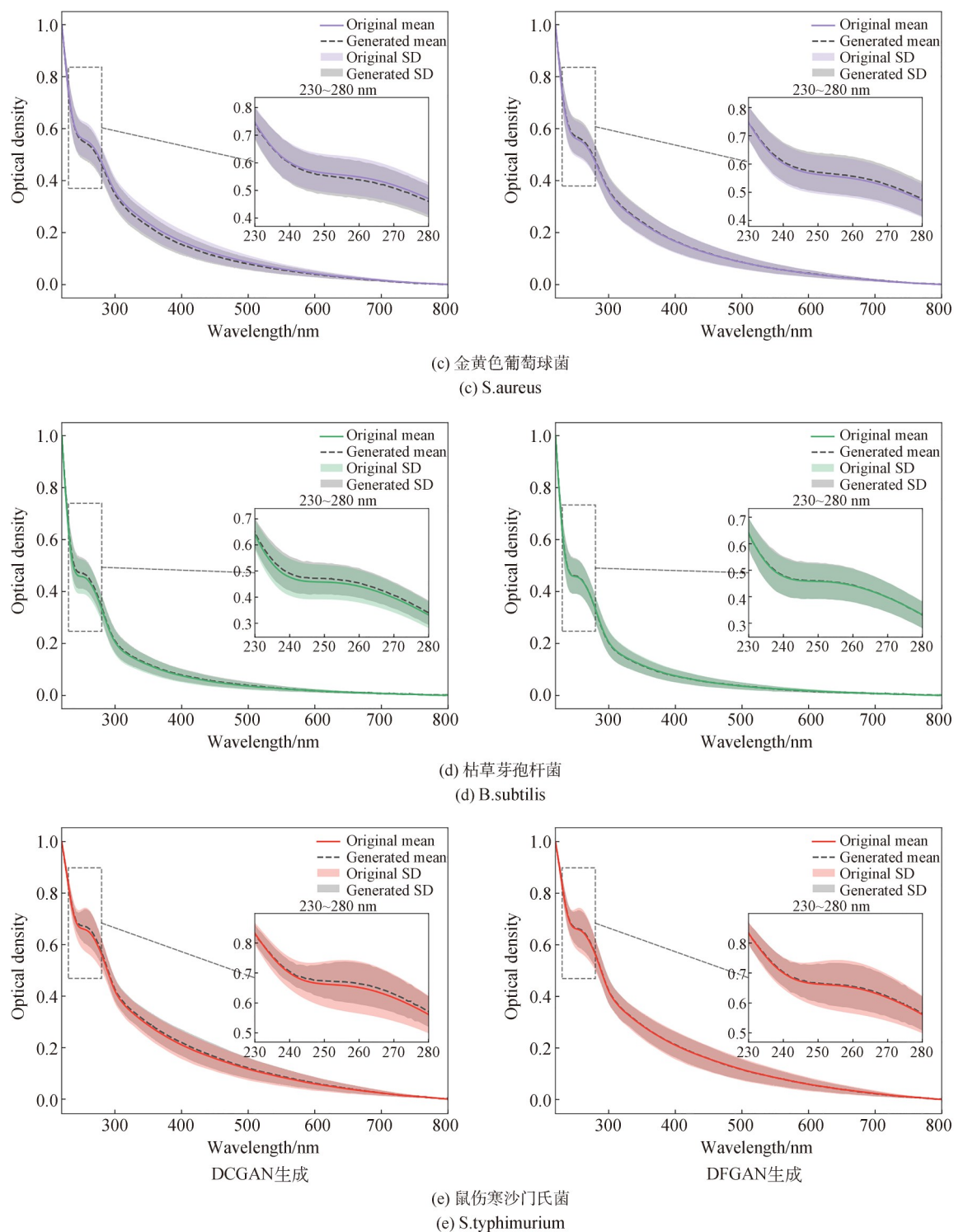


图 6 原始样本与生成样本的均值谱,误差范围为 ± 1 个标准差谱带

Fig. 6 Mean spectra of the original samples and the generated samples, with an error range of ± 1 standard deviation spectral band

由图 6 可知,从 220~800 nm 全波段光谱特征来看,DCGAN 与 DFGAN 两种数据增强方法生成样本的均值谱,均能有效复现原始细菌透射光谱的核心特征。在 240~300 nm 波长范围内,

准确还原了 260 nm 处细菌对应的特征吸收峰;在 400~800 nm 波段,也完整复现了光密度值随波长增加而单调递减的规律。对于肺炎克雷伯菌、粪链球菌和金黄色葡萄球菌,两种数据增强

