Optoelectronics 光电子, 2025, 15(4), 83-92
Published Online December 2025 in Hans. https://www.hanspub.org/journal/oe
https://doi.org/10.12677/oe.2025.154009

Hans汉斯

基于近红外光谱与机器学习的圈枝/驳枝陈皮鉴别

郑惠萍，郑成勇*

五邑大学数学与计算科学学院，广东 江门

收稿日期：2025年9月19日；录用日期：2025年11月26日；发布日期：2025年12月4日

摘　要

圈枝和驳枝是园艺生产中两种重要的育苗方法。这两种方法生产的陈皮在市场上表现出明显的差异，准确区分它们至关重要。然而，传统的识别方法存在主观性强、效率低、准确性差等问题，难以满足大规模识别的需求。为了解决这些问题，本研究比较了三种传统机器学习算法(支持向量机(SVM)、随机森林(RF)和K近邻(KNN))和五种前沿的时间序列深度学习算法(TSMixer、MSDMixer、TimesNet、PatchMixer和TimeMixer)在基于近红外光谱数据区分圈枝陈皮和驳枝陈皮方面的性能。同时，本研究选取了最小最大归一化等多种数据预处理方法，深入探讨了不同预处理方法对各种算法性能的影响。实验结果表明，传统算法适用于对计算资源和时间要求较低的场景，而深度学习算法在数据量充足、计算资源丰富的条件下可以实现更准确的识别。此外，不同的数据预处理方法对算法的性能有显著影响。深度学习算法(如PatchMixer和TimeMixer)和传统算法(如SVM和KNN)在特定的预处理下可以达到或接近100%的平均准确率。本研究不仅为近红外光谱在植物栽培方法识别中的应用提供了实证支持，也为实际应用中的算法选择和数据预处理方法提供了重要参考。

关键词

近红外光谱，圈枝，驳枝，陈皮鉴别，SVM，深度学习，TimeMixer

Identification of Chenpi from Cutting and Grafting via Near-Infrared Spectroscopy and Machine Learning

Huiping Zheng, Chengyong Zheng*

School of Mathematics and Computational Science, Wuyi University, Jiangmen Guangdong

Received: September 19, 2025; accepted: November 26, 2025; published: December 4, 2025

*通讯作者。

文章引用: 郑惠萍, 郑成勇. 基于近红外光谱与机器学习的圈枝/驳枝陈皮鉴别[J]. 光电子, 2025, 15(4): 83-92.
DOI: 10.12677/oe.2025.154009

Abstract

Cutting and grafting are two important seedling cultivation methods in horticultural production. The dried tangerine peels (Chenpi) produced by these two methods show obvious differences in the market, and it is crucial to accurately distinguish between them. However, traditional identification methods have problems such as strong subjectivity, low efficiency, and poor accuracy, making it difficult to meet the needs of large-scale identification. To address these issues, this study compared the performance of three traditional machine learning algorithms (Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN)) and five state-of-the-art time series deep learning algorithms (TSMixer, MSDMixer, TimesNet, PatchMixer, and TimeMixer) in distinguishing between cutting and grafting based on near-infrared spectroscopy data. At the same time, this study selected various data preprocessing methods such as normalization and conducted an in-depth exploration of the impact of different preprocessing methods on the performance of various algorithms. The experimental results show that traditional algorithms are suitable for scenarios with low requirements for computing resources and time, while deep learning algorithms can achieve more accurate identification under the conditions of sufficient data volume and abundant computing resources. In addition, different data preprocessing methods have a significant impact on the performance of the algorithms. Deep learning algorithms (such as PatchMixer and TimeMixer) and traditional algorithms (such as SVM and KNN) can reach or approach an average accuracy of 100% under specific preprocessing. This study not only provides empirical support for the application of near-infrared spectroscopy in the identification of plant cultivation methods but also offers important references for algorithm selection and data preprocessing methods in practical applications.

