

基于 NIRS 和 PSO-LightGBM 的番茄营养成分快速检测

施利春¹ 王松伟² 王治忠²

(1. 河南职业技术学院, 河南 郑州 450046; 2. 郑州大学, 河南 郑州 450001)

摘要: [目的] 实现番茄干物质、可溶性糖、总酸、维生素 C、可溶性固形物及番茄红素 6 项营养成分的快速无损高精度检测。[方法] 提出一种近红外光谱(NIRS)结合粒子群优化算法(PSO)改进轻量梯度提升机(LightGBM)的番茄营养成分检测模型(PSO-LightGBM)。以 6 项营养成分为检测指标, 采集 471 份番茄样品的近红外光谱数据(扫描范围 400~2 500 nm); 对比 7 种预处理方法, 筛选出标准正态变量变换(SNV)与一阶导数结合的最优方案以消除噪声与基线漂移; 采用 PSO 算法优化 LightGBM 关键超参数(叶子节点数、学习率、树深度、特征采样比例、Bagging 比例), 构建多成分同步预测模型; 同时构建最小二乘回归(PLSR)、未优化 LightGBM、网格搜索优化 LightGBM(GS-LightGBM)作为对比模型。[结果] PSO-LightGBM 模型对 6 项营养成分预测精度均最优: 干物质、可溶性固形物的预测集决定系数(R^2)分别达 0.987, 0.973, 残差预测偏差(E_{RPD})分别为 6.732, 5.115; 总酸、可溶性糖、维生素 C 的 R^2 均超过 0.95, E_{RPD} 均大于 2.5; 番茄红素 R^2 提升至 0.956, 较未优化 LightGBM 提升 12.4%。PSO-LightGBM 预测集均方根误差(E_{RMSEP})较 PLSR 平均降低 34.7%, 较 GS-LightGBM 平均降低 15.2%。[结论] 该方法解决了 LightGBM 参数随机设定导致的泛化能力不足问题, 能够实现番茄多营养成分同步精准检测。

关键词: 近红外光谱; 粒子群算法; 番茄; 营养成分; 无损检测; 轻量梯度提升机

Rapid detection of tomato nutritional components based on NIRS and PSO-LightGBM

Shi Lichun¹ Wang Songwei² Wang Zhizhong²

(1. Henan Polytechnic, Zhengzhou, Henan 450046, China; 2. Zhengzhou University, Zhengzhou, Henan 450001, China)

Abstract: [Objective] To achieve rapid, non-destructive, and high-precision detection of six nutritional components in tomatoes, including dry matter, soluble sugar, total acid, vitamin C, soluble solids, and lycopene. [Methods] A tomato nutritional component detection model based on near-infrared spectroscopy (NIRS) combined with light gradient boosting machine (LightGBM) optimized using the particle swarm optimization (PSO) algorithm (PSO-LightGBM) was proposed. Using the six nutritional components as detection indicators, near-infrared spectral data (scanning range: 400~2 500 nm) were collected from 471 tomato samples. Seven preprocessing methods were compared, and the optimal combination of standard normal variate (SNV) transformation and first derivative was selected to eliminate noise and baseline drift. The PSO algorithm was used to optimize the key hyperparameters of LightGBM (number of leaves, learning rate, tree depth, feature sampling ratio, and Bagging ratio) to construct a synchronous multi-component prediction model. Meanwhile, partial least squares regression (PLSR), unoptimized LightGBM, and grid search-optimized LightGBM (GS-LightGBM) models were established as comparison models. [Results] The PSO-LightGBM model achieved the highest prediction accuracy for all six nutritional components. The coefficients of determination for the prediction set (R_p^2) for dry matter and soluble solids reached 0.987 and 0.973, respectively, with residual prediction

基金项目: 河南省科学技术厅科技攻关项目(编号: 242102211036)

通信作者: 施利春(1971—), 男, 河南职业技术学院副教授, 硕士。E-mail: shilichun@sohu.com

收稿日期: 2026-03-11 改回日期: 2026-07-28

引用格式: 施利春, 王松伟, 王治忠. 基于 NIRS 和 PSO-LightGBM 的番茄营养成分快速检测[J]. 食品与机械, 2026, 42(9): 74-82.

Citation: Shi Lichun, Wang Songwei, Wang Zhizhong. Rapid detection of tomato nutritional components based on NIRS and PSO-LightGBM[J]. Food & Machinery, 2026, 42(9): 74-82.

deviation (E_{RPDP}) values of 6.732 and 5.115, respectively. The R_p^2 values for total acid, soluble sugar, and vitamin C all exceeded 0.95, and their E_{RPDP} values were all greater than 2.5. The R_p^2 for lycopene increased to 0.956, representing a 12.4% increase compared with that of the unoptimized LightGBM model. The prediction-set root mean square error (E_{RMSEP}) of PSO-LightGBM was, on average, 34.7% lower than that of PLSR and 15.2% lower than that of GS-LightGBM. **[Conclusion]** This method addresses the insufficient generalization ability caused by the random setting of LightGBM parameters and enables synchronous and accurate detection of multiple nutritional components in tomatoes.

Keywords: near-infrared spectroscopy; particle swarm optimization algorithm; tomato; nutritional components; non-destructive testing; light gradient boosting machine

番茄作为全球广泛消费的果蔬品种,其干物质、可溶性糖、维生素C、番茄红素等营养成分不仅决定了风味与营养价值,也是品质分级与市场定价的核心依据^[1-2]。2024年中国番茄年产量超6 000万t,在生鲜消费与加工产业中占据重要地位^[3]。传统检测方法(如高效液相色谱法测定番茄红素、滴定法测定总酸)虽精度较高,但存在样品预处理复杂、耗时久、破坏性强等局限,无法满足现代化生产中“实时监测、批量筛查”的技术需求^[4-5]。

近红外光谱(NIRS)能够快速、无损地同步检测多项指标,因此在食品品质分析中得到了广泛应用^[6-7]。该技术通过检测样品中C—H、O—H、N—H等官能团的吸收特征,构建光谱数据与营养成分之间的定量关系模型,从而实现间接测定^[8]。在番茄检测领域,已有研究采用偏最小二乘回归(PLSR)、支持向量机(SVR)等算法构建预测模型,但存在明显不足。PLSR为线性模型,难以拟合光谱与成分间强非线性关系^[9];SVR虽适合小样本、泛化能力优良,但对核函数与参数敏感,高维光谱下调参复杂^[10];传统神经网络(如BP)结构固定、易陷入局部最优,大网络规模易过拟合,且训练效率低^[11]。

