张晋恒,柏孝燚,伍金凤,等. 基于高光谱和集成学习的人参果维生素 C 含量无损检测方法[J]. 江苏农业学报,2025,41(9):1771-1780.
doi:10.3969/j.issn.1000-4440.2025.09.012

基于高光谱和集成学习的人参果维生素 C 含量无损检测方法

张晋恒[1], 柏孝燚[1], 伍金凤[2], 周 兵[1]
(1.云南农业大学理学院,云南 昆明 650201; 2.云南农业大学食品科学技术学院,云南 昆明 650201)

摘要: 维生素 C 含量是评价人参果品质的重要指标,本研究通过获取人参果的高光谱数据,对人参果维生素 C 含量进行快速无损检测。为有效消除数据噪声的影响,采用移动平均平滑、多元散射校正、一阶导数、最小二乘平滑滤波进行光谱预处理,通过支持向量回归建模并对比预测效果,确定最优光谱预处理方法。针对高光谱数据特征降维的问题,采用竞争性自适应重加权算法、连续投影算法、轻量级梯度提升机算法提取与人参果维生素 C 含量高度相关的特征波长。将选定的特征波长结合支持向量回归、随机森林回归、多层感知机和 Stacking 方法进行建模并对比预测性能,确定最佳预测模型。结果表明,Stacking 方法具有最佳预测性能,其验证集决定系数(R^2)为 0.917 2,均方根误差(*RMSE*)为 15.053,相对预测误差(*RPD*)为3.595 3,该方法能够快速、准确地预测人参果维生素 C 含量,为人参果品质评价和分级分选提供技术支持。

关键词: 人参果; 维生素 C; 高光谱; 机器学习; 集成学习
中图分类号: TP181;TS255.7 **文献标识码:** A **文章编号:** 1000-4440(2025)09-1771-10

A non-destructive detection method for vitamin C content in ginseng fruit based on hyperspectral data and ensemble learning

ZHANG Jinheng[1], BAI Xiaoyi[1], WU Jinfeng[2], ZHOU Bing[1]
(1.*College of Science*, *Yunnan Agricultural University*, *Kunming* 650201, *China*; 2.*College of Food Science and Technology*, *Yunnan Agricultural University*, *Kunming* 650201, *China*)

Abstract: Vitamin C content serves as a crucial index for evaluating the quality of ginseng fruits. In this study, rapid and non-destructive detection of vitamin C content in ginseng fruits was conducted by acquiring their hyperspectral data. To effectively eliminate the impact of data noise, spectral preprocessing methods such as moving average smoothing, multivariate scatter correction, first-order derivative, and least squares smoothing filter were employed. The optimal spectral preprocessing method was determined by comparing the predictive performance of models built using support vector regression. Addressing the issue of dimension reduction in hyperspectral data, competitive adaptive reweighted sampling, successive projections algorithm, and light gradient boosting machine algorithm were utilized to extract feature wavelengths highly correlated with the vitamin C content in ginseng fruits. The selected feature wavelengths were then combined with support vector regression, random forest regression, multi-layer perceptron, and Stacking methods for

收稿日期:2025-01-25
基金项目:云南省重大科技专项(202302AE09002003)
作者简介:张晋恒(1979-),男,云南建水人,硕士,讲师,主要从事高光谱技术在农业食品中的应用研究。(E-mail)zhang_zhw@163.com
通讯作者:周 兵,(E-mail)bingzhoukm@126.com

modeling and comparison to identify the best predictive model. The results indicated that the Stacking method exhibited the best prediction performance, with coefficient of determination (R^2) of 0.917 2, root mean square error (*RMSE*) of 15.053, and relative percent deviation (*RPD*) of 3.595 3 for the validation set. This method enables rapid and accurate prediction of vitamin C content in ginseng fruits, providing technical support for the evaluation, grading, and sorting of ginseng fruit quality.

Key words: ginseng fruit; vitamin C; hyperspectral; machine learning; ensemble learning

人参果是原产于南美洲的茄科类多年生水果和蔬菜两用型作物[1],在中国南方地区分布广泛[2]。人参果果肉清香多汁、风味独特,并且富含人体所需的维生素 C、蛋白质、氨基酸以及多种矿物质元素,营养价值高[3-4]。人参果维生素 C(V_C)含量远高于大多数茄科品种,维生素 C 含量是人参果的重要品质指标之一[5],因此,快速无损检测人参果果实中的维生素 C 含量对于人参果品质评价十分重要。

果蔬内部品质的传统检测方法虽然准确性高[6-7],但破坏性强,且操作复杂、成本高、效率低。近年来,研究人员基于高光谱分析技术快速、无损的特点,广泛开展了果蔬内部营养物质含量的检测研究。班兆军等[8]利用高光谱和深度学习技术预测苹果的糖酸比,通过使用多元散射校正(MSC)和竞争性自适应重加权采样算法(CARS)分别对原始光谱数据进行预处理和特征波长提取后,基于深度学习的定量光谱分析模型预测准确率达0.937 0。Lan 等[9]利用标准正态变量变换(SNV)进行预处理,使用主成分分析算法进行特征波长提取,构建偏最小二乘回归模型预测苹果的干物质含量和总糖含量,准确率分别达到了 0.83 和 0.81。此外,研究者还针对猕猴桃[10]、苹果[11]、桃[12]等的可溶性固形物(SSC)含量进行预测,准确率分别达到0.981 2、0.913 2、0.902 6。为了基于高光谱数据和机器学习方法预测果蔬的维生素 C 含量,He 等[13]基于高光谱数据,利用最小角回归结合最小绝对收缩和选择算子来获取甘薯中维生素 C 的特征波长,确定了 9 个与维生素 C 含量强烈相关的波长,通过偏最小二乘回归达到了良好的预测效果。Fatchurrahman 等[14]对可见光-近红外和近红外高光谱预测枸杞营养品质的模型性能进行比较,其中维生素 C 在400~1 000 nm 波长获得了最佳模型性能,通过应用平滑、对数、一阶导数的组合进行预处理,选择有效波长并去除异常值样本后,提高了校准和预测性能,训练集和验证集的决定系数(R^2)分别为 0.96 和 0.91。上述研究通过比较和优化数据预处理方法、特征提取方法,构建的模型实现了样品维生素 C 含量的预测,但利用高光谱技术针对人参果维生素 C 含量进行无损检测的研究未见报道。