Keywords

Near-Infrared Spectroscopy, Cutting, Grafting, Chenpi Identification, SVM, Deep Learning, TimeMixer

1. 引言

在园艺生产中，圈枝和驳枝是两种关键的育苗方式，对植物生长、果实品质及抗病虫害能力影响显著。圈枝柑通过环切母树枝条诱导生根移植，能最大程度保留母树基因，果实品质优，制成的陈皮在高端市场和药用领域备受青睐，但产量较低。驳枝柑则利用嫁接技术，以抗病虫害能力强的红柠檬为砧木，嫁接茶枝柑枝条，其产量高、抗病虫害能力强，适合大规模种植，市面上的新会陈皮多为驳枝陈皮。

精准鉴别圈枝陈皮和驳枝陈皮意义重大，一方面，能够满足不同市场和消费者的多样化需求，提升种植效益，推动园艺产业的可持续发展；另一方面，可以有效杜绝市场上产品混淆现象，保障消费者权益，维护公平有序的市场竞争环境。文献[1]采用气相色谱－质谱法比较了圈枝陈皮和驳枝陈皮的综合品质，结果表明五种挥发性化合物可以作为潜在的化学标记物，用于区分圈枝陈皮和驳枝陈皮。然而，气相色谱－质谱法耗时且具有破坏性，需要复杂的样品制备程序。此外，基于形态特征和经验的其他传统鉴定方法也存在主观性强、效率低、准确性差等问题，难以满足大规模鉴定的需求。

近年来，近红外光谱(NIRS)技术因其快速、高效和无损检测的优势，在材料鉴定领域得到了快速发展。该技术的核心原理是利用不同物质分子在近红外光带中的特定吸收特性，并通过对近红外光与物质

之间相互作用的分析实现精准分析[2]。这种技术已在农业、食品等领域得到了广泛应用[3]-[5]。在新会陈皮的鉴定中，它也可以快速准确地识别年龄和产地的差异[6]-[8]，为圈枝和驳枝的鉴定提供了新的思路。

鉴于近红外光谱数据具有时间序列特征，本研究系统地选取了五种先进的深度学习架构，包括TSMixer、MSDMixer、TimesNet、PatchMixer 和 TimeMixer。这些架构基于时间序列深度学习算法在处理时间序列依赖建模和特征提取方面的显著优势。研究还与经典的机器学习算法(SVM、RF、KNN)进行了比较，以构建一个稳健高度的圈枝和驳枝鉴定模型。此外，研究重点关注了数据预处理的关键作用，并将系统地探讨最小最大归一化、Z 标准化和均值中心化等预处理策略对模型性能的调节效应。

通过构建传统算法和深度学习算法的对照实验体系，并结合对标准化策略的精细研究，本研究旨在建立一个二维评估框架：从算法架构层面分析时间序列特征提取机制，并从数据表示层面优化输入分布特征。本研究旨在突破现有近红外光谱分析的技术瓶颈，为园艺产品的智能鉴定提供具有理论创新性和实际有效性的技术解决方案。

2. 材料与方法

2.1. 近红外光谱仪

本研究采用了一款 NIR-R210 手持式反射近红外光谱仪(如图 1 所示)，其核心参数如下：光谱响应范围为 900~1700 nm(覆盖了特征近红外波段)，光学分辨率为 10 nm (典型值)/14 nm (最大值)，波长精度为 ±1 nm (典型值)/±2 nm (最大值)，信噪比为 6000:1。

Figure 1. Near-infrared spectrometer NIR-R210
图 1. 近红外光谱仪 NIR-R210

2.2. 光谱数据采集

在本研究中，采集了圈枝和驳枝来源的陈皮样本，其陈化时间均在三年以内。具体而言，共采集了 9 批次圈枝陈皮(产生 54 个光谱数据点)和 12 批次驳枝陈皮(产生 72 个光谱数据点)。所有陈皮样本均来自不同的种植户，以确保样本空间涵盖三个维度：加工差异、年份差异和产地环境差异。这种方法旨在增强模型泛化能力验证的可靠性。