LightGBM作为基于梯度提升决策树(GBDT)的高效集成学习框架,具有计算复杂度低、学习速度快、非线性拟合能力强、抗过拟合等优势,适合高维光谱数据的定量预测。但LightGBM性能高度依赖叶子节点数、学习率、树深度等超参数,人工或随机设定易导致性能波动、泛化能力不足^[12-14]。粒子群优化(PSO)算法全局搜索能力强、参数少、收敛快,可高效完成模型超参数寻优,已成功用于光谱模型优化^[15-17]。

针对番茄红素光谱响应弱、预测精度低,以及现有模型未充分优化、多成分同步检测精度不足的问题,研究拟融合NIRS与PSO-LightGBM,构建番茄干物质、可溶性糖、总酸、维生素C、可溶性固形物及番茄红素6项营养成分同步检测模型,通过光谱预处理优化、PSO超参数寻优以及多模型对比,旨在为番茄品质快速无损检测提供技术支持。

1 材料与方法

1.1 番茄样品准备

2025年8—9月,于北京、河北、浙江、山东、江苏等主

产区共采集番茄样品471份,包括常规红果番茄434份和黄色番茄37份,覆盖圣女果、大粉果、硬果番茄等不同品种及多个成熟度等级。样品采集后立即置于4℃保温箱中运输至实验室,并于2d内完成光谱采集与营养成分分析。样品制备采用四分法取样,去除果柄及腐烂部分后切块,使用破壁机匀浆2min,制成均匀浆状(无可见果皮与籽粒)。将匀浆分为3份,其中1份用于光谱采集,剩下2份用于营养成分测定。

1.2 主要试剂与仪器

1.2.1 试剂

氢氧化钠:优级纯,西陇科学股份有限公司;

乙醇、盐酸、碳酸氢钠:分析纯,国药集团化学试剂有限公司。

1.2.2 仪器

电子天平:PL203型,瑞士梅特勒—托利多公司;

近红外光谱仪:NIRS DS2500型,丹麦FOSS公司;

超高效液相色谱仪:Acquity UPLC型,美国Waters公司;

手持折射仪:H028354型,日本ATAGO公司;

数显恒温水浴锅:HH-4型,澳华仪器有限公司;

破壁机:AUX-PB920型,奥克斯集团有限公司。

1.3 营养成分测定

(1) 干物质:依据GB 5009.3—2016《食品安全国家标准 食品中水分的测定》中的直接干燥法。

(2) 总酸:依据GB 12456—2021《食品安全国家标准 食品中总酸的测定》中的酸碱指示剂滴定法。

(3) 可溶性糖:依据NY/T 1278—2007《蔬菜及其制品中可溶性糖的测定 铜还原碘量法》。

(4) 维生素C:依据GB 5009.86—2025《食品安全国家标准 食品中抗坏血酸的测定》中的二氯酚酚滴定法。

(5) 可溶性固形物:依据NY/T 2637—2014《水果和蔬菜可溶性固形物含量的测定 折射仪法》。

(6) 番茄红素:依据NY/T 1651—2008《蔬菜及其制品中番茄红素的测定 高效液相色谱法》。

每样品平行测定3次取平均值。

1.4 近红外光谱采集

在光谱采集前,先将光谱仪开启并预热30min。随

后使用聚四氟乙烯标准白板进行基线校正,以消除环境光干扰和仪器漂移影响。将番茄匀浆注入样品杯,装样量控制在杯容积的 1/2~2/3,再将样品杯置于样品台上,环境温度维持 25 ℃。每个样品连续扫描 3 次,取平均光谱作为原始光谱,以吸光度形式保存。为保证数据一致性,每测定 10 个样品后重新进行白板校正。

1.5 数据处理与模型构建

1.5.1 样本集划分 为确保校正集与预测集样本分布的一致性并降低模型过拟合风险,采用肯纳德—斯通(KS)方法,依据样本在光谱空间中的欧氏距离进行代表性样本选择。按照 3:1 的比例,将全部 471 份样品划分为校正集(353 份)与预测集(118 份)^[18]。

1.5.2 光谱预处理 对比原始光谱、标准正态变量变换(SNV)、多元散射校正(MSC)、一阶导数、二阶导数、MSC+一阶导数、SNV+一阶导数 7 种方案的预处理效果。

1.5.3 模型评价指标 以校正集的均方根误差(E_{RMSEC})、相关系数(R_c)、决定系数(R_c^2),以及预测集的均方根误差(E_{RMSEP})、相关系数(R_p)、决定系数(R_p^2)、残差预测偏差(E_{RPDP}) 7 项指标综合评价模型性能。其中, E_{RMSEC} 与 E_{RMSEP} 越小, R_c^2 、 R_p^2 越接近 1, $E_{\text{RPDP}} > 3.0$ 表明模型性能优异, $2.0 < E_{\text{RPDP}} < 3.0$ 表明模型可定量预测^[19]。各评价指标计算式如式(1)~式(7)所示。

$$E_{\text{RMSEC}} = \sqrt{\frac{1}{n_c} \sum_{i=1}^{n_c} (y_{ci} - \hat{y}_{ci})^2}, \quad (1)$$

$$E_{\text{RMSEP}} = \sqrt{\frac{1}{n_p} \sum_{i=1}^{n_p} (y_{pi} - \hat{y}_{pi})^2}, \quad (2)$$

$$R_c = \frac{\sum_{i=1}^{n_c} (y_{ci} - \bar{y}_c)(\hat{y}_{ci} - \bar{\hat{y}}_c)}{\sqrt{\sum_{i=1}^{n_c} (y_{ci} - \bar{y}_c)^2} \sqrt{\sum_{i=1}^{n_c} (\hat{y}_{ci} - \bar{\hat{y}}_c)^2}}, \quad (3)$$