方法生成光谱的均值谱、±1 个标准差谱带,均与原始光谱的对应统计量几乎完全重合,这主要是由于上述三种细菌的原始光谱特征辨识度高、数据分布相对集中,生成模型更易学习到其稳定的光谱分布规律。对于枯草芽孢杆菌与鼠伤寒沙门氏菌,与 DCGAN 生成结果相比,DFGAN 模型生成光谱的均值谱与原始光谱均值谱的吻合度更优,具体体现在 230~280 nm 特征吸收峰的峰高、峰宽拟合度更高,300~340 nm 波段的基线斜率偏差更小,全波段范围内的均值偏差均小于 DCGAN 生成结果;同时,DFGAN 生成样本的标准差谱带,与原始样本标准差谱带的重合范围更广。这表明基于 DFGAN 模型生成的样本光谱,在全波段的波动范围更贴合原始光谱的固有离散特性,能够完整复现原始数据集的分布规律;而 DCGAN 模型生成样本的标准差谱带在 240~280 nm 存在收窄(如:图 6(a)(b)(d)(e)),直接反映出其生成样本分布相对集中,个体间特征差异不足。

为客观比较最终确定的 DFGAN 与 DCGAN 生成样本的质量,验证本文所提动态筛选策略的有效性,本文采用 MMD 和 SWD 作为独立的外部对比评价指标,来评价两个模型生成光谱的效果。需要说明的是,虽然 DFGAN 在训练阶段的动态筛选过程中也使用了 MMD 和 SWD,但彼时二者仅作为模型内部的中间筛选准则,用于择优候选生成器。而在本节的质量评价阶段,其评价对象已转变为不同数据增强方法(DFGAN 与 DCGAN)在相同训练集下生成的最终定型样本,评价目标属于跨算法的模型性能公平验证,二者的评价场景与目的相互独立。

针对 5 种细菌样本,在两种数据增强方法下,相似度分析对比如表 1 所示。基于 DFGAN 数据增强方法生成的光谱与原始光谱相比,其最大均值差异值整体优于 DCGAN 增强结果。除枯草芽孢杆菌和鼠伤寒沙门氏菌的 MMD 值在两种方法中相同外,其余三种细菌(肺炎克雷伯菌、粪链球菌和金黄色葡萄球菌)的 MMD 值均低于 DCGAN 方法,其中金黄色葡萄球菌的改善最为显著,其 MMD 值从 0.059 8 降至 0.034 9。在 SWD 指标方面,枯草芽孢杆菌和鼠伤寒沙门氏菌,在两种增强方法下所得 SWD 值相同,而在粪

表 1 基于 DCGAN 和 DFGAN 增强方法所生成光谱与原始光谱相似性评价

Tab. 1 Similarity evaluation between the spectra generated by the DCGAN and DFGAN methods and the original spectra

Bacterial	DCGAN		DFGAN	
	MMD	SWD	MMD	SWD
<i>K. pneumoniae</i>	0.028 1	0.004 8	0.021 9	0.005 2
<i>E. faecalis</i>	0.040 3	0.008 0	0.027 7	0.006 7
<i>S. aureus</i>	0.059 8	0.007 5	0.034 9	0.005 9
<i>B. subtilis</i>	0.020 9	0.004 1	0.020 9	0.004 1
<i>S. typhimurium</i>	0.028 2	0.010 7	0.028 2	0.010 7

链球菌、金黄色葡萄球菌上,DFGAN 方法得到的 SWD 值低于 DCGAN 方法的计算结果。与 DCGAN 相比,肺炎克雷伯菌的 SWD 值在 DFGAN 中略有上升,增幅仅为 0.000 4。综合五种细菌样本整体结果来看,DFGAN 生成数据与原始光谱数据相似性更高。结果表明,DFGAN 模型通过动态筛选不同训练阶段的生成器,在实现细菌透射光谱增强的同时,有效提升了生成光谱与原始光谱的全局及局部分布相似度。

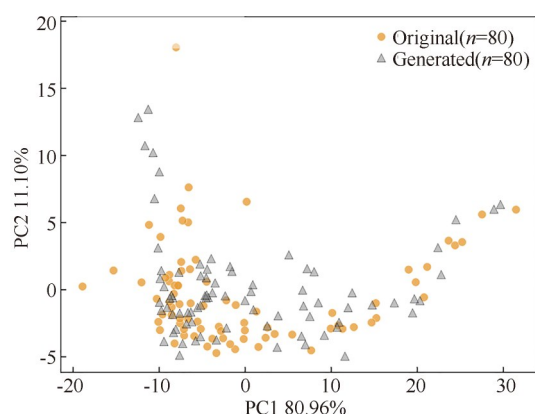
考虑到不同细菌本身存在固有光谱差异,会干扰特征空间分布的评价结果。为直观对比 DCGAN、DFGAN 生成样本与同类原始细菌样本的特征分布一致性,本文针对各类细菌分别开展 PCA 分析。具体流程为:先对每类细菌的原始光谱样本和生成光谱样本进行归一化处理,随后将两类样本合并,并在该类别独立数据集内完成 PCA 降维,提取前两个主成分 PC1 和 PC2 并构建二维特征空间。通过对比 DCGAN 和 DFGAN 各类细菌生成样本与原始样本在二维特征空间中的方差贡献率及分布范围,综合评估两类模型样本的分布一致性和特征多样性,其结果如图 6 所示。