本研究拟以来自云南省石林县的人参果为研究对象,通过试验获得不同成熟度人参果的高光谱和维生素 C 含量数据,使用移动平均平滑(MA)、一阶导数(D1)、多元散射校正(MSC)、最小二乘平滑滤波(SG)4 种预处理方法对原始光谱进行预处理,采用轻量级梯度提升机(LGBM)、连续投影算法(SPA)、竞争性自适应重加权算法(CARS)提取预处理光谱的特征波长,将机器学习中的随机森林(RF)、支持向量回归(SVR)和多层感知机(MLP)作为基础模型,采用 Stacking 融合策略,建立基于集成学习的人参果维生素 C 含量无损检测模型,验证集成模型对人参果维生素 C 含量预测的准确性与有效性。本研究结果可为果蔬维生素 C 含量无损检测提供技术支撑和参考。

1 材料与方法

1.1 试验材料与试验仪器

试验开始于 2024 年 11 月,试验材料来自云南省昆明市石林县人参果种植区(103.513 6° E,24.822 4°N)。选取不同成熟度、无病虫害、无损伤的人参果 120 个,果重为80~140 g。将人参果样品表面擦洗干净并依次编号。

试验试剂:2,6-二氯靛酚钠盐($C_{12}H_6Cl_2NNaO_2$,福州飞净生物有限公司产品)、偏磷酸[$(HPO_3)_n$,山东科源生化有限公司产品]、碳酸氢钠($NaHCO_3$,汕头市西陇化工厂有限公司产品)、高岭土(苏州信清科技有限公司产品)、抗坏血酸(天津市风船化学试剂科技有限公司产品)。

非成像地物光谱仪(FieldSpec ® HandHeld2TM

型),美国 Analytical Spectral Devices 公司产品,光谱探测范围为325~1 075 nm,光谱分辨率为 1 nm;电子天平(FA2104N 型),上海菁海仪器有限公司产品,精度 0.1 mg;粉碎机(M82 型),九阳股份有限公司产品。

1.2 数据采集

使用非成像地物光谱仪按编号依次采集人参果的实际光谱数据,用化学方法测定各编号果实的维生素 C 含量,最后利用采集到的人参果光谱数据和维生素 C 含量实测值进行建模。

1.2.1 *高光谱数据采集与校正* 为确保数据采集的稳定性和准确性,采集光谱数据前将仪器预热 30 min,测量过程中每隔 10 min 采集 1 次暗光谱和参考光谱并重新进行黑白校正。使用设备配套软件 IndicoProVersion3.1 采集光谱,在样品果腹赤道处等间隔采集 5 个点,每个点以采集 5 次的平均光谱结果作为该点的光谱数据,每个样品共获得 5 条平均光谱曲线。通过 ViewSpecPro 软件把采集的光谱数据导入计算机并进行分析,处理后得到400~1 000 nm 波长范围的平均光谱反射强度曲线。按照公式(1)计算果实反射率值[15-16]。

$$R(\lambda)=\frac{I(\lambda)-I_{\text{dark}}(\lambda)}{I_{\text{white}}(\lambda)-I_{\text{dark}}(\lambda)} \tag{1}$$

式中,R 为校正后的果实光谱数据;I 为未经校正的果实原始光谱数据;I_{dark} 为暗光谱;I_{white} 为参考光谱;λ 为波长。

1.2.2 *维生素 C 含量的测定* 参考《食品安全国家标准 食品中抗坏血酸的测定》(GB 5009.86-2016),采用 2,6-二氯靛酚滴定法测定人参果中维生素 C 含量。从人参果赤道面选取果肉 50 g,加入等量偏磷酸溶液后迅速在粉碎机中制成匀浆。精确称量 10.00 g 匀浆样品至烧杯中,用偏磷酸溶液溶解后转移至 100 mL 容量瓶并稀释至刻度线,摇匀后过滤。若滤液有色,可按 1 g 样品添加 0.4 g 高岭土进行脱色处理后再过滤。取 10 mL 滤液置于 50 mL 锥形瓶中,用已标定的 2,6-二氯靛酚溶液进行滴定,直至溶液保持粉红色 15 s 不褪色[17]。维生素 C 含量依据公式(2)[18]计算得出,同时进行空白试验校准。

$$\text{维生素 C 含量(mg/kg)}=\frac{T\times n\times(V-V_0)}{m}\times 1\,000 \tag{2}$$

式中,T:2,6-二氯靛酚溶液滴定度;n:稀释倍数;V:试验滴定消耗的 2,6-二氯靛酚溶液体积,mL;V_0:滴定空白消耗的 2,6-二氯靛酚溶液体积,mL;m:样本重量,g。