$$R_p = \frac{\sum_{i=1}^{n_p} (y_{pi} - \bar{y}_p)(\hat{y}_{pi} - \bar{\hat{y}}_p)}{\sqrt{\sum_{i=1}^{n_p} (y_{pi} - \bar{y}_p)^2} \sqrt{\sum_{i=1}^{n_p} (\hat{y}_{pi} - \bar{\hat{y}}_p)^2}}, \quad (4)$$

$$R_c^2 = \frac{\left(n_c \sum_{i=1}^{n_c} \hat{y}_{ci} y_{ci} - \sum_{i=1}^{n_c} \hat{y}_{ci} \sum_{i=1}^{n_c} y_{ci} \right)^2}{\left[n_c \sum_{i=1}^{n_c} \hat{y}_{ci}^2 - \left(\sum_{i=1}^{n_c} \hat{y}_{ci} \right)^2 \right] \left[n_c \sum_{i=1}^{n_c} y_{ci}^2 - \left(\sum_{i=1}^{n_c} y_{ci} \right)^2 \right]}, \quad (5)$$

$$R_p^2 = \frac{\left(n_p \sum_{i=1}^{n_p} \hat{y}_{pi} y_{pi} - \sum_{i=1}^{n_p} \hat{y}_{pi} \sum_{i=1}^{n_p} y_{pi} \right)^2}{\left[n_p \sum_{i=1}^{n_p} \hat{y}_{pi}^2 - \left(\sum_{i=1}^{n_p} \hat{y}_{pi} \right)^2 \right] \left[n_p \sum_{i=1}^{n_p} y_{pi}^2 - \left(\sum_{i=1}^{n_p} y_{pi} \right)^2 \right]}, \quad (6)$$

$$E_{\text{RPDP}} = \frac{p_{\text{SD}}}{E_{\text{RMSEP}}}, \quad (7)$$

式中:

$y_{ci}, \hat{y}_{ci}$ ——校正集第 i 个样本的实测值与预测值;

$y_{pi}, \hat{y}_{pi}$ ——预测集第 i 个样本的实测值与预测值;

$\bar{y}_c, \bar{\hat{y}}_c$ ——校正集实测值与预测值平均值;

$\bar{y}_p, \bar{\hat{y}}_p$ ——预测集实测值与预测值平均值;

n_c, n_p ——校正集与预测集样本数;

p_{SD} ——预测集实测值标准差。

1.5.4 PSO-LightGBM 模型构建 在获得最优光谱预处理方案(SNV+一阶导数)的基础上,构建基于粒子群优化算法(PSO)改进轻量梯度提升机(LightGBM)的番茄营养成分同步预测模型(PSO-LightGBM)。

LightGBM 是基于梯度提升决策树(GBDT)的高效集成学习框架,采用叶节点优先生长策略,能够有效拟合高维光谱数据与营养成分之间的非线性关系。设训练集为 $\{(x_i, y_i)\}_{i=1}^n$, 其中 x_i 为第 i 个样本的近红外光谱特征向量, y_i 为对应的营养成分实测值,则 LightGBM 的模型预测值可表示为^[19-20]:

$$\hat{y}_i = \sum_{k=1}^K \eta f_k(x_i), \quad (8)$$

式中:

K ——决策树数量;

η ——学习率(控制迭代步长);

$f_k(x_i)$ ——第 k 棵决策树对样本 x_i 的输出值。

在第 t 次迭代中,模型通过最小化损失函数(式(9))来学习新的决策树。

$$L^{(t)} = \sum_{i=1}^n \left(l(y_i, \hat{y}_i^{(t-1)}) + f_t(x_i) \right) + \Omega(f_t), \quad (9)$$

式中:

$l(\cdot)$ ——损失函数,用于衡量预测值与实测值之间的偏差;

$\hat{y}_i^{(t-1)}$ ——前 $t-1$ 棵树对第 i 个样本的累积预测值;

$f_t(x_i)$ ——第 t 棵决策树对样本 x_i 的输出值;

$\Omega(f_t)$ ——正则化项,用于控制模型复杂度以防止过拟合。

其中,正则化项 $\Omega(f_t)$ 为:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2, \quad (10)$$

式中:

T ——单棵树的叶节点数;

λ —— L_2 正则系数;

γ ——叶节点最小增益阈值(控制树复杂度);

w_j ——第 j 个叶节点的输出权重。

在缺乏先验经验的情况下,LightGBM 模型的性能主

要受叶子节点数、学习率、特征采样比例、Bagging比例等超参数影响。为避免人工设定参数的盲目性,以式(11)校正集均方根误差(E_{RMSEC})为目标函数,运用粒子群优化算法对LightGBM模型关键超参数进行迭代寻优。

$$E_{\text{RMSEC}} = \sqrt{\frac{1}{n_c} \sum_{i=1}^{n_c} (y_{ci} - \hat{y}_{ci})^2}, \quad (11)$$

式中:

n_c ——校正集样本数;

y_{ci} ——校正集中第*i*个样本的营养成分实测值;

$\hat{y}_{ci}$ ——校正集中第*i*个样本的营养成分预测值。

PSO算法的关键参数设置^[21]:种群规模 $N=50$,最大迭代次数 $T=100$,学习因子 $c_1=c_2=2$ 。各超参数的搜索范围分别为:叶子节点数 $[5, 50]$ 、学习率 $[0.01, 0.3]$ 、树最大深度 $[3, 15]$ 、特征采样比例 $[0.5, 1]$ 、Bagging比例 $[0.5, 1]$ 。在每次迭代中,PSO更新各粒子的速度和位置,并依据适应度函数评价当前超参数组合的优劣,最终输出全局最优解作为LightGBM模型的最优超参数配置。PSO-LightGBM模型的整体构建流程如图1所示。同时,为验证PSO-LightGBM模型的性能优势,设置3种对比模型:①PLSR模型(采用留一法交叉验证确定最优主成分数);②未优化LightGBM模型(超参数随机初始化);③GS-LightGBM模型(网格搜索优化LightGBM参数,搜索步长0.1,交叉验证折数 $k_{\text{fold}}=5$)。

2 试验结果与分析

2.1 样本集营养成分统计

校正集与预测集的6项营养成分含量统计结果如表1所示。两组样本的含量范围、平均值、标准差及变异系数(CV)均较为接近,CV均超过30%,覆盖了番茄从青涩到完全成熟的全阶段(如番茄红素5.32~212.00 mg/kg),数

