



冷君娇,许亚男,董学会,孙群. 基于高光谱数据结合域适应算法的北柴胡种子跨批次真实性鉴定[J]. 中国农业大学学报,2025,30(10):286-298.

LENG Junjiao, XU Yanan, DONG Xuehui, SUN Qun. Cross-batch authenticity identification of *Bupleurum chinense* DC. seeds based on hyperspectral data combined with domain adaptation algorithms[J]. *Journal of China Agricultural University*, 2025, 30(10): 286-298.

DOI: 10.11841/j.issn.1007-4333.2025.10.22

基于高光谱数据结合域适应算法的北柴胡种子跨批次 真实性鉴定

冷君娇^{1,2} 许亚男^{1,2} 董学会^{1,2} 孙群^{1,2*}

(1. 中国农业大学 农学院, 北京 100193;

2. 中国农业大学 中药材研究中心/农业农村部农作物种子全程技术研究北京创新中心, 北京 100193)

摘要 为构建基于高光谱数据和域适应算法的不同批次北柴胡种子跨年份精准鉴定模型,采集10个批次北柴胡(*Bupleurum chinense* DC.)及5种近缘物种400~1 000 nm波段的高光谱数据,利用随机森林(RF)、支持向量机(SVM)、多层感知器(MLP)、偏最小二乘判别分析(PLS-DA)及集成学习算法建立北柴胡种子跨批次鉴定模型,通过竞争性自适应权重(CARS)结合协方差对齐(CORAL)算法探索提高模型对不同批次种子的跨年份数据预测能力的可行性,最后使用SHAP算法对模型进行可解释性分析。结果表明:SNV-PLS-DA模型的预测准确率为85.6%。采用CARS算法筛选出97个关键特征波段,构建的SNV-CARS-PLS-DA模型预测准确率为91.4%。2024和2025年采集同一批次(2024年10月)收获的万荣柴胡的跨年份高光谱数据间存在以线性偏移为主导的混合偏移(集中于830~900 nm)。经CORAL算法校准后,模型预测2024年采集的光谱数据准确率从79.7%提升到94.3%。SHAP算法表明673.5~746.6、810.6~842.7和907.8~984.7 nm是模型预测的关键波段。综上,CORAL算法适用于校正北柴胡不同批次跨年份光谱数据的偏移,能够提高模型预测的鲁棒性,可用于不同批次北柴胡种子跨年份精准鉴定。

关键词 北柴胡; 真实性; 高光谱; 协方差对齐

中图分类号 R282.5

文章编号 1007-4333(2025)10-0286-13

文献标志码 A

Cross-batch authenticity identification of *Bupleurum chinense* DC. seeds based on hyperspectral data combined with domain adaptation algorithms

LENG Junjiao^{1,2}, XU Yanan^{1,2}, DONG Xuehui^{1,2}, SUN Qun^{1,2*}

(1. College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China;

2. Chinese Medicinal Herbs Research Center/The Innovation Center (Beijing) of Crop Seeds Whole-Process Technology

Research of Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100193, China)

Abstract To construct a cross-year accurate identification model for different batches of *Bupleurum chinense* DC. seeds based on hyperspectral data and domain adaptation algorithms, hyperspectral data in the 400 to 1 000 nm band were collected from 10 batches of *B. chinense* DC. and 5 related species. Machine learning models for cross-batch discrimination were established using random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), partial least squares discriminant analysis (PLS-DA), and ensemble

收稿日期: 2025-07-14

基金项目: 新疆维吾尔自治区重点研发计划(2024B040039)

第一作者: 冷君娇(ORCID:0009-0004-9735-6181),硕士研究生,E-mail:s20233010116@cau.edu.cn

通讯作者: 孙群(ORCID:0000-0002-1303-0431),副教授,主要从事种子活力无损检测研究,E-mail:sunqun@cau.edu.cn

learning algorithms. The feasibility of improving the model's prediction ability on the cross-year data of different seed batches was explored by combining the competitive adaptive reweighted sampling (CARS) with CORrelation ALignment (CORAL) algorithm. Finally, the SHAP algorithm was used to conduct interpretability analysis on the model. The results showed that: The prediction accuracy of the SNV-PLS-DA model reached 85.6%. Using the CARS algorithm, a total of 97 key characteristic bands were identified, and the prediction accuracy of the constructed SNV-CARS-PLS-DA model was 91.4%. There existed a mixed shift dominated by linear shift (concentrated within 830 to 900 nm) between the hyperspectral data collected in 2024 and 2025 for the same batch of Wanrong *Bupleurum* harvested in October 2024. After calibration by the CORAL algorithm, the accuracy of the model in predicting the spectral data in 2024 increased from 79.7% to 94.3%. The SHAP algorithm indicates that the bands of 673.5-746.6, 810.6-842.7, and 907.8-984.7 nm were the key bands for the model prediction. In conclusion, the CORAL algorithm is suitable for correcting the shift of spectral data of cross-year for different batches of *Bupleurum chinense* DC. seeds, and can improve the robustness of model prediction.

