

基于高光谱成像技术结合机器学习 鉴别茶叶的储藏年份

余静孝, 韦可鑫, 韦庆益*

(华南理工大学食品与工程学院, 广东广州 510640)

摘要: 茶叶品质与价值受储藏年份显著影响, 建立快速、准确的年份鉴别方法具有重要意义。为探索高光谱成像 (HSI) 技术用于茶叶储藏年份识别的可行性, 本研究采用基于可见光 (VIS) 和近红外 (NIR) 波段的 HSI 技术结合机器学习 (ML) 方法, 对 2009 年、2012 年、2015 年、2018 年、2021 年和 2025 年 6 个储藏年份的茶叶样品进行分类鉴别。采集了样品在 VIS (400~1 000 nm) 和 NIR (1 000~2 400 nm) 波段的光谱数据, 并采用一阶导数 (1-st der)、二阶导数 (2-nd der)、Savitzky-Golay 平滑 (SGS)、标准正态变量变换 (SNV) 及多元散射校正 (MSC) 进行预处理; 进一步构建支持向量机分类器 (SVC)、随机森林分类器 (RFC) 和 K 近邻分类器 (KNNC) 模型进行茶叶储藏年份分类, 并基于最优模型对高光谱图像进行可视化分析。结果表明, VIS 波段中 MSC-RFC 模型表现最佳, 分类准确率为 97.22%; NIR 波段中 1-st der-RFC 模型性能最优, 分类准确率达到 100.00%。研究表明, 基于本研究样品条件, HSI 结合光谱预处理与 ML 方法在茶叶储藏年份识别中表现出一定的可行性, 但其在多来源样品中的泛化能力仍需通过独立外部验证进一步评估。

关键词: 茶叶; 储藏年份; 高光谱成像; 机器学习; 图像可视化

DOI: 10.13982/j.mfst.1673-9078.2027.6.0408

Identifying the Storage Years of Tea Leaves Using Hyperspectral Imaging Technology Combined with Machine Learning

YU Jingxiao, WEI Kexin, WEI Qingyi*

(School of Food Science and Engineering, South China University of Technology, Guangzhou 510640, China)

Abstract: The quality and value of tea leaves are significantly influenced by their different storage years, thus establishing a rapid and accurate method for determining storage years is crucial. To explore the feasibility of hyperspectral imaging (HSI) technology for identifying tea storage years, this study used HSI technology in the visible (VIS) and near-infrared (NIR) regions, combined with machine learning (ML) methods, to classify and identify tea samples from six storage years: 2009, 2012, 2015, 2018, 2021, and 2025. Spectral data were collected from the VIS (400~1 000 nm) and NIR (1 000~2 400 nm) regions, and spectral preprocessing was performed using first-order derivative (1-st der), second-order derivative (2-nd der), Savitzky-Golay smoothing (SGS), standard normal variable transformation (SNV), and multiplicative scatter correction (MSC). Support vector machine classifier (SVC), random forest classifier (RFC), and K-nearest neighbor classifier (KNNC) models were then constructed for tea storage year classification, and the optimal model was used to perform visualization analysis of the hyperspectral images. The results showed that the MSC-RFC model performed best in the VIS region, with a classification accuracy of 97.22%. In the NIR region, the 1-st der-RFC model demonstrated the best performance, achieving a classification accuracy of 100.00%. The study indicates that HSI technology combined with spectral preprocessing methods and ML models shows certain feasibility for identifying tea storage years based on the sample conditions of this study. However, the generalizability of the method across samples from multiple sources still requires further evaluation through independent external validation.

Key words: tea leaves; storage years; hyperspectral imaging; machine learning; image visualization

收稿日期: 2026-03-17; 修回日期: 2026-06-16; 接受日期: 2026-06-16

基金项目: 广州市科技计划项目 (2025B03J0124)

作者简介: 余静孝 (1998-), 男, 博士研究生, 研究方向: 食品品质智能无损检测, E-mail: Jingxiaoyu1230@163.com

通讯作者: 韦庆益 (1977-), 女, 博士, 副教授, 研究方向: 农产品质量安全检测及控制技术, E-mail: Feweiqingyi@scut.edu.cn

茶作为全球广泛消费的重要饮品,其品质特征和经济价值会受到品种、加工方式、储藏条件和储藏时间等因素的共同影响。不同茶类在储藏过程中的品质演变规律存在明显差异。白茶^[1]、普洱茶^[2]等茶类在适宜条件下具有一定的陈化价值,而绿茶属于典型的不发酵茶,通常以新鲜品质作为主要评价基础,其品质价值更多依赖于鲜爽香气、翠绿色泽、滋味协调性以及活性成分的保持程度^[3]。随着储藏时间延长,绿茶中的叶绿素、茶多酚、氨基酸、挥发性风味物质等成分会发生不同程度的氧化、降解和转化,进而导致色泽变暗、香气减弱、滋味劣变以及抗氧化活性下降等问题^[4,5]。因此,对于绿茶而言,储藏年份鉴定的核心意义并非评价其陈化增值潜力,而是客观判断其新鲜度、储藏状态和年份标识的真实性,并防止以陈充新、年份虚标和品质误判等问题。建立一种客观、可靠且高效的绿茶储藏年份鉴定方法,对于保障消费者权益、规范茶叶市场秩序和促进茶产业标准化发展具有重要意义。

目前,茶叶年份鉴定主要依赖感官评审和仪器化学分析^[6]。感官评审方法直观简便,能够综合反映茶叶色泽、香气和滋味等整体品质特征,但该方法易受评审人员经验、主观判断和环境条件的影响,存在重复性差、标准化程度不足等问题^[7,8]。仪器化学分析方法,如高效液相色谱和气相色谱-质谱联用等,能够对儿茶素、氨基酸、咖啡碱、挥发性成分等与储藏年份相关的化学指标进行准确测定,但通常需要复杂前处理,且存在检测周期长、成本高、破坏性强和难以现场应用等不足^[9,10]。因此,开发一种快速、客观且适合批量检测的绿茶储藏年份鉴定技术,已成为茶叶品质评价和市场监管中的重要需求。

高光谱成像(Hyperspectral Imaging, HSI)技术融合了成像技术与光谱分析技术,能够同时获取样品的空间信息与光谱信息,在食品品质检测和农产品真实性鉴别中展现出良好的应用前景^[11-13]。对于绿茶样品而言,HSI技术不仅能够表征茶叶外观形态和色泽分布等空间特征,还能够反映与内部化学组成相关的光谱响应。绿茶中叶绿素、茶多酚、水分、氨基酸及含氢基团相关物质的变化会引起反射光谱差异,从而为储藏年份识别提供潜在信息依据^[14]。然而,高光谱数据通常具有维度高、冗余变量多、相邻波段相关性高以及噪声干扰明显等特点,建模过程容易受到维度过高、变量共线性和无关信息干扰的影响^[15]。因此,从高维光谱数据中提取与储藏年份相关的有效特征,并构建具有良好泛化能力的分类模型,是实现绿茶年份准确鉴别的关键。

近年来,机器学习(Machine Learning, ML)方法的发展为高光谱数据解析提供了有效工具^[16,17]。与传统依赖人工经验和单一指标判别的方法相比,ML方法能够从复杂、高维和非线性数据中自动挖掘与目标类别相关的潜在特征模式,并建立光谱响应与储藏年份之间的映射关系^[18]。在高光谱分类任务中,ML方法不仅能够提高复杂光谱信息的利用效率,还可通过特征筛选、模式识别和模型优化等方式降低冗余变量和噪声信息对分类结果的影响^[19]。对于储藏年份差异较为复杂的绿茶样品而言,光谱响应往往受到色素降解、多酚氧化、水分变化和含氢基团振动差异等多种因素的共同影响,其类别边界可能并非简单线性关系^[20]。因此,引入ML方法有助于增强模型对细微光谱差异的识别能力,提高绿茶储藏年份判别的客观性和准确性。系统比较不同ML分类策略在高光谱数据中的适用性,可为筛选稳定、可靠且适合绿茶年份鉴定的建模方法提供依据。

基于此,本研究以客家绿茶为对象,构建融合HSI技术与ML方法的储藏年份快速鉴定策略。本研究建立了包括高光谱信息采集、高维数据处理、模型参数优化和分类性能比较在内的完整分析流程。通过比较不同ML分类模型对储藏年份相关光谱特征的挖掘能力,筛选适用于客家绿茶年份鉴定的建模策略,以实现针对不同储藏年份茶叶样品的准确识别。该研究可为客家绿茶新鲜度评价、年份标识监管和茶叶品质快速筛查提供技术支撑,并为HSI技术在茶叶品质标准化评价中的应用提供参考。

1 材料与方法

1.1 样品制备

不同储藏年份(2009年、2012年、2015年、2018年、2021年和2025年)的客家绿茶样品均购自广州芳村茶叶市场(广州,中国),样品如图1所示。本研究所选茶叶样品的年份覆盖了2009年至2025年较宽的储藏时间范围,年份设置总体上以约3年为间隔,以形成较明显的储藏年份梯度。在茶叶样品选择时,尽量保证样品属于同一品类、外观状态相近、无霉变和无异味,以减少品类和样品状态差异对实验结果的影响。考虑到市场样品可能存在批次、加工条件、初始含水状态及储藏环境等差异,本研究采用统一的样品前处理方法以提高样品间的可比性。首先,采用粉碎机将茶叶样品粉碎,并经200目筛网过筛,获得粒径均一的茶叶粉末。然后,利用压片机

(HY-12, 天津天光光学仪器科技有限公司, 天津, 中国) 将茶叶粉末压制成药片样本 (30 MPa, 30 s)。最后, 将制备好的圆片样品置于干燥箱 (DHG-9035A, 上海益恒科学仪器有限公司, 上海, 中国) 中, 于 30℃ 环境下干燥保存备用。