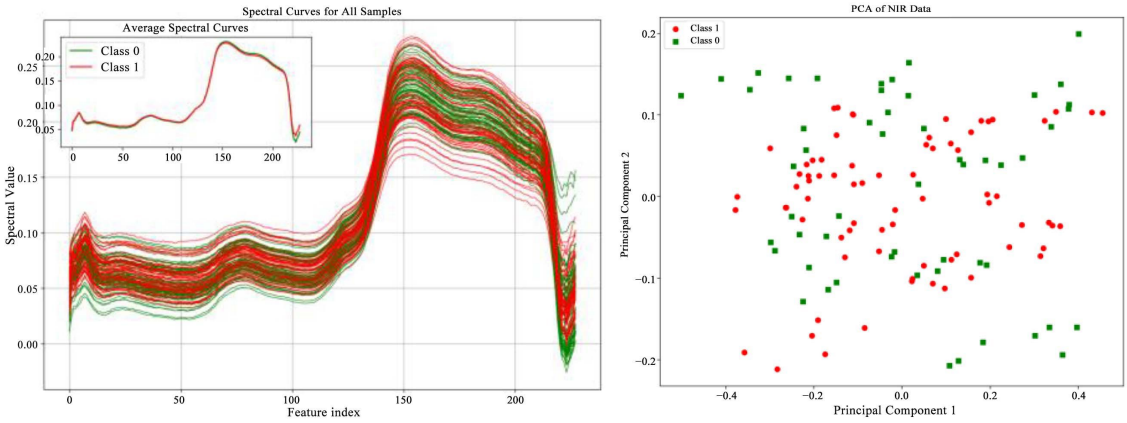
陈皮的近红外光谱数据采集方法如下：

(1) 样品准备：江门新会陈皮样本按种植方式和批次保存。对于每个批次，随机抽取 2~3 片样本，粉碎后通过 100 目筛筛选。精确称量 100~200 mg 的粉末，采用压片法制备直径为 13 mm、厚度为 1~4 mm 的圆形片剂。每个批次平行制备 6 个样品。

(2) 光谱采集：在实验前，使用标准白板对 NIR-R210 手持近红外光谱仪进行校准。将样品放置在仪器的检测窗口处，确保样品表面与检测窗口紧密接触。在消除环境光干扰后进行光谱扫描。

(3) 数据标注：在获取光谱数据后，系统记录了样本对应的种植方式(圈枝或驳枝)和批次代码等元数据信息，建立了一个完整的可追溯性鉴定系统。

图 2 展示了采集的圈枝陈皮和驳枝陈皮的近红外光谱数据特征。左图显示了所有样本的原始光谱曲线(附图显示了圈枝和驳枝种植方式的平均光谱对比)，右图呈现了使用主成分分析(PCA)降维后样本在 PC1 与 PC2 空间中的分布情况。结果表明，在 900~1700 nm 范围内，两种种植方式的原始光谱并未表现出显著差异，且其 PCA 投影点集在二维空间中高度重叠。这一现象揭示了传统光谱分析方法在区分陈皮种植方式方面的能力有限，需要进一步构建具有高维特征提取能力的机器学习模型以实现有效区分。



(a) 两类的原始光谱曲线及其均值对比　　(b) PCA 降维后的散点图

Figure 2. Near-infrared spectral characteristics of Chenpi from cutting and grafting
图 2. 圈枝和驳枝陈皮的近红外光谱特征

2.3. 数据预处理

设原始光谱数据集为 $X \in R^{n\times p}$，其中 n 表示样本数量，p 表示波长点数量。本研究系统地评估了以下六种数据预处理方法：

(1) 最小最大归一化(Min-Max Scaling, MM)：

所有特征值被缩放到[0, 1]区间，适用于具有明确边界的特征数据。其计算公式为：

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{1}$$

(2) Z 标准化 (Z-Score Normalization, ZS)：

将数据转换为均值为 0、标准差为 1 的分布，有助于提高模型的收敛速度。其计算公式为：

$$X_{std} = \frac{X - \mu}{\sigma} \tag{2}$$

其中，μ 表示特征的均值，σ 表示标准差。

(3) 均值中心化(Mean Centering, MC)：

通过减去均值来消除数据的偏移，适用于需要保持数据分布不变的场景。其公式为：

$$X_{mean} = X - \mu \tag{3}$$

(4) L2 范数归一化 (L2-norm Normalization, L2N)：

通过调整特征向量的尺度进行特征缩放。L2 范数是向量元素平方和的平方根，定义为：

$$x_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{4}$$

(5) 鲁棒标准化(Robust Standardization, RS)：

使用中位数和四分位距计算数据的中心和尺度，以减少异常值的影响。其公式为：

$$X_{norm} = \frac{X - \text{Median}(X)}{\text{IQR}(X)} \tag{5}$$

其中 Median 表示中位数，IQR 表示四分位距。

(6) 幂变换标准化(Power Transformation Standardization, PTS)：

通过调整数据的幂次来改善数据的分布特性，使其更接近正态分布。本研究采用 Yeo-Johnson 变换[9]，该方法允许数据包含负值，适用于本研究。Yeo-Johnson 变换的公式为：

$$Y = \begin{cases} \left(\dfrac{X^{\lambda}-1}{\lambda}\right), \lambda \neq 0 \\ \log(X+1), \lambda = 0 \end{cases}, X \geq 0 \tag{6}$$

或

$$Y = \begin{cases} \left(\dfrac{(1-X)^{2-\lambda}-1}{2-\lambda}\right), \lambda \neq 2 \\ -\log(1-X), \lambda = 2 \end{cases}, X \leq 0 \tag{7}$$

