

基于改进 CycleGAN 的玉米蛋白质含量近红外光谱检测方法

王文娟^{1,2} 吕思豆¹ 陈明^{1,2} 邹一波^{1,2} 葛艳^{1,2}

(1. 上海海洋大学信息学院, 上海 201306; 2. 农业农村部渔业信息重点实验室, 上海 201306)

摘要: 玉米富含的蛋白质对人类健康具有重要意义, 精准测定其蛋白质含量是实现玉米优质品种选育与营养含量评估的基础。利用近红外光谱技术可以实现对玉米蛋白质含量无损检测, 然而, 近红外光谱仪器及生化指标检测成本高昂, 机器学习在特征提取方面又存在不足。为此, 本文提出了一种基于改进循环生成对抗网络 (Cycle-constraint generative adversarial network, CycleGAN) 的玉米蛋白质含量近红外光谱检测方法。使用基于 CycleGAN 的数据降噪模型 (NIRDM_CycleGAN) 对生成的数据进一步降噪, 将处理后的数据用于后续实验。NIRDM_CycleGAN 模型引入特征增强模块 (ISE), 以增强其降噪能力。在 NIRDM_CycleGAN 模型降噪处理后, 采用中值滤波抑制光谱数据末端的异常峰值。实验结果表明, 在完成数据增强后, 采用 NIRDM_CycleGAN 模型对样本进行降噪可进一步提高预测模型性能。本文构建的预测模型 (NIRPM_CNN) 取得最佳预测结果, 预测均方根误差 (Root mean square error of prediction, RMSEP)、预测集决定系数 (R_p^2)、相对分析误差 (Relative percent deviation, RPD) 分别为 0.032 1、0.980 2 和 7.118 3。针对近红外光谱数据样本不足的情况, 本文提出的方法可对原始近红外光谱数据进行数据增强与降噪处理, 有效扩充数据集规模, 从而显著降低检测成本并缩短检测时间, 实现了低成本、快速且精准的玉米蛋白质含量近红外光谱无损检测目标。

关键词: 玉米蛋白质含量; 近红外光谱; NIREM_WGAN; NIRDM_CycleGAN; NIRPM_CNN

中图分类号: S513; TP183 **文献标识码:** A **文章编号:** 1000-1298(2026)18-0365-08

OSID:



Near Infrared Spectral Detection Method for Maize Protein Content Based on Improved CycleGAN

Wang Wenjuan^{1,2} Lü Sidou¹ Chen Ming^{1,2} Zou Yibo^{1,2} Ge Yan^{1,2}

(1. College of Information Technology, Shanghai Ocean University, Shanghai 201306, China

2. Key Laboratory of Fisheries Information, Ministry of Agriculture and Rural Affairs, Shanghai 201306, China)

Abstract: The abundant protein in maize is important to human health, and accurate determination of protein content is fundamental to variety breeding and nutritional evaluation. Near-infrared spectroscopy (NIRS) was used for the non-destructive measurement of maize protein content. However, NIRS instruments and biochemical reference assays were costly. In addition, the feature-extraction capability of conventional machine-learning methods was limited. Accordingly, a near-infrared spectroscopic method for maize protein determination based on an improved cycle-constraint adversarial network (CycleGAN) was proposed. Unlike most studies centred on generation tasks, a CycleGAN-based data-denoising model (NIRDM_CycleGAN) was innovatively adopted to further denoise the generated spectra, and the processed data were used in subsequent experiments. A feature-enhancement module (ISE) was introduced into the NIRDM_CycleGAN model to strengthen its denoising capability. After denoising by NIRDM_CycleGAN, a median filter was applied at the spectral end to suppress abnormal peaks. Experimental results showed that after data augmentation, prediction performance was further improved by denoising with the NIRDM_CycleGAN model. Comparative trials demonstrated that the prediction model developed (NIRPM_CNN) exhibited the best performance, delivering a root mean square error of

收稿日期: 2025-06-05 修回日期: 2025-08-25

基金项目: 广东省重点领域研发计划项目(2021B0202070001)

作者简介: 王文娟(1983—), 女, 副教授, 博士, 主要从事水产品质量溯源和撮合交易机制研究, E-mail: wangwj@shou.edu.cn

通信作者: 陈明(1966—), 男, 教授, 博士, 主要从事智慧水产养殖和水产品追溯技术研究, E-mail: mchen@shou.edu.cn

prediction (RMSEP) of 0.032 1, a coefficient of determination (R_p^2) of 0.980 2, and a relative percent deviation (RPD) of 7.118 3 on the test set. To address the limited NIR sample size, the proposed approach was applied separately for data augmentation and denoising of the raw spectra. The dataset was expanded. As a result, detection cost was reduced and testing time was shortened. Thus low-cost, rapid, and accurate non-destructive NIR detection was achieved.