图 1 不同储藏年份茶叶的实际样品

Fig.1 Real samples of tea leaves from different storage years

1.2 理化性质

采用 Folin-Ciocalteu 法测定茶叶中的总酚含量^[21]。具体方法如下: 向试管中依次加入 1 mL 95%乙醇、5 mL 蒸馏水、0.5 mL 50% (V/V) Folin-Ciocalteu 试剂和 1 mL 茶叶样品萃取液, 充分混匀后反应 5 min, 再加入 1 mL 5% Na_2CO_3 溶液。随后, 将反应混合液在室温下静置 1 h, 并在 760 nm 波长处测定其吸光度。以不同浓度的没食子酸标准溶液绘制标准曲线, 总酚含量以茶叶干重计 ($\text{mg}\cdot\text{g}^{-1}$)。茶叶样品对 DPPH 自由基的清除活性采用参考 Ye 和 Huang 并略作调整方法进行测定^[22]。取 0.2 mL 预先稀释 10 倍的茶叶提取液, 与 3 mL $0.1\text{ mmol}\cdot\text{L}^{-1}$ DPPH 乙醇溶液混合均匀, 在 25℃ 避光条件下孵育 30 min 后, 于 517 nm 波长处测定吸光度。DPPH 自由基清除率计算公式如下:

$$\text{DPPH}(\%) = (1 - A_s/A_c) * 100 \quad (1)$$

式中:

DPPH——DPPH 清除率;

A_s ——样品组的吸光度;

A_c ——对照组的吸光度。

1.3 高光谱成像系统

利用 HSI 系统采集不同储藏年份茶叶样品的高光谱图像数据。该系统由 VIS 波段 (400~1 000 nm) 和 NIR 波段 (1 000~2 500 nm) 两套数据采集系统组成。VIS 数据采集系统主要包括: CCD 相机 (DL-604M, Andor Technology Ltd., Belfast, UK)、光谱仪 (Inspector V10E, Spectral Imaging Ltd., Oulu, Finland)、镜头 (OLE23, Schneider Ltd., Bad Kreuznach, Germany) 以及与移动传送带成 45° 角布置的 150W 卤素灯光源。NIR 数据采集系统主要包括: CCD 相机 (XC403, Xenics Infrared Solutions, Leuven, Belgium)、光谱仪 (Specim V25E, Spectral Imaging Ltd., Oulu, Finland)、镜头 (OLES30, Xenics Infrared Solutions, Leuven, Belgium) 以及由两只 150 W 钨卤素灯组成的照明单元 (2900-ER, Illumination Technologies Inc., New York, USA)。两套系统的共用部分包括平移台 (IRCP0076-1COMB, Isuzu Optics Corp., Taiwan, China) 和装有数据采集软件 (Spectral Image software, Isuzu Optics Corp., Taiwan, China) 的计算机^[23]。数据采集软件可用于控制波长范围、电机转速、曝光时间和图像采集过程^[24]。

1.4 高光谱数据采集与光谱提取

将茶叶压片样品置于黑色背景实验台上, 并采用塑料盒固定样品。采集时, 相机与样品表面的距离设为 30 cm, 相机内部电机速度设为 $1.1\text{ mm}\cdot\text{s}^{-1}$, 曝光时间设为 20 ms^[25], 在上述条件下获取样品的反射高光谱图像。由于成像系统中存在暗电流及环境噪声等因素, 原始高光谱图像通常会受到一定程度的系统误差干扰^[26]。为降低这些因素对图像质量和后续分析结果的影响, 需对原始图像进行黑白校正。校正后的高光谱图像的方法如公式 (2) 所示^[27]。随后, 将校正后的高光谱图像导入 ENVI 5.1 软件, 在茶叶压片样品区域内选取感兴趣区域 (Region of interest,

ROI), 并提取 ROI 内所有像素点的平均反射光谱, 作为后续分类建模的分析数据。

$$R = (I - B)/(W - B)$$
 (2)

式中:

R——校正后的高光谱图像数据;

I——原始高光谱图像数据;

B——黑板的高光谱图像数据;

W——白板的高光谱图像数据。

1.5 数据划分

为提高模型评估结果的可靠性并增强模型的泛化能力, 本研究利用分层划分策略将数据集划分为训练集、验证集和测试集, 具体结果如表 1 所示。此外, 样本量设置主要综合考虑了不同储藏年份茶叶样品的可获得性、类别平衡性以及模型训练与评估的需求。本研究在保证各年份样本数量一致的基础上, 每个年份均选取 60 个样本, 以尽可能覆盖同一储藏年份内茶叶样品的个体差异, 并避免类别不均衡对模型训练和分类结果造成影响。数据集划分时, 各储藏年份样品均按照相同数量分配至训练集、验证集和测试集, 即每个年份分别包含 38 个训练样本、10 个验证样本和 12 个测试样本。训练集(共 228 个样本)用于模型构建和特征学习, 验证集(共 60 个样本)用于模型参数优化及训练过程监控, 测试集(共 72 个样本)不参与模型训练和参数调节, 仅用于评估模型对未知样本的识别性能^[28]。该分层划分策略能够保证不同储藏年份样品在各数据子集中均具有代表性, 有助于在模型充分学习数据特征的同时, 通过相对独立的验证集和测试集降低过拟合风险, 并更客观地反映模型的泛化能力^[29]。

表 1 不同储藏年份茶叶样本数据划分结果

Table 1 Division results of tea leaf sample data from different storage years

年份	总数	训练集	验证集	测试集
2009	60	38	10	12
2012	60	38	10	12
2015	60	38	10	12
2018	60	38	10	12
2021	60	38	10	12
2025	60	38	10	12
总计	360	228	60	72

1.6 光谱预处理

在高光谱数据采集过程中, 受环境条件、仪器噪声及人为操作等因素的影响, 原始光谱往往包含噪声、基线漂移及散射干扰等无效信息^[30]。为提高光谱数据质量并增强后续建模的稳定性, 通常需要对原始光谱进行预处理^[10]。因此, 本研究采用一阶导数(First-order Derivative, 1-st der)、二阶导数(Second-order Derivative, 2-nd der)、Savitzky-Golay 平滑(Savitzky-Golay Smoothing, SGS)、标准正态变量变换(Standard Normal Variable Transformation, SNV)和多元散射校正(Multiplicative Scatter Correction, MSC)五种方法对原始光谱进行预处理, 并对比不同预处理方法对模型性能的影响。其中, 1-st der 可有效减弱基线漂移和背景变化带来的干扰^[31]。2-nd der 可进一步消除线性和二次基线效应, 增强谱峰分辨能力。SGS 能够在保留光谱主要特征的同时抑制高频噪声。SNV 主要用于校正由颗粒尺寸差异、表面散射及光程变化引起的误差^[32]。MSC 则以平均光谱为参考, 对各条光谱的加性和乘性散射效应进行校正, 从而减少样品物理状态差异对光谱基线的影响^[33]。

1.7 无监督模型

为探索高光谱数据的内在结构及样本分布特征, 本研究采用主成分分析(Principal Component Analysis, PCA)和 t-分布随机邻域嵌入(T-distributed Stochastic Neighbor Embedding, t-SNE)两种无监督学习方法对光谱数据进行分析。PCA 通过线性变换将原始高维数据映射到若干相互正交的主成分上, 在保留主要数据信息的同时实现数据降维, 可用于初步观察不同年份茶叶样本的分布特征和聚类趋势。t-SNE 则是一种非线性降维方法, 能够将高

维光谱数据映射到二维空间, 并尽可能保留样本之间的局部邻域关系, 从而更直观地展示复杂数据的非线性分布特征^[29]。

本研究进一步引入轮廓系数、类间/类内距离比和聚类纯度对无监督降维及聚类结果进行量化评价。轮廓系数用于表征类内聚集和类间分离程度, 类间/类内距离比用于衡量不同年份样品之间的差异是否大于同一年份样品内部差异, 聚类纯度则用于评价无监督聚类结果与真实年份标签的一致性。各指标的计算方法如下所示:

$$S = \frac{1}{N} \sum_{i=1}^N \frac{b(i)-a(i)}{\max\{a(i), b(i)\}} \quad (3)$$

$$D_{\text{intra}} = \frac{1}{N} \sum_{k=1}^K \sum_{x_i \in G_k} \|x_i - \mu_k\| \quad (4)$$

$$D_{\text{inter}} = \frac{2}{K(K-1)} \sum_{p=1}^{K-1} \sum_{q=p+1}^K \|\mu_p - \mu_q\| \quad (5)$$

$$R = D_{\text{inter}}/D_{\text{intra}} \quad (6)$$

$$\text{Purity} = \frac{1}{N} \sum_{m=1}^M \max_k |C_m \cap L_k| \quad (7)$$

式中:

$a(i)$ ——第*i*个样品与其所属年份类别内其他样品之间的平均距离;

$b(i)$ ——第*i*个样品与最近的其他年份类别中所有样品之间的平均距离;

N ——样品总数;

S ——所有样品轮廓系数的平均值;

x_i ——第*i*个样品在特征空间中的向量;

μ_k ——第*k*个年份类别中所有样品的中心;

$\|x_i - \mu_k\|$ ——样品 x_i 到其所属年份类别中心的距离;

μ_q ——第*q*个年份类别的中心;

μ_p ——第*p*个年份类别的中心;

$\|\mu_p - \mu_q\|$ ——两个年份类别中心之间的距离;

D_{intra} ——类内距离;

D_{inter} ——类间距离;

R ——类间分离程度与类内离散程度之间的相对关系;