3. 算法

3.1. 传统算法概述

SVM 是一种监督学习算法，通过最大化类别间隔来构建最优超平面以实现分类。它可以通过使用核函数(如径向基函数(RBF))来处理非线性问题。RBF 核函数能够将数据映射到高维空间，从而在高维空间中找到一个线性可分的超平面，进而解决原始空间中的非线性分类问题。

随机森林是一种基于 Bootstrap 抽样的集成学习方法，通过构建多个决策树来提高分类的准确性和稳定性[10]。在节点分裂过程中，随机选择特征，并通过投票机制进行决策，从而增强模型的泛化能力，有效减少过拟合现象。随机森林能够处理高维数据，并且对数据中的噪声和异常值具有较强的鲁棒性。

KNN 是一种基于实例的学习算法，通过计算测试样本与训练样本之间的局部距离(如欧几里得距离或曼哈顿距离)来选择 K 个最近邻样本，并根据这些最近邻样本的投票结果进行决策[11]。KNN 算法简单且具有很强的适应性，但其性能高度依赖于 K 值的选择和距离度量方法。较小的 K 值可能导致模型对噪声数据过于敏感，而较大的 K 值可能会使模型过于平滑，从而降低分类精度。

3.2. 深度学习算法概述

TSMixer 是一种具有时空混合机制的时间序列分析模型[12]。它通过线性层和 ReLU 激活函数在时间和特征维度上进行双重混合，能够有效捕捉长期和短期依赖关系以及复杂的时间序列模式。该模型特别适用于高维数据。

MSDMixer 提出了一种多尺度卷积框架，通过并行使用不同大小的卷积核来提取短期、中期和长期特征，并将这些特征融合以实现多时间粒度的特征建模[13]。这种方法非常适合分析复杂的时间序列结构。

TimesNet 结合了时间卷积网络(Temporal Convolutional Networks, TCN)和周期卷积模块(Periodic Convolutional Modules, PCN)，分别用于捕捉短期依赖关系和长期周期性特征[14]。该模型在处理具有显著周

期性或趋势成分的时间序列数据方面表现出色。

PatchMixer 借鉴了图像分块技术，将时间序列划分为固定长度的块[15]。通过线性投影和激活函数，它促进了局部与全局特征之间的交互，从而增强了对复杂模式的解析能力。

TimeMixer 构建了一个时间和空间的双重混合架构[16]。它首先在时间维度上进行特征重组，然后在通道维度上进行信息融合，形成了一个分层的时间依赖关系建模机制。该模型在各种场景中都表现出广泛的适用性。

3.3. 模型评估与实验配置

在本研究中，采用标准化分类准确率(Accuracy)作为核心评估指标，其定义为正确预测样本数与总样本数之比。计算公式如下：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

其中，TP (真正例)和 TN (真负例)分别表示正确预测的正类和负类样本的数量，而 FP (假正例)和 FN (假负例)则表示被错误分类的样本数量。该指标通过混淆矩阵计算得出，并全面反映了模型的整体分类性能。

所有模型均通过 10 次独立重复实验进行性能评估。每次实验中，样本集按 8:2 的比例随机划分为训练集与测试集，划分基于样本个体进行。模型在训练集上采用 5 折交叉验证结合网格搜索进行超参数优化，并在独立测试集上计算分类准确率。模型性能以 10 次实验的准确率的均值±标准差表示，以确保结果的统计显著性。根据算法特性采用了不同的配置：

(1) 传统机器学习算法：实施了 5 折交叉验证网格搜索，以遍历预定义的超参数空间(例如 SVM 的核函数类型包括线性核、RBF 核和多项式核；RF 的树深度 $\in [10, 20, 30, \text{None}]$，等)。优化目标是验证集的准确率，最终选择具有最佳泛化性能的参数组合。

(2) 深度学习算法：设置最大训练轮数为 100 轮，并采用 Adam 优化器(base learning rate $\eta = 0.001$)。根据硬件资源动态调整批量大小(batch_size $\in [32, 128]$)。为了控制变量，重点关注时间混合机制的有效性验证，网络架构保持默认配置。所有实验均在配备 32 GB 内存和 NVIDIA RTX A4000 GPU 的 Windows 11 操作系统上进行。

4. 结果与分析

4.1. 传统机器学习算法结果

在本节中，我们全面分析了三种传统机器学习算法——SVM、RF 和 KNN 在不同预处理方法下的性能表现。研究旨在准确评估每种算法在识别圈枝陈皮和驳枝陈皮方面的有效性，并深入探讨预处理方法对算法性能的影响机制。