由图 7 可知(左列为 DCGAN 生成,右列为 DFGAN 生成),五种细菌透射光谱的 PC1 方差贡献率集中在 68.76%~85.53% 之间,PC2 方差贡献率集中在 6.46%~12.76% 之间,两者的累积方差贡献率均稳定超过了 81%。其中,与 DCGAN 生成方法相比,DFGAN 在肺炎克雷伯菌、粪链球菌、金黄色葡萄球菌和鼠伤寒沙门氏菌的 PC1 贡献率上都有所提升,分别提升了 0.37%,

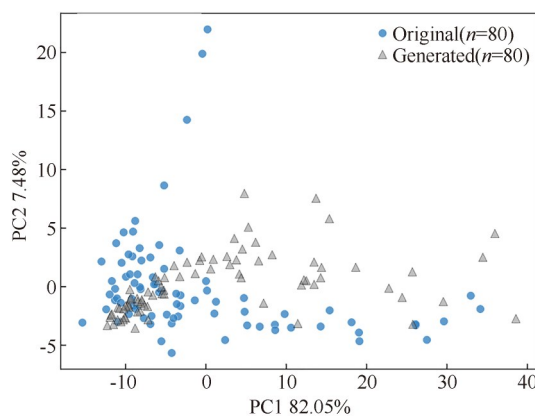
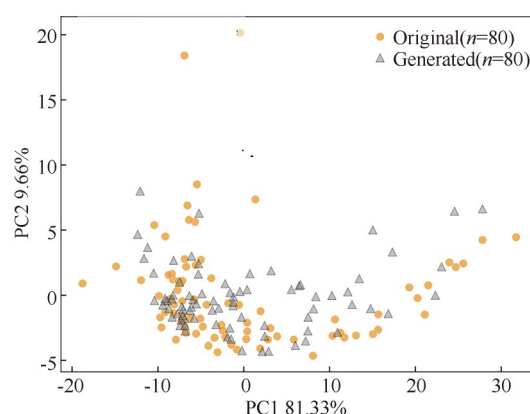
1.35%, 1.75% 和 1.81%, 说明 DFGAN 生成的样本分布更加贴近原始数据主要特征的分布, 分布一致性更强。

两种方法生成的样本在 PC1-PC2 二维特征空间中, 均与原始样本形成明显重叠区域, 表明两种方法均可有效学习细菌透射光谱的核心特

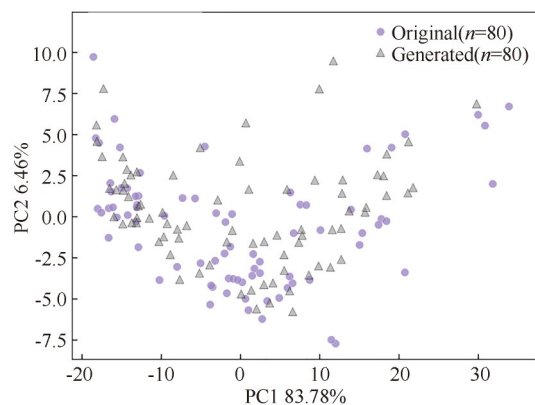
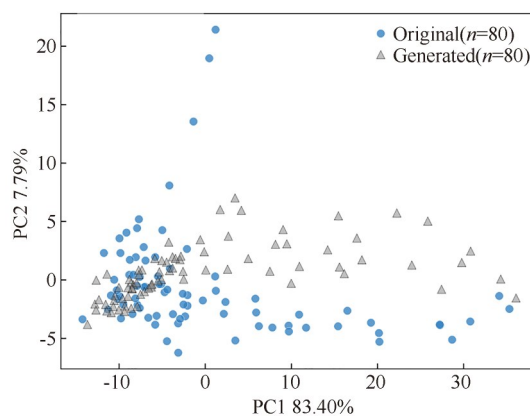
征分布。然而, 进一步的对比显示, 引入动态筛选机制的 DFGAN 所生成的样本在分布范围上与原始光谱更加吻合。以肺炎克雷伯菌、金黄色葡萄球菌和鼠伤寒沙门氏菌为例, DFGAN 生成的样本不仅精准锚定了原始数据集的高密度中心区, 还能有效延展至特征空间的边缘分布带,