C_m ——第*m*个聚类簇;

L_k ——第*k*个真实年份类别;

$|C_m \cap L_k|$ ——表示第*m*个聚类簇中属于真实年份类别*k*的样品数量;

Purity——聚类纯度。

1.8 有监督模型

为构建并比较茶叶储藏年份判别模型, 本研究选用支持向量分类器 (Support Vector Classifier, SVC)、随机森林分类器 (Random Forest Classifier, RFC) 和 K 近邻分类器 (K-nearest Neighbor Classifier, KNNC) 三种监督学习模型。SVC 的基本思想是在特征空间中构建最优分类超平面, 使不同类别样本之间的间隔最大化。该方法能够借助核函数处理非线性分类问题, 因而具有较强的泛化能力^[34]。RFC 是一种基于 Bagging 策略的集成学习方法, 通过构建多棵决策树并对其分类结果进行集成, 实现对样本类别的判别。该方法能够通过样本扰动和特征扰动增强模型多样性, 从而降低过拟合风险并提升泛化性能^[35]。KNNC 则基于距离度量, 通过计算待分类样本与训练样本之间的距离, 选取最近的 K 个邻居并依据多数投票原则确定类别, 该方法简单直观^[36]。

为提高模型性能并保证方法的可重复性, 本研究对 SVC、RFC 和 KNNC 模型的关键超参数进行了优化。不同模型的主要优化参数包括 SVC 的核函数类型和正则化参数 C, RFC 的决策树数量和最大深度, 以及 KNNC 的邻居数 K 和距离度量方式。表 2 总结了各模型关键超参数的优化范围。

表 2 三种分类模型关键超参数的优化范围

Table 2 Optimization ranges for key hyperparameters of three classification models		
模型	超参数	范围
SVC	核函数类型	['linear', 'poly', 'rbf']
	正则化参数 C	在 0.01 和 10 之间等距取 50 个数值
RFC	决策树数量	在 10 和 1 000 之间等距取 10 个数值
	最大深度	[10, 50, 100, 200, 500]
KNNC	邻居数 K	[1, 3, 5, 7, 9]
	距离度量方式	['Manhattan', 'Euclidean']

为综合评价各分类模型的分类性能，本研究采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1-score 作为评价指标。其中，Accuracy 表示正确分类的样本数占总样本数的比例；Precision 表示模型预测为某一类别的样本中预测正确的比例；Recall 表示某一真实类别样本中被正确识别的比例；F1-score 为 Precision 和 Recall 的调和平均值，可综合反映模型在精确率和召回率之间的平衡关系。各指标的计算公式如下^[37]：

$$Accuracy = \frac{\sum_{k=1}^K C_{kk}}{\sum_{k=1}^K \sum_{j=1}^K C_{kj}}$$

(8)

$$Precision = \frac{1}{K} \sum_{i=1}^K \frac{C_{ii}}{\sum_{k=1}^K C_{ki}}$$

(9)

$$Recall = \frac{1}{K} \sum_{i=1}^K \frac{C_{ii}}{\sum_{j=1}^K C_{ij}}$$

(10)

$$F1-score = \frac{1}{K} \sum_{i=1}^K \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}$$

(11)

式中：

Accuracy——分类模型的准确率；

Precision——分类模型的宏平均精确率；

Recall——分类模型的宏平均召回率；

F1_score——分类模型的宏平均 F1-score；

K——类别总数；

C_{ij} ——混淆矩阵中实际类别为第*i*类、预测类别为第*j*类的样本数；

C_{kk} ——混淆矩阵中第*k*类被正确分类的样本数；

$\sum_{k=1}^K C_{kk}$ ——所有类别中被正确分类的样本总数；

$\sum_{k=1}^K \sum_{j=1}^K C_{kj}$ ——测试样本总数；

C_{ii} ——第*i*类样本被正确预测为第*i*类的样本数；

$\sum_{k=1}^K C_{ki}$ ——所有被预测为第*i*类的样本总数；

$\sum_{j=1}^K C_{ij}$ ——实际属于第*i*类的样本总数。

1.9 图像可视化

基于已构建的茶叶储藏年份高光谱判别模型，本研究对茶叶样本高光谱图像中各像素点进行逐像素判别分析，以实现样本年份信息的可视化表达^[38]。具体流程如下：首先，对制备好的茶叶压片样品进行高光谱扫描，获取原始高光谱图像。随后，经黑白校正和背景去除，提取茶叶压片区域的有效像素。然后将区域内每个像素点的完整光谱数据输入训练好的判别模型，获得对应的年份预测结果。最后，根据不同预测类别赋予相应颜色编码，生成茶叶年份分布的伪彩色图像。

为进一步客观评价像素级预测结果，本研究补充引入像素级准确率、同一压片内部预测一致性指数以及多数投票样品级预测结果三种指标。像素级准确率表示某一样品 ROI 内被正确预测为真实年份的像素比例；预测一致性指数表示同一样品内占比最高的预测类别所对应的像素比例，用于反映像素预测结果的空间一致性；多数投票样品级预测结果则将样品内像素数量最多的预测年份作为该样品的最终预测结果，用于评价像素级结果整合后的

样品级判别能力。通过这些指标,可以更加直观地比较像素级预测结果与区域平均光谱分类结果之间的差异。

$$Acc_{pixel} = \frac{1}{M} \sum_{i=1}^M \left(\frac{1}{N_i} \sum_{j=1}^{N_i} I(\hat{y}_{ij} = y_i) \right) \times 100\% \quad (12)$$

$$PCI_{avg} = \frac{1}{M} \sum_{i=1}^M \frac{\max_{c \in C} n_{ic}}{N_i} \times 100\% \quad (13)$$

$$Acc_{MV} = \frac{1}{M} \sum_{i=1}^M I(\arg \max_{c \in C} n_{ic} = y_i) \times 100\% \quad (14)$$

式中:

Acc_{pixel} ——像素级准确率;

M ——样品数量;

N_i ——第*i*个样品 ROI 内的像素数量;

y_i ——真实年份类别;

$\hat{y}_{ij}$ ——第*i*个样品中第*j*个像素的预测类别;

PCI_{avg} ——预测一致性指数;

C ——年份类别集合;

n_{ic} ——第*i*个样品中被预测为类别*c*的像素数量;

Acc_{MV} ——多数投票样品级结果;

$I(\cdot)$ ——为指示函数。

1.10 统计学分析

本研究中各项理化指标均独立测定 3 次,结果以均值±标准差表示。采用 IBM SPSS Statistics 软件对不同储藏年份茶叶样品的理化指标进行单因素方差分析(One-way ANOVA),并通过 Duncan 多重比较检验分析组间差异的显著性($P < 0.05$)。此外,光谱数据提取、数据划分、光谱预处理、无监督建模、有监督建模及图像可视化等分析过程,均在 Python 3.8 环境下基于 Scikit-learn 库实现,开发平台为 Jupyter Notebook。图表绘制采用 Origin 2021 软件完成。

2 结果与讨论

2.1 理化性质分析

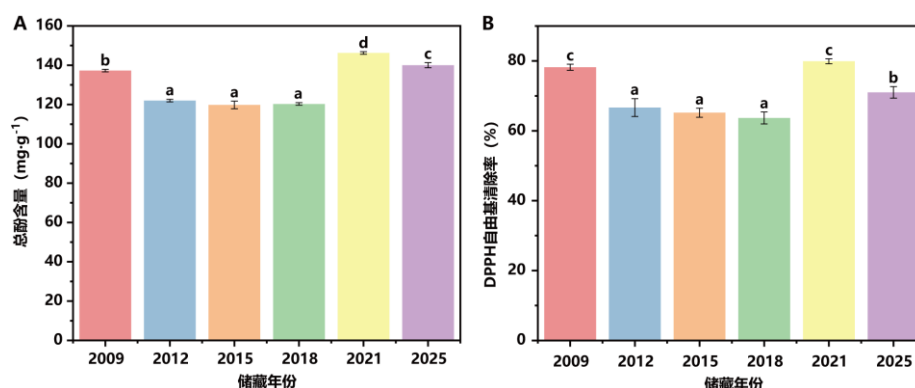


图 2 不同储藏年份茶叶的理化指标

Fig.2 Physicochemical indicators of tea leaves from different storage years

注: 图中不同字母表示不同储藏年份茶叶样品间差异显著 ($P < 0.05$), 相同字母表示差异不显著 ($P > 0.05$)。

图 2A 和 2B 表明,随着储藏年份的增加,茶叶样品的总酚含量和 DPPH 自由基清除率总体呈下降趋势。其中,2025 年新制茶样的总酚含量和 DPPH 自由基清除率均高于储藏年份较长的样品,说明随着储藏时间的延长,茶叶中部分活性成分可能发生降解或转化,从而导致其抗氧化能力减弱。需要指出的是,2021 年和 2025 年样品在变

化趋势上表现出一定异常,说明茶叶理化指标的变化并非完全由储藏年份这一单一因素决定,还可能受到储藏条件,包括温度、湿度、光照及氧气暴露程度等因素的共同影响。此外,茶叶储藏过程中涉及多种化学成分的协同变化,而现有有限的理化指标难以全面反映这些变化与年份特征之间的关系,光谱特征与年份相关化学变化之间的系统性关联仍有待进一步分析。基于此,本研究将进一步引入 HSI 技术采集茶叶样品的光谱信息,并结合 ML 方法对高维数据进行分析,以挖掘与储藏年份相关的更多特征信息,从而实现对茶叶储藏年份的更准确鉴别。