Keywords *Bupleurum chinense* DC.; authenticity; hyperspectral; CORrelation ALignment

柴胡是我国常用大宗中药材,具有解热、抗炎及免疫调节等作用^[1]。依据《中国药典》(2015版)^[1],其法定基原为北柴胡(*Bupleurum chinense* DC.)和狭叶柴胡(*Bupleurum scorzonerifolium* Willd.)的干燥根。目前柴胡主要依赖种子繁殖,但面临种源混杂、种子活力低以及种子寿命短等问题^[2]。尽管部分地区已建立柴胡种子质量分级体系,如内蒙古 DB14/T 1491—2017《柴胡种子质量标准》^[3]、山西 DB15/T 1296—2017《北柴胡种子质量分级》^[4],实际生产中仍普遍存在陈种翻新、净度不达标等问题^[5],严重制约柴胡药材的标准化生产。种子真实性是中药材质量控制的基础,核心在于判定供检种子与其标注信息的一致性^[6]。中药材种源混杂会导致中药药性改变,引发用药安全风险。但传统鉴定方法(如表型性状、显微观察、理化分析及分子鉴定^[7])存在依赖专业经验且破坏样品的局限性。因此,北柴胡真实性快速无损的鉴定,对于提高北柴胡药材质量、保障用药安全具有重要的意义。

高光谱成像技术的快速发展突破了中药材传统鉴定方式的局限,通过融合图像与光谱信息,同步无损获取样本外部形态特征与内部化学成分信息^[8],已在农业、环境、医学等多领域广泛应用^[9]。大量研究证实,高光谱成像技术结合机器学习算法在农作物种子,如玉米^[10]、小麦^[11]、大豆^[12]等品种、种质鉴定方面具有巨大潜力。Huang 等^[13]基于跨年份玉米高光谱数据建立的更新式 LSSVM 模型,鉴定准确率达 94.4%;Wang 等^[14]在甜玉米种子分类中使用 CNN-LSTM 模型,准确率超 95.0%。现有高光谱种子真实性鉴定研究多基于同时期采集不同批次数据^[10,13],对不同批次跨年份数据的适应性研究不足。以北柴胡为例,其与近缘

种三岛柴胡(*Bupleurum falcatum*)形态及光谱特征相似,且跨年份数据因仪器老化、环境因素及种子自然老化等干扰,导致基于单年采集数据建立的模型应用到同一批次种子的次年数据适用性降低。尽管领域适应算法,如迁移成分分析(TCA)、协方差对齐(CORAL)已在农业跨场景数据对齐中展现潜力,例如用不同近红外仪的玉米检测^[15]及近红外定量建模^[16],但主要应用于跨仪器校准,中药材种子不同批次跨年份光谱偏移问题未得到解决。TCA 基于最大均值差异(MMD)将不同年份数据映射至再生核希尔伯特空间(RKHS),达到减小数据分布差异的目的^[17]。CORAL 通过对齐协方差矩阵,有效缓解时间漂移导致的模型性能衰退^[18]。

目前基于光谱数据建立北柴胡真实性鉴定模型并使用域适应算法提高模型跨批次跨年份数据预测鲁棒性的研究鲜有报道。本研究基于高光谱数据建立北柴胡真实性鉴定模型,并通过域适应算法提高模型对于跨批次跨年份数据预测的准确率,结合 SHAP 算法为模型提供可解释性分析,旨在构建不同批次的北柴胡种子跨年份精准鉴定模型,以为中药材种子真实性鉴定提供参考。

1 材料与方法

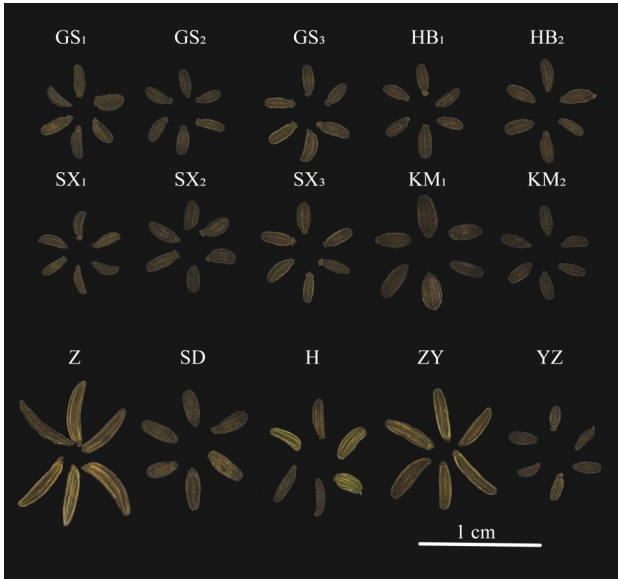
1.1 试验材料

10 个批次北柴胡及 5 个批次近缘种柴胡(表 1),均收集于 2023 年(图 1)。其中,北柴胡与非北柴胡样品各 600 粒。另设独立外部测试集,含 2024 年采自山西省运城市万荣县西乡村的北柴胡 300 粒(后称万荣柴胡)。所有建模材料收集于河北