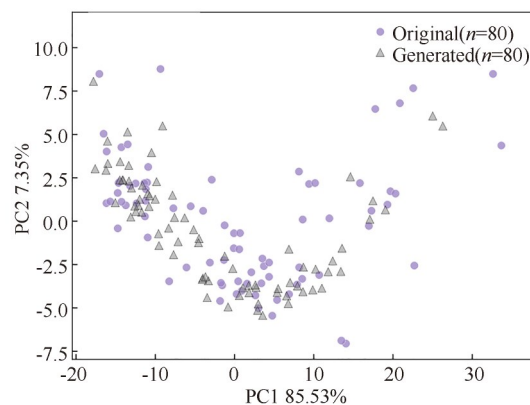
(a) 肺炎克雷伯菌
(a) *K. pneumoniae*



(b) 粪链球菌
(b) *E. faecalis*



(c) 金黄色葡萄球菌
(c) *S. aureus*



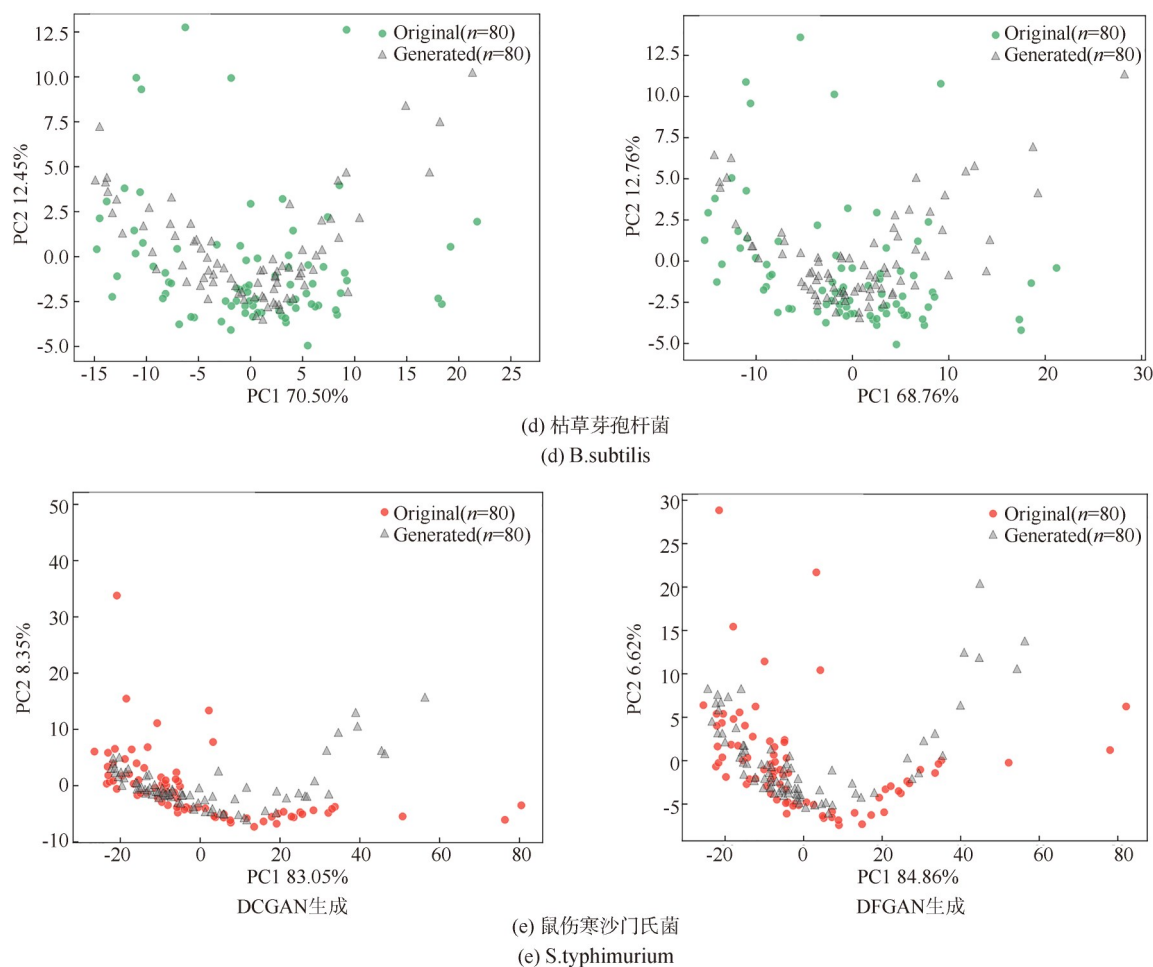


图7 DCGAN和DFGAN的生成样本与原始样本的PCA分布对比

Fig. 7 Comparison of PCA Distributions between generated samples from DCGAN and DFGAN and original samples