2.2 平均光谱分析

图 3 展示了从校正后高光谱图像中提取的 6 个不同储藏年份茶叶样品的平均光谱。总体来看,各年份样品的光谱曲线整体轮廓较为相似,且均在 VIS 波段约 660 nm 处和 NIR 波段约 1 950 nm 处出现明显特征峰,表明不同储藏年份茶叶在主要化学组成上具有一定共性。然而,不同年份样品在部分特征波段仍表现出明显差异,表明储藏过程中茶叶色素、含水状态及内部化学组分的变化会引起 VIS 和 NIR 光谱响应的改变。

对于 VIS 波段(图 3A),在 400~550 nm,各年份茶叶样品的平均光谱反射率基本一致,说明其在该波段内的光谱差异较小,难以据此实现有效区分。而在 550~700 nm,各样品光谱反射率整体呈上升趋势,但不同年份之间的差异仍不明显。值得注意的是,在 700~1 000 nm,2021 年与 2025 年样品的光谱曲线与其他年份样品表现出较明显的分离,说明储藏年份较近的茶叶在该波段具有一定的可辨识度。相比之下,2009 年、2012 年、2015 年与 2018 年样品的光谱曲线重叠较多,表明相邻年份之间的光谱特征较为接近,区分难度较大。

对于 NIR 波段(图 3B),不同年份样品之间的光谱差异较 VIS 波段更为明显,尤其是在 1 500~1 800 nm 范围内表现更为突出。NIR 光谱中的 1 450 nm 和 1 950 nm 附近通常与水分中 O-H 基团的倍频和合频吸收有关,因此该区域的差异可能反映了不同储藏年份茶叶含水状态及水分结合形式的变化。1 500~1 800 nm 范围内的光谱响应可能与 C-H、N-H 和 O-H 等基团的振动吸收有关,能够反映蛋白质、多酚类物质及部分有机组分的变化。2 100~2 400 nm 范围内的吸收特征则可能与 C-H、O-H 和 N-H 基团的合频吸收有关,进一步反映茶叶中多酚、纤维素、蛋白质及碳水化合物等组分在储藏过程中的变化。由此可见,NIR 波段包含了更多与茶叶内部化学组成相关的信息,因此相较于 VIS 波段更有利于茶叶储藏年份的区分。

总体而言,VIS 波段主要反映茶叶储藏过程中外观颜色和色素相关变化,NIR 波段则更敏感于水分状态及内部化学组分变化。二者在信息来源上具有明显互补性,而 HSI 技术能够进一步结合光谱特征与空间分布信息,为茶叶储藏年份鉴别提供更全面的数据基础。尽管不同年份茶叶样品在部分敏感波段已表现出一定差异,但相邻年份之间仍存在明显的光谱重叠现象,说明仅依靠平均光谱的直观比较仍难以实现精确判别。因此,仍需进一步结合 ML 方法,对高光谱数据进行特征提取与模式识别,以提高茶叶储藏年份鉴别的准确性。

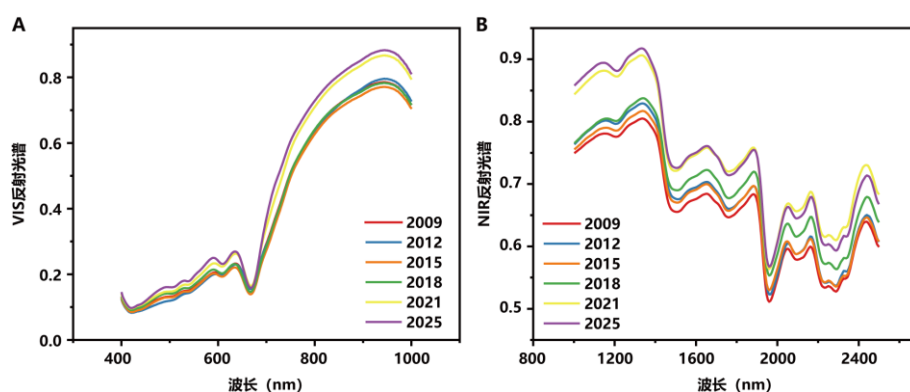


图 3 不同储藏年份茶叶的平均光谱

Fig.3 Average spectra of tea leaves from different storage years

2.3 预处理光谱数据分析

图 4 展示了 6 个储藏年份的茶叶样品在 VIS 和 NIR 波段的原始平均光谱及经 5 种预处理方法处理后的光谱结果。总体来看,不同年份样品的原始光谱曲线轮廓较为相似,但在部分特征波段存在光谱重叠和交叉现象,这在一定程度上会降低基于原始光谱建立模型的判别精度。经 1-st der 方法处理后(图 4B 和图 4H),原始光谱中的基

线漂移得到有效消除,有助于增强光谱变化特征。2-nd der 方法处理后(图 4C 和图 4I),虽然能够进一步削弱线性和二次基线的影响,但同时也会放大高频噪声,从而可能对后续建模稳定性产生不利影响。SGS 方法处理结果如图 4D 和图 4J 所示,该方法主要用于抑制随机噪声。平滑处理后的光谱曲线与原始光谱整体变化趋势基本一致,仅在局部区域表现出一定的平滑效果。SNV(图 4E 和图 4K)及 MSC(图 4F 和图 4L)处理结果表明,这两种方法主要用于校正由样品颗粒大小、表面状态及散射效应差异引起的光谱干扰。经过 SNV 和 MSC 方法处理后,光谱曲线在保留原始整体形状和特征峰位置的同时,不同年份样品在纵向上的离散程度明显减小,说明这两种方法能够有效降低由样品物理性质差异引起的非目标变异。综合比较不同预处理方法可以看出,1-st der、SNV 和 MSC 在消除基线漂移、减弱散射干扰及突出光谱特征方面表现较好,有利于提高光谱数据质量,并为后续模型构建与年份判别提供更可靠的数据基础。

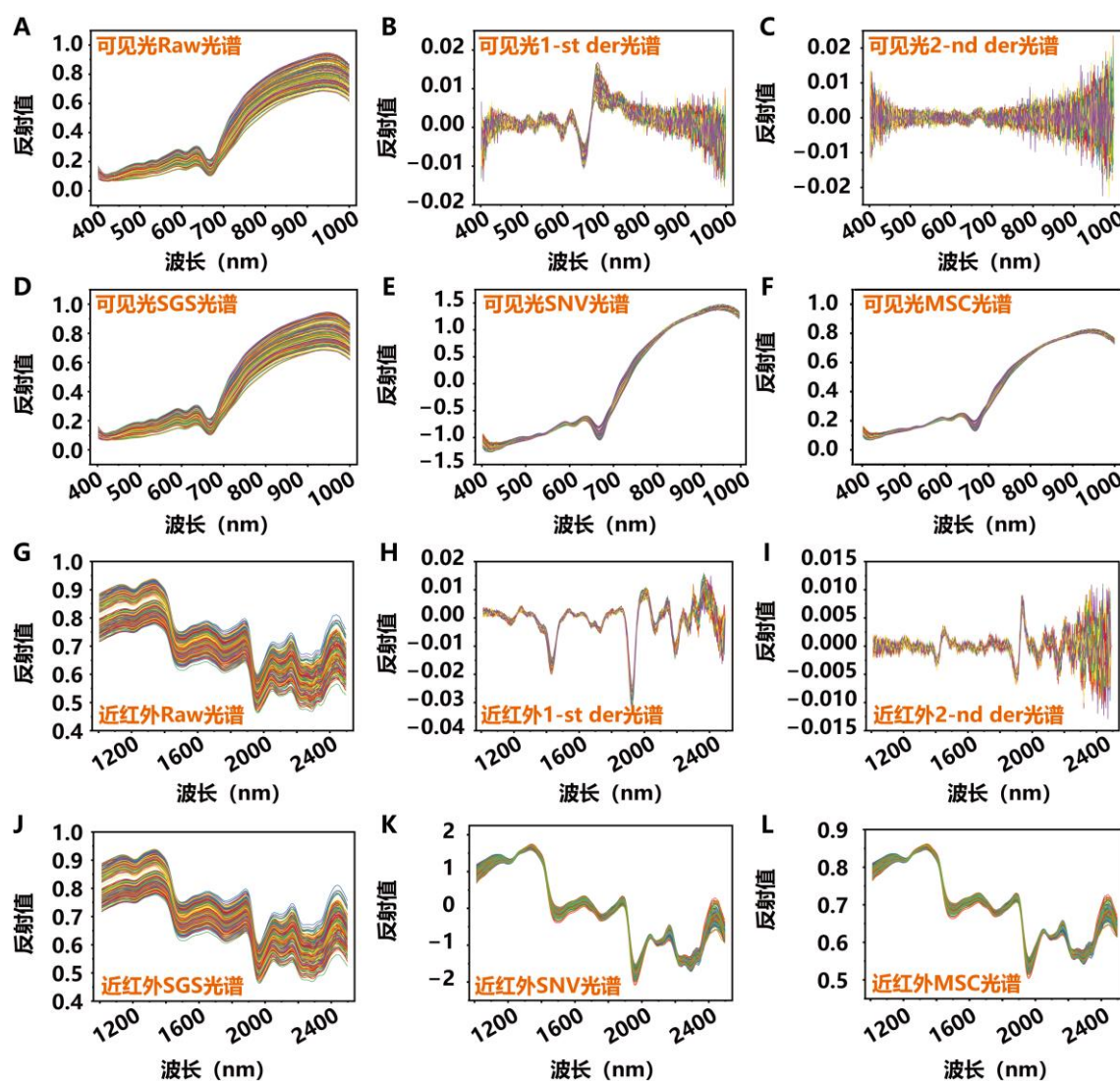


图4 不同储藏年份茶叶光谱预处理结果

Fig.4 Spectral preprocessing results of tea leaves from different storage years

2.4 无监督模型鉴别效果分析

图5展示了6个不同储藏年份茶叶样品的PCA聚类结果及载荷分布,以及基于t-SNE的降维可视化结果。总体来看,两种无监督分析方法所得结果具有较好的一致性,均能够在一定程度上反映不同年份茶叶样品之间的分布特征。在VIS波段(图5A)中,2021年和2025年样品分别形成较为紧密的聚类,并与其他年份样品表现出较明显的分离,说明这两个年份样品具有较好的类内一致性和类间可分性。然而,2009年、2012年、2015年和2018年样品在降维后的特征空间中存在较为明显的重叠现象,难以实现有效区分,表明相邻年份样品之间的光谱差异