Key words: maize protein content; near infrared spectroscopy; NIREM_WGAN; NIRD_M_CycleGAN; NIRPM_CNN

0 引言

玉米是我国重要的粮食作物,在农业和工业生产中占有重要地位。玉米中富含的蛋白质不仅是维持人体健康的重要营养物质,还有助于减少配制营养全面的牲畜饲料的生产成本。蛋白质含量是评价玉米品质的重要指标^[1]。因此,玉米蛋白质含量的测定具有重要意义。传统的蛋白质含量测定方法主要有凯氏定氮法和杜马斯法^[2],这些方法不仅耗时费力,且通常仅能在实验室环境实施,很难实时有效地测量大规模田间玉米作物,难以满足企业在检测环节中对高效性的要求。因此,构建快速、无损、精准的玉米蛋白质含量测定方法具有重要的研究价值与应用意义。

近年来,随着各类光学传感器的快速发展,近红外光谱技术逐渐成熟,体系日益完善,已广泛应用于农业领域^[3-7],具有高效、无损的特点^[8]。现有的机器学习模型可对光谱数据及其对应的目标变量或类别标签进行建模分析^[9-12]。目前这些研究大多围绕光谱预处理、特征波段筛选或机器学习模型展开,旨在提升预测或分类的准确率,但没有解决获取大量近红外光谱和生化指标数据难度较大、成本较高的问题。数据量不足会在一定程度上影响建模的精度和鲁棒性,近红外光谱仪器和生化指标检测费用高昂。这些因素将制约近红外光谱技术在无损检测领域的广泛应用。

数据增强方法能够生成贴近真实样本分布的新样本,有效提升模型泛化能力,并降低过拟合风险^[13],减少检测样本光谱和生化指标数据所需要的成本,缩短检测时间。目前针对光谱数据的传统数据增强方法有随机偏移法、光谱混合法和叠加噪声法^[14]等。这些方法没有提取光谱数据的内部特征,并且可能产生与原始数据分布不一致的数据。近年来,生成对抗网络(Generative adversarial network, GAN)^[15]通过生成器与判别器之间的对抗训练实现数据增强,能提取光谱数据的内部特征。GAN 已被成功应用于图像生成领域^[16]。在光谱无损检测领域,GAN 已展现出卓越的应用潜力。近年来,文献[17-20]采用不同类型的 GAN 模型对光谱数据进行数据增强,有效提升了模型性能。上述研究均

采用不同类型的 GAN 模型对真实样本进行数据增强,但未对增强后的数据实施进一步降噪处理。在生成过程,GAN 通常以固定维度的随机潜在向量作为输入,并通过生成器将其映射为生成样本,在训练过程中使生成样本逐渐接近真实样本分布,从而引入随机性和多样性。因此,对增强后的数据进行降噪处理可进一步提升数据质量。

针对上述问题,本文提出基于 Wasserstein 距离与梯度惩罚生成对抗网络(Wasserstein generative adversarial network with gradient penalty, WGAN - GP)^[21]的数据增强模型(NIREM_WGAN)、基于循环生成对抗网络(Cycle-constraint generative adversarial network, CycleGAN)^[22]的降噪模型(NIRD_M_CycleGAN)与基于一维卷积神经网络(One-dimensional convolutional neural network, 1D-CNN)和多层感知机(Multilayer perceptron, MLP)的预测模型(NIRPM_CNN)。对原始样本进行预处理得到真实样本;采用 NIREM_WGAN 模型生成高质量样本;利用 NIRD_M_CycleGAN 模型对 NIREM_WGAN 模型生成的样本进行降噪,以保证生成数据更加接近真实数据;使用 NIRPM_CNN、1D-CNN、偏最小二乘回归(Partial least squares regression, PLSR)和支持向量回归(Support vector regression, SVR)模型对样本进行建模。

1 材料与方法

1.1 数据来源与预处理

Feng 等^[23]发现,采用大规模生成对抗网络(Big generative adversarial networks, BigGAN)与改进的基于风格的生成对抗网络(Style-based generative adversarial network 2, StyleGAN2)模型生成图像的感知质量随原始训练样本的规模呈明显的“U”形变化。当训练样本极少时,生成模型易过拟合,几乎直接复制训练样本,因而生成的图像在视觉上逼近真实图像,并获得较高的主观评分。然而,此类近乎复制的样本缺乏多样性,难以视为高质量生成样本。当样本数量略有增加但仍不足以覆盖真实样本分布时,生成器无法学习完整的样本特征,复制率与生成图像质量同步下降。随着样本规模进一步扩大,生

成模型能够获取更丰富的多样性信息,生成图像的质量再次提高,逐渐逼近真实样本分布,且复制率维持在较低水平。Zhu 等^[24]发现,随着训练数据集规模增大,cDCGAN 生成的伪光谱数据质量总体提高。在基于光谱技术与数据增强方法研究^[17-20,25-27]中,原始样本数量均不少于 80 个。选取玉米近红外光谱数据集^[28]作为原始样本,数据集共 80 个样本。该样本规模既可验证本文所提方法在小样本条件下的有效性,又能保证较低的检测成本和较短的检测时间,同时在一程度保证生成样本的多样性。原始数据集来源于 EVRI 网站公开发布的玉米近红外光谱数据集。选取其中 m5 仪器采集的光谱数据及其对应的蛋白质含量数据作为研究对象,该光谱数据光谱曲线如图 1 所示。m5 近红外光谱仪波长为 1 100~2 498 nm,间隔为 2 nm。

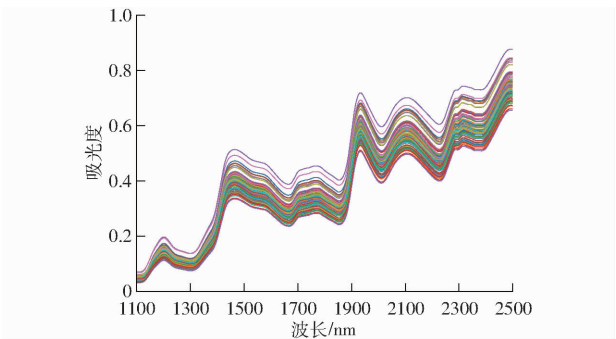


图 1 原始样本近红外光谱曲线

采用最大最小归一化、基于标准分数 (Z-Score) 的异常值检测、基于 PLSR 与蒙特卡罗交叉验证 (MCCV) 的残差分析和多元散射校正 (Multiplicative scatter correction, MSC) 对原始样本进行预处理,所得样本简称为真实样本。提出的基于改进 CycleGAN 的玉米蛋白质含量近红外光谱检测方