无明显的异常离群现象。特别是在金黄色葡萄球菌的分布投影中,DFGAN生成样本的边界拓扑结构与真实样本几乎完全等效。这一空间多样性的优越表现,与表1中该菌全局与局部分布相似度(MMD与SWD)等统计指标的优化形成了交叉印证,表明DFGAN生成样本的特征分布更贴切原始样本光谱数据的真实分布规律。

综上所述,从视觉形态、统计分布相似度和特征空间多样性三个维度评估,DCGAN和DFGAN生成的光谱样本和原始光谱相比,均表现出良好的可信度和多样性。其中,DFGAN生成样本的质量整体优于DCGAN,这主要得益于其采用动态筛选机制,该机制有效缓解了DCGAN生成样本多样性不足的问题,使模型能够生成既符合原始光谱特征,又具备丰富个体差异的高质量样本。

4.2 基于DCGAN和DFGAN光谱增强的细菌种类识别

首先将5种细菌共600条原始光谱,采用分层抽样法划分训练集和测试集,其过程是:先将每个培养时间光谱分成两层,即高梯度浓度(菌液:去离子水=1:1,1:3,1:7)和低梯度浓度(菌液:去离子水=1:15,1:31,1:63),随机抽取每种细菌每个培养时间(1~20 h,间隔为1 h,共20个时间段)的高梯度浓度和低梯度浓度中的3条光谱各1条,即每种细菌的每个时期分层随机抽取2条光谱作为测试样本,每种细菌20个培养时间共有40条光谱测试样本,剩余的80条光谱作为训练样本,最终,训练集光谱为400条,测试集样本为200条。数据集的划分如表2所示。

采用K近邻(K-nearest Neighbor, KNN)、SVM(Support Vector Machine, SVM)和卷积神

表 2 原始数据集和增强数据集的划分

Tab. 2 Partition of the original dataset and augmented dataset

Bacterial	Original		Augmented	
	Training set	Test set	Training set	Test set
<i>K. pneumoniae</i>	80	40	160	40
<i>E. faecalis</i>	80	40	160	40
<i>S. aureus</i>	80	40	160	40
<i>B. subtilis</i>	80	40	160	40
<i>S. typhimurium</i>	80	40	160	40

神经网络 (Convolution Neural Network, CNN) 三种分类算法,建立细菌的识别模型。细菌识别结果以整体识别准确率 (Accuracy, Acc)、精确率 (Precision, Pre)、召回率 (Recall, Rec) 和 F1 分数 (F1-Score, F1) 为评价指标。其计算公式分别为:

$$\begin{aligned} Acc &= \frac{TP + TN}{TP + TN + FP + FN} \\ Pre &= \frac{TP}{TP + FP} \\ Rec &= \frac{TP}{TP + FN} \\ F1 &= 2 \times \frac{Pre \times Rec}{Pre + Rec}, \end{aligned} \quad (5)$$

其中: TP 表示真正例,即真实类别为正类、模型预测类别也为正类的样本数量; TN 表示真负例,即真实类别为负类、模型预测类别同样为负类的样本数量; FP 表示假正例,即真实类别为负类、模型预测类别为正类的样本数量; FN 表示假负例,即真实类别为正类、模型预测类别为负类的

样本数量。

原始透射光谱结合三种分类模型的细菌识别结果如表 3 所示。在原始数据集上, SVM 模型在准确率 (91.50%)、精确率 (91.87%)、召回率 (91.50%) 和 F1 分数 (91.33%) 均显著优于 KNN 和 CNN 模型的识别结果,表明 SVM 模型在细菌识别任务中具有更优的准确性和鲁棒性。这是因为 KNN 模型对高维光谱数据的非线性特征拟合能力较差,易受样本分布波动影响,识别准确率仅为 73.50%, CNN 模型虽具有优异的深层特征提取能力,但受训练样本数量不足的限制,无法充分发挥深度学习模型的优势,其识别准确率为 87.50%。

表 3 原始透射光谱结合 KNN, SVM 和 CNN 模型的细菌识别结果

Tab. 3 Bacterial identification results of the original transmitted spectra based on the KNN, SVM and CNN models (%)

Model	Acc	Pre	Rec	F1
KNN	73.50	73.68	73.50	73.35
SVM	91.50	91.87	91.50	91.33
CNN	87.50	87.92	87.50	89.50