1.3 光谱预处理

为了消除或最小化噪声、光散射、基线漂移、仪器误差及环境变化等因素的影响,采集的光谱信息需经过光谱预处理方法进行校正[19]。原始光谱采用移动平均平滑(MA)[20]、多元散射校正(MSC)[21]、一阶导数(D1)[22]、最小二乘平滑滤波(SG)[23]4 种方法进行预处理,选取决定系数(R^2)较大、均方根误差(*RMSE*)较小的方案作为最优方案。

1.4 特征波长的筛选

为优化预处理后的光谱数据,采用竞争性自适应重加权算法(CARS)、连续投影算法(SPA)以及轻量级梯度提升机算法(LGBM)进行特征波长的筛选。CARS 使用蒙特卡洛采样技术,通过迭代过程剔除权重较低的波长点,保留权重绝对值较大的点,以此寻找具有最低交叉验证均方根误差(*RMSECV*)的特征子集,从而确定最佳的特征组合[24]。SPA 依据矢量空间共线性最小化原理,有效地从冗长的光谱数据中精炼出既具代表性又有最小冗余的特征波长[25]。LGBM 模型能够有效捕捉光谱数据中的细微复杂特征,通过量化各特征在模型训练时对损失函数降低的贡献度,评估每个特征对模型效能的提升效果,进而筛选出最具影响力的特征波长[26]。

1.5 模型的建立与评价

集成学习是一种综合多个模型优势形成最优模型的方法,该方法能有效提升模型性能,其核心在于基础模型、元模型的选择和融合策略的设计。本研究对比和分析了 SVR、MLP 和 RF 等常见光谱定量分析模型的优势,其中 SVR 是通过寻找最优超平面来最小化误差,对异常值具有鲁棒性并能有效处理高维空间中的非线性回归问题[27]。MLP 作为一种包含一个或多个隐藏层的前馈神经网络,能够自适应学习并精确拟合复杂的非线性数据[28]。RF 通过聚合各决策树的预测结果来捕捉复杂的非线性关系,同时能有效降低过拟合风险[29]。

因此,本研究基于采集到的人参果维生素 C

含量实测数据和对应的高光谱数据,选用 SVR 和 MLP 为基础模型, RF 作为元模型,采用 Stacking 集成学习策略,构建人参果维生素 C 含量无损检测模型,并与单一模型进行比较,以决定系数(R^2)、均方根误差(*RMSE*)和相对预测误差(*RPD*)作为评价指标,评估集成模型性能。评价指标计算方法如公式(3)、公式(4)、公式(5)所示:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (4)$$

$$RPD = (\text{标准差}/\text{均方根误差}) \times 100\% \quad (5)$$

式中,$\bar{y}_i$代表人参果维生素 C 含量实测值的平均值;y_i代表人参果维生素 C 含量的实测值;$\hat{y}_i$代表人参果维生素 C 含量的预测值;n 为人参果样本数。其中,R^2越接近 1、*RMSE* 越小、*RPD* 越大表明模型性能越优[30]。

2 结果与分析

2.1 人参果原始光谱曲线

人参果原始光谱图像如图 1 所示。光谱整体曲线趋势基本一致,450~580 nm 区间光谱反射率急剧升高,与果实表面色素吸收性质变化有关[31];581~940 nm 区间变化趋缓,在 740 nm 附近有较小吸收峰;在941~980 nm 区间反射率下降,其中在 960 nm 波长附近有明显的吸收峰,该特征峰与人参果中的含水量有关,可能与 O-H 键伸缩振动引起光谱反射率下降有关[32-34]。

2.2 维生素 C 含量的统计描述

采用 2,6-二氯靛酚滴定法测定 120 个人参果样品的维生素 C 含量,测定结果见表 1。采用基于联合 X-Y 距离的样本划分算法(SPXY)[35]将人参果维生素 C 含量数据以4∶1 的比例划分为校正集和验证集,其中校正集人参果维生素 C 含量数据 96 个,验证集人参果维生素 C 含量数据 24 个,分别统计分析校正集和验证集的维生素 C 含量特征。由表 2 可知,3 组样本的维生素 C 含量变化范围相对一致,用变异系数评估数据集的离散程度[36],本次划分的数据集变异系数大小为中等(变异系数为10.00%~30.00%表示中等程度的变异,小于 30.00%表明适合用于建模;小于 10.00%为较好;大于 30.00%为较差)且数据集之间差异不大,表明数据集样本划分一致且数据分布合理,适合用于建模。

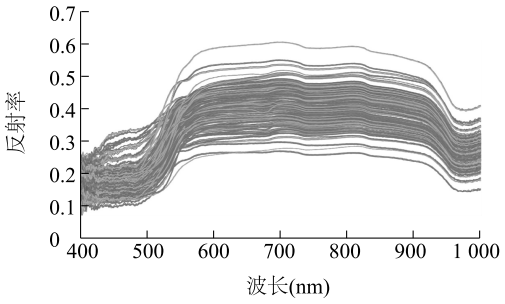


图 1 人参果原始光谱曲线

Fig.1 Original spectral curves of ginseng fruit

2.3 光谱数据预处理

采用支持向量回归(SVR)分别对 MA、D1、MSC、SG 预处理后的光谱数据进行建模对比,以验证集决定系数(R^2)、*RMSE* 和 *RPD* 来确定最佳预处理方法,得到不同预处理方法的维生素 C 含量预测建模结果。结果(表 3)表明,MSC 建模的维生素 C 含量预测模型精度较高,并优于原始光谱,验证集 R^2 为0.838 0,*RMSE* 为19.810,*RPD* 为2.484 8,光谱反射率和果实维生素 C 含量的相关性显著增强,预处理效果最佳。从预处理后光谱图像(图 2)可以看出,MSC 处理后的光谱曲线更加平滑,有效抑制了噪声并校正了散射效应。