表1 北柴胡种子真实性建模样品信息

Table 1 *Bupleurum chinense* DC. seeds authenticity modeling sample information

类别 Category	编号 Number	来源 Source	数量/粒 Quantity
北柴胡 <i>Bupleurum chinense</i> DC.	GS ₁	甘肃 Gansu	65
	GS ₂		65
	GS ₃		65
	HB ₁	河北 Hebei	65
	HB ₂		65
	SX ₁	山西 Shanxi	65
	SX ₂		65
	SX ₃		65
	KM ₁	中国西南野生生物种质资源库 The Germplasm Bank of Wild Species	40
	KM ₂		40
藏柴胡 <i>Bupleurum Marginatum</i> var. <i>stenophyllum</i>	Z	河北 Hebei	150
三岛柴胡 <i>Bupleurum falcatum</i>	SD	河北 Hebei	300
红柴胡 <i>Bupleurum scorzonerifolium</i>	H	中国西南野生生物种质资源库 The Germplasm Bank of Wild Species	50
竹叶柴胡 <i>Bupleurum marginatum</i>	ZY	中国西南野生生物种质资源库 The Germplasm Bank of Wild Species	50
银州柴胡 <i>Bupleurum yinchowense</i>	YZ	中国西南野生生物种质资源库 The Germplasm Bank of Wild Species	50



供试材料编号见表1。
The test materials is shown in Table 1.
图1 建模样品不同批次种子

Fig. 1 Modeling samples of different batches of seeds

省保定市安国中药材市场以及中国西南野生生物种质资源库。

1.2 高光谱数据采集

高光谱成像系统主要包括:CCD相机(EM285 CL型,Raptor Photonics Limited,爱尔兰)、150 W 卤素灯(IT 3900e型,Illumination Technologies Inc.,美国)、移动平台(IRCP0076-1COMB型,五铃光学股份有限公司,中国)、计算机和暗箱等(图2)。测定种子400~1 000 nm波段的高光谱数据。系统参数为:电控平台移动速度1.71 mm/s,相机曝光时间7 ms,物距21.5 cm,采集软件为Spectral Image software(五铃光学股份有限公司)。高光谱数据采集全程在暗箱内进行,先采集黑板图像,再将种子分散置于移动平台,随平台移动完成摄像头全平台扫描。采用HSI Analyzer软件(五铃光学股份有限公司)依据公式:

$I=(I_0-B)/(W-B)$

式中： I 为校正后光谱图像； I_0 为原始图像； B 为黑板图像； W 为白板图像。进行黑板白板校正，利用 Matlab 2022a 对光谱图像进行分割，并提取 ROI 中

的平均光谱数据。2025 年 5 月采集所有建模样品的高光谱数据，2024 年 11 月和 2025 年 5 月分别采集万荣柴胡的高光谱数据。

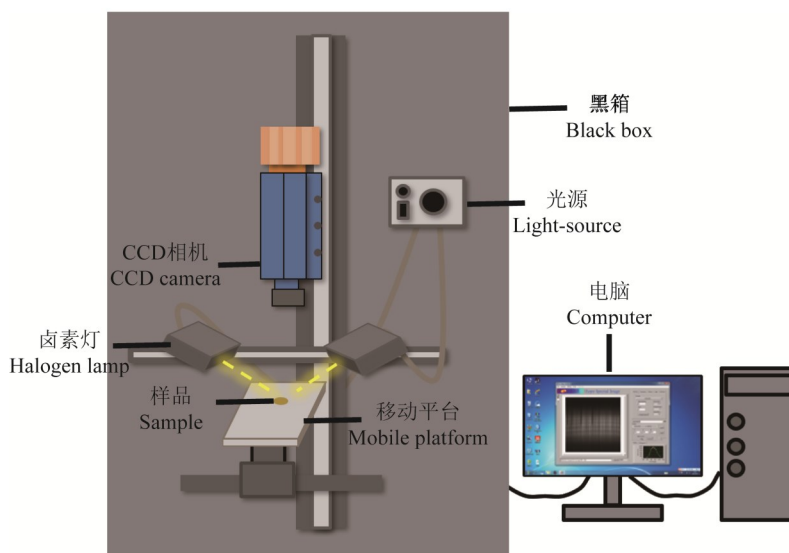


图 2 高光谱采集平台示意图

Fig. 2 Schematic diagram of hyperspectral acquisition platform

1.3 光谱数据处理

1.3.1 高光谱预处理

原始光谱含较多噪声，需进行预处理以提高数据信噪比与可靠性。预处理方法包括 SG 平滑 (savitzky-golay smoothing, SG)、一阶导数 (first derivative, FD)、多元散射校正 (multiplicative scatter correction, MSC)、标准正态变换 (standard normal variate, SNV) 及其组合。