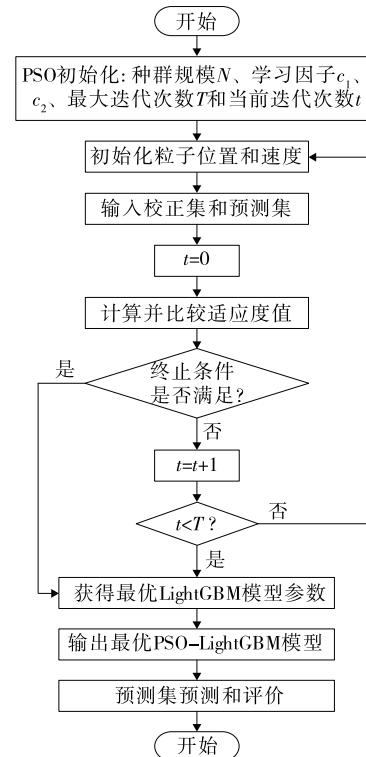


图1 PSO-LightGBM流程图

Figure 1 Flowchart of PSO-LightGBM

据分布合理,满足模型构建与验证要求。

2.2 光谱预处理效果分析

不同预处理方法的效果对比结果如表2所示,包括SNV和MSC。SNV+一阶导数组合预处理的 P_{SNR} 最高(42.8),校正集 E_{RMSEC} 最低(干物质0.215%、总酸0.321 g/kg、可溶性糖0.382%、维生素C 3.612 mg/kg、可溶性固形物0.243%、番茄红素9.875 mg/kg),校正集 R_c^2 平均达0.902,

表1 校正集与预测集番茄样品营养成分统计

Table 1 Statistics of nutritional components in calibration and prediction sets of tomato samples

样本集	指标	单位	样本数	最小值	最大值	平均值	标准差	变异系数/%
校正集	干物质	%	353	4.08	12.70	6.62	1.72	26.0
	总酸	g/kg	353	1.50	8.26	4.13	1.21	29.3
	可溶性糖	%	353	0.91	6.34	3.01	1.09	36.2
	维生素C	mg/kg	353	68.80	493.00	205.12	98.76	48.1
	可溶性固形物	%	353	3.30	10.70	5.63	1.57	27.9
	番茄红素	mg/kg	353	5.32	212.00	47.89	27.34	57.1
预测集	干物质	%	118	4.11	11.20	6.75	1.70	25.2
	总酸	g/kg	118	2.17	8.25	4.29	1.25	29.1
	可溶性糖	%	118	1.42	6.72	3.16	1.15	36.4
	维生素C	mg/kg	118	88.00	493.00	203.45	92.68	45.6
	可溶性固形物	%	118	3.50	10.30	5.78	1.55	26.8
	番茄红素	mg/kg	118	7.22	111.30	46.98	26.12	55.6

显著优于单一预处理方法(如原始光谱 $P_{\text{SNR}}=29.5$, 平均 $R_c^2=0.786$)。这是因为 SNV 可消除样品颗粒度与表面散射差异, 一阶导数能有效校正基线漂移, 二者结合可最大程度保留有效光谱信息。预处理前后的光谱对比如图 2 所示, 可见 SNV+一阶导数处理后, 光谱噪声显著降低, 特征峰更清晰。

表 2 光谱预处理效果分析
Table 2 Analysis of spectral preprocessing effects

预处理	P_{SNR}	E_{RMSEC}						R_c^2
		干物质/ %	总酸/ ($\text{g}\cdot\text{kg}^{-1}$)	可溶性糖/ %	维生素 C/ ($\text{mg}\cdot\text{kg}^{-1}$)	可溶性固 形物/%	番茄红素/ ($\text{mg}\cdot\text{kg}^{-1}$)	
原始光谱	29.5	0.582	0.765	0.893	8.745	0.621	22.361	0.786
MSC	34.2	0.415	0.532	0.678	6.892	0.456	18.425	0.831
SNV	35.7	0.389	0.498	0.635	6.531	0.412	17.258	0.845
一阶导数	32.6	0.458	0.576	0.721	7.324	0.489	19.632	0.822
二阶导数	31.8	0.472	0.593	0.745	7.568	0.503	20.147	0.817
MSC+一阶导数	40.3	0.238	0.356	0.415	3.987	0.267	10.532	0.889
SNV+一阶导数	42.8	0.215	0.321	0.382	3.612	0.243	9.875	0.902