从表 1 的数据可以看出，不同的预处理方法对每种算法的性能产生了显著不同的影响。对于 SVM 算法，基于原始数据进行分类时的平均准确率为 0.89 ± 0.06。经过 MC 预处理后，准确率显著提高至 0.98

Table 1. Average accuracy of traditional algorithms under different preprocessing methods
表 1. 不同预处理方法下传统算法的平均准确率

Classifier	Raw Data	MM	ZS	MC	L_2N	RS	PTS
SVM	0.89 ± 0.06	0.95 ± 0.05	0.97 ± 0.04	**0.98 ± 0.03**	0.86 ± 0.06	0.96 ± 0.05	**0.98 ± 0.03**
RF	0.78 ± 0.09	0.83 ± 0.10	0.80 ± 0.11	0.79 ± 0.07	**0.90 ± 0.09**	0.82 ± 0.10	0.81 ± 0.10
KNN	**0.94 ± 0.05**	0.92 ± 0.05	0.93 ± 0.06	**0.94 ± 0.05**	0.89 ± 0.06	**0.94 ± 0.05**	0.93 ± 0.06

±0.03，类似地，在经过 PTS 预处理后，准确率也达到 0.98±0.03。MM 预处理后的准确率为 0.95±0.05，而 RS 预处理后的准确率为 0.96±0.05。这清楚地表明，MC、PTS、MM 和 RS 等预处理方法都能显著提升 SVM 的分类性能，其中 MC 和 PTS 的效果最为显著。

RF 算法的性能在不同的预处理方法下有所不同。在基于原始数据进行训练时，准确率为 0.78±0.09。在各种预处理方法中，L2N 最为有效，将准确率提高到 0.90±0.09。经过 MM、ZS、RS 和 PTS 预处理后的准确率分别为 0.83 ± 0.10、0.80 ± 0.11、0.82 ± 0.10 和 0.81 ± 0.10。这表明 L2N 预处理方法在提升 RF 算法识别圈枝陈皮和驳枝陈皮的能力方面最为有效。

对于 KNN 算法，基于原始数据训练得到的准确率为 0.94±0.05。经过 MM 预处理后的准确率为 0.92 ± 0.05，经过 ZS 预处理后的准确率为 0.93 ± 0.06，与原始数据的性能相当。此外，在不同的预处理方法中，KNN 算法的性能波动相对较小。尽管如此，仍可以看出 MM 和 ZS 预处理在一定程度上维持了 KNN 算法的性能。

4.2. 深度学习算法结果

在本节中，我们系统地分析了五种深度学习算法——TSMixer、MSDMixer、TimesNet、PatchMixer 和 TimeMixer 在不同预处理方法下的性能，重点关注每种算法的平均准确率的动态变化。

从表 2 呈现的数据可以看出，当处理原始数据时，不同深度学习算法的平均准确率表现出显著差异。使用原始数据时，TSMixer 和 MSDMixer 的平均准确率仅为 0.54 和 0.55，相对较低。这一现象表明，原始数据中可能存在噪声干扰和数据分布不均匀等问题，阻碍了这些模型有效提取圈枝陈皮和驳枝陈皮的关键特征。相比之下，使用原始数据时 TimesNet 的平均准确率达到 0.63，略高于 TSMixer 和 MSDMixer，表明该模型在处理原始数据方面具有一定优势，但仍存在较大的优化空间。PatchMixer 和 TimeMixer 在处理原始数据时表现相对较好，平均准确率分别高达 0.86 和 0.93，充分展示了这两种算法在捕捉原始数据特征方面的强大能力。

Table 2. Average accuracy of deep learning algorithms under different preprocessing methods
表 2. 不同预处理方法下深度学习算法的平均准确率

Model	Raw Data	MM	ZS	MC	L_2N	RS	PTS
TSMixer	0.54 ± 0.11	**0.96 ± 0.04**	0.88 ± 0.10	0.95 ± 0.05	0.60 ± 0.13	0.94 ± 0.11	0.91 ± 0.12
MSDMixer	0.55 ± 0.12	0.92 ± 0.11	0.94 ± 0.05	**0.90 ± 0.06**	0.52 ± 0.09	0.91 ± 0.08	0.93 ± 0.05
TimesNet	0.63 ± 0.10	0.92 ± 0.05	**0.93 ± 0.06**	0.69 ± 0.10	0.48 ± 0.07	0.92 ± 0.09	0.92 ± 0.07
PatchMixer	0.86 ± 0.09	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.01**	0.92 ± 0.08	**1.00 ± 0.01**	**1.00 ± 0.01**
TimeMixer	0.93 ± 0.07	0.99 ± 0.03	**1.00 ± 0.00**	0.99 ± 0.02	0.80 ± 0.14	**1.00 ± 0.01**	0.95 ± 0.05