法由上述预处理方法、NIREM_WGAN 模型、NIRDM_CycleGAN 模型和 NIRPM_CNN 模型构成。

1.2 模型构建

1.2.1 NIREM_WGAN 模型

提出的 NIREM_WGAN 模型可用于小样本近红外光谱数据及其对应蛋白质含量的联合生成任务。将 NIREM_WGAN 模型批量大小设定为 8,结构如图 2 所示。与基于 MLP 的 WGAN-GP 模型相比,NIREM_WGAN 模型在生成器主干网络中使用一维转置卷积层代替线性层,在判别器主干网络中使用一维卷积层代替线性层。为进一步提高模型收敛速度和训练稳定性,并改善生成样本质量,引入贝叶斯优化方法对学习率和梯度惩罚系数等关键超参数进行调优。卷积结构利用局部连接与权重共享机制,在保证特征提取能力的同时显著减少可训练参数量,从而降低计算与存储开销并提高计算效率。生成器接收一批随机噪声向量,先将其调整为 (8, 100, 1) 的张量,其中 8、100 和 1 分别表示批量大小、通道数和序列长度。随后,经包含一维转置卷积层、批归一化 (Batch normalization, BN) 层、线性整流单元 (Rectified linear unit, ReLU) 层和双曲正切 (Hyperbolic tangent, Tanh) 层的生成器主干网络处理后,其形状变为 (8, 1, 701)。判别器接收生成样本与真实样本,并输出形状为 (8, 1) 的判别分数,其中 8 表示批量大小,1 表示判别器对每个输入样本输出 1 个判别分数。

1.2.2 NIRDM_CycleGAN 模型

提出的 NIRDM_CycleGAN 模型可对 NIREM_WGAN 模型生成的样本进行降噪处理。将 NIRDM_CycleGAN 模型批量大小设置为 16,结构如图 3 所示。相较于原始 CycleGAN 模型,该模型在生成器

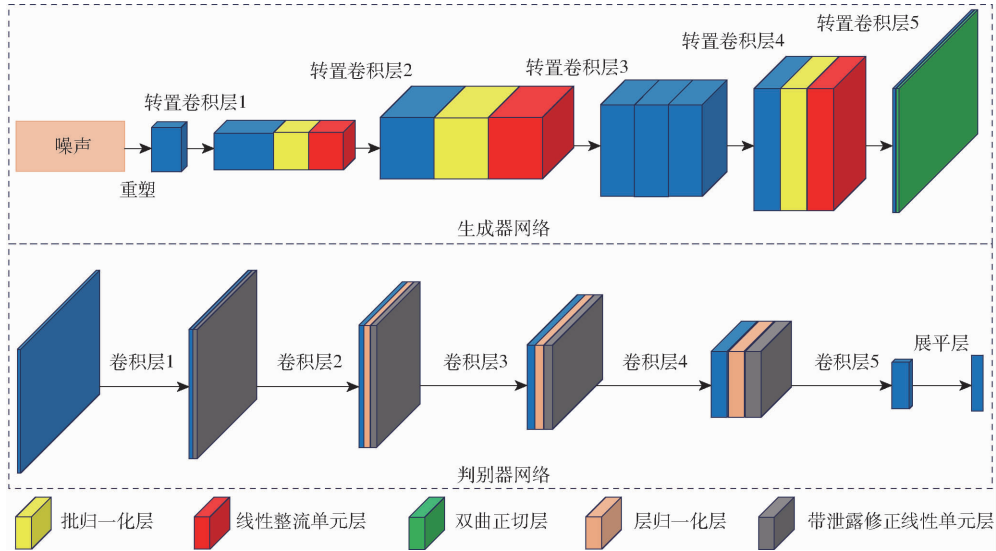


图 2 NIREM_WGAN 模型结构

Fig. 2 Architecture of NIREM_WGAN model

与判别器中引入了一种新型注意力模块 ISE。ISE 模块对压缩和激励模块 (Squeeze and excitation, SE) 进行改进,能够自适应地提升重要通道的权重,降低次要通道的权重。NIRDM_CycleGAN 模型训练阶段综合采用最小二乘对抗损失、循环一致性损失和恒等映射损失,既保证域间转换的真实性,又保持原始光谱的整体形态。模型训练完成后,对生成光谱末端进行中值滤波,进一步去除噪声并抑制异常峰值。与原始 SE 模块相比,ISE 模块先自适应选择压缩率,以生成全局通道权重;随后,按预设子带长度沿

通道方向将特征向量的所有通道划分为 3 个子带,并以固定的最小压缩率对各子带进行通道重标定;最后,将各子带输出进行拼接,再与全局权重相乘。形状为 (16,700) 的特征张量经过生成器的第 1 个线性层,其形状变为 (16,256),16 表示批量大小,700 表示特征维度。随后,特征张量依次经 ReLU 层、ISE 模块、第 2 个线性层、ReLU 层、ISE 模块和第 3 个线性层后恢复至最初特征维度。生成样本和真实样本特征张量经过判别器后得到形状为 (16,1) 的判别结果。