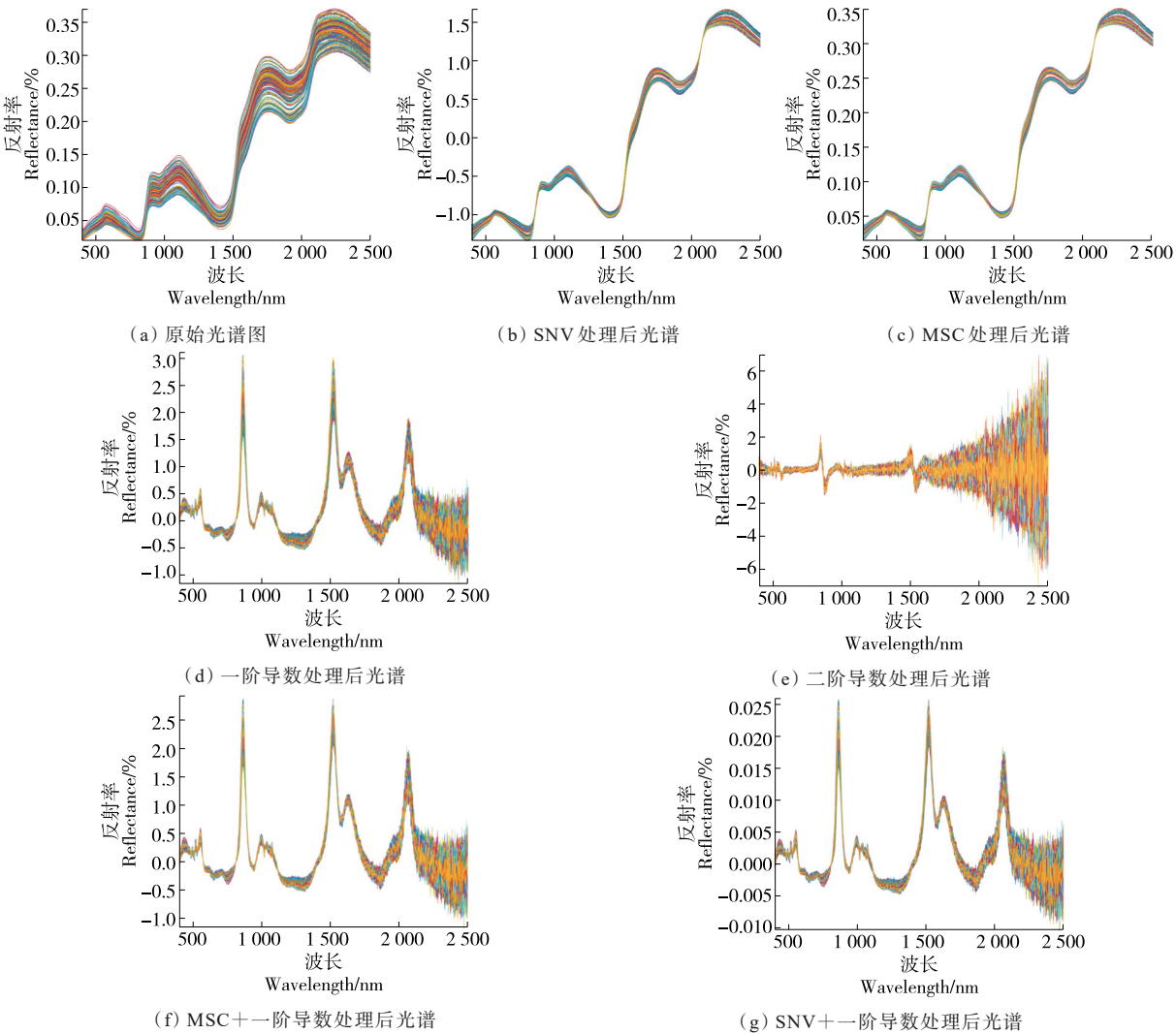


图 2 番茄样品近红外光谱预处理结果
Figure 2 Results of near-infrared spectral preprocessing of tomato samples

2.3 不同模型预测性能对比

PLSR、原始 LightGBM、GS-LightGBM 以及 PSO-LightGBM 4 种模型对番茄干物质、总酸、可溶性糖、维生

素 C、可溶性固形物、番茄红素 6 项核心营养成分进行定量预测,各模型预测性能不同评价指标如表 3~表 5 所示,模型预测值与实测值的拟合效果如图 3 所示。

表 3 不同模型对番茄营养成分的预测性能对比 (E_{RMSEP})

Table 3 Comparative prediction performance of different models for tomato nutritional components (E_{RMSEP})						
模型	干物质/%	总酸/($\text{g}\cdot\text{kg}^{-1}$)	可溶性糖/%	维生素 C/($\text{mg}\cdot\text{kg}^{-1}$)	可溶性固形物/%	番茄红素/($\text{mg}\cdot\text{kg}^{-1}$)
PLSR	0.400	0.695	0.782	6.893	0.458	18.762
未优化 LightGBM	0.312	0.523	0.596	5.347	0.342	14.325
GS-LightGBM	0.273	0.475	0.513	4.862	0.301	12.876
PSO-LightGBM	0.231	0.412	0.435	4.215	0.265	11.243

表 4 不同模型对番茄营养成分的预测性能对比 (R_p^2)

Table 4 Comparative prediction performance of different models for tomato nutritional components (R_p^2)						
模型	干物质/%	总酸/($\text{g}\cdot\text{kg}^{-1}$)	可溶性糖/%	维生素 C/($\text{mg}\cdot\text{kg}^{-1}$)	可溶性固形物/%	番茄红素/($\text{mg}\cdot\text{kg}^{-1}$)
PLSR	0.921	0.935	0.918	0.942	0.903	0.901
未优化 LightGBM	0.957	0.902	0.915	0.913	0.938	0.903
GS-LightGBM	0.971	0.936	0.928	0.937	0.952	0.908
PSO-LightGBM	0.982	0.967	0.959	0.974	0.965	0.956

表 5 不同模型对番茄营养成分的预测性能对比 (E_{RPDP})

Table 5 Comparative prediction performance of different models for tomato nutritional components (E_{RPDP})						
模型	干物质/%	总酸/($\text{g}\cdot\text{kg}^{-1}$)	可溶性糖/%	维生素 C/($\text{mg}\cdot\text{kg}^{-1}$)	可溶性固形物/%	番茄红素/($\text{mg}\cdot\text{kg}^{-1}$)
PLSR	4.256	1.793	1.725	1.864	3.382	1.392
未优化 LightGBM	5.478	2.145	2.036	2.178	4.521	1.896
GS-LightGBM	6.125	2.387	2.294	2.395	4.987	2.013
PSO-LightGBM	6.892	2.678	2.583	2.745	5.347	5.114

由表 3 可知,PSO-LightGBM 在所有成分上均取得最低误差,预测偏差最小;PLSR 误差最大,原始 LightGBM 居中,网格搜索与粒子群优化均有效降低预测误差。由表 4 可知,PSO-LightGBM 拟合优度最优,多数成分 R_p^2 超 0.95,远高于 PLSR;GS-LightGBM 次之,原始 LightGBM 对部分成分拟合略弱。由表 5 可知,PSO-LightGBM 值最高,预测可靠性最强。

综上所述,传统 PLSR 性能最弱,LightGBM 经优化后预测能力显著提升,其中 PSO 优化效果优于 GS 优化,更适合番茄多营养成分同步精准预测,为番茄品质快速检测提供优选模型。

2.4 PSO-LightGBM 模型稳定性验证

2.4.1 交叉验证 为评估所建近红外定量模型的稳定性和泛化能力,对校正集进行 5 折交叉验证。各营养成分对应的 E_{RMSEC} 及 R_c^2 的平均值与标准差结果见表 6。

由表 6 可知,各营养成分的 E_{RMSEC} 平均值均处于较低水平,且对应的标准差较小,说明模型在不同验证折次间的预测误差波动小、稳定性好。同时,所有指标的 R_c^2 平均值稳定在 0.978~0.985,标准差不超过 0.004,表明模型对