1.3.2 提取特征光谱

针对光谱波段冗余的问题，采用竞争自适应重加权采样 (competitive adaptive reweighted sampling, CARS) 算法和连续投影算法 (successive projections algorithm, SPA) 提取特征光谱。筛选出对建模性能贡献最大、冗余度最小且代表性强的特征波段组合。

1.3.3 光谱校准

针对不同批次跨年份光谱数据存在的数据偏移问题，采用迁移成分分析 (transfer component analysis, TCA)、协方差对齐 (CORrelation ALignment, CORAL) 以及组合算法对数据进行校准。

1.4 建立北柴胡种子真实性鉴定模型

建模数据集含 1 200 粒种子的高光谱数据，按

7:3 分为训练集 (840 粒) 与测试集 (360 粒)，用于构建北柴胡与非北柴胡的二分类模型。采用 Python 3.9 对高光谱数据预处理，建立随机森林 (random forest, RF)、支持向量机 (support vector machine, SVM)、多层感知器 (multi-layer perceptron, MLP)、偏最小二乘判别分析 (partial least squares-discriminant analysis, PLS-DA) 等机器学习模型，以及投票 (voting ensemble model, Voting)、堆叠 (stacking ensemble model, Stacking) 集成模型。通过随机化搜索优化超参数 (表 2)，并采用 5 折交叉验证，以准确率评估模型性能。采用 SHAP (Shapley Additive exPlanations) 算法对所建最优模型进行可解释性分析，量化各特征对模型预测的贡献。

1.5 试验流程

首先采集不同批次北柴胡及近缘种柴胡的高光谱图像，提取单粒种子 ROI 区域 400~1 000 nm 共 765 个波段的平均光谱反射率；对提取的高光谱数据进行预处理后，采用 CARS 与 SPA 算法提取特征光谱；基于预处理及特征光谱提取后的数据，构建机器学习与集成学习模型；另取 1 个批次外部数据集用于模型测试，分析 2 年的光谱数据差异，通过域适应算法实现数据对齐并对其进行预测；最后对最优模型进行 SHAP 可解释性分析，见图 3。

表2 超优参数搜索范围

Table 2 Super-optimal parameter search range

模型 Model	参数 Parameter	范围 Range
RF	树的数量 n_estimators	50~200
	树的最大深度 Max_depth	3~10
	节点可分的最小样本数 Min_samples_split	2~10
	叶子节点最小样本数 Min_samples_leaf	1~5
SVM	惩罚系数 C	0.01、0.1、1、10
	核函数参数 Gamma	0.01、0.1、1、10
MLP	隐藏层的数量和每个隐藏层中神经元的数量 Hidden_layer_sizes	(10,)、(18,)、(20,)、(30,)、(50,)
	激活函数 Activation	relu、tanh、logistic
	求解器 Solver	adam、sgd
	正则化强度 Alpha	0.0001、0.001、0.01
	学习率 Learning_rate	constant、adaptive
PLS-DA	潜变量数量 n_components	5~30

