玉米的近红外光谱数据

1 材料与方法

1.1 数据集

本数据集包含 80 个玉米样本的近红外光谱数据，波长范围为 1100-2498 nm（间隔 2 nm，共 700 个通道），以及对应的水分、油脂、蛋白质和淀粉含量值。数据存储于 Excel 文件中，其中光谱数据位于第 5 列至最后一列，成分含量位于前 4 列。玉米样本的近红外光谱特征重要性分析如图 1 所示。

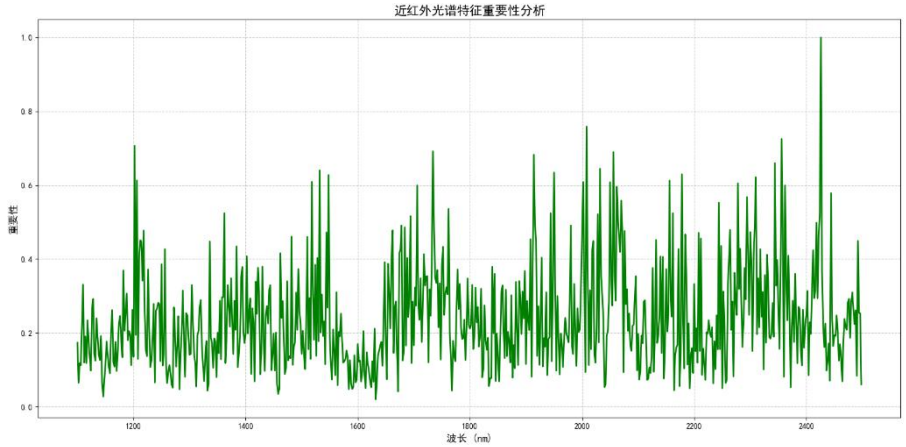

图 1 玉米样本的近红外光谱数据特征重要性分析

1.2 数据预处理

1.2.1 Savitzky-Golay 平滑

使用窗口长度为 7，多项式阶数为 3 的 Savitzky-Golay 滤波器对光谱数据进行平滑处理，以降低噪声。

1.2.2 标准正态变量变换(SNV)

对每个样本的光谱进行标准化处理，消除样本间由于散射效应带来的差异，使不同样本的光谱数据具有可比性。

1.2.3 归一化处理

对光谱和成分数据分别进行 Z-score 标准化，使所有特征处于同一量级，加速模型收敛。

1.2.4 数据集划分

将数据集按 7:3 比例划分为训练集和测试集。

1.3 评估指标

使用均方误差（MSE）和决定系数（R^2）作为评估指标。MSE 用于衡量预测值与真实值之间误差的平方均值，值越小表示预测结果越准确；R^2 用于评估模型的拟合优度，取值范围在 0-1 之间，越接近 1 表示模型对数据的拟合效果越好。

2 模型设计

2.1 模型架构

本研究设计了一个 1D CNN 模型，用于从光谱数据中提取特征并预测成分含量。

2.1.1 卷积层模块

（1）Conv1：1 维卷积层（1→32 通道，核大小 11，padding=5），捕捉局部光谱特征。通过该层，能够提取光谱数据中较小范围内的特征信息。

（2）Conv2：1 维卷积层（32→64 通道，核大小 7，padding=3），进一步提取更高级特征，对 Conv1 提取的特征进行整合和深化。

（3）Conv3：1 维卷积层（64→128 通道，核大小 5，padding=2），结合最大池化（核大小 2），降低特征维度，同时保留重要的特征信息。

2.1.2 全连接层模块

（1）FC1：256 节点，带 BatchNorm 和 Dropout（0.5），防止过拟合，增强模型的泛化能力。

（2）FC2：128 节点，同样带 BatchNorm 和 Dropout（0.5），进一步对特征进行处理和筛选。

（3）FC3：4 节点（对应 4 种成分），输出最终的预测值。

2.1.3 激活与正则化

全网络使用 ReLU 激活函数，缓解梯度消失问题，使网络能够更好地学习数据特征；各层间添加 BatchNorm 和 Dropout，增强模型的泛化能力，避免模型在训练过程中出现过拟合现象。