校正集样本的解释能力高且一致性强。综合误差与决定系数的表现,所建模型在交叉验证中未出现显著过拟合现象,具备良好的拟合稳定性和预测可靠性,可为外部验证及实际应用提供支持。

2.4.2 重复性分析 为评估样本划分对模型稳定性的影响,重复训练 PSO-LightGBM 模型 10 次,每次均按 3:1 的比例随机划分校正集与预测集,结果见表 7。

由表 7 可知,各营养成分对应的 E_{RMSEP} 变异系数均低于 2.5%,其中番茄红素最高(1.98%),干物质最低

表 6 交叉验证结果

Table 6 Cross-validation results			
指标	单位	E_{RMSEC}	R_c^2
干物质	%	0.230±0.004	0.983±0.002
总酸	g/kg	0.411±0.009	0.980±0.003
可溶性糖	%	0.433±0.008	0.981±0.002
维生素 C	mg/kg	4.205±0.065	0.978±0.004
可溶性固形物	%	0.264±0.005	0.985±0.002
番茄红素	mg/kg	11.241±0.087	0.979±0.003

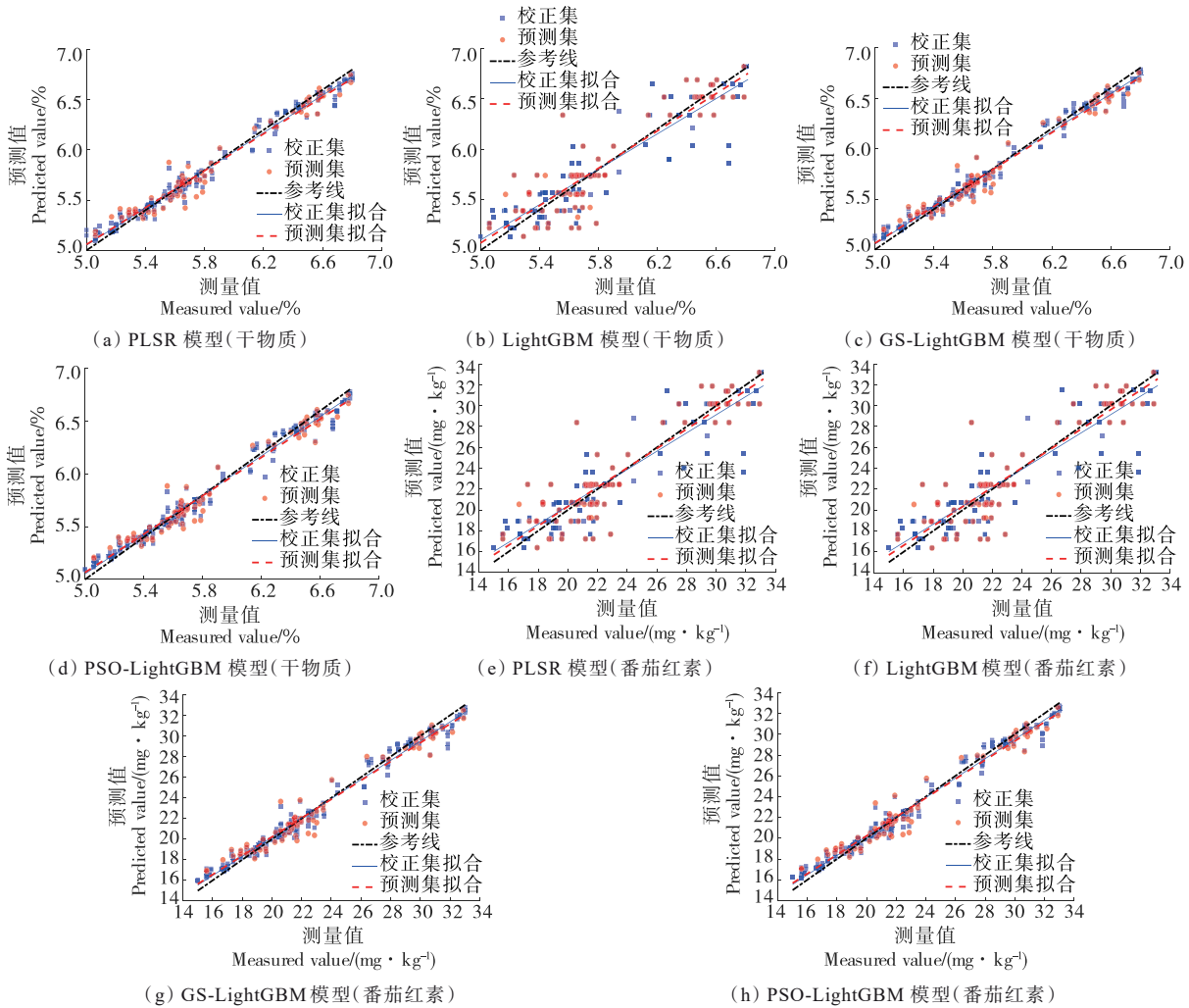


图3 预测值与实测值的拟合图(以干物质和番茄红素为例)

Figure 3 Fitting diagrams of predicted and measured values (taking DM and lycopene as examples)

表7 PSO-LightGBM模型重复性分析结果

Table 7 Results of repeatability analysis for the PSO-LightGBM model

指标	E_{RMSEP}			R_p^2		
	平均值	标准差	变异系数/%	平均值	标准差	变异系数/%
干物质	0.235%	0.004%	1.70	0.980	0.003	0.31
总酸	0.418 g/kg	0.008 g/kg	1.91	0.964	0.006	0.70
可溶性糖	0.442%	0.008%	1.81	0.956	0.007	0.82
维生素C	4.287 mg/kg	0.075 mg/kg	1.75	0.971	0.005	0.57
可溶性固形物	0.271%	0.005%	1.85	0.963	0.004	0.42
番茄红素	11.476 mg/kg	0.227 mg/kg	1.98	0.956	0.007	0.86

(1.70%);同时, R_p^2 的变异系数均小于1.0%,可溶性糖最高也仅为0.82%。上述结果说明,样本划分的随机性对模型预测性能影响较小,模型在预测精度(E_{RMSEP})与解释能力(R_p^2)方面均表现出良好的稳定性和一致性,具备较强的实际应用潜力。