2.4 特征波长提取

2.4.1 *使用 CARS 提取特征波长* 使用 CARS 筛选特征波长,结果如图 3 所示。设置蒙特卡洛采样次数为 50 次,单次采样比例为总数的 70%。全光谱数据维度 601 个,当采样次数为 19 次时,交叉验证均方根误差(*RMSE*)最小,对应特征变量数为 66 个,表明之前的采样运算中,大量光谱数据中与预测维生素 C 含量无关的信息被去除,之后随着采样次数的增加,模型性能开始变差。

2.4.2 *使用 SPA 提取特征波长* 设置特征变量个数选择范围为1~50 个,使用 SPA 对光谱数据进行特征波长筛选,结果如图 4 所示,当特征变量数量为 30 个时,交叉验证均方根误差(*RMSE*)最小,此时提取的 30 个特征波长为最终提取结果。

表 1 人参果样本维生素 C 含量

Table 1 Vitamin C content values in ginseng fruit samples

样本编号	维生素 C 含量 (mg/kg)	样本编号	维生素 C 含量 (mg/kg)	样本编号	维生素 C 含量 (mg/kg)	样本编号	维生素 C 含量 (mg/kg)
1	199.9	31	107.7	61	233.2	91	241.6
2	188.2	32	146.1	62	216.6	92	166.6
3	182.3	33	146.1	63	183.3	93	133.3
4	147.0	34	100.0	64	191.6	94	199.9
5	158.8	35	123.0	65	208.3	95	183.3
6	170.5	36	146.1	66	258.2	96	149.9
7	152.9	37	161.5	67	274.9	97	158.3
8	107.7	38	153.8	68	191.6	98	224.9
9	184.6	39	169.2	69	249.9	99	199.9
10	100.0	40	192.3	70	283.2	100	191.6
11	92.3	41	183.3	71	291.6	101	216.6
12	84.6	42	183.3	72	208.3	102	233.2
13	123.0	43	291.6	73	183.3	103	166.6
14	176.9	44	216.6	74	224.9	104	199.9
15	100.0	45	199.9	75	208.3	105	166.6
16	184.6	46	216.6	76	216.6	106	158.3
17	92.3	47	224.9	77	274.9	107	133.3
18	107.7	48	208.3	78	224.9	108	174.9
19	123.0	49	299.9	79	233.2	109	116.6
20	84.6	50	233.2	80	208.3	110	208.3
21	176.9	51	266.6	81	116.6	111	149.9
22	138.4	52	241.6	82	158.3	112	174.9
23	299.9	53	224.9	83	291.6	113	183.3
24	84.6	54	208.3	84	174.9	114	191.6
25	169.2	55	299.9	85	233.2	115	258.2
26	107.7	56	191.6	86	224.9	116	274.9
27	161.5	57	233.2	87	133.3	117	283.2
28	123.0	58	249.9	88	208.3	118	224.9
29	146.1	59	199.9	89	149.9	119	216.6
30	123.0	60	266.6	90	249.9	120	174.9

表 2 人参果维生素 C 含量统计特征

Table 2 Statistical characteristics of V_C content in ginseng fruit

样本	数量 (个)	最大值 (mg/kg)	最小值 (mg/kg)	平均值 (mg/kg)	标准差 (mg/kg)	变异系数 (%)
总数据	120	299.9	84.6	188.7	53.8	28.51
校正集	96	299.9	84.6	183.9	53.5	29.09
验证集	24	299.9	100.0	207.7	50.8	24.46