2.2 模型结构图

CNN 模型结构图如图 2 所示。

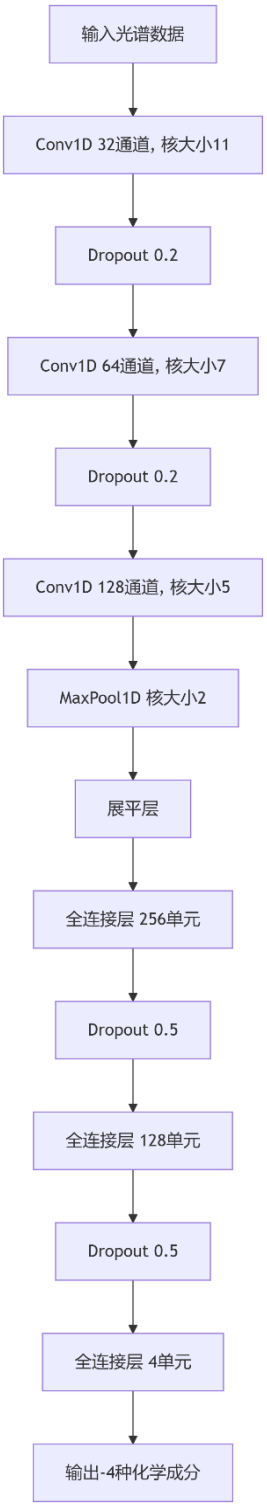

图 2 CNN 模型结构图

3 模型训练

3.1 训练配置

（1）损失函数：采用均方误差（MSE），作为回归任务常用的损失函数，能

够有效衡量预测值与真实值之间的差异。

（2）优化器：选择 Adam 优化器，设置学习率为 0.0005，weight_decay 为 1e-5。Adam 优化器结合了 Adagrad 和 RMSProp 的优点，能够自适应地调整学习率，在训练过程中具有较好的收敛速度和稳定性。

（3）学习率调度器：使用 ReduceLROnPlateau，当训练损失在 10 个 epoch 内没有下降时，将学习率降低为原来的 0.5 倍，有助于模型在训练后期更好地收敛。

（4）训练轮次：设定训练轮数为 200，并采用早停机制，当连续 20 个 epoch 损失没有下降时停止训练，防止模型过拟合，同时提高训练效率。

3.2 训练过程

在训练过程中，使用 PyTorch 构建数据加载器，将训练集以 batch_size=8 的大小输入模型进行训练。每一轮训练结束后，记录训练损失，并根据学习率调度器调整学习率。通过早停机制，在第 136 轮时触发停止条件，此时模型在训练集上的损失趋于稳定，不再有明显下降，如图 3 所示。

图 3 损失函数

3.3 训练结果

图 4 展示了模型训练过程中的损失曲线。从曲线中可以看出，在训练初期，损失值快速下降，这是因为模型在初始阶段能够迅速学习到数据中的一些简单特征。随着训练的进行，损失下降速度逐渐变缓，在接近早停轮次时，损失基本稳定在一个较低的水平，说明模型已经较好地拟合了训练数据。