实验中采用 DCGAN 和 DFGAN 两种数据增强方法,将各类细菌原始数据集中训练集透射光谱 (80 条) 进行增强,单类细菌训练集各增加 80 条光谱,扩增后每类细菌光谱训练集光谱扩充到 160 条,测试集保持不变,仍为 40 条光谱 (增强后数据集划分见表 2)。最终 5 种细菌共包含 1 000 条光谱数据,其中训练集 800 条,测试集 200 条。

表 4 基于 DCGAN 和 DFGAN 数据增强结合 KNN, SVM 和 CNN 模型识别结果对比

Tab. 4 Comparison of recognition results using DCGAN and DFGAN data augmentation combined with KNN, SVM and CNN models (%)

Model	DCGAN				DFGAN			
	Acc	Pre	Rec	F1	ACC	Pre	Rec	F1
KNN	80.00	79.71	80.00	79.06	80.00	79.64	80.00	79.15
SVM	93.00	93.30	93.00	92.97	93.50	93.67	93.50	93.48
CNN	91.50	93.02	91.50	91.04	95.00	95.45	95.00	95.00

由表 4 可知,采用 DCGAN 和 DFGAN 方法对细菌透射光谱数据进行扩充后,结合三种分类

算法,整体识别性能均得到显著提升。经 DCGAN 增强后, SVM 模型识别性能优于 KNN 和

CNN;而经DFGAN 增强后,CNN 模型的识别性
能达到最优,其准确率、精确率、召回率和 F1 分
数 分 别 为 95.00%, 95.45%, 95.00% 和
95.00%,为全部对比模型中的最佳水平。结果
表明,两种数据增强方法可有效丰富细菌透射光
谱的样本特征,改善样本量不足所引发的模型泛
化能力偏弱问题,其中 DFGAN 表现更优。

为进一步验证本文所提 DFGAN 数据增强
方法在细菌识别任务中性能优势,将该方法与近
邻插值增强 (Synthetic Minority Oversampling
Technique, SMOTE)、高斯噪声增强两类传统经
典数据增强方法进行对比实验。实验控制变量
保持统一:各增强方法对每类细菌训练样本扩增
80 条,测试集沿用原始 200 条光谱,分类器选用
上述实验中性能最优的 CNN 模型。不同增强方
法下结合 CNN 模型的细菌识别性能对比结果如
表 5 所示。

表 5 基于 SMOTE、高斯噪声增强和 DFGAN 结合
CNN 模型的识别结果对比

Tab.5 Comparison of recognition results based on
SMOTE, Gaussian noise enhancement, and DF-
GAN combined with CNN model (%)

Enhancement methods	Acc	Pre	Rec	F1
SMOTE	87.00	90.10	87.00	86.47
Gaussian noise	92.00	93.15	92.00	91.79
DFGAN	95.00	95.45	95.00	95.00

由表 5 可知,与原始数据集相比,SMOTE 近
邻插值增强对模型识别准确率仅为 87.00%,识
别准确率未得到提升。高斯噪声增强可将模型
准确率由 87.50% 提升至 92.00%,表明通过添
加随机扰动能够适度增强模型的准确性,但该方
法仅能为原始样本引入随机波动,无法生成具备
新特征分布的有效样本,因此性能提升幅度较为
有限。

相比之下,本文所提 DFGAN 方法取得了最
优识别效果,相较于无增强的原始数据集,准确
率提升了 7.50%。这一结果充分验证了 DF-
GAN 数据增强方法的有效性,同时证明了在水
体细菌识别任务中,DFGAN-CNN 模型可实现
高准确性、高鲁棒性的识别。

5 结 论

基于透射光谱分析的水体细菌识别方法具
有无损、非侵入和测量速度快等优势,在水体细
菌微生物检测上有广阔应用前景。针对细菌透
射光谱数据量有限的问题,提出基于 DCGAN 和
DFGAN 数据增强方法,并结合 KNN, SVM 和
CNN 三种分类算法建立水体细菌识别模型。结
果表明:(1)基于 DFGAN 数据增强方法扩增细
菌透射光谱数据与原始光谱数据相比,在视觉形
态、统计分布相似性和特征空间多样性方面均优
于 DCGAN 的数据增强方法;(2)与原始数据集
细菌识别结果相比,采用 DCGAN 和 DFGAN 数
据增强后的透射光谱,结合上述三种分类算法,
对细菌识别的准确率、精确率和 F1 分数都有明
显提升,尤其是 DFGAN-CNN 方法所建立细菌
识别模型表现最优,准确率、精确率、召回率与
F1 分数分别达到 95.00%, 95.45%, 95.00% 与
95.00%。本文所提 DFGAN 数据增强方法,能
够有效提升水体细菌的识别性能,为小样本场
景下光谱数据增强研究提供思路。将透射光谱
技术与 DFGAN-CNN 模型相结合,可实现水体
细菌的无损、准确识别,同时为 CNN 等深度学
习算法在透射光谱分析领域的应用提供可行
方案。