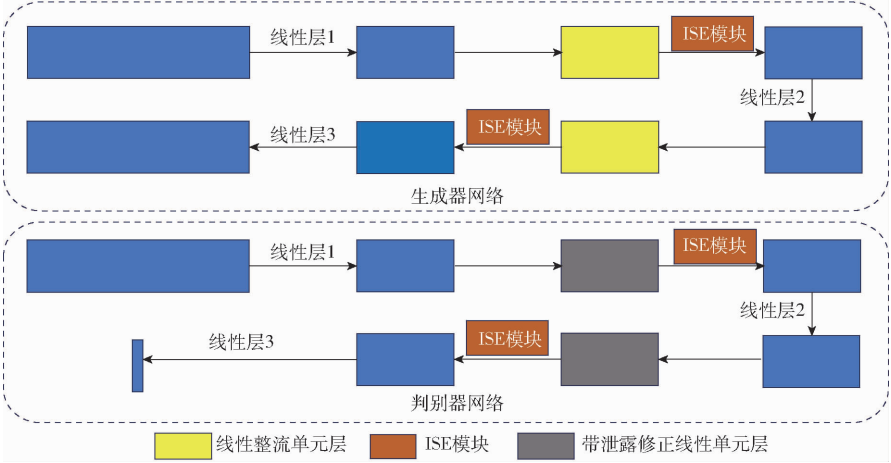


图 3 NIRDM_CycleGAN 模型结构

Fig. 3 Architecture of NIRDM_CycleGAN model

1.2.3 NIRPM_CNN 模型

采用 NIRPM_CNN 模型对样本进行建模。将 NIRPM_CNN 模型批量大小设置为 32,结构如图 4 所示。与原始 1D-CNN 模型相比,NIRPM_CNN 模型在特征提取子网络后引入联合注意力模块 (CSG),该模块包括改进的通道注意力、改进的空间注意力和多头自注意力。此外,NIRPM_CNN 模型使用贝叶斯优化对自适应矩估计 (Adaptive moment estimation, Adam) 优化器的学习率和权重衰减、空间

注意力中可变形卷积核尺寸进行了优化,并在全连接网络中加入随机丢弃层 (Dropout),从而有效抑制过拟合。改进的通道注意力首先引入可学习的通道关系矩阵,通过矩阵乘积显式建模通道间的统计相关结构;随后,基于输入特征动态生成双分支 MLP 权重,实现对通道映射的自适应分配,并利用快速傅里叶变换提取频域信息,构建频域特征增强模块;最后,经 Sigmoid 函数门控对双分支输出进行特征融合。改进的空间注意力采用可变形卷积,通过可学

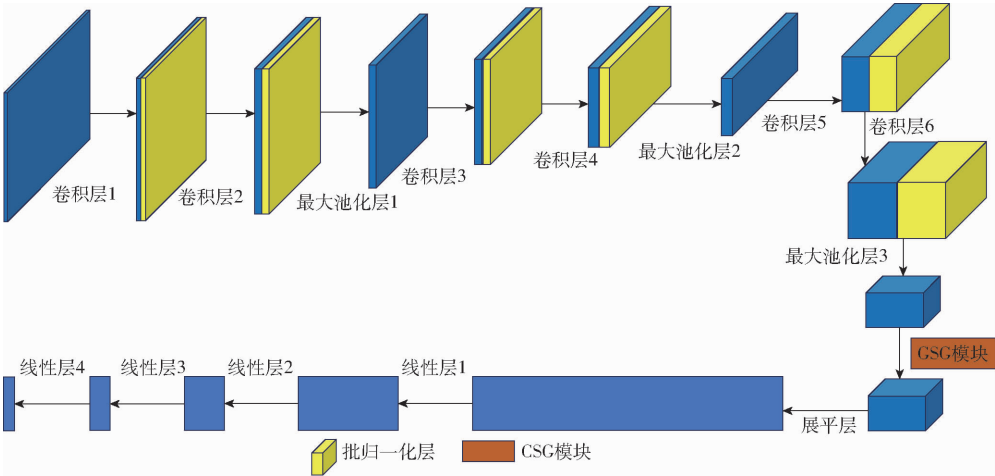


图 4 NIRPM_CNN 模型结构

Fig. 4 Architecture of NIRPM_CNN model

习偏移改变采样位置,从而自适应调整感受野。形状为(32,1,700)的张量经包含一维卷积层、BN 层、池化层和 CSG 模块的特征提取子网络处理后,其形状变为(32,256,17)。随后,经展平层后其形状变为(32,4 352),并依次经过 4 层全连接网络。最后,得到形状为(32,1)的回归预测值。

1.3 评价指标

1.3.1 样本质量评价指标

采用 Fréchet Inception 距离^[29](Fréchet inception distance,FID)、峰值信噪比(Peak signal to noise ratio,PSNR)、皮尔逊相关系数(Pearson correlation coefficient,PCC)和结构相似性(Structural similarity,SSIM)4 种评价指标评估生成模型和降噪模型性能。其中,FID 越小,PSNR 越大,PCC 和 SSIM 越接近 1,表明生成模型或降噪模型性能越好。

1.3.2 预测模型评价指标

采用预测集均方根误差(RMSEP)、预测集决定系数(R_p^2)和相对分析误差(Relative percent deviation,RPD)3 种评价指标衡量模型预测能力。其中,RMSEP 越接近 0, R_p^2 越接近 1,RPD 越大,表明预测模型性能越好。