相对较小。在 NIR 波段（图 5B）中，样品的聚类分布趋势与 VIS 波段基本一致，即 2021 年和 2025 年样品仍具有较好的分离效果，而其余 4 个年份样品之间依然存在较大程度的交叠。与此同时，t-SNE 可视化结果（图 5C 和图 5D）所呈现的样品分布规律与 PCA 分析结果高度一致，进一步验证了不同储藏年份茶叶样品在无监督特征空间中的分布特征。

为进一步分析光谱差异与茶叶储藏过程中理化成分变化之间的潜在联系，本研究对 VIS 和 NIR 波段的 PCA 主成分载荷进行了分析。由图 5E 可知，VIS 波段 PC1 和 PC2 的累计贡献率达到 95.31%，其中 PC1 贡献率为 86.52%，说明 VIS 光谱的大部分变异信息可由 PC1 解释。VIS 波段载荷在约 600~700 nm 区域出现明显变化，该区域通常与茶叶色泽变化、叶绿素吸收及色素降解有关，表明不同储藏年份茶叶样品在可见光区的差异可能主要来源于储藏过程中色素成分的变化。由图 5F 可知，NIR 波段 PC1 和 PC2 的累计贡献率达到 99.07%，其中 PC1 贡献率为 94.52%，说明 NIR 光谱能够较集中地表征样品间主要差异。NIR 波段载荷在约 1 400~1 500 nm、1 900~2 000 nm 和 2 200~2 400 nm 附近变化较为明显，这些波段通常与 O-H、C-H 和 N-H 等含氢基团的倍频和合频吸收有关，可能反映水分、茶多酚、氨基酸及其他有机成分在储藏过程中的变化。这也在一定程度上与图 2 中茶多酚含量和 DPPH 抗氧化活性随储藏年份变化的结果相呼应。

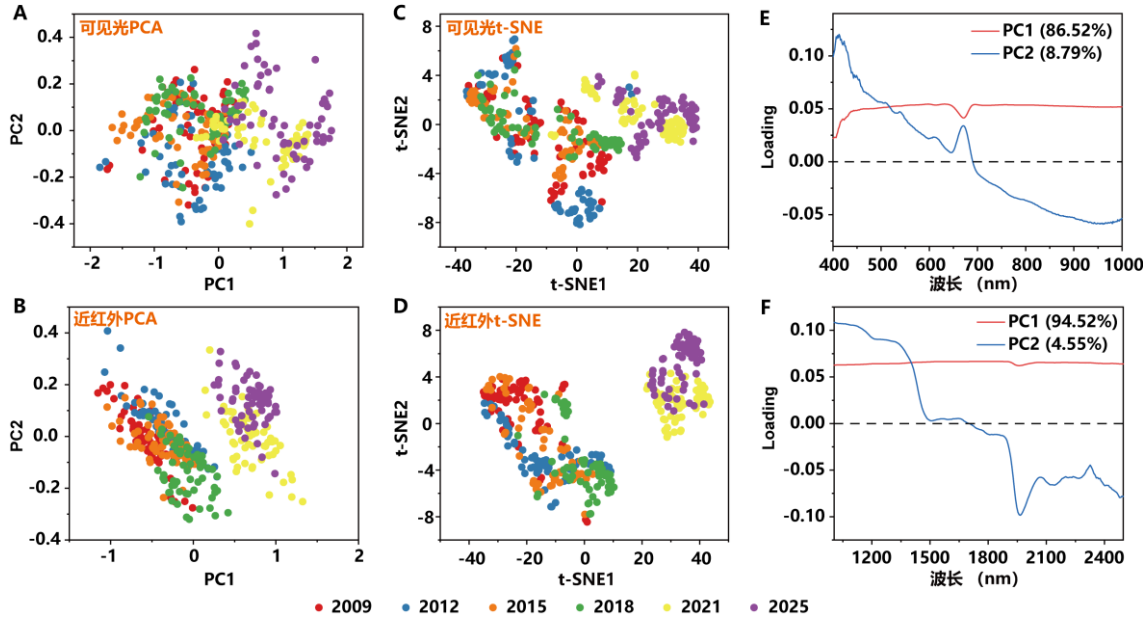


图 5 不同储藏年份茶叶的无监督分析

Fig.5 Unsupervised analysis of tea leaves from different storage years

由表 3 可知，基于 VIS 光谱的 PCA 和 t-SNE 结果的轮廓系数分别为-0.057 5 和-0.068 6，均接近 0 且为负值，说明不同储藏年份茶叶样品在 VIS 降维空间中存在明显重叠。同时，其聚类纯度分别仅为 0.372 2 和 0.377 8，进一步表明 VIS 波段对储藏年份的区分能力有限。相比之下，NIR 光谱的 PCA 和 t-SNE 的类间/类内距离比分别为 3.197 9 和 3.976 4，聚类纯度分别提高至 0.463 9 和 0.505 6，说明 NIR 波段能够提供更多与储藏年份相关的化学信息。然而，基于 NIR 光谱的 PCA 和 t-SNE 结果的轮廓系数分别仅为 0.016 1 和 0.030 1，仍接近 0，表明不同年份样品之间仍存在一定重叠。因此，单纯依靠 PCA 和 t-SNE 等无监督方法难以实现准确区分，仍需进一步结合 ML 模型进行特征提取和模式识别。

表 3 不同储藏年份茶叶的无监督量化分析

Table 3 Unsupervised quantitative analysis of tea leaves from different storage years

无监督方法	轮廓系数	类间/类内距离比	聚类纯度
可见光 PCA	-0.057 5	1.693 6	0.372 2
可见光 t-SNE	-0.068 6	1.688 6	0.377 8
近红外 PCA	0.016 1	3.197 9	0.463 9
近红外 t-SNE	0.030 1	3.976 4	0.505 6

综合来看，无监督分析结果与前述平均光谱分析结果相互印证，即 2021 年和 2025 年样品能够与其他年份样

品较好地区分,而 2009 年、2012 年、2015 年和 2018 年等相邻年份样品之间仍存在明显的光谱重叠现象,导致其难以仅依靠无监督方法实现准确鉴别。因此,为进一步提高茶叶储藏年份判别的准确性和可靠性,有必要引入有监督学习模型进行深入分析。

2.5 有监督模型鉴别效果分析

表 4 和表 5 分别比较了基于 VIS 和 NIR 光谱,并结合原始光谱(Raw)、5 种预处理方法(1-st der、2-nd der、SGS、SNV 和 MSC)以及 3 种有监督模型(SVC、RFC 和 KNNC)所得的不同储藏年份茶叶分类性能。

在 VIS 波段中,不同预处理方法对模型性能的影响较为明显。SVC 模型对光谱预处理方法较为敏感,当采用 Raw 或 SGS 处理时,测试集 Accuracy 均为 54.17%,F1-score 均为 50.98%,说明模型对不同年份茶叶样品的区分能力有限。经 1-st der 预处理后,SVC 模型性能明显提升,测试集 Accuracy、Precision、Recall 和 F1-score 分别达到 90.28%、90.16%、90.28%和 89.88%,表明 1-st der 处理有助于增强 VIS 波段中与储藏年份相关的判别特征。相比之下,2-nd der-SVC 组合虽然训练集 Accuracy 达到 100%,但验证集和测试集 Accuracy 分别仅为 15.26%和 19.44%,F1-score 也仅为 16.93%,说明该组合存在明显过拟合现象。这可能是因为 2-nd der 在增强光谱局部差异的同时,也放大了 VIS 波段中的噪声和微小波动。由于 VIS 波段主要反映茶叶表面色泽和色素相关信息,易受样品颗粒状态和光散射影响,SVC 模型可能学习到训练集中的偶然差异,而未能提取稳定的年份判别特征。

对于 RFC 和 KNNC 模型,多个组合在训练集上的 Accuracy 均达到 100%,说明模型具有较强的拟合能力。然而,训练集准确率较高并不必然代表模型具有良好的泛化能力,因此仍需重点关注验证集和测试集结果。VIS 波段下,RFC-MSC 组合在测试集上表现最佳,Accuracy、Precision、Recall 和 F1-score 分别为 97.22%、98.83%、97.22%和 97.12%;KNNC-SNV 和 KNNC-MSC 组合也取得了较好的测试集结果,Accuracy 和 F1-score 均达到 95.83%。这些结果表明,合适的 SNV 或 MSC 处理能够提高 VIS 光谱分类模型的判别性能。

表 4 VIS 光谱数据不同预处理方法与建模方法组合的建模效果

Table 4 Modeling effects of different combinations of preprocessing and modeling methods for VIS data