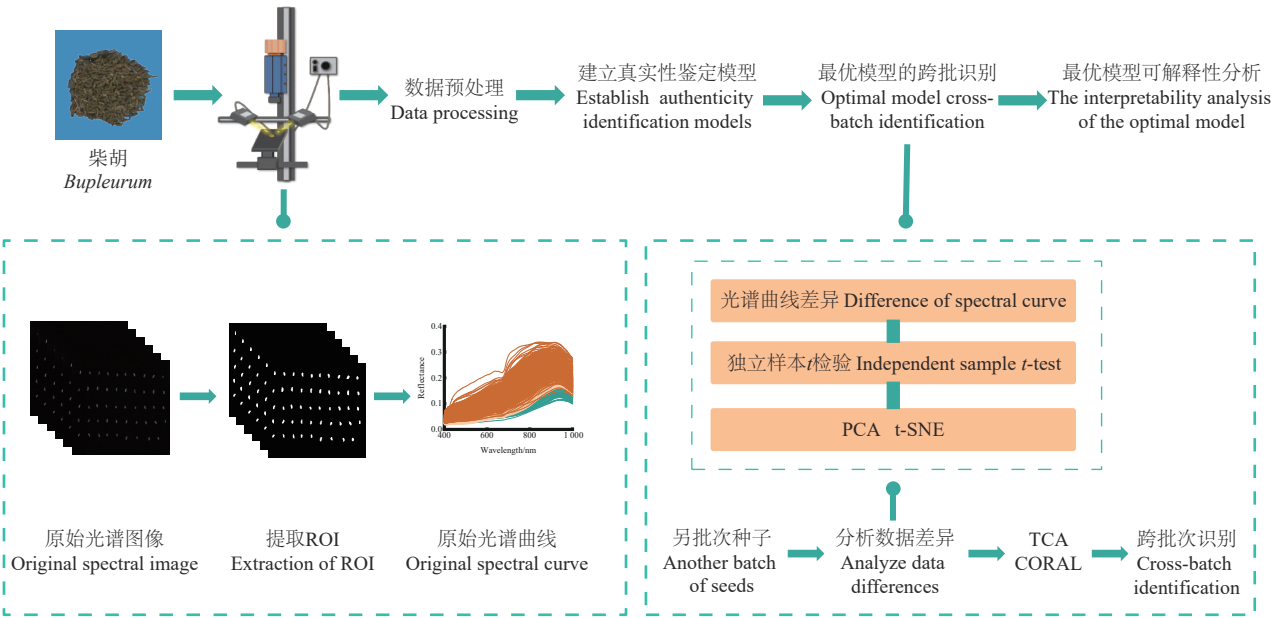


图3 试验流程图

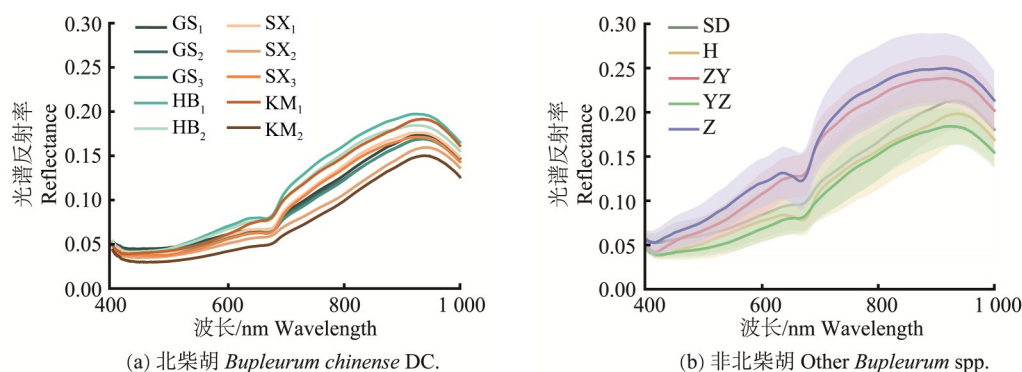
Fig. 3 Flow chart of experiment

2 结果与分析

2.1 基于高光谱数据建立北柴胡种子真实性鉴定模型

由图4可知,10个批次的北柴胡平均光谱曲线趋势相似,均在680 nm处呈现特征吸收谷,约920 nm处呈现特征吸收峰。其中,河北省第1批次

北柴胡(HB₁)光谱反射率最高,中国西南野生生物种质资源库第2批北柴胡(KM₂)最低。非北柴胡样品的平均光谱曲线因种类不同差异较大(图4(b)),680和920 nm处均存在特征吸收谷与吸收峰,但银州柴胡(YZ)吸收谷不明显,竹叶柴胡(ZY)和藏柴胡(Z)的光谱反射率均显著高于其他3个近缘种,且920 nm吸收峰峰形较平缓。



GS₁、GS₂、GS₃为采集于甘肃省的第1~3批次,HB₁和HB₂为采集于河北省的第1、2批次,SX₁、SX₂和SX₃为采集于山西省的第1~3批次,KM₁和KM₂为来自于中国西南野生生物种质资源库的第1、2批次。SD,三岛柴胡;H,红柴胡;ZY,竹叶柴胡;YZ,银州柴胡;Z,藏柴胡。

GS₁、GS₂ and GS₃ are collected as the 1st to 3rd batches from Gansu Province; HB₁ and HB₂ are collected as the 1st and 2nd batches from Hebei Province; SX₁, SX₂ and SX₃ are collected as the 1st to 3rd batches from Shanxi Province; KM₁ and KM₂ are collected as the 1st and 2nd batches from the Germplasm Bank of Wild Species. SD, *Bupleurum falcatum*; H, *Bupleurum scorzonnerifolium*; ZY, *Bupleurum marginatum*; YZ, *Bupleurum yinchowense*; Z, *Bupleurum marginatum* var. *stenophyllum*.

图4 建模样品北柴胡(a)与非北柴胡(b)平均光谱曲线

Fig. 4 The average spectral curves of modeling samples of *Bupleurum chinense* DC. (a) and other *Bupleurum* spp. (b)

由图5可知,不同预处理方法对北柴胡与非北柴胡样品的平均光谱曲线影响显著。原始光谱经FD、MSC及SNV处理后,曲线形态发生明显改变:FD在680 nm处形成正负双峰,且非北柴胡正峰强度高于北柴胡,920 nm处曲线呈负值;MSC与SNV可消除基线漂移,清晰呈现两类样品光谱曲线的细微差异,且经过处理后仍存在680与920 nm的吸收

谷和吸收峰(图5(c))。此外,北柴胡在400~680 nm的光谱曲线较非北柴胡平缓,而680~920 nm波段的光谱反射率增速更快(图5(d))。经SG平滑结合其他预处理得到的平均光谱曲线与单独采用其他预处理的平均光谱曲线相比无显著差异,表明SG平滑对单个样品光谱曲线的降噪作用未在平均光谱曲线中体现。