作者贡献声明:

胡玉霞:开展水体细菌光谱检测技术调研,
提出将透射光谱结合数据增强策略用于水体细
菌光谱数据处理与样本扩充,指导实验样品制
备、透射光谱采集与模型构建,统筹实验资源与
研究流程,监督实验全过程,进行论文的审阅与
修改;

黄书龙:参与水体细菌菌悬液制备与透射光
谱实验,完成细菌透射光谱数据采集;实现对抗
生成网络数据增强方法扩充数据,构建分类模型
并进行性能验证,开展光谱特征分析与实验结果
分析,负责论文初稿撰写与修改;

徐亚玲:参与实验样品制备与透射光谱实
验,进行光谱预处理;

薛巧玲:参与实验样品制备与透射光谱实
验,参与光谱分析与结果分析;

付金超:参与实验样品制备与透射光谱实验,负责不同模型性能对比实验;

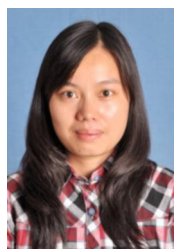
胡敦:负责实验结果准确性与可靠性验证,进行论文审阅与修改。

参考文献:

- [1] POKHAREL N, NEUPANE P, KARKI A, *et al.* Physicochemical factors Influencing *E. coli* Contamination in kathmandu valley ponds: public health and environmental implications [J]. *Environmental Health Insights*, 2025, 19: 11786302251327938.
- [2] 邓闯,王昱人,刘水晶,等. 武汉市湖泊水体三种致腹泻病原菌的分布特征[J]. *中国环境科学*, 2024, 44(10): 5875-5884.
- DENG M, WANG Y R, LIU W J, *et al.* The distribution characteristics of three types of diarrheal pathogens in the lakes of Wuhan city[J]. *China Environmental Science*, 2024, 44 (10): 5875-5884. (in Chinese)
- [3] 曾强, 邹堂春. 中国水污染状况及其引起的健康危害与对策[J]. *中华疾病控制杂志*, 2023, 27(5): 503-507.
- ZENG Q, WU T C. The occurrence of water pollution and its adverse health effects and future countermeasures in China [J]. *Chinese Journal of Disease Control & Prevention*, 2023, 27 (5): 503-507 (in Chinese).
- [4] OON Y L, OON Y S, AYAZ M, *et al.* Waterborne pathogens detection technologies: advances, challenges, and future perspectives[J]. *Frontiers in Microbiology*, 2023, 14: 1286923.
- [5] 邓闯, 李露, 宋康. 水环境中致病菌快速检测技术及应用研究进展[J]. *环境科学研究*, 2023, 36 (10): 1835-1844.
- DENG M, LI L, SONG K. Rapid molecular biology detection and risk assessment of waterborne pathogens: a review[J]. *Research of Environmental Sciences*, 2023, 36(10): 1835-1844. (in Chinese)
- [6] GAO M, TAN F Y, SHEN Y, *et al.* Rapid detection method of bacterial pathogens in surface waters and a new risk indicator for water pathogenic pollution[J]. *Scientific Reports*, 2024, 14: 1614.
- [7] MUHSIN S A, AL-AMIDIE M, SHEN Z Y, *et al.* A microfluidic biosensor for rapid simultaneous detection of waterborne pathogens [J]. *Biosensors and Bioelectronics*, 2022, 203: 113993.
- [8] 任显宗, 王巧贞, 杨玲, 等. 基于多波长透射光谱对多种细菌的鉴别方法研究[J]. *中国卫生检验杂志*, 2021, 31(24): 2945-2950, 2957.
- REN X Z, WANG Q Z, YANG L, *et al.* Study on the identification method of multiple bacteria based on multi-wavelength transmission spectrum[J]. *Chinese Journal of Health Laboratory Technology*, 2021, 31(24): 2945-2950, 2957. (in Chinese)
- [9] GAN T T, ZHAO N J, HU Y X, *et al.* Spectral features analysis of multi-wavelength transmission spectra of pathogenic bacterial microbes in water [J]. *Spectroscopy and Spectral Analysis*, 2018, 38 (5): 1610-1619. (in Chinese)
- [10] 冯春, 赵南京, 殷高方, 等. 多波长透射光谱特征提取结合支持向量机的水体细菌识别方法研究[J]. *光谱学与光谱分析*, 2021, 41(9): 2940-2944.
- FENG C, ZHAO N J, YIN G F, *et al.* Study on multi-wavelength transmission spectral feature extraction combined with support vector machine for bacteria identification[J]. *Spectroscopy and Spectral Analysis*, 2021, 41(9): 2940-2944. (in Chinese)
- [11] HU Y X, ZHU Y H, HU D, *et al.* Rapid identification of bacteria in water by multi-wavelength transmittance spectroscopy and the artificial neural network[J]. *Frontiers in Environmental Science*, 2024, 11: 1274605.
- [12] SMITH J M, HUFFMAN D E, ACOSTA D, *et al.* Reagent-free bacterial identification using multivariate analysis of transmission spectra[J]. *Journal of Biomedical Optics*, 2012, 17(10): 1070021.
- [13] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, *et al.* Generative Adversarial Networks [EB/OL]. 2014: *arXiv*: 1406.2661. <https://arxiv.org/abs/1406.2661>
- [14] LU Y Z, CHEN D, OLANIYI E, *et al.* Generative adversarial networks (GANs) for image augmentation in agriculture: a systematic review[J]. *Computers and Electronics in Agriculture*, 2022, 200: 107208.
- [15] HUANG G F, JAFARI A H. Enhanced balancing GAN: minority-class image generation[J]. *Neural Computing and Applications*, 2023, 35(7): 5145-5154.
- [16] LORENCIN I, BARESSI ŠEGOTA S,