2.5 不同品种与贮藏条件下模型适用性验证

将预测集按番茄品种(红果、黄果)与贮藏时间(0,3,6 d)分组,验证 PSO-LightGBM 模型的适用性,结果如表8和表9所示。不同品种与贮藏条件下,模型的 E_{RMSEP} 与 R_p^2 波动较小, E_{RPDP} 均大于2.0,表明模型对不同品种番茄及贮

表 8 不同品种与贮藏条件下的性能分析(E_{RMSEP})

Table 8 Performance analysis under different varieties and storage conditions (E_{RMSEP})

分组条件	干物质/%	总酸/(g·kg ⁻¹)	可溶性糖/%	维生素 C/(mg·kg ⁻¹)	可溶性固形物/%	番茄红素/(mg·kg ⁻¹)
红果番茄($n=96$)	0.228	0.407	0.429	4.182	0.261	11.175
黄果番茄($n=22$)	0.243	0.429	0.456	4.357	0.278	11.532
贮藏 0 d($n=40$)	0.225	0.398	0.423	4.125	0.257	11.086
贮藏 3 d($n=39$)	0.232	0.415	0.437	4.236	0.265	11.273
贮藏 6 d($n=39$)	0.238	0.421	0.445	4.301	0.271	11.387

表 9 不同品种与贮藏条件下的性能分析(R_p^2)

Table 9 Performance analysis under different varieties and storage conditions (R_p^2)

分组条件	干物质	总酸	可溶性糖	维生素 C	可溶性固形物	番茄红素
红果番茄($n=96$)	0.983	0.969	0.961	0.976	0.966	0.953
黄果番茄($n=22$)	0.979	0.958	0.952	0.967	0.961	0.965
贮藏 0 d($n=40$)	0.984	0.972	0.963	0.978	0.968	0.971
贮藏 3 d($n=39$)	0.981	0.966	0.958	0.973	0.964	0.963
贮藏 6 d($n=39$)	0.978	0.962	0.955	0.969	0.960	0.973

藏过程中的营养成分变化均具有良好的预测能力,鲁棒性较强,可满足多样化应用场景需求。

3 结论

针对传统番茄营养成分检测方法耗时、破坏性强、多成分同步检测精度不足的问题,提出基于近红外光谱结合粒子群优化算法改进轻量梯度提升机的番茄营养成分快速检测方法。通过对 471 份番茄样品的 6 项营养成分(干物质、可溶性糖、总酸、维生素 C、可溶性固形物、番茄红素)进行建模分析,首先在光谱预处理方面,对比 7 种方法后发现标准正态变量变换与一阶导数组合方案最优,信噪比达 42.8,平均决定系数达 0.902,能有效消除噪声与基线漂移;在模型性能方面,粒子群优化算法改进轻量梯度提升机模型对全部 6 项指标的预测精度均优于偏最小二乘回归、未优化轻量梯度提升机及网格搜索优化轻量梯度提升机模型,其中干物质和可溶性固形物的预测集决定系数分别高达 0.982 和 0.965,残差预测偏差分别为 6.892 和 5.347,总酸、可溶性糖、维生素 C 的决定系数均超过 0.95、残差预测偏差均大于 2.5,特别是针对预测难度较大的番茄红素,决定系数提升至 0.956、较未优化轻量梯度提升机提高 12.4%。同时,粒子群优化算法改进轻量梯度提升机模型的预测集均方根误差较偏最小二乘回归模型平均降低 34.7%、较网格搜索优化轻量梯度提升机模型平均降低 15.2%,有效解决了轻量梯度提升机参数随机设定导致的泛化能力不足问题。此外,通过 5 折交叉验证和重复性试验验证,模型预测性能稳定,且在不同番茄品种(红果、黄果)及不同贮藏时间(0,3,6 d)条件下均表现出良好的鲁棒性。所提出的粒子群优化算法改进轻量梯度提升机方法实现了番茄多营养成分的同步精准检测,为

果蔬品质无损检测提供了新方案,可满足食品加工与流通环节的快速检测需求。该研究的不足之处在于模型对番茄红素的预测精度仍低于其他指标,且仅针对单一果蔬品种,未验证模型在其他果蔬中的泛化能力,后续可引入深度学习或集成学习以进一步提升弱信号成分的检测精度,拓展模型至樱桃、西瓜等不同果蔬的品质检测,并开发便携式近红外设备,实现田间实时检测应用。

参考文献

[1] 魏子朝, 卢苗, 雷文晔, 等. 一种融合叶绿素荧光技术与可见—近红外光谱的番茄幼苗热胁迫无损检测方法[J]. 光谱学与光谱分析, 2024, 44(6): 1 613-1 619.
Wei Z C, Lu M, Lei W Y, et al. A nondestructive method combined chlorophyll fluorescence with visible-NIR spectroscopy for detecting the severity of heat stress on tomato seedlings[J]. Spectroscopy and Spectral Analysis, 2024, 44(6): 1 613-1 619.
[2] 吴至境, 刘富强, 李志刚, 等. 基于 DCGAN 数据增强的樱桃番茄可溶性固形物含量光谱检测方法[J]. 食品科学, 2025, 46(2): 214-221.
Wu Z J, Liu F Q, Li Z G, et al. Spectroscopic method for detection of soluble solid content in cherry tomato using deep convolutional generative adversarial network-based data augmentation[J]. Food Science, 2025, 46(2): 214-221.
[3] 黄彦发, 张瑶, 朱彦哲, 等. 基于深度学习的樱桃番茄 γ -氨基丁酸含量无损检测研究[J]. 分析测试学报, 2025, 44(12): 2 554-2 563.
Huang Y F, Zhang Y, Zhu Y Z, et al. A study on non-destructive detection of GABA content in cherry tomatoes based on deep learning[J]. Journal of Instrumental Analysis, 2025, 44(12):