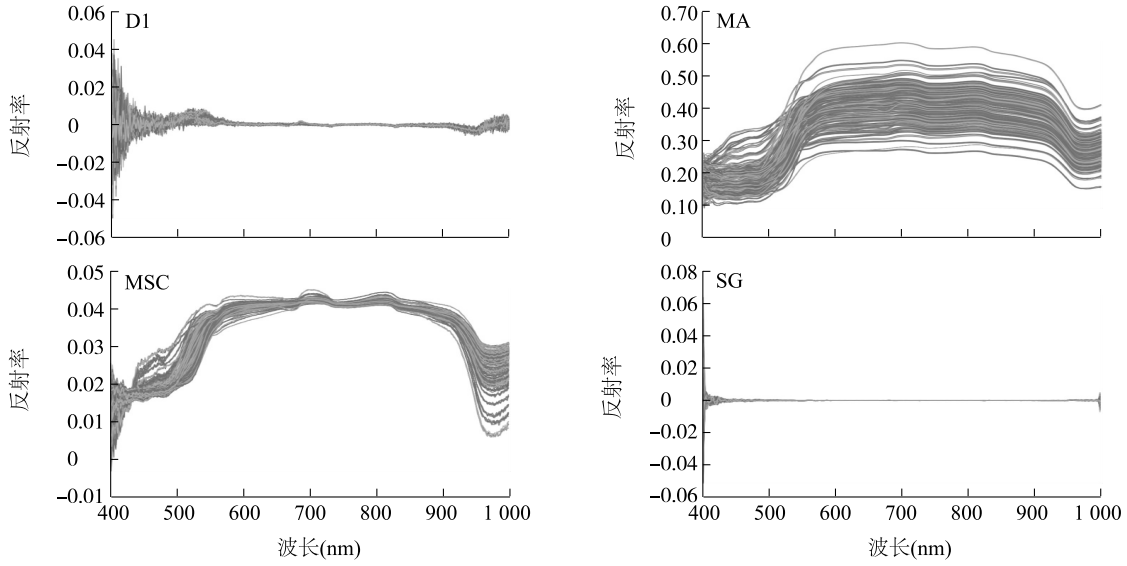
变异系数=(标准差/平均值)×100%。

表 3 人参果维生素 C 含量不同预处理方法与支持向量回归(SVR)建模结果

Table 3 Modeling results of V_C content in ginseng fruit by different pretreatment methods with support vector regression (SVR)

预处理方法	校正集		验证集		*RPD*
	R^2	*RMSE*	R^2	*RMSE*	
原始光谱	0.884 2	18.663	0.809 1	21.505	2.288 9
MA	0.884 3	18.653	0.821 9	20.776	2.369 3
D1	0.616 7	33.955	0.073 0	47.393	1.038 6
MSC	0.892 2	18.006	0.838 0	19.810	2.484 8
SG	0.580 9	35.504	−0.000 7	49.242	0.999 6

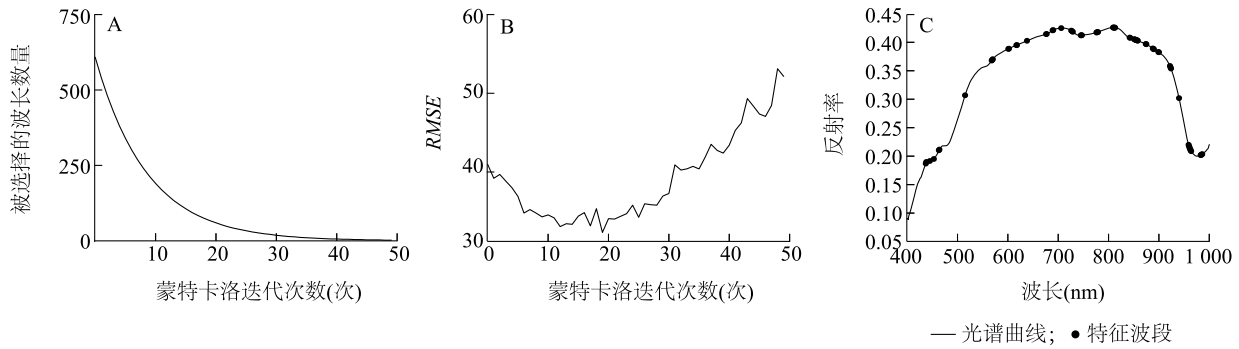
MA:移动平均平滑;D1:一阶导数;MSC:多元散射校正;SG:最小二乘平滑滤波。R^2:决定系数;*RMSE*:均方根误差;*RPD*:相对预测误差。



D1:一阶导数;MA:移动平均平滑;MSC:多元散射校正;SG:最小二乘平滑滤波。

图 2 不同方法预处理后的光谱图像

Fig.2 Spectral images after preprocessing with different methods

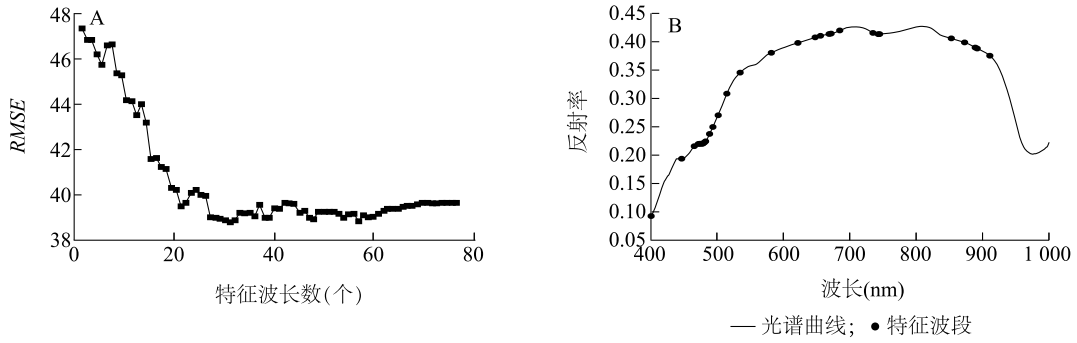


A:被选择的波长数量随蒙特卡洛迭代次数增加的变化情况;B:交叉验证均方根误差(*RMSE*)随蒙特卡洛迭代次数增加的变化情况;C:特征波长选择结果。

图 3 使用竞争性自适应重加权采样算法(CARS)选择特征波长

Fig.3 Feature wavelength selection by competitive adaptive reweighted sampling (CARS)

2.4.3 使用 LGBM 算法提取特征波长 使用 LGBM 算法内置的特征重要性评估方法进行特征波长筛



A:交叉验证均方根误差(*RMSE*)随特征波长提取数量增加的变化情况;B:特征波长选择结果。

图 4 基于连续投影算法(SPA)的特征波长选择

Fig.4 Feature wavelength selection by successive projections algorithm (SPA)