在预处理方法对模型性能的影响方面，MM、ZS、RS 和 PTS 等预处理技术显著提升了大多数模型的性能。以 TSMixer 为例，经过 MM 预处理后，其平均准确率从原始数据的 0.54 显著跃升至 0.96。MSDMixer 在经过 ZS 预处理后，平均准确率达到了 0.94。在上述预处理方法下，PatchMixer 和 TimeMixer 的平均准确率几乎接近或直接达到了 1.00，展现出卓越的分类性能。然而，L2 范数归一化预处理方法对某些模型的效果较差。例如，经过 L2 范数归一化预处理后，TimesNet 的平均准确率不仅未能提升，反而比原始数据有所下降。

具体而言，在 MM、ZS、MC、RS 和 PTS 等多种预处理方法的影响下，TSMixer 的平均准确率实现了显著增长，其中经过 MM 预处理后的最高平均准确率达到了 0.96。然而，L2 范数归一化预处理方法对 TSMixer 的提升效果非常有限，仅将平均准确率提高到 0.60。在接受了 MM、ZS、RS 和 PTS 预处理后，

MSDMixer 的平均准确率超过了 0.90，其中经过 ZS 预处理后的最高准确率达到了 0.94。相比之下，L2 范数归一化预处理方法对 MSDMixer 性能的提升几乎毫无效果，甚至导致准确率下降。在接受了 MM、ZS、RS 和 PTS 预处理后，TimesNet 的平均准确率提高到了 0.92 以上。然而，经过 L2 范数归一化预处理后，其平均准确率仅为 0.48，处于较低水平。在各种预处理方法下，PatchMixer 和 TimeMixer 均展现出了极高的平均准确率。经过 MM、ZS 和 RS 预处理后，PatchMixer 的平均准确率达到了或接近 1.00。经过 ZS 预处理后，TimeMixer 的平均准确率达到了 1.00。在其他预处理方法(如 MM 和 RS 预处理)下，其准确率也分别达到了 0.99 和 1.00。

综上所述，不同的预处理方法对深度学习算法的性能有着显著不同的影响。MM、ZS、RS 和 PTS 等预处理方法显著提升了大多数模型的性能，而 L2 范数归一化预处理方法对某些模型的效果却不尽如人意。因此，在实际应用场景中，优先采用高效的预处理方法(如 MM)来处理 TSMixer 和 MSDMixer 等模型，将有助于提升模型的整体性能。

4.3. 传统算法与深度学习算法的比较分析

在识别圈枝陈皮和驳枝陈皮的任务中，传统算法和深度学习算法展现出不同的特点和优势。本节通过仔细比较这两种算法在最佳预处理条件下的核心性能指标(如平均准确率和标准差)，深入分析它们的性能差异，旨在为该领域的算法选择和优化提供有力依据。

从表 3 的数据可以看出，在各自的最佳预处理方法下，PatchMixer 和 TimeMixer 均达到了 1.00 的平均准确率。这一卓越表现充分表明，这两种深度学习算法能够准确捕捉圈枝陈皮和驳枝陈皮数据中的关键特征，从而实现极高精度的分类，并在同类算法中脱颖而出。在传统算法中，SVM 在采用 MC 和 PTS 两种预处理方法后，也成功达到了 0.98 的平均准确率。这一结果与某些尖端深度学习算法的准确率相当，有力地证明了 SVM 在经过适当预处理后能够有效地、准确地对圈枝陈皮和驳枝陈皮进行分类。

值得注意的是，KNN 算法即使在没有标准化预处理的情况下，仍然达到了 0.97 的平均准确率。这一结果突显了 KNN 算法对原始数据的强大适应性，能够在没有复杂繁琐的预处理程序的情况下，基于原始数据实现高准确率，这在一定程度上反映了该算法的简单性和高效性。

不同算法对预处理方法的依赖程度存在显著差异。以深度学习算法为例，TSMixer 在采用 RS 预处理方法后，平均准确率提升至 0.95。这一数据直观地反映了合适的预处理方法对其性能提升具有显著的促进作用。MSDMixer 在经过 ZS 预处理后，平均准确率达到了 0.94，表明该算法对 ZS 预处理方法高度敏感，并在这种预处理条件下表现出良好的稳定性。相比之下，在传统算法中，SVM 高度依赖特定的预处理方法(MC/PTS)才能实现高准确率，这清楚地表明 SVM 对预处理的依赖程度较高。与此形成鲜明对比的是，KNN 在没有标准化预处理的情况下表现良好，明确表明其对预处理的依赖程度较低。RF 在 L2 范数归一化预处理下能够达到较高的准确率，这表明 RF 对特定预处理方法也有一定要求，但其依赖程度不如 SVM 显著。