2 结果与讨论

2.1 NIREM_WGAN 模型数据增强结果

2.1.1 蛋白质含量数据增强结果

采用最大最小归一化对 NIREM_WGAN 模型训练过程中生成样本和真实样本的蛋白质含量进行处理,并对比了不同训练轮次(Epoch)下两组样本的蛋白质含量箱型图。实验结果表明,在训练轮次为 1 215 时,两组数据最小值、下四分位数、中位数、上四分位数和最大值总体最为接近。因此,将 NIREM_WGAN 模型在训练轮次为 1 215 时生成的样本作为最佳生成数据,并用于后续实验分析。

统计了真实样本和 NIREM_WGAN 模型最终生成样本蛋白质含量最小值、下四分位数、中位数、平均数、上四分位数和最大值,结果如表 1 所示。真实样本蛋白质含量平均数为 8.682 5%,NIREM_WGAN 模型最终生成样本蛋白质含量平均数为 8.704 4%,二者非常接近。除平均数外,真实样本和 NIREM_WGAN 模型最终生成样本蛋白质含量其余 5 个指标也都比较接近。由表 1 可知,两组样本蛋白质含量的下四分位数、中位数、平均数和上四分位数均较为接近,整体差异较小。

2.1.2 近红外光谱数据增强结果

绘制生成样本与真实样本光谱数据的平均光谱

表 1 真实样本和 NIREM_WGAN 模型生成样本蛋白质含量描述性统计量

Tab.1 Descriptive statistics of protein content in real samples and samples generated by

描述性统计量	NIREM_WGAN model		%
	真实样本蛋白 质含量	NIREM_WGAN 模型 生成样本蛋白质含量	
最小值	7.654 0	7.556 0	
下四分位数	8.338 0	8.342 9	
中位数	8.552 0	8.576 3	
平均数	8.682 5	8.704 4	
上四分位数	9.021 0	9.110 5	
最大值	9.711 0	9.650 9	

曲线,如图 5 所示。由图 5 可知,2 条曲线整体趋势一致,变化规律基本相同,但在局部波段仍存在细微起伏差异。

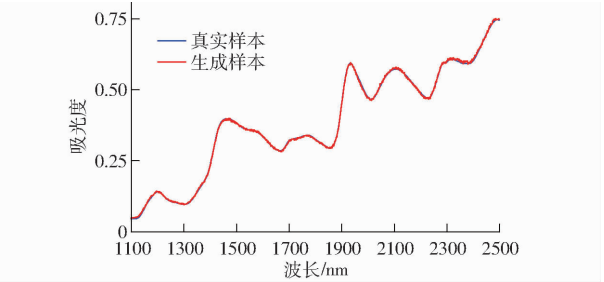


图 5 生成样本和真实样本平均光谱曲线
Fig.5 Average near-infrared spectral curves of generated and real samples

采用 GAN、深度卷积生成对抗网络(Deep convolutional generative adversarial network,DCGAN)和基于 Wasserstein 距离的生成对抗网络(Wasserstein generative adversarial networks,WGAN)对真实样本进行扩增。计算 GAN、DCGAN、WGAN 和 NIREM_WGAN 4 种生成模型生成样本与真实样本间的 FID、PSNR、PCC 和 SSIM 4 项指标,结果如表 2 所示。由表 2 可知,NIREM_WGAN 模型生成样本与真实样本光谱数据分布一致性最佳,其 FID、PSNR、PCC 和 SSIM 分别为 0.011 3、43.381 5 dB、0.999 7 和 0.988 4,验证了该模型在光谱数据生成中的显著优势。

表 2 4 种生成模型生成样本质量评估结果
Tab.2 Quality evaluation results of samples generated by four generative models

模型	FID	PSNR/dB	PCC	SSIM
GAN	0.045 7	31.245 6	0.995 1	0.953 1
DCGAN	0.030 1	34.865 3	0.997 9	0.969 1
WGAN	0.022 4	37.439 6	0.998 8	0.979 9
NIREM_WGAN	0.011 3	43.381 5	0.999 7	0.988 4

2.2 NIRDm_CycleGAN 模型降噪结果

采用 NIRDm_CycleGAN 模型对 NIREM_WGAN

模型生成样本进行降噪,并绘制降噪后样本与真实样本平均光谱曲线,结果如图 6 所示。由图 6 可知,NIRDM_CycleGAN 模型降噪后样本与真实样本平均光谱曲线几乎完全重合。与 NIREM_WGAN 模型生成样本平均光谱曲线相比(图 5),NIRDM_CycleGAN 模型降噪后样本平均光谱曲线更加平滑,与真实样本平均曲线在整体趋势上更一致。

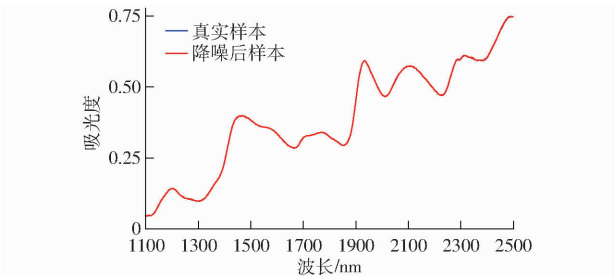


图 6 降噪后样本和真实样本平均光谱曲线
Fig. 6 Average near-infrared spectral curves of denoised samples and real samples