选,先计算400~1 000 nm 内每个波段在决策树构建过程中对分裂增益的贡献,分裂增益贡献越大的特征,重要性越高,然后按照重要性排序,筛选出前 50 个特征波段,这些特征波段的累积增益已达 95%,表明所提取的特征波段包含了大部分有效信息。全波段的增益重要性得分分布如图 5 所示,LGBM 算法提取特征显著的光谱波段主要集中在 400~500 nm 和 700~900 nm。增益重要性得分从高到低排名前 50 的特征波段依次为 485 nm、472 nm、811 nm、467 nm、719 nm、415 nm、734 nm、540 nm、958 nm、830 nm、953 nm、589 nm、602 nm、671 nm、815 nm、405 nm、545 nm、621 nm、650 nm、440 nm、718 nm、829 nm、622 nm、547 nm、435 nm、828 nm、698 nm、701 nm、947 nm、450 nm、826 nm、543 nm、400 nm、735 nm、441 nm、860 nm、471 nm、823 nm、687 nm、584 nm、552 nm、560 nm、613 nm、559 nm、673 nm、855 nm、601 nm、733 nm、588 nm、663 nm。

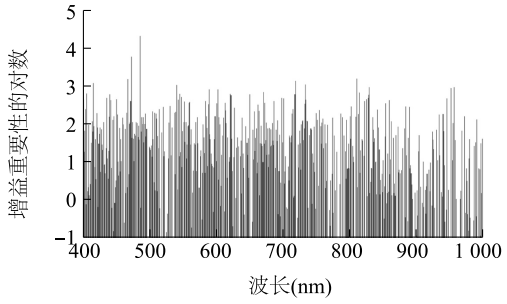


图 5 特征波长增益重要性得分分布

Fig. 5 Gain-based importance score distribution of feature wavelengths

2.4.4 特征波长提取效果与分析 比较使用全波段光谱数据和 3 种特征提取方法处理后的光谱数据所构建的 SVR 模型性能,结果如表 4 所示,LGBM-SVR 模型预测效果最佳,SPA-SVR 模型预测效果最差。SPA 模型提取特征变量为 30 个时,校正集和验证集的决定系数均为最低,均方根误差均为最高,表明特征提取过程丢失的高度相关区域较多,导致模型预测效果较差。LGBM 模型提取的特征变量为 50 个时,校正集和验证集的决定系数均为最高值,均方根误差均为最低值,*RPD* 达到2.881 7,表明该模型能够有效识别并保留关键特征波长,减少冗余信息,更适合作为人参果光谱数据的特征提取方法,为后续的建模提供高质量的数据。

2.5 预测模型的构建

将经过 LGBM 处理后的光谱数据作为输入变量,以化学法测定得到的人参果维生素 C 含量为标签建立光谱数据集,采用 SPXY 算法将数据集划分为校正集和验证集后,分别建立 RF、SVR、MLP 和 Stacking 4 种预测模型,预测人参果维生素 C 含量,预测效果评价如表 5 所示。根据验证集决定系数(R^2)从高到低排序依次为 Stacking、MLP、SVR、RF;根据验证集 *RMSE* 从低到高排序依次为 Stacking、MLP、SVR、RF;根据 *RPD* 从高到低排序依次为 Stacking、MLP、RF、SVR。

从建模结果(表 5)可以看出,Stacking 模型预测效果最好,验证集 R^2、*RPD* 均为所有模型最高值,分别为0.917 2、3.595 3,*RMSE* 最低,为 15.053。随机森林(RF)模型的预测效果最差,验证集 R^2最低,为 0.878 6,原因可能与光谱数据样本数量少且特征复杂有关,证明随机森林(RF)模型对小样本数据的学习能力较强而预测能力较弱,这一结果与其他研究者的结果[37-38]一致。

表 4 不同特征波长提取方法的支持向量回归(SVR)模型性能对比

Table 4 Comparison of support vector regression (SVR) models based on characteristic wavelengths extracted by different extraction methods

模型	特征数	校正集		验证集		*RPD*
		R^2	*RMSE*	R^2	*RMSE*	
SVR	601	0.892 2	18.006	0.838 0	19.810	2.484 8
CARS-SVR	66	0.932 8	14.029	0.852 1	20.114	2.600 1
SPA-SVR	30	0.837 7	21.086	0.736 2	26.863	1.946 9
LGBM-SVR	50	0.954 5	11.542	0.879 6	18.149	2.881 7

SVR:支持向量回归;CARS:竞争性自适应重加权算法;SPA:连续投影算法; LGBM:轻量级梯度提升机算法;R^2:决定系数;*RMSE*:均方根误差;*RPD*:相对预测误差。

表 5 不同预测模型结果对比

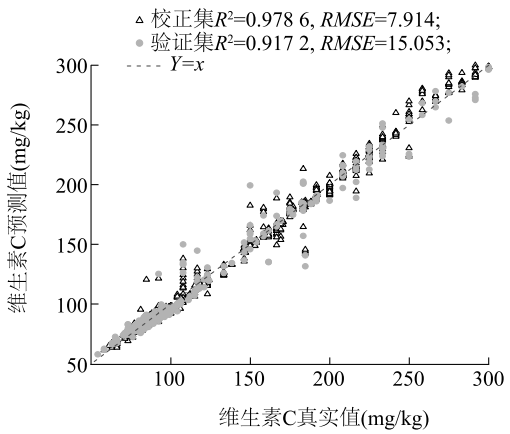
Table 5 Comparison of different prediction models