- 2 554-2 563.
- [4] 高燕飞, 徐雪峰, 黄余, 等. 基于物联网和机器学习的番茄内外品质在线检测[J]. 食品与机械, 2025, 41(7): 78-85.
Gao Y F, Xu X F, Huang Y, et al. Online detection of tomato internal and external quality based on IoT and machine learning[J]. Food & Machinery, 2025, 41(7): 78-85.
- [5] 高翔堃, 董璇, 刘超, 等. 近红外光谱联合机器学习测定樱桃番茄中的番茄红素[J]. 量子电子学报, 2025, 42(3): 313-323.
Gao X K, Dong X, Liu C, et al. Determination of lycopene in cherry tomatoes using near infrared spectroscopy combined with machine learning[J]. Chinese Journal of Quantum Electronics, 2025, 42(3): 313-323.
- [6] 田文强, 王巧华, 徐步云, 等. 基于近红外光谱的腌制期咸鸭蛋理化指标无损检测[J]. 食品科学, 2023, 44(2): 319-326.
Tian W Q, Wang Q H, Xu B Y, et al. Non-destructive detection of physical and chemical indicators of salted duck eggs during salting using near-infrared spectroscopy[J]. Food Science, 2023, 44(2): 319-326.
- [7] 王冬, 栾云霞, 王欣然, 等. 近红外光谱无损分析肉类品质的研究进展[J]. 肉类研究, 2024, 38(5): 61-70.
Wang D, Luan Y X, Wang X R, et al. Progress in non-destructive analysis of meat quality by near-infrared spectroscopy[J]. Meat Research, 2024, 38(5): 61-70.
- [8] 康明月, 王成, 孙鸿雁, 等. 基于改进的 WOA-LSSVM 樱桃番茄内部品质检测方法研究[J]. 光谱学与光谱分析, 2023, 43(11): 3 541-3 550.
Kang M Y, Wang C, Sun H Y, et al. Research on internal quality detection method of cherry tomatoes based on improved WOA-LSSVM[J]. Spectroscopy and Spectral Analysis, 2023, 43(11): 3 541-3 550.
- [9] 周宇坤, 陈孝敬, 谢忠好, 等. 一种近红外光谱数据预处理组合优化策略[J]. 光谱学与光谱分析, 2025, 45(1): 52-58.
Zhou Y K, Chen X J, Xie Z H, et al. A combinatorial optimization strategy for near-infrared spectral data preprocessing[J]. Spectroscopy and Spectral Analysis, 2025, 45(1): 52-58.
- [10] 左晓佳, 再努热·吐尔孙. 肉品新鲜度评价及保鲜技术研究进展[J]. 肉类研究, 2023, 37(12): 69-75.
Zuo X J, Tuerxun Z. Research progress in meat freshness evaluation and preservation technologies[J]. Meat Research, 2023, 37(12): 69-75.
- [11] 方瑶, 谢天铎, 郭渭, 等. 基于近红外光谱的金鲟鱼新鲜度快速检测技术[J]. 江苏农业学报, 2021, 37(1): 213-218.
Fang Y, Xie T H, Guo W, et al. Rapid detection technology of pomfret freshness based on near infrared spectroscopy[J]. Jiangsu Journal of Agricultural Sciences, 2021, 37(1): 213-218.
- [12] 陈晶, 温雅君, 张延国, 等. 利用近红外光谱技术快速分析番茄 6 项常规营养成分[J]. 中国蔬菜, 2024(11): 106-113.
Chen J, Wen Y J, Zhang Y G, et al. Rapid analysis of six nutritional components of tomatoes using near-infrared spectroscopy[J]. China Vegetables, 2024(11): 106-113.
- [13] 刘艳群, 肖付刚, 陈彩虹. 基于随机森林回归模型的小麦粉灰分含量快速测定[J]. 食品与机械, 2024, 40(9): 79-83.
Liu Y Q, Xiao F G, Chen C H. Study on rapid determination method of ash content in wheat flour based on stochastic forest regression model[J]. Food & Machinery, 2024, 40(9): 79-83.
- [14] 刘瑜明, 王巧华, 陈远哲, 等. 猪肉理化指标的近红外光谱无损检测[J]. 光谱学与光谱分析, 2024, 44(5): 1 346-1 353.
Liu Y M, Wang Q H, Chen Y Z, et al. Non-destructive near-infrared spectroscopy of physical and chemical indicator of pork meat[J]. Spectroscopy and Spectral Analysis, 2024, 44(5): 1 346-1 353.
- [15] 任智磊, 赵霄霄, 冯景, 等. 基于近红外光谱结合网格搜索—随机森林—自适应提升算法无损检测牛肉新鲜度[J]. 肉类研究, 2025, 39(11): 1-8.
Ren Z L, Zhao X X, Feng J, et al. Non-destructive detection of beef freshness using near infrared spectroscopy combined with grid search-random forest-adaptive boosting algorithm[J]. Meat Research, 2025, 39(11): 1-8.
- [16] 姜小刚, 朱明旺, 姚金良, 等. 基于近红外光谱技术苹果尺寸差异对糖度模型适用性的影响[J]. 华中农业大学学报, 2024, 43(1): 242-248.
Jiang X G, Zhu M W, Yao J L, et al. Effects of apple size on applicability of model for predicting content of sugar based on near infrared spectroscopy[J]. Journal of Huazhong Agricultural University, 2024, 43(1): 242-248.
- [17] Guo M Q, Lin H, Wang K Q, et al. Data fusion of near-infrared and Raman spectroscopy: an innovative tool for non-destructive prediction of the TVB-N content of salmon samples[J]. Food Research International, 2024, 189: 114564.
- [18] Lan C F, Song B W, Zhang L, et al. State prediction of hydroturbine based on WOA-RF-adaboost[J]. Energy Reports, 2022, 8: 13 129-13 137.
- [19] Liu T, Li J L, Xia R, et al. Experimental analysis of coal gangue identification using visible-near infrared spectroscopy under different working conditions with local and global model[J]. International Journal of Coal Preparation and Utilization, 2024, 44(8): 1-22.
- [20] Zhang Y, Zhang C Y, Duan X X, et al. Visual/near-infrared spectroscopy combined with linear discriminant analysis and machine learning for classification of apple damage[J]. Food Science, 2024, 45(22): 255-261.
- [21] 韩东, 孙晓荣, 李建伟. 基于粒子群优化的深度极限学习机建模方法[J]. 控制工程, 2022, 29(4): 721-727.
Han D, Sun X R, Li J W. Deep extreme learning machine improved by particle swarm optimization algorithm[J]. Control Engineering, 2022, 29(4): 721-727.