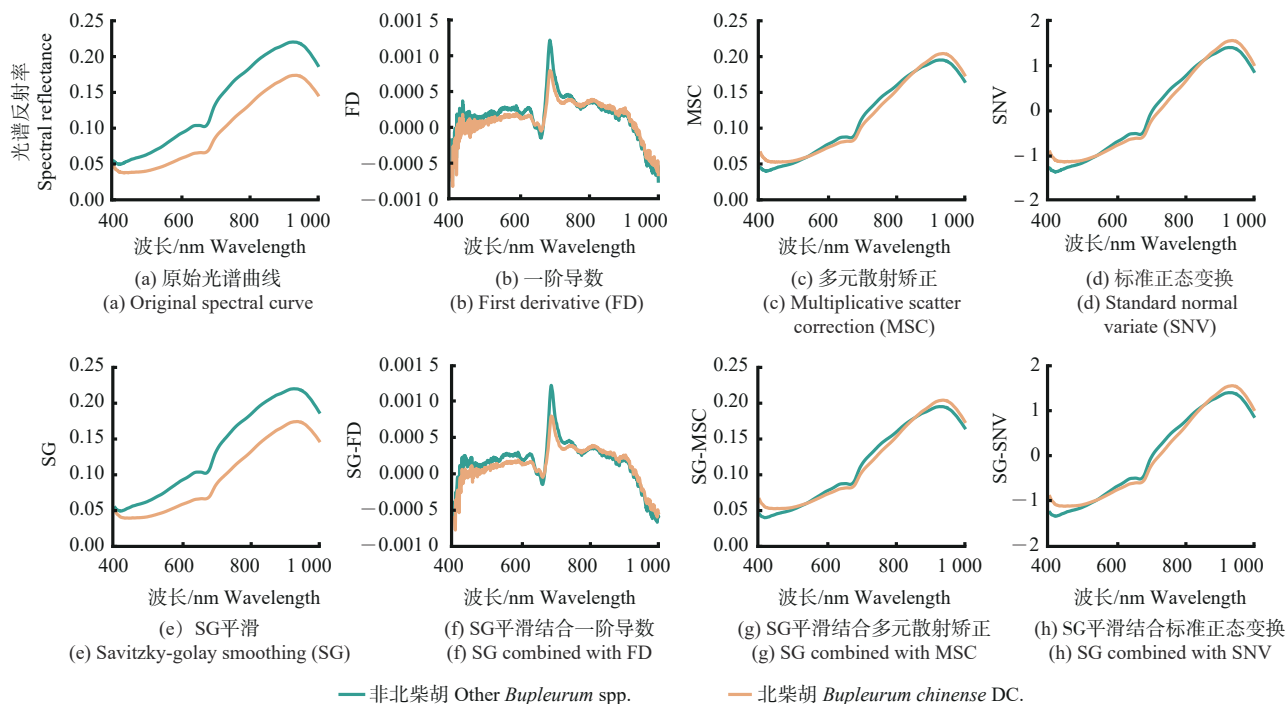


图5 建模样品北柴胡与非北柴胡不同预处理平均光谱曲线

Fig. 5 Different preprocessing average spectral curves of modeling samples of *Bupleurum chinense* DC. and other *Bupleurum* spp.

基于不同预处理的高光谱数据建立机器学习模型,除SG-PLS-DA外,其余模型训练集准确率均高于测试集(表3),其中RF与SVM模型训练集准确率高于测试集5.0%以上,存在过拟合现象。对于RF、SVM和MLP,采用SG平滑与其他预处理组合时,其测试集准确率较单独预处理有部分提升,其中SNV-PLS-DA模型建模准确率最高为85.6%。进一步建立的集成学习模型仍存在过拟合现象(表4),最优模型为SG-MS-Stacking预测准确率为81.7%。

表3 基于高光谱数据机器学习模型对北柴胡种子真实性的预测准确率
Table 3 Prediction accuracy of *Bupleurum chinense* DC. seeds authenticity using machine learning model based on hyperspectral data %

预处理 Preprocessing	RF		SVM	
	训练集	测试集	训练集	测试集
	Training set	Testing set	Training set	Testing set
RAW	95.5	75.8	88.1	79.4
SG	98.8	74.7	87.4	78.1
FD	99.8	73.9	100.0	81.9
MSC	91.2	76.7	94.0	81.7
SNV	93.7	76.4	93.0	83.1
SG-FD	99.5	75.8	100.0	83.3
SG-MS	91.2	77.5	92.9	81.4
SG-SNV	93.5	76.9	92.9	83.3

预处理 Preprocessing	MLP		PLS-DA	
	训练集	测试集	训练集	测试集
	Training set	Testing set	Training set	Testing set
RAW	79.8	76.7	82.9	81.9
SG	79.6	76.7	83.1	83.3
FD	84.3	77.2	84.4	78.1
MSC	78.0	75.3	85.5	85.0
SNV	73.5	72.5	85.7	85.6
SG-FD	85.0	80.8	82.7	78.1
SG-MS	77.9	75.3	86.2	84.2
SG-SNV	76.7	75.8	84.2	83.9