模型	预处理	训练集	验证集	测试集			
		Accuracy	Accuracy	Accuracy	Precision	Recall	F1-score
SVC	Raw	49.65	47.56	54.17	49.33	54.17	50.98
	1-st der	99.04	79.17	90.28	90.16	90.28	89.88
	2-nd der	100.00	15.26	19.44	17.46	19.44	16.93
	SGS	49.57	47.56	54.17	49.33	54.17	50.98
	SNV	59.38	58.29	60.85	66.00	60.85	57.30
	MSC	59.38	58.29	60.85	66.00	60.85	57.30
RFC	Raw	100.00	80.84	87.50	89.32	87.50	86.79
	1-st der	100.00	93.04	95.83	95.94	95.83	95.83
	2-nd der	100.00	31.62	40.27	43.44	40.27	41.28
	SGS	100.00	80.84	87.50	89.10	87.50	87.18
	SNV	100.00	96.87	88.88	89.28	88.88	88.86
	MSC	100.00	81.25	97.22	98.83	97.22	97.12
KNNC	Raw	100.00	77.04	87.50	88.22	87.50	87.68
	1-st der	100.00	38.20	40.27	45.54	40.27	41.95
	2-nd der	100.00	15.63	19.44	16.74	19.44	17.73
	SGS	100.00	77.39	87.50	88.22	87.50	87.68
	SNV	100.00	96.53	95.83	95.92	95.83	95.83
	MSC	100.00	96.18	95.83	95.92	95.83	95.83

注: 表中建模结果的单位均为%。

在 NIR 波段中,不同模型与预处理方法组合的分类性能总体优于 VIS 波段。尤其是 1-st der-RFC 模型在当前数据划分下取得了较高的分类性能,训练集、验证集和测试集 Accuracy 均达到 100%,同时 Precision、Recall 和

F1-score 也均为 100%。此外, SNV-KNNC 和 MSC-KNNC 组合在测试集上 Accuracy、Precision、Recall 和 F1-score 均达到 100%。上述结果表明, NIR 波段包含了更多与茶叶储藏年份相关的内部化学信息, 1-st der、SNV 和 MSC 预处理方法能够进一步增强有效光谱特征差异, 提高模型分类性能。

在本研究样本量有限的条件下, 测试集共包含 72 个样本, 其中每个储藏年份类别仅包含 12 个样本。对于包含 6 个年份类别的多分类任务而言, 1-st der-RFC 模型在当前测试集上取得 100%的准确率, 表明该模型在本研究样本条件和当前数据划分下对客家绿茶储藏年份具有较好的识别能力。然而, 由于每个年份类别的测试样本数量相对有限, 且样品主要来源于同一市场体系, 模型性能仍可能受到样本规模、批次差异、储藏环境差异以及数据划分偶然性的影响。因此, 后续研究仍需进一步扩大样本规模, 纳入不同产地、不同加工批次、不同储藏环境和不同市场来源的茶叶样品, 并构建更大规模、多批次的独立测试集, 以系统评估模型在实际应用条件下的稳定性、鲁棒性和泛化能力。

总体来看, 相比于 VIS 光谱, NIR 光谱结合适当预处理和 ML 模型能够更有效地区分不同储藏年份茶叶样品。对于训练集准确率较高但验证集和测试集表现较差的模型组合, 提示其可能存在过拟合风险; 而对于验证集和测试集均表现较好的模型组合, 则可认为其在本研究数据条件下具有较好的年份识别能力。

表 5 NIR 光谱数据不同预处理方法与建模方法组合的建模效果 (%)

Table 5 Modeling effects of different combinations of preprocessing and modeling methods for NIR data (%)

模型	预处理	训练集	验证集	测试集			
		Accuracy	Accuracy	Accuracy	Precision	Recall	F1-score
SVC	Raw	79.17	78.48	80.55	82.19	80.55	80.41
	1-st der	99.04	96.87	97.22	97.44	97.22	97.22
	2-nd der	99.91	82.65	76.38	78.43	76.38	76.94
	SGS	79.25	78.47	80.55	82.19	80.55	80.41
	SNV	87.59	85.78	86.11	87.12	86.11	86.14
	MSC	86.46	85.80	86.11	86.67	86.11	85.93
RFC	Raw	100.00	81.60	80.55	80.73	80.55	80.54
	1-st der	100.00	100.00	100.00	100.00	100.00	100.00
	2-nd der	100.00	86.82	83.33	84.62	83.33	83.51
	SGS	100.00	81.25	83.33	83.51	83.33	83.32
	SNV	100.00	97.92	95.83	96.23	95.83	95.88
	MSC	100.00	92.05	95.83	96.05	95.83	95.89
KNNC	Raw	100.00	85.06	86.11	87.64	86.11	86.57
	1-st der	100.00	90.28	94.44	94.84	94.44	94.43
	2-nd der	100.00	49.30	48.61	50.09	48.61	48.82
	SGS	100.00	85.06	86.11	87.64	86.11	86.57
	SNV	100.00	98.28	100.00	100.00	100.00	100.00
	MSC	100.00	98.08	100.00	100.00	100.00	100.00

注: 表中建模结果的单位均为%。

根据不同波段的建模结果, 本研究选取各模型对应的最优预处理方法用于茶叶储藏年份分类, 测试集混淆矩阵结果如图 6 所示。在 VIS 波段中, 经 MSC 处理后的 RFC 模型表现最佳, 仅有 2 个 2018 年样本被误判为相邻年份的 2015 年, 其余各年份样本均被正确分类, 表明该模型能够较好地区分不同储藏年份茶叶, 尤其对年份差异较明显的样本具有较高识别精度。在 NIR 波段中, 经 1-st der 预处理的 RFC 模型和经 SNV 预处理的 KNNC 模型在测试集上均取得 100%的准确率, 实现了对 6 个储藏年份茶叶样品的正确分类。

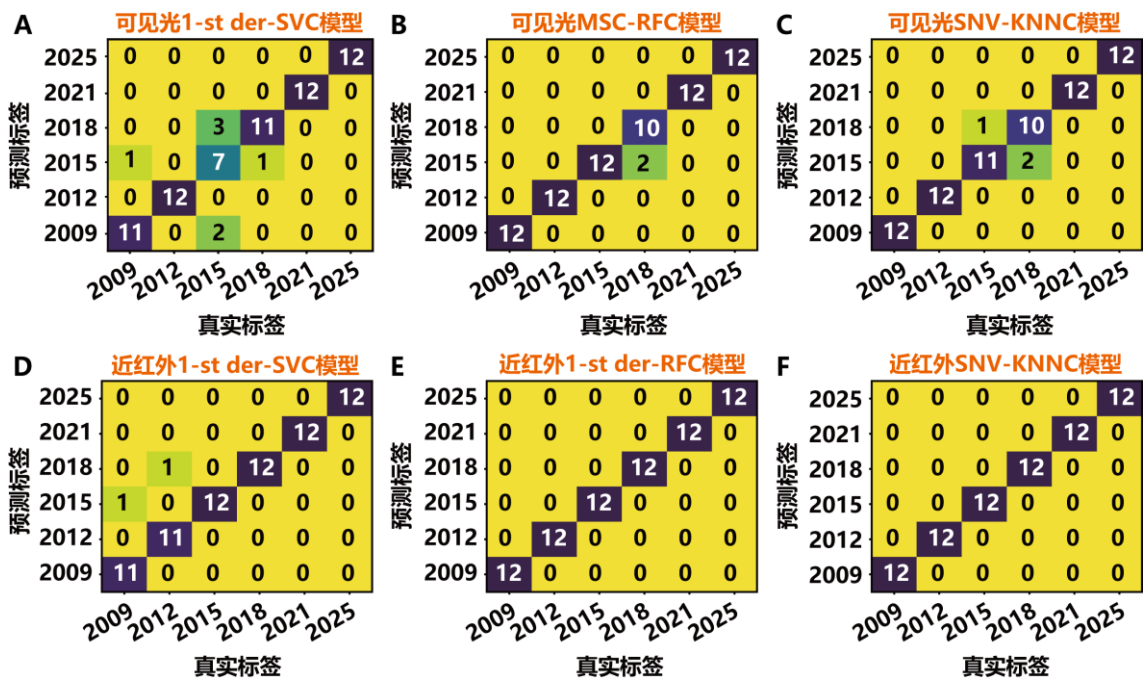


图 6 不同储藏年份茶叶有监督模型的混淆矩阵

Fig.6 Confusion matrices of supervised models for tea stored for different years

RFC 模型在不同波段和预处理组合中整体表现较优。为进一步增强模型的可解释性，本研究基于 RFC 模型对 VIS 和 NIR 波段的光谱波长重要性进行了分析，结果如图 7 所示。VIS 波段（图 7A）的重要波长主要集中在约 600~700 nm 范围内，该区域与茶叶表面色泽及叶绿素等色素成分的光谱响应密切相关。随着储藏时间的延长，绿茶中色素成分发生降解和转化，可能导致该区域反射光谱产生明显差异。相比之下，NIR 波段（图 7B）的重要波长分布范围更广，主要集中在约 1 400~1 500 nm、1 700~1 900 nm 和 2 000 nm 附近区域，这些波段通常与 O-H、C-H 和 N-H 等含氢基团的倍频和合频吸收有关，能够反映水分、茶多酚、氨基酸及其他有机成分变化所引起的光谱差异。该结果说明，RFC 模型并非仅依赖随机噪声进行分类，而是在一定程度上利用了与茶叶储藏过程中理化成分变化相关的关键光谱区域，从而进一步支持了 HSI 结合 ML 用于茶叶储藏年份鉴别的可行性和合理性。