对比 NIRDM_CycleGAN、U 型卷积神经网络(U-architecture convolutional networks, U-Net)^[30]和 CycleGAN 模型降噪性能,结果如表 3 所示。NIRDM_CycleGAN 模型降噪性能最好,4 项指标分别为 0.006 5、48.260 9 dB、0.999 9 和 0.998 7。结合表 2 和表 3 可知,NIRDM_CycleGAN 模型降噪后,FID 由 0.011 3 下降至 0.006 5,PSNR 由 43.381 5 dB 上升至 48.260 9 dB,PCC 由 0.999 7 上升至 0.999 9,SSIM 由 0.988 4 上升至 0.998 7。上述结果表明,NIRDM_CycleGAN 模型具有明显的降噪效果,能够有效抑制噪声。降噪后样本在分布与结构相似性上更接近真实样本,整体质量更高。

表 3 3 种降噪模型质量评估结果 Tab.3 Quality evaluation results of three denoising models				
模型	FID	PSNR/dB	PCC	SSIM
U-Net	0.011 2	45.895 5	0.999 8	0.997 8
CycleGAN	0.009 4	46.670 8	0.999 9	0.998 1
NIRDM_CycleGAN	0.006 5	48.260 9	0.999 9	0.998 7

2.3 预测结果

2.3.1 不同模型在不同样本上的预测结果

对比 NIRPM_CNN、1D-CNN、PLSR 和 SVR 模型在真实样本、真实样本和 NIREM_WGAN 生成样本(简称为“真实+生成”)、真实样本和 NIRDM_CycleGAN 降噪后样本(简称为“真实+降噪”)中的预测性能指标,结果如表 4 所示。由表 4 可知,在只用真实样本进行建模时,PLSR 模型预测性能优于 NIRPM_CNN、1D-CNN 和 SVR 模型,其 RMSEP、 R_p^2 和 RPD 分别为 0.069 2、0.921 2 和 3.646 2。在真实

样本和 NIREM_WGAN 模型生成样本、真实样本和 NIRDM_CycleGAN 模型降噪后样本联合训练时,NIRPM_CNN 模型预测性能最佳,RMSEP 为 0.032 1, R_p^2 和 RPD 分别为 0.980 2 和 7.118 3。同时,由表 4 可知,数据增强可以提高 NIRPM_CNN、1D-CNN、PLSR 和 SVR 模型性能。

表 4 不同模型在不同样本上的预测结果
Tab.4 Prediction results of different models on different sample sets

样本	模型	RMSEP	R_p^2	RPD
真实	PLSR	0.069 2	0.921 2	3.646 2
	SVR	0.112 2	0.793 0	2.249 7
	1D-CNN	0.097 8	0.759 2	2.109 4
	NIRPM_CNN	0.097 6	0.845 1	2.630 0
真实+生成	PLSR	0.066 5	0.927 3	3.712 8
	SVR	0.106 2	0.814 4	2.323 7
	1D-CNN	0.053 2	0.945 4	4.286 6
	NIRPM_CNN	0.039 9	0.969 3	5.716 6
真实+降噪	PLSR	0.061 1	0.938 6	4.040 1
	SVR	0.100 1	0.835 1	2.465 3
	1D-CNN	0.047 1	0.957 1	4.835 9
	NIRPM_CNN	0.032 1	0.980 2	7.118 3

2.3.2 原始样本数量对 4 种预测模型的影响

从 80 个原始样本中分别随机选取 70 个和 60 个样本作为后续实验的两组原始样本。首先,采用上述的预处理方法对两组原始样本进行预处理,得到两组真实样本;然后分别使用 NIREM_WGAN 和 NIRDM_CycleGAN 模型对两组真实样本进行数据增强和数据降噪;最后,使用 NIRPM_CNN、1D-CNN、PLSR 和 SVR 模型对降噪后数据进行预测,结果如表 5 所示,其中“真实 1”和“真实 2”分别表示根据 60 个原始样本得到的真实样本、根据 70 个原始样本得到的真实样本。表 5 为在两组真实样本训练集持续加入 NIRDM_CycleGAN 模型降噪样本后,4 种

表 5 不同模型在玉米近红外光谱数据集不同原始样本上预测结果
Tab.5 Prediction results of different models on different original samples in maize near-infrared spectral dataset

样本	模型	RMSEP	R_p^2	RPD
真实 1+降噪	PLSR	0.058 9	0.913 9	3.411 8
	SVR	0.095 5	0.773 4	2.103 0
	1D-CNN	0.052 9	0.927 5	3.720 1
	NIRPM_CNN	0.047 4	0.941 6	4.144 9
真实 2+降噪	PLSR	0.062 0	0.922 6	3.598 4
	SVR	0.098 2	0.805 9	2.272 3
	1D-CNN	0.053 0	0.934 3	3.907 8
	NIRPM_CNN	0.047 0	0.948 3	4.405 2

预测模型在预测集上取得的最佳预测指标。实验结果表明,使用 80 个原始样本经本文方法处理后得到的预测指标最佳(见表 4, $R_p^2 = 0.980\ 2$),而使用 60 个原始样本经本文方法处理后得到的预测指标较差($R_p^2 = 0.941\ 6$)。当原始样本数量低于 80 个时,经本文方法处理后,模型预测性能下降。将 80 个原始样本用本文提出的方法进行处理,可获得较好的预测效果。因此,80 是比较合适的原始样本规模。