模型	校正集		验证集		*RPD*
	R^2	*RMSE*	R^2	*RMSE*	
RF	0.983 8	6.888	0.878 6	18.222	2.970 2
SVR	0.954 5	11.542	0.879 6	18.149	2.881 7
MLP	0.970 3	9.333	0.904 8	16.138	3.353 8
Stacking	0.978 6	7.914	0.917 2	15.053	3.595 3

R^2:决定系数;*RMSE*:均方根误差;*RPD*:相对预测误差。RF:随机森林;SVR:支持向量回归;MLP:多层感知机。

2.6 最优模型的预测效果与分析

综合数据预处理、特征波长提取和建模的研究结果可知,在预测人参果维生素 C 含量时,采用 MSC-LGBM-Stacking 方式构建的模型表现最好。该模型校正集的 R^2 和 *RMSE* 分别为0.978 6和7.914,验证集的 R^2、*RMSE* 和 *RPD* 分别为0.917 2、15.053 和3.595 3。验证集表现出更高的精度和稳定性,R^2 和 *RPD* 相较于其他 3 种单一模型(MLP、SVR、RF)平均分别提高了 3.2%和 14.6%,*RMSE* 平均低 14.7%。图 6 为 MSC-LGBM-Stacking 模型的拟合散点图,直观展示最优模型预测值和真实值之间的接近程度。从拟合情况来看,模型总体拟合效果较好,在低含量区域的数据点更贴近 $Y=x$ 线,预测更准确;在高含量区域数据点虽然略有分散,但整体上仍保持较好的一致性。这一结果表明,MSC-LGBM-Stacking 集成学习模型有效融合了 MLP、SVR、RF 的优势,能够利用光谱中复杂的成分信息预测人参果的维生素 C 含量。



R^2:决定系数;*RMSE*:均方根误差。

图 6 人参果维生素 C 含量预测最优模型性能

Fig.6 Performance of the optimal model for predicting vitamin C content in ginseng fruit

3 结 论

以云南石林人参果为研究对象,采集不同成熟度人参果样品的高光谱数据,通过光谱数据预处理、特征提取,获取降噪后的最优波段组合,进而构建人参果维生素 C 含量预测模型并进行比较,结论如下:

(1)比较 4 种预处理方法(MA、D1、MSC、SG)对预测结果的影响,确定 MSC 为最佳预处理方法,该方法有效消除了散射效应所导致的光谱基线漂移和幅度变化,显著提高了光谱数据的准确性和可靠性,校正集和验证集的 R^2 分别为0.892 2和0.838 0。

(2)对比分析全波段 SVR 模型和 3 种特征波长提取方法(CARS、SPA、LGBM)提取的特征波长所构建的 SVR 模型的效果可以看出,LGBM 算法提取的特征波长 SVR 模型对人参果维生素 C 含量预测的准确性优于其他方法,该方法有效提取了对分裂增益贡献较大的特征波段,校正集和验证集的 R^2 分别为0.954 5和0.879 6。

(3)基于上述 2 点结论,对比分析 4 种预测模型,Stacking 方法校正集和验证集的 R^2 分别为0.978 6和0.917 2,验证集 *RMSE* 最小,*RPD* 最大,分

别为 15.053 和3.595 3,具有相对更高的精度和稳定性,更好的泛化能力。

综上,由 MSC-LGBM-Stacking 构建的人参果维生素 C 含量预测模型可以更好地预测人参果维生素 C 含量,本研究结果可为果蔬维生素 C 含量无损检测提供技术支撑和参考。

参考文献:

[1] RODRÍGUEZ-BURRUEZO A, KOLLMANNSBERGER H, PROHENS J, et al. Analysis of the volatile aroma constituents of parental and hybrid clones of pepino (*Solanum muricatum*)[J]. Journal of Agricultural and Food Chemistry,2004,52(18):5663-5669.

[2] 王 琰,张文刚,杨希娟,等. 人参果片干制方式及其品质特性研究[J]. 食品工业,2024,45(10):22-28.

[3] 杜丽娟,黄兴龙,毕亚楠,等. 石林人参果的品质差异及综合性评价[J]. 食品安全质量检测学报,2023,14(23):107-114.

[4] 张子琛,王玉英,张晚秋,等. 8 个人参果品种(系)的果实品质评价[J]. 热带作物学报,2024,45(3):524-532.

[5] 杨世鹏,蒋晓婷,许盼盼,等. 人参果营养成分、采后生理及贮藏保鲜方式研究进展[J]. 西北农业学报,2020,29(10):1447-1456.

[6] 罗 弦,王晓莉,罗 娅,等.基于主成分分析的海藻肥对枇杷果实品质影响的综合评价[J].四川农业大学学报,2024,42(6):1212-1219.

[7] 徐朝阳. 2,6-二氯酚靛酚滴定法与碘量法测定蔬菜水果中维生素 C 方法的准确度比较[J]. 食品安全导刊,2021(25):100-101.

[8] 班兆军,高喧翔,马肄恒,等. 基于高光谱和深度学习的苹果品质无损检测方法[J]. 江苏农业学报,2024,40(8):1446-1454.

[9] LAN W J, JAILLAIS B, RENARD C M G C, et al. A method using near infrared hyperspectral imaging to highlight the internal quality of apple fruit slices[J]. Postharvest Biology and Technology,2021,175:111497.

[10] ZHU H Y, CHU B Q, FAN Y Y, et al. Hyperspectral imaging for predicting the internal quality of kiwifruits based on variable selection algorithms and chemometric models[J]. Scientific Reports, 2017,7(1):7845.