表4 基于高光谱数据集集成学习模型对北柴胡种子真实性的预测准确率
Table 4 Prediction accuracy of *Bupleurum chinense* DC. seeds authenticity using ensemble learning model based on hyperspectral data %

预处理 Preprocessing	投票 Voting		堆叠 Stacking	
	训练集	测试集	训练集	测试集
	Training set	Testing set	Training set	Testing set
RAW	85.1	76.4	84.8	76.9
SG	85.1	76.7	84.5	76.7
FD	99.9	80.3	100.0	82.2
MSC	85.8	77.8	85.5	81.1
SNV	86.3	77.8	86.2	81.1
SG-FD	99.8	81.7	100.0	83.6
SG-MS	85.6	77.2	85.0	81.7
SG-SNV	86.2	77.8	86.8	81.1

2.2 基于特征光谱数据建立北柴胡种子真实性鉴定模型

由表 5 可知,经 CARS、SPA 及 CARS-SPA 筛选后,分别得到 97、15 和 24 个关键特征波段,筛选率依次为 12.7%、2.0% 和 3.1%。基于特征波段构建的 SNV-CARS-PLS-DA 模型测试集准确率达 91.4%,较全波段建模提升 5.8%。CARS 筛选了 87.3% 的特征波段,但模型性能提升,表明基于全

波段构建的北柴胡真实性鉴定模型中存在冗余信息;而 SPA 筛选 95.0% 以上波段,可能同时去除了部分与目标预测相关的波段。基于 CARS 提取的特征波段利用已建立的 SNV-PLS-DA 模型预测万荣柴胡于 2024 年 11 月和 2025 年 5 月采集的高光谱数据预测准确率分别为 79.7% 和 89.5%。可能由于同一批次种子不同时间采集的高光谱数据发生偏移导致了模型预测准确率下降。

表 5 基于特征光谱 SNV-PLS-DA 模型的预测准确率

Table 5 Prediction accuracy of SNV-PLS-DA model based on characteristic spectrum

提取算法 Extraction algorithm	波段数量 Number of bands	筛选率/% Rate	波段分布/nm Waveband distribution	训练集/% Training sets	测试集/% Testing set
CARS	97	12.7	400~1 000	90.8	91.4
SPA	15	2.0	400.4, 401.1, 406.3, 407.8, 409.2, 427.8, 451.7, 567.0, 627.0, 669.6, 683.8, 695.7, 813.8, 875.6, 999.1	82.4	82.2
CARS-SPA	24	3.1	418.9, 432.2, 455.4, 456.2, 483.4, 496.3, 514.6, 533.8, 553.8, 579.4, 600.4, 634.1, 649.9, 665.6, 673.5, 733.1, 821.1, 864.4, 876.4, 931.8, 966.3, 976.7, 984.7, 998.3	85.0	85.6