2.4 多源数据集实验结果与分析

小麦数据集^[31]和谷物数据集^[32]均属于近红外光谱数据集。为进一步验证本文方法在不同数据集上的可行性,将本文的整套方法分别应用于来源于制造商 A、B 的 2 个小麦数据集,以及预测目标为酪蛋白含量的谷物数据集,计算该方法在 3 种数据集上的 RMSEP、 R_p^2 和 RPD 指标,结果如表 6 所示。其中,“小麦 A”和“小麦 B”分别表示来源于制造商 A 的小麦数据集和来源于制造商 B 的小麦数据集,二者均包含近红外光谱数据及其对应的蛋白质含量数据。以酪蛋白含量为预测目标的谷物数据集包含

231 个近红外光谱数据及其对应的酪蛋白含量。由表 6 可知,本文方法在 3 种数据集上均取得了较好的预测效果,其中来源于制造商 A 和制造商 B 小麦数据集中光谱对蛋白质预测 R_p^2 分别为 0.982 4 和 0.983 9,谷物数据集中光谱对酪蛋白预测 R_p^2 为 0.984 5,均高于 0.98。由此可见,本文方法在农业领域多种近红外光谱数据集上均表现出较好的可行性,取得了良好的预测结果,为近红外光谱在农业领域的无损检测提供了可靠的技术支持。

3 结论

(1)NIREM_WGAN 模型生成样本与真实样本光谱数据的平均光谱曲线整体趋势一致,二者的蛋白质含量整体差异较小。

(2)NIRDM_CycleGAN 模型可进一步提升生成样本质量和预测模型性能。

(3)当真实样本和 NIREM_WGAN 模型生成样本、真实样本和 NIRDM_CycleGAN 模型降噪后样本联合训练时,NIRPM_CNN 模型预测性能优于 1D-CNN、PLSR 和 SVR 模型。数据增强在一定程度上可提高 NIRPM_CNN、1D-CNN、PLSR 和 SVR 模型性能。

(4)将本文的预处理方法、NIREM_WGAN、NIRDM_CycleGAN 和 NIRPM_CNN 模型分别应用于本文所提 2 个小麦数据集和谷物数据集后得到的 R_p^2 均高于 0.98。本文方法在这些数据集上都取得了良好的预测结果,表明其在不同的近红外光谱数据集上均表现出较好的可行性。本文方法无须额外采集样本,能够在一定程度上减少样本采集与理化指标测定成本,缩短检测样本时间。同时,该方法为小样本近红外光谱数据增强提供了一种有效的降噪思路,提高了小样本条件下近红外光谱无损检测预测精度。

表 6 不同模型在小麦数据集和谷物数据集预测结果

Tab.6 Prediction results of different models on wheat and cereal datasets

数据	样本	模型	RMSEP	R_p^2	RPD
小麦 A	真实	PLSR	0.546 4	0.932 5	3.852 3
	真实	SVR	0.880 1	0.824 9	2.391 8
	真实 + 降噪	NIRPM_CNN	0.281 1	0.982 4	7.550 6
小麦 B	真实	PLSR	0.047 8	0.946 4	4.323 0
	真实	SVR	0.083 7	0.835 4	2.466 9
	真实 + 降噪	NIRPM_CNN	0.026 4	0.983 9	7.894 5
谷物	真实	PLSR	0.052 4	0.952 2	4.578 4
	真实	SVR	0.074 6	0.902 8	3.168 6
	真实 + 降噪	NIRPM_CNN	0.025 9	0.984 5	8.048 1

参 考 文 献

[1] Yu Y, Qiao Y, Fan C, et al. Machine learning and near-infrared fusion-driven quantitative characterization and detection of protein content in maize kernels[J]. Frontiers in Nutrition, 2025, 12: 1719661.

[2] Sezer B, Bilge G, Boyaci I H. Laser-induced breakdown spectroscopy based protein assay for cereal samples[J]. Journal of Agricultural and Food Chemistry, 2016, 64(49): 9459 – 9463.

[3] 冀荣华, 李常昊, 郑立华, 等. 基于稀疏自注意力和可见-近红外光谱的土壤氮含量预测[J]. 农业机械学报, 2024, 55(10): 392 – 398,409.

Ji Ronghua, Li Changhao, Zheng Lihua, et al. Prediction of soil nitrogen content based on sparse self-attention and visible-near-infrared spectroscopy[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(10): 392 – 398,409. (in Chinese)

[4] 李禧龙, 韩亚芬, 潘宇轩, 等. 基于改进 ResNet 的马铃薯黑心病近红外光谱检测方法[J]. 农业机械学报, 2024, 55(12): 470 – 479.

Li Xilong, Han Yafen, Pan Yuxuan, et al. Detection on potato black heart disease by near infrared spectroscopy based on improved ResNet[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(12): 470 – 479. (in Chinese)

[5] 王懂, 杨玮, 曹永研, 等. 基于近红外光谱信息的土壤电导率预测模型研究[J]. 农业机械学报, 2022, 53(增刊 1): 218 – 223.