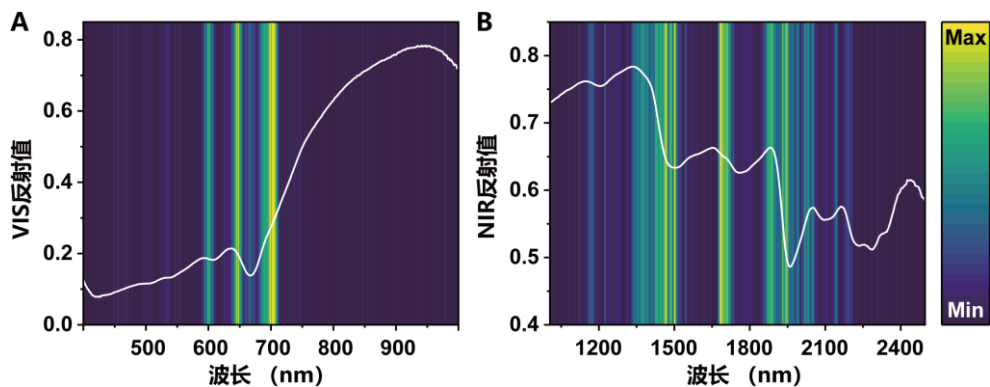


图 7 基于 RFC 模型 VIS 和 NIR 光谱波长重要性分析

Fig.7 Analysis of the importance of VIS and NIR spectral wavelengths based on the RFC model

综合比较可见，NIR 波段在茶叶年份鉴别中表现出更优的分类能力，这与前述平均光谱分析和无监督分析结果相一致。结合 RFC 模型在训练集、验证集和测试集上的稳定表现，可以认为，RFC 模型不仅具有较强的茶叶储藏年份判别能力，而且在模型泛化能力和鲁棒性方面也表现较优，是本研究中综合性能较好的分类模型。

2.6 高光谱图像可视化分析

基于不同分类模型对 VIS 和 NIR 波段光谱数据的分析结果，本研究进一步对不同储藏年份茶叶样品高光谱图

像进行逐像素分类可视化,结果如图8所示。图中每列茶叶压片样品从左至右分别对应2009年、2012年、2015年、2018年、2021年和2025年6个储藏年份,不同颜色表示模型预测得到的储藏年份类别。

从图8可以看出,不同模型在逐像素预测中存在明显差异。VIS 1-st der-SVC模型几乎将所有像素预测为2021年,NIR SNV-KNNC模型则几乎将所有像素预测为2012年,说明二者在单像素尺度下存在明显的系统性偏差。相比之下,VIS MSC-RFC、VIS SNV-KNNC和NIR 1-st der-SVC能够在一定程度上反映不同年份样品之间的差异,但同一压片内部仍存在多个年份类别并存的现象。NIR 1-st der-RFC模型对2021年和2025年样品的预测结果相对集中,表现出较好的年份识别趋势,但对2009年、2012年、2015年和2018年样品仍存在预测混杂。上述结果说明,逐像素分类不仅受局部噪声影响,还与单像素光谱的空间非均质性、散射差异和区域平均光谱训练模型的尺度差异有关。

表4和表5中的模型训练与测试主要基于茶叶压片区域的平均光谱,而图8中的分类图像是将训练好的模型应用于单像素光谱后得到的逐像素预测结果。因此,区域级预测和像素级预测在输入尺度上存在明显差异。区域平均光谱能够在一定程度上降低局部噪声、表面不均一性、散射差异和光照波动的影响,更适合反映样品整体光谱特征;而单像素光谱保留了更多局部空间异质性,容易受到茶叶压片内部颗粒差异、表面粗糙度、厚度不均、边缘效应和局部阴影等因素的影响。因此,部分模型即使在区域平均光谱测试集中取得较高分类准确率,在逐像素预测中仍可能出现系统性偏差或同一样品内部多个预测年份并存的现象。

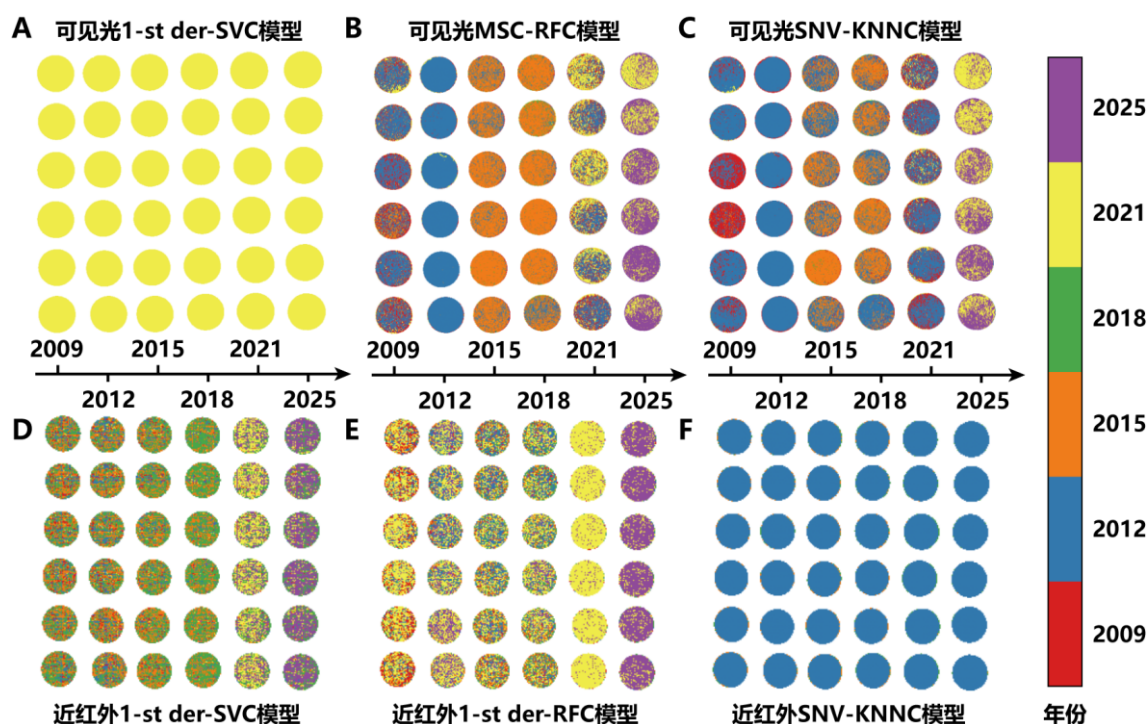


图8 不同储藏年份茶叶高光谱图像可视化结果

Fig.8 Visualization results of hyperspectral images of tea leaves from different storage years

为进一步定量比较区域级预测与逐像素预测之间的差异,本研究补充计算了不同模型的像素级准确率、平均预测一致性指数和多数投票样品级准确率,结果如表6所示。总体来看,不同模型的像素级准确率均明显低于基于区域平均光谱获得的测试集准确率,说明由区域平均光谱训练得到的模型直接应用于单像素光谱时,存在明显的输入尺度差异。其中,VIS 1-st der-SVC和NIR SNV-KNNC模型的平均预测一致性指数分别达到100.00%和97.11%,但其多数投票样品级准确率仅为16.67%,说明这类模型虽然在同一压片内部预测结果较为一致,但存在系统性偏向某一错误类别的问题。相比之下,VIS MSC-RFC和NIR 1-st der-RFC模型的平均预测一致性指数分别为71.36%和72.00%,多数投票样品级准确率均为33.33%,表明其能够在一定程度上保留不同样品之间的类别差异,但仍受到单像素光谱局部波动和空间非均质性的影响。

表 6 不同模型的像素级准确率、平均预测一致性指数和多数投票样品级准确率

Table 6 Pixel-level accuracy, average prediction consistency index and majority voting sample-level accuracy of different models

模型	像素级准确率	平均预测一致性指数	多数投票样品级准确率
可见光 1-st der-SVC	16.99	100.00	16.67
可见光 MSC-RFC	25.32	71.36	33.33
可见光 SNV-KNNC	21.41	70.25	20.83
近红外 1-st der-SVC	24.66	50.13	33.33
近红外 1-st der-RFC	28.82	72.00	33.33
近红外 SNV-KNNC	15.90	97.11	16.67

注：表中结果的单位均为%。

上述结果说明，逐像素分类结果表现不佳并不仅由随机噪声引起，还可能与区域平均光谱和单像素光谱之间的输入尺度差异有关。区域平均光谱能够削弱颗粒差异、表面粗糙度、厚度不均和局部散射等影响，而单像素光谱则更容易受到局部空间异质性和边缘效应的干扰。因此，后续研究可进一步引入空间平滑滤波、多数投票后处理、基于像素级或块级样本的模型训练，以及融合空间上下文信息的 CNN 模型，以提高高光谱图像逐像素预测的稳定性和鲁棒性。

3 结论

本研究基于 VIS 和 NIR 的 HSI 技术，系统比较了 VIS 和 NIR 两个波段光谱信息、5 种预处理方法（1-st der、2-nd der、SGS、SNV 和 MSC）以及 3 种 ML 模型（SVC、RFC 和 KNNC）在茶叶储藏年份鉴别中的应用效果。结果表明，不同波段、预处理方法和分类模型对茶叶年份判别性能均具有显著影响。总体来看，NIR 波段较 VIS 波段表现出更强的年份区分能力，说明 NIR 光谱中包含了更多与茶叶储藏年份相关的有效信息。在光谱预处理方面，1-st der、SNV 和 MSC 能够较好地减弱基线漂移和散射干扰，提升模型对年份特征的提取能力，从而提高分类精度。进一步比较不同预处理和模型组合发现，各模型在不同波段下均存在相应的最优组合，其中 VIS 波段的较优组合为 1-st der-SVC、MSC-RFC 和 SNV-KNNC，NIR 波段的较优组合为 1-st der-SVC、1-st der-RFC 和 SNV-KNNC。其中，NIR 波段下 1-st der-RFC 模型表现最佳，在测试集上准确率达到 100%，且训练集、验证集和测试集结果保持一致，表明该模型在本研究数据条件下具有较强的年份识别能力和较好的稳定性。与此同时，逐像素可视化结果显示，区域平均光谱模型直接应用于单像素预测时仍存在一定尺度差异，提示后续仍需进一步优化像素级建模和空间信息融合方法。因此，NIR 波段结合的 1-st der-RFC 组合能够实现对茶叶储藏年份较为准确和稳定的预测。值得注意的是，本研究采用粉碎压片后的茶叶粉末样品进行分析，虽然有助于提高样品均一性和光谱采集稳定性，但该前处理过程会改变茶叶原有物理形态，因此后续研究仍需进一步开展完整茶叶样品的直接检测，以提高方法的实际应用性。本研究样品来源于单一市场体系，且尚未引入独立外部验证集，因此模型在不同来源、批次及储藏条件样品中的泛化能力仍需进一步验证。此外，后续研究应进一步扩大不同来源、不同批次和不同储藏条件下的样本，并引入外部独立测试集，以提高模型的泛化能力和实际应用可靠性。总的来说，VIS 和 NIR 的 HSI 技术结合 ML 方法在茶叶储藏年份快速鉴别中表现出一定可行性，可为茶叶品质评价、年份识别及市场溯源提供潜在的技术支撑。

参考文献

[1] 蒲璐璐,罗金龙,方仕茂,等.不同年份贵州白茶品质差异分析[J/OL].食品工业科技,2025,1-17.