[11] TIAN X, LI J B, WANG Q Y, et al. A multi-region combined model for non-destructive prediction of soluble solids content in apple,based on brightness grade segmentation of hyperspectral imaging[J]. Biosystems Engineering,2019,183:110-120.

[12] PAN L Q, ZHANG Q, ZHANG W, et al. Detection of cold injury in peaches by hyperspectral reflectance imaging and artificial neural network[J]. Food Chemistry,2016,192:134-141.

[13] HE H J, ZHANG C, BIAN X H, et al. Improved prediction of vitamin C and reducing sugar content in sweetpotatoes using hyperspectral imaging and LARS-enhanced LASSO variable selection [J]. Journal of Food Composition and Analysis, 2024, 132: 106350.

[14] FATCHURRAHMAN D, NOSRATI M, AMODIO M L, et al. Comparison performance of visible-NIR and near-infrared hyperspectral imaging for prediction of nutritional quality of goji berry (*Lycium barbarum* L.)[J]. Foods,2021,10(7):1676.

[15] 张 伏,曹炜桦,崔夏华,等. 基于 SG-CARS-IBP 的圣女果可溶性固形物可见/近红外光谱无损检测[J]. 光谱学与光谱分析,2023,43(3):737-743.

[16] 袁旭林. 基于高光谱成像技术的苹果糖度无损检测系统研究[D]. 济南:山东大学,2021.

[17] 高梦蕾. 三种赏食两用植物的无土栽培技术研究[D]. 哈尔滨:东北农业大学,2018.

[18] 郭林鸽,殷 勇,于慧春,等. 基于 Fisher 判别分析可分性信息融合的马铃薯 VC 含量高光谱检测方法[J]. 食品科学,2024,45(7):164-171.

[19] MISHRA P, KARAMI A, NORDON A, et al. Automatic de-noising of close-range hyperspectral images with a wavelength-specific shearlet-based image noise reduction method[J]. Sensors and Actuators B:Chemical,2019,281:1034-1044.

[20] 李奇辰,李民赞,杨 玮,等. 基于拉曼光谱的水溶性磷定量分析[J]. 光谱学与光谱分析,2023,43(12):3871-3876.

[21] GUO W, LI X X, XIE T H. Method and system for nondestructive detection of freshness in *Penaeus vannamei* based on hyperspectral technology[J]. Aquaculture,2021,538:736512.

[22] DÖPPER V, ROCHA A D, BERGER K, et al. Estimating soil moisture content under grassland with hyperspectral data using radiative transfer modelling and machine learning[J]. International Journal of Applied Earth Observation and Geoinformation,2022, 110:102817.

[23] MA L, ZHANG Y, ZHANG Y Y, et al. Rapid nondestructive detection of chlorophyll content in muskmelon leaves under different light quality treatments[J]. Agronomy,2022,12(12):3223.

[24] LI H D, LIANG Y Z, XU Q S, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration[J]. Analytica Chimica Acta,2009,648(1):77-84.

[25] ARAUJO M C U, SALDANHA T C B, GALVAO R K H, et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis[J]. Chemometrics and Intelligent Laboratory Systems,2001,57(2):65-73.

[26] 吴继忠,时艺丹,黄 慧,等. 基于改进堆叠自编码器结合 LightGBM 的近红外光谱回归算法研究[J]. 分析测试学报,2023,42(9):1112-1118.

[27] DRUCKER H, BURGES C J C, KAUFMAN L, et al. Support vector regression machines[C]//MOZER M C, JORDAN M, PETSCHE T. Advances in Neural Information Processing Systems 9. Cambridge:MIT Press,1996:155-161.

[28] 张 驰,郭 媛,黎 明. 人工神经网络模型发展及应用综述[J]. 计算机工程与应用,2021,57(11):57-69.

[29] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45

(1):5-32.

[30] VISCARRA ROSSEL R A, MCGLYNN R N, MCBRATNEY A B. Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy[J]. Geoderma,2006,137(1/2):70-82.

[31] 刘子涵,李 明,赵峙尧,等. 基于高光谱成像技术和机器学习的猕猴桃果实可溶性固形物含量预测[J]. 果树学报,2024,41(12):2606-2620.

[32] MISHRA P, WOLTERING E, BROUWER B, et al. Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with variable selection and model updating approach[J]. Postharvest Biology and Technology,2021,171:111348.

[33] LI X L, WEI Y Z, XU J, et al. SSC and pH for sweet assessment and maturity classification of harvested cherry fruit based on NIR hyperspectral imaging technology[J]. Postharvest Biology and Technology,2018,143:112-118.

[34] 郑艺蕾. 基于高光谱和太赫兹光谱的甘薯品质检测方法研究[D]. 南昌:华东交通大学,2020.

[35] 王世芳,韩 平,崔广禄,等. SPXY 算法的西瓜可溶性固形物近红外光谱检测[J]. 光谱学与光谱分析,2019,39(3):738.

[36] 宋子怡,常庆瑞,郑智康,等. 基于高光谱和连续投影算法的猕猴桃叶片氮平衡指数的估测[J]. 江苏农业学报,2024,40(7):1260-1267.

[37] 陶惠林,冯海宽,杨贵军,等. 基于无人机数码影像和高光谱数据的冬小麦产量估算对比[J]. 农业工程学报,2019,35(23):111-118.

[38] 段丹丹,刘仲华,赵春江,等. 基于特征光谱参数的叶片和冠层尺度茶多酚含量估算[J]. 光谱学与光谱分析,2024,44(3):814-820.

(责任编辑:陈海霞)