Wang Dong, Yang Wei, Cao Yongyan, et al. Prediction model of bulk soil electrical conductivity based on near-infrared spectral

- information[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(S1): 218 – 223. (in Chinese)
- [6] El Maouardi M, De Braekeleer K, Bouklouze A, et al. Comparison of near-infrared and mid-infrared spectroscopy for the identification and quantification of argan oil adulteration through PCA, PLS – DA and PLS[J]. Food Control, 2024, 165: 110671.
- [7] Li L, Li L, Li J, et al. A strategy to improve the detection accuracy and universality of highland barley multi-quality attributes based on near infrared spectroscopy combined with model transfer method[J]. Food Chemistry, 2025, 480: 143887.
- [8] Beć K B, Grabska J, Huck C W. In silico NIR spectroscopy: a review. Molecular fingerprint, interpretation of calibration models, understanding of matrix effects and instrumental difference [J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2022, 279: 121438.
- [9] 赵娟, 沈懋生, 浦育歌, 等. 基于近红外光谱与多品质指标的苹果出库评价模型研究[J]. 农业机械学报, 2023, 54(2): 386 – 395.
- Zhao Juan, Shen Maosheng, Pu Yuge, et al. Out-of-warehouse evaluation and prediction model of apple based on near-infrared spectroscopy combined with multiple quality indexes[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(2): 386 – 395. (in Chinese)
- [10] 汪六三, 黄子良, 王儒敬. 基于近红外光谱和机器学习的大豆种皮裂纹识别研究[J]. 农业机械学报, 2021, 52(6): 361 – 368.
- Wang Liusan, Huang Ziliang, Wang Rujing. Identification of soybean seed coat crack based on near infrared spectroscopy and machine learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(6): 361 – 368. (in Chinese)
- [11] Wu X, Zeng S, Fu H, et al. Determination of corn protein content using near-infrared spectroscopy combined with A – CARS – PLS[J]. Food Chemistry: X, 2023, 18: 100666.
- [12] Zheng R, Jia Y, Ullagaddi C, et al. Optimizing feature selection with gradient boosting machines in PLS regression for predicting moisture and protein in multi-country corn kernels via NIR spectroscopy[J]. Food Chemistry, 2024, 456: 140062.
- [13] 许德龙, 林民, 王玉荣, 等. 基于大语言模型的 NLP 数据增强方法综述[J]. 计算机科学与探索, 2025, 19(6): 1395 – 1413.
- Xu Delong, Lin Min, Wang Yurong, et al. A survey of NLP data augmentation methods based on large language models[J]. Journal of Frontiers of Computer Science and Technology, 2025, 19(6): 1395 – 1413. (in Chinese)
- [14] 卢青, 牟涛涛, 陈少华. 基于拉曼光谱的药品智能分析[J]. 光散射学报, 2023, 35(1): 16 – 23.
- Lu Qing, Mu Taotao, Chen Shaohua. Intelligent analysis of drugs based on Raman spectroscopy[J]. Chinese Journal of Light Scattering, 2023, 35(1): 16 – 23. (in Chinese)
- [15] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C] // Advances in Neural Information Processing Systems 27 (NIPS 2014), 2014: 2672 – 2680.
- [16] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139 – 144.
- [17] Zhang L, Wang Y, Wei Y, et al. Near-infrared hyperspectral imaging technology combined with deep convolutional generative adversarial network to predict oil content of single maize kernel[J]. Food Chemistry, 2022, 370: 131047.
- [18] Bao X, Huang D, Yang B, et al. Combining deep convolutional generative adversarial networks with visible-near infrared hyperspectral reflectance to improve prediction accuracy of anthocyanin content in rice seeds[J]. Food Control, 2025, 174: 111218.
- [19] Huang Y, Chen Z, Liu J. Limited agricultural spectral dataset expansion based on generative adversarial networks[J]. Computers and Electronics in Agriculture, 2023, 215: 108385.
- [20] 杨森, 张新昇, 王振民, 等. 基于近红外光谱和深度学习数据增强的大米品种检测[J]. 农业工程学报, 2023, 39(19): 250 – 257.
- Yang Sen, Zhang Xin'ao, Wang Zhenmin, et al. Rice variety detection based on near-infrared spectroscopy and deep learning data augmentation[J]. Transactions of the CSAE, 2023, 39(19): 250 – 257. (in Chinese)
- [21] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein GANs[C] // Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017: 5769 – 5779.
- [22] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C] // 2017 IEEE International Conference on Computer Vision (ICCV), 2017: 2242 – 2251.
- [23] Feng Q, Guo C, Benitez-Quiroz F, et al. When do GANs replicate? On the choice of dataset size[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 6701 – 6710.
- [24] Zhu F, Wang J, Lv P, et al. Generating labeled samples based on improved cDCGAN for hyperspectral data augmentation: a case study of drought stress identification of strawberry leaves[J]. Computers and Electronics in Agriculture, 2024, 225: 109250.
- [25] Lyu H, 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.
- [26] Qi H, Huang Z, Jin B, et al. SAM – GAN: an improved DCGAN for rice seed viability determination using near-infrared hyperspectral imaging[J]. Computers and Electronics in Agriculture, 2024, 216: 108473.
- [27] Hu H, Mei Y, Zhou Y, et al. Optimizing starch content prediction in kudzu: integrating hyperspectral imaging and deep learning with WGAN – GP[J]. Food Control, 2024, 166: 110762.
- [28] The NIRS data set of corn[EB/OL]. [2024 – 12 – 20]. <https://eigenvector.com/resources/data-sets/>.
- [29] Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium[C] // Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017: 6627 – 6638.
- [30] Ronneberger O, Fischer P, Brox T. U – Net: convolutional networks for biomedical image segmentation[C] // Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015: 234 – 241.
- [31] The NIRS data set of wheat[EB/OL]. [2024 – 12 – 27]. https://www.cnirs.org/content.aspx?page_id=22&club_id=409746&module_id=239453.
- [32] The NIRS data set of grain[EB/OL]. [2025 – 02 – 23]. <https://eigenvector.com/resources/data-sets/>.