[2] YANG J Y, LI D Q, CHEN J Y, et al. Rapid quality evaluation of Pu-erh raw tea storage years via hyperspectral imaging coupled with machine learning and microbial profiling [J]. Food Chemistry, 2026, 517: 149484.

[3] 唐林,张艳诚,李家华,等.基于近红外光谱的普洱茶年份检测研究[J].广东农业科学,2014,41(17):93-96.

[4] 陈书媛,张友超,杨杰,等.基于高光谱成像技术的白茶储藏年份判别[J].食品工业科技,2021,42(18):276-283.

[5] YANG Z, ZHANG Y X, SONG N, et al. Undescribed phenylpropanoid-substituted ester-type catechins from green tea exert antiaging effects in Caenorhabditis elegans via activating DAF-16, HSF-1, and heat shock proteins [J]. Journal of Agricultural and Food Chemistry, 2025, 73(25): 15673-15692.

- [6] HU Y, WU Y L, SUN J, et al. Distinguishing different varieties of oolong tea by fluorescence hyperspectral technology combined with chemometrics [J]. *Foods*, 2022, 11(15): 2344.
- [7] 马梦君,胡新龙,邱首哲,等.基于代谢组学的不同年份青砖茶主要品质成分分析[J].*茶叶科学*,2025,45(1):133-144.
- [8] LI Z Q, YIN X L, GU H W, et al. Discrimination and prediction of Qingzhuan tea storage year using quantitative chemical profile combined with multivariate analysis: Advantages of MRM^{HR} based targeted quantification metabolomics [J]. *Food Chemistry*, 2024, 448: 139088.
- [9] PU H B, YU J X, SUN D-W, et al. Distinguishing pericarpium citri reticulatae of different origins using terahertz time-domain spectroscopy combined with convolutional neural networks [J]. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2023, 299: 122771.
- [10] 孙若璠,高健健,周丹阳,等.基于代谢组学的不同贮藏年份紧压寿眉白茶化学成分分析[J].*食品科学*,2025,46(7):217-228.
- [11] SUN D-W, PU H B, YU J X. Applications of hyperspectral imaging technology in the food industry [J]. *Nature Reviews Electrical Engineering*, 2024, 1(4): 251-263.
- [12] WANG Z H, PENG W J, TIAN Z R, et al. Rapid detection of synthetic pigments in black tea using hyperspectral imaging technology and machine learning [J]. *Food Chemistry-X*, 2025, 31: 103192.
- [13] LONG T, TANG X Y, LIANG C J, et al. Detecting bioactive compound contents in Dancong tea using VNIR-SWIR hyperspectral imaging and KRR model with a refined feature wavelength method [J]. *Food Chemistry*, 2024, 460: 140579.
- [14] LIU Y W, PU H B, SUN D-W. Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications [J]. *Trends in Food Science & Technology*, 2017, 69: 25-35.
- [15] FENG Z H, WANG L Y, YANG Z Q, et al. Hyperspectral monitoring of powdery mildew disease severity in wheat based on machine learning [J]. *Frontiers in Plant Science*, 2022, 13: 828454.
- [16] YANG Z F, LU X P, CHEN L C. Discriminating the adulteration of varieties and misrepresentation of vintages of Pu'er tea based on Fourier transform near infrared diffuse reflectance spectroscopy [J]. *Frontiers in Chemistry*, 2025, 13: 1546702.
- [17] TIAN J J, XU S F, WU Y J, et al. Authenticating vintage in white tea: Appearance-taste-aroma-based three-in-one non-invasive anticipation [J]. *Food Research International*, 2025, 199: 115394.
- [18] CHENG J H, SUN D-W, PU H B, et al. Integration of classifiers analysis and hyperspectral imaging for rapid discrimination of fresh from cold-stored and frozen-thawed fish fillets [J]. *Journal of Food Engineering*, 2015, 161: 33-39.
- [19] TIAN J J, WU Y J, XU S F, et al. Non-invasive anticipation of infusion taste in fine-manipulated green teas through hyperspectral appearance analysis guided by ECG content [J]. *Food Chemistry*, 2024, 458: 140254.
- [20] LI F L, SHEN J F, YANG Q F, et al. Monitoring quality changes in green tea during storage: A hyperspectral imaging method [J]. *Food Chemistry-X*, 2024, 23: 101538.
- [21] SHETTY K, CURTIS O F, LEVIN R E. Specific interaction of mucoid strains of *Pseudomonas* spp. with oregano (*Origanum vulgare*) clones and the relationship to prevention of hyperhydricity in tissue culture [J]. *Journal of Plant Physiology*, 1996, 149: 605-611.
- [22] YE C L, HUANG Q. Extraction of polysaccharides from herbal *Scutellaria barbata* D. Don (Ban-Zhi-Lian) and their antioxidant activity [J]. *Carbohydrate Polymers*, 2012, 89(4): 1131-1137.
- [23] CHENG W W, SUN D-W, CHENG J H. Pork biogenic amine index (BAI) determination based on chemometric analysis of hyperspectral imaging data [J]. *LWT-Food Science and Technology*, 2016, 73: 13-19.
- [24] CHENG W W, SUN D W, PU H B, et al. Heterospectral two-dimensional correlation analysis with near-infrared hyperspectral imaging for monitoring oxidative damage of pork myofibrils during frozen storage [J]. *Food Chemistry*, 2018, 248: 119-127.
- [25] BAI G R, ZHANG T S, LIU Z Y, et al. Research on rapid and non-destructive detection of adulterated wheat flour using hyperspectral imaging and machine learning [J]. *Journal of Food Composition and Analysis*, 2025, 147: 108050.
- [26] PHILLIPS T, ABDULLA W. A new honey adulteration detection approach using hyperspectral imaging and machine learning [J]. *European Food Research and Technology*, 2023, 249(2): 259-272.
- [27] SHU Z, LI X, LIU Y D. Detection of chili foreign objects using hyperspectral imaging combined with chemometric and target detection algorithms [J]. *Foods*, 2023, 12(13): 2618.

- [28] PAN Y Y, SUN D W, CHENG J H, et al. Non-destructive detection and screening of non-uniformity in microwave sterilization using hyperspectral imaging analysis [J]. Food Analytical Methods, 2018, 11(6): 1568-1580.
- [29] MA J, CHENG J H, SUN D W, et al. Mapping changes in sarcoplasmatic and myofibrillar proteins in boiled pork using hyperspectral imaging with spectral processing methods [J]. LWT-Food Science and Technology, 2019, 110: 338-345.
- [30] WANG T, CHEN Y K, HUANG Y Y, et al. Prediction of the quality of anxi tieguanyin based on hyperspectral detection technology [J]. Foods, 2024, 13(24): 4126.
- [31] WANG J J, ZAREEF M, HE P H, et al. Evaluation of matcha tea quality index using portable NIR spectroscopy coupled with chemometric algorithms [J]. Journal of the Science of Food and Agriculture, 2019, 99(11): 5019-5027.
- [32] GAO W J, CHENG X, LIU X H, et al. Apple firmness detection method based on hyperspectral technology [J]. Food Control, 2024, 166: 110690.
- [33] LI J, WU G F, GUO F, et al. Detection of protein content in alfalfa using visible/ near-infrared spectroscopy technology [J]. Bioresources, 2024, 19(2): 3808-3825.
- [34] HONG Z Q, ZHANG C, KONG D D, et al. Identification of storage years of black tea using near-infrared hyperspectral imaging with deep learning methods [J]. Infrared Physics & Technology, 2021, 114(6): 103666.
- [35] ZHANG S P, TAN Z L, LIU J, et al. Determination of the food dye indigotine in cream by near-infrared spectroscopy technology combined with random forest model [J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 227: 117551.
- [36] HE C Y, JI Y, WU B, et al. Non-destructive classification of chrysanthemum tea using near-infrared spectroscopy (NIRS) and fuzzy improved pseudoinverse linear discriminant analysis (FIPLDA) [J]. Analytical Letters, 2024, 57(6): 905-919.
- [37] WANG P, FAN E, WANG P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning [J]. Pattern Recognition Letters, 2021, 141: 61-67.
- [38] JIANG M W, LI Y T, SONG J, et al. Study on black spot disease detection and pathogenic process visualization on winter jujubes using hyperspectral imaging system [J]. Foods, 2023, 12(3): 435.