- ANĐELIĆ N, *et al.* On urinary bladder cancer diagnosis: utilization of deep convolutional generative adversarial networks for data augmentation [J]. *Biology*, 2021, 10(3): 175.
- [17] LI H, ZHANG L, SUN H, *et al.* Discrimination of unsound wheat kernels based on deep convolutional generative adversarial network and near-infrared hyperspectral imaging technology [J]. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2022, 268: 120722.
- [18] 李灵巧, 李彦晖, 殷琳琳, 等. 基于 DCGAN 的拉曼光谱样本扩充及应用研究 [J]. *光谱学与光谱分析*, 2021, 41(2): 400-407.
- LIL Q, LI Y H, YIN L L, *et al.* Data augmentation of Raman spectral and its application research based on DCGAN [J]. *Spectroscopy and Spectral Analysis*, 2021, 41(2): 400-407. (in Chinese)
- [19] LYU H Y, GRAFTON M, RAMILAN T, *et al.* Synthetic hyperspectral reflectance data augmentation by generative adversarial network to enhance grape maturity determination [J]. *Computers and Electronics in Agriculture*, 2025, 235: 110341.
- [20] 胡玉霞, 薛芙蓉, 张瑞祥, 等. 基于透射光谱的水体混合细菌快速识别方法 [J]. *光子学报*, 2025, 54(7): 0730004.
- HU Y X, XUE F R, ZHANG R X, *et al.* Rapid identification of mixed bacteria in water based on transmission spectroscopy [J]. *Acta Photonica Sinica*, 2025, 54(7): 0730004. (in Chinese)
- [21] LEE W, LENFERINK A T M, OTTO C, *et al.* Classifying Raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection [J]. *Journal of Raman Spectroscopy*, 2020, 51(2): 293-300.
- [22] 赵勇, 荣康, 谈爱玲. 基于一维卷积神经网络的雌激素粉末拉曼光谱定性分类 [J]. *光谱学与光谱分析*, 2019, 39(12): 3755-3760.
- ZHAO Y, RONG K, TAN A L. Qualitative analysis method for Raman spectroscopy of estrogen based on one-dimensional convolutional neural network [J]. *Spectroscopy and Spectral Analysis*, 2019, 39(12): 3755-3760. (in Chinese)
- [23] SHANG L W, MA D Y, FU J J, *et al.* Fluorescence imaging and Raman spectroscopy applied for the accurate diagnosis of breast cancer with deep learning algorithms [J]. *Biomedical Optics Express*, 2020, 11(7): 3673.
- [24] YANG Y, QIU C L, ZHOU D X, *et al.* NIR-GAN: a spectral data augmentation framework for medicine-food homologous herb identification [J]. *Journal of Food Composition and Analysis*, 2025, 148: 108328.

作者简介:



胡玉霞(1983—),女,博士,副教授,硕士研究生导师,主要从事细菌微生物光谱检测技术和人工智能算法等方面的研究。E-mail: yxhuhf@163.com

通讯作者:



胡敦(1986—),男,博士,高级工程师,主要从事检测技术与自动化装置方面的研究。E-mail: hudun@mail.ustc.edu.cn