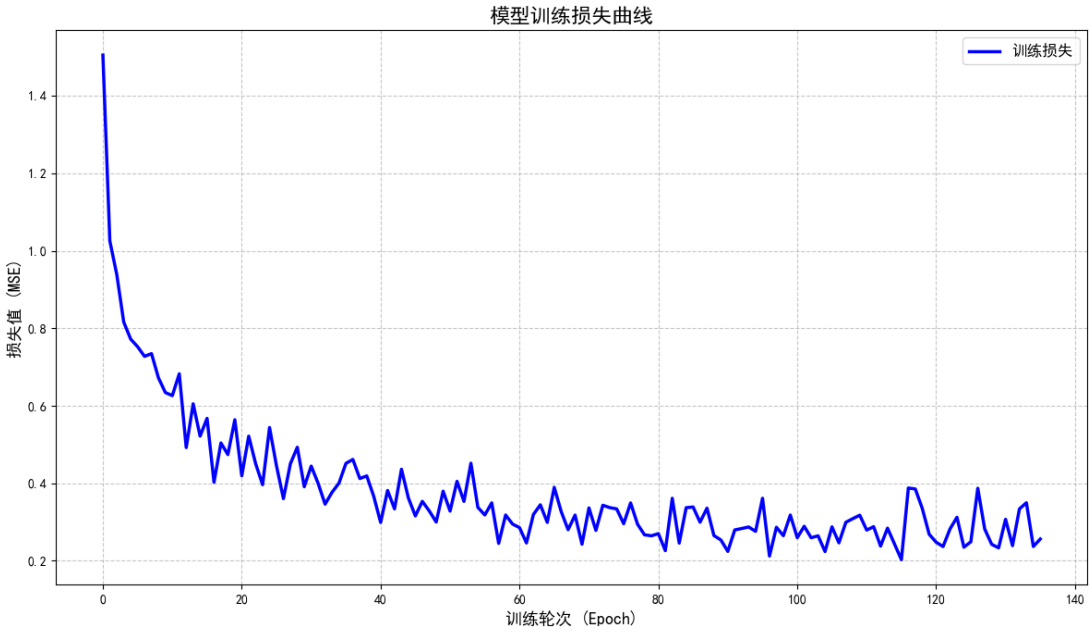

图 4 损失曲线

4 模型评估

在测试集上，模型的评估结果如表 1 所示。

表 1 预测性能指标

成分	MSE	R^2
水分	0.039	0.698
油脂	0.009	0.576
蛋白质	0.039	0.836
淀粉	0.222	0.727

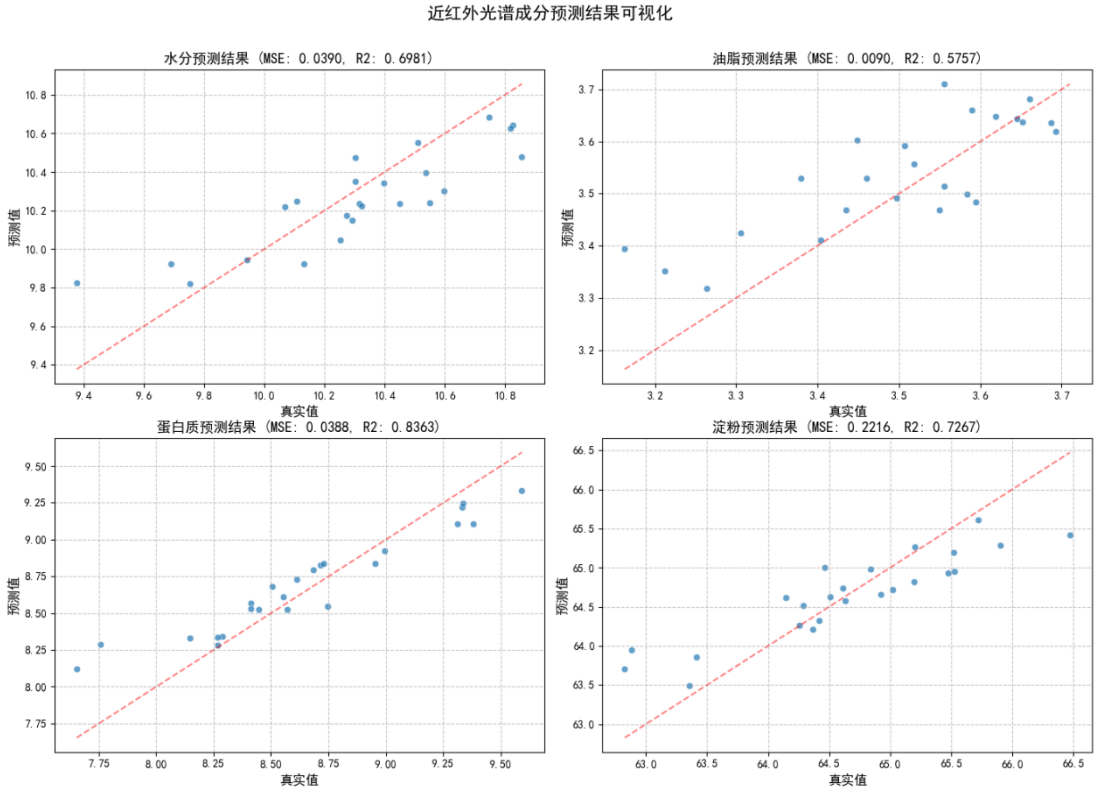

图 5 测试集上各成分预测值与真实值的对比散点图

图 5 展示了测试集上各成分预测值与真实值的对比散点图。从图中可以看出，大部分数据点集中分布在对角线附近，说明模型在一定程度上能够准确预测玉米各成分的含量，但仍有部分数据点偏离对角线，尤其是在油脂含量的预测上，表明模型在某些成分的预测上还存在一定的误差。

5 模型优化

5.1 优化方向

5.1.1 数据增强

由于样本数量有限，可通过数据增强技术扩充数据集。例如，对光谱数据进行平移、缩放、添加噪声等操作，生成更多的训练样本，提高模型的泛化能力。

5.1.2 网络结构调整

尝试调整卷积层和全连接层的节点数量、层数，或者更换不同类型的神经网络架构，如使用 Transformer 架构，探索更适合近红外光谱数据的模型结构。

5.1.3 超参数调优

进一步优化超参数，如学习率、batch_size、Dropout 比例等。可以采用网格

搜索、随机搜索或贝叶斯优化等方法，找到最优的超参数组合，提高模型性能。

5.2 优化实验

在后续研究中，计划开展优化实验。首先，对数据进行增强处理，生成两倍数量的训练样本，重新训练模型并评估性能；其次，尝试将模型中的部分卷积层替换为 Transformer 块，构建 CNN-Transformer 混合模型；最后，使用贝叶斯优化对超参数进行全面调优，对比不同优化方案下模型在测试集上的表现，确定最佳的优化策略。

6 模型优缺点分析

6.1 优点

基于 CNN 的模型能够自动提取光谱数据的特征，相比传统的化学计量学方法，减少了人工特征工程的工作量，提高了分析效率。通过添加 BatchNorm 和 Dropout 等正则化手段，模型在一定程度上避免了过拟合，具有较好的泛化能力，能够适应不同样本的光谱数据。

6.2 不足

模型的性能受限于数据集的大小和质量，在样本数量较少的情况下，模型的泛化能力可能会受到影响。深度学习模型的训练需要较高的计算资源，包括 GPU 等硬件设备，限制了模型在一些计算资源有限环境中的应用。