根据对各种数据的综合分析，深度学习算法 PatchMixer 和 TimeMixer 在准确率方面表现出色，在识别圈枝陈皮和驳枝陈皮的任务中展现出无与伦比的强大优势。在传统算法中，SVM 和 KNN 在适当的条件下也能实现高准确率。更重要的是，与深度学习算法相比，SVM 和 KNN 具有更低的模型复杂性，对计算资源的需求更为精简，并且显著缩短了训练时间。例如，在经过精心选择的预处理后，SVM 不仅可以确保高准确率，还具有相对高效的计算性能，能够在实际应用场景中实现更快的操作。由于其对原始数据的良好适应性，KNN 在实际部署和使用过程中可以更方便、快速地投入使用，为圈枝陈皮和驳枝陈皮的识别任务提供高效的解决方案。

Table 3. Comparison of optimal accuracy of traditional and deep learning algorithms in identifying Chenpi from cutting and grafting
表 3. 传统算法与深度学习算法在圈枝陈皮和驳枝陈皮识别中的最优准确率比较

Algorithm	Optimal Preprocessing Method	Average Accuracy
TimeMixer	MM	**1.00 ± 0.00**
PatchMixer	MM/ZS	**1.00 ± 0.00**
SVM	MC/PTS	0.98 ± 0.03
TSMixer	MM	0.96 ± 0.04
MSDMixer	ZS	0.94 ± 0.05
KNN	—/MC/RS	0.94 ± 0.05
TimesNet	ZS	0.93 ± 0.06
RF	L_2N/ MC/ RS	0.90 ± 0.09

5. 结论

本研究深入探讨了近红外光谱技术在识别圈枝陈皮和驳枝陈皮中的应用，对传统算法和深度学习算法的性能进行了全面且详细的比较，获得了具有重要科学价值和实践意义的一系列发现。

在数据预处理阶段，研究揭示了不同标准化方法对不同算法性能的影响存在明显差异。在传统算法领域，均值中心化和幂变换标准化等预处理技术有效地将 SVM 的平均准确率提升至 0.98，充分展示了它们在优化 SVM 算法性能方面的积极作用。相比之下，KNN 算法即使没有任何标准化处理，仍能凭借其对原始数据的卓越适应性达到 0.97 的高准确率。这一现象深刻地说明了不同传统算法在预处理需求和适应性方面的显著差异。

在深度学习算法框架中，最小最大归一化和 Z 标准化等预处理方法显著提升了某些模型的性能。具体而言，在特定的预处理条件下，PatchMixer 和 TimeMixer 的平均准确率成功突破并稳定在 1.00，凸显了这些预处理方法在增强深度学习模型性能方面的重要作用。

从算法性能比较的角度来看，传统算法和深度学习算法各有优势且相互补充。传统算法由于其较低的计算复杂性、较短的训练时间和便捷的部署与应用过程，在实际操作中具有显著优势。以 SVM 和 KNN 为例，在经过适当的预处理步骤或直接使用原始数据后，它们能够高效且快速地完成圈枝陈皮和驳枝陈皮的识别任务，同时确保一定的准确率，满足对计算资源和时间敏感的应用场景的需求。

另一方面，深度学习算法的核心优势在于其强大的特征学习能力，能够自动且深入地挖掘数据中隐藏的复杂模式和关键特征。PatchMixer 和 TimeMixer 在圈枝陈皮和驳枝陈皮识别任务中展现出的卓越分类能力，不仅为该领域提供了高精度的技术解决方案，也突显了深度学习算法在处理复杂数据特征方面的巨大潜力。

本研究的发现为圈枝陈皮和驳枝陈皮的识别提供了丰富多样的技术路径和策略选择。在计算资源有限、时间紧迫且数据规模较小的实际应用场景中，传统算法以其高效性和便捷性无疑是更理想的选择。而在追求极致准确率、数据丰富且计算资源充足的情境中，深度学习算法能够充分发挥其优势，实现对圈枝陈皮和驳枝陈皮更准确、更可靠的识别，为相关产业的高质量发展提供有力的技术保障。

总之，本研究不仅为近红外光谱技术在圈枝陈皮和驳枝陈皮识别中的应用奠定了坚实的理论和实践基础，还为后续相关研究和技术革新提供了宝贵的参考和启示，需要指出的是，尽管部分模型在本研究中达到了接近或达到 100%的准确率，但该性能是在特定仪器(NIR-R210)与受控实验室环境下取得的，且样本在批次分布与陈化时间上仍存在局限，因此其在真实工业场景中的泛化能力仍需进一步验证。未来工作将探索域自适应、多中心协同建模等策略，以提升模型在跨设备、跨批次条件下的鲁棒性，并通过构建独立外部验证集推动技术向实际应用转化，有望实现近红外光谱在陈皮品质鉴别中的标准化与智能化发展。

基金项目

本研究得到了五邑大学港澳联合研发基金(项目编号：2022WGALH16)的支持。

参考文献

[1] Tan, E., Li, F., Lin, X., Ma, S., Zhang, G., Zhou, H., *et al*. (2022) Comparative Study on Comprehensive Quality of Xinhui Chenpi by Two Main Plant Propagation Techniques. *Food Science & Nutrition*, **11**, 1104-1112. https://doi.org/10.1002/fsn3.3148

[2] Pasquini, C. (2018) Near Infrared Spectroscopy: A Mature Analytical Technique with New Perspectives—A Review. *Analytica Chimica Acta*, **1026**, 8-36. https://doi.org/10.1016/j.aca.2018.04.004

[3] Cozzolino, D. (2016) Near Infrared Spectroscopy and Food Authenticity. In: Espiñeira, M. and Santaclara, F.J., Eds., *Advances in Food Traceability Techniques and Technologies*, Elsevier, 119-136. https://doi.org/10.1016/b978-0-08-100310-7.00007-7

[4] Prananto, J.A., Minasny, B. and Weaver, T. (2020) Near Infrared (NIR) Spectroscopy as a Rapid and Cost-Effective Method for Nutrient Analysis of Plant Leaf Tissues C. In: Sparks, D.L., Ed., *Advances in Agronomy*, Academic Press, 1-49.

[5] Tsuchikawa, S., Ma, T. and Inagaki, T. (2022) Application of Near-Infrared Spectroscopy to Agriculture and Forestry. *Analytical Sciences*, **38**, 635-642. https://doi.org/10.1007/s44211-022-00106-6

[6] 余梅, 李嘉仪, 范伟, 等. 基于近红外光谱仪与模式识别方法的不同年份陈皮无损鉴别研究[J]. 食品研究与开发, 2021, 42(19): 171-178.

[7] 杨济齐, 沈婉莹, 魏晓芳, 等. 基于二维相关红外光谱的陈皮快速鉴别研究[J]. 中南药学, 2022, 20(3): 544-550.

[8] Zhang, X., Gao, Z., Yang, Y., Pan, S., Yin, J. and Yu, X. (2022) Rapid Identification of the Storage Age of Dried Tangerine Peel Using a Hand-Held near Infrared Spectrometer and Machine Learning. *Journal of Near Infrared Spectroscopy*, **30**, 31-39. https://doi.org/10.1177/09670335211057232

[9] Yeo, I. (2000) A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, **87**, 954-959. https://doi.org/10.1093/biomet/87.4.954

[10] Biau, G. and Scornet, E. (2016) A Random Forest Guided Tour. *Test*, **25**, 197-227. https://doi.org/10.1007/s11749-016-0481-7

[11] Patgiri, C. and Ganguly, A. (2021) Adaptive Thresholding Technique Based Classification of Red Blood Cell and Sickle Cell Using Naïve Bayes Classifier and K-Nearest Neighbor Classifier. *Biomedical Signal Processing and Control*, **68**, Article ID: 102745. https://doi.org/10.1016/j.bspc.2021.102745

[12] Feld, S.M. (2023) TSMixer in PyTorch.

[13] Zhong, S., Song, S., Zhuo, W., Li, G., Liu, Y. and Chan, S.G. (2024) A Multi-Scale Decomposition MLP-Mixer for Time Series Analysis. *Proceedings of the VLDB Endowment*, **17**, 1723-1736. https://doi.org/10.14778/3654621.3654637

[14] Wu, H., Hu, T., Liu, Y., *et al*. (2022) TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. *The* 11*th International Conference on Learning Representations*, Kigali, 1-5 May 2023.

[15] Gong, Z., *et al*. (2023) PatchMixer: A Patch-Mixing Architecture for Long-Term Time Series Forecasting.

[16] Wang, S., Wu, H., Shi, X., *et al*. (2024) TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. *The* 12*th International Conference on Learning Representations*, Vienna, 7-11 May 2024.