2.3 跨年份万荣柴胡种子光谱数据差异分析

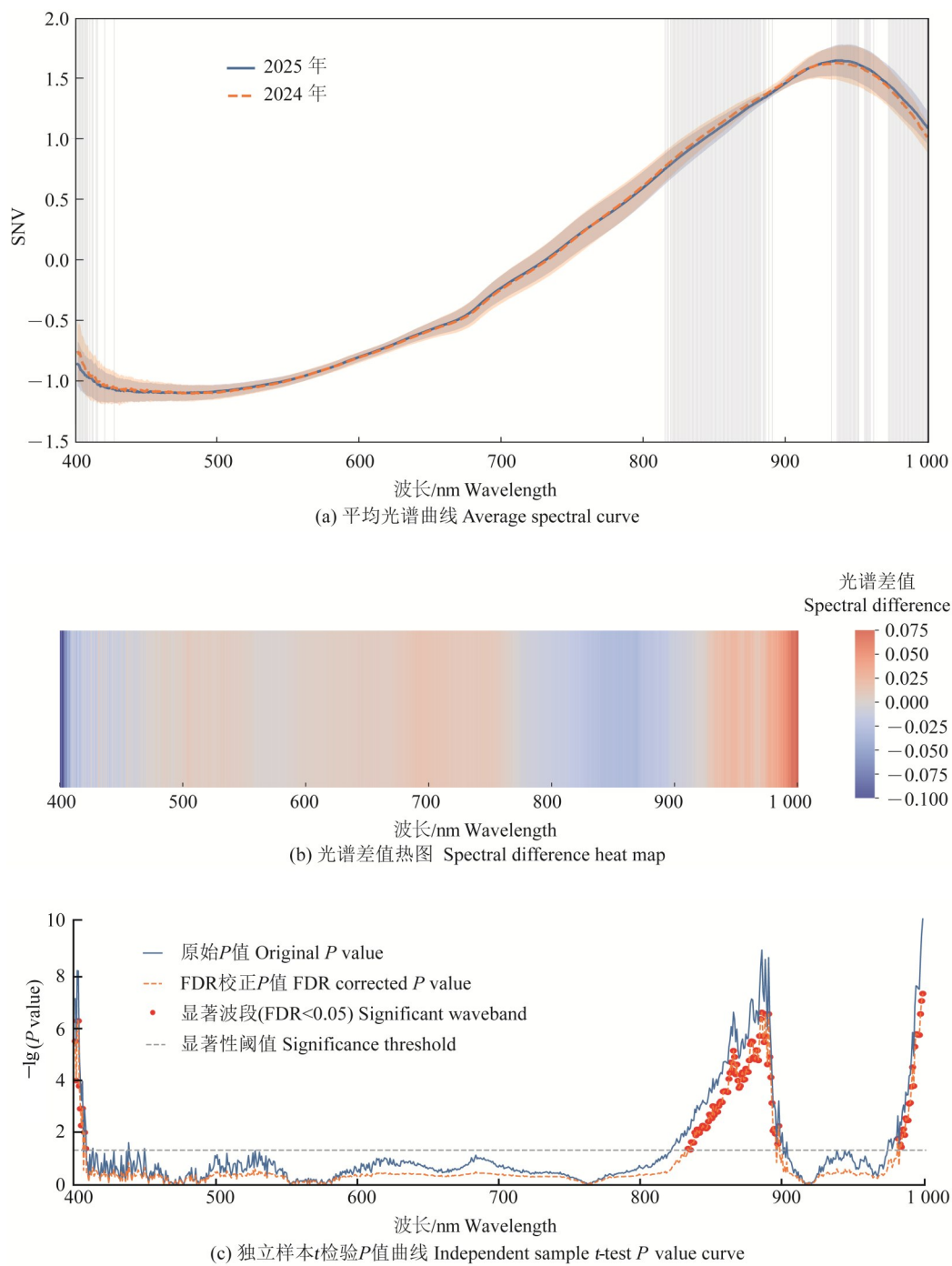
由图 6 可知,2024 和 2025 年光谱曲线总体趋势基本一致,但在 400~420、810~890 及 940~1 000 nm 存在明显差异。全波段反射率变化方向不一致:2024 年 400~450 与 750~900 nm 光谱反射率均高于 2025 年,930~1 000 nm 附近则相反,且不同区域光谱形状变化不同,表明 2 年的光谱存在非线性差异(图 6(b))。独立样本 t 检验显示,随时间变化部分波段光谱数据存在显著差异,其中 400、830~900 及 1 000 nm 波段反射率差异显著,CARS 算法亦提取到该范围的 19 个波段(图 6(c))。PCA 与 t-SNE 得分图中,数据点分散且无明显聚集区域,整体呈高维度复杂性,线性主成分与非线性成分均无法区分 2 年的数据(图 7)。2 年的数据经 PCA 和 t-SNE 降维后的质心距离分别为 0.28 和 0.08,表明数据随时间发生了混合偏移,且线性偏移程度大于非线性偏移。

鉴于 2 年的光谱数据存在以线性偏移为主导的混合偏移,采用 TCA 与 CORAL 算法进行校正。2

种算法均未导致模型过拟合(表 6),其中 CORAL 算法性能最优,训练集与测试集准确率均超 90.0%;而 2 种算法的组合未进一步提升模型性能,推测因 CORAL 已通过对齐协方差矩阵实现良好预测效果,此时引入 TCA 最小化 MMD 会与已对齐的数据分布产生目标冗余,且 TCA 降维可能丢失部分光谱信息。针对 2 年的光谱数据,CORAL 算法对 2024 年采集数据的预测准确率达 94.3%,对 2025 年与建模样品同时期采集的光谱数据预测准确率达 100.0%。可见,该算法适用于校准年份间的数据差异。

2.4 SHAP 解释 PLS-DA 模型

由图 8 可知,基于 SHAP 可解释性分析显示,重要性排前十的波段集中于红光与近红外区域,其中 810.6 nm 对模型决策的重要性最大,平均 $|\text{SHAP}| > 0.12$,显著高于其他特征,推测该波段可能是区分北柴胡与非北柴胡种子的主要差异波段,且与种子中的特异性化学成分相关;而 924.6、842.7、733.1、616.9 和 907.8 nm 的平均 $|\text{SHAP}|$



(a)中灰色竖线为该波段2024与2025年光谱反射率的绝对差值大于所有波段差值的标准差。(c)中灰色虚线显著性阈值为0.05。

The gray vertical line indicates that the absolute difference between the 2025 and 2024 spectra of this band is greater than the standard deviation of the difference between all bands in (a). The significance threshold of gray dotted line is 0.05 in (c).

图6 2024和2025年万荣柴胡平均光谱曲线、光谱差值热图及独立样本t检验P曲线

Fig. 6 Wanrong *Bupleurum* average spectral curve in 2024 and 2025, spectral difference heat map and independent sample *t*-test *P* value curve

均 ≤ 0.1 。SHAP蜂群图显示810.6、984.7、924.6及907.8 nm对模型预测为正效应,即这些波段反射率较高时更易预测为北柴胡种子;其余波段则呈现

负效应(图8(b))。

由图9可知,5个波段对模型预测的影响均呈现近似线性关系,推测因PLS-DA为线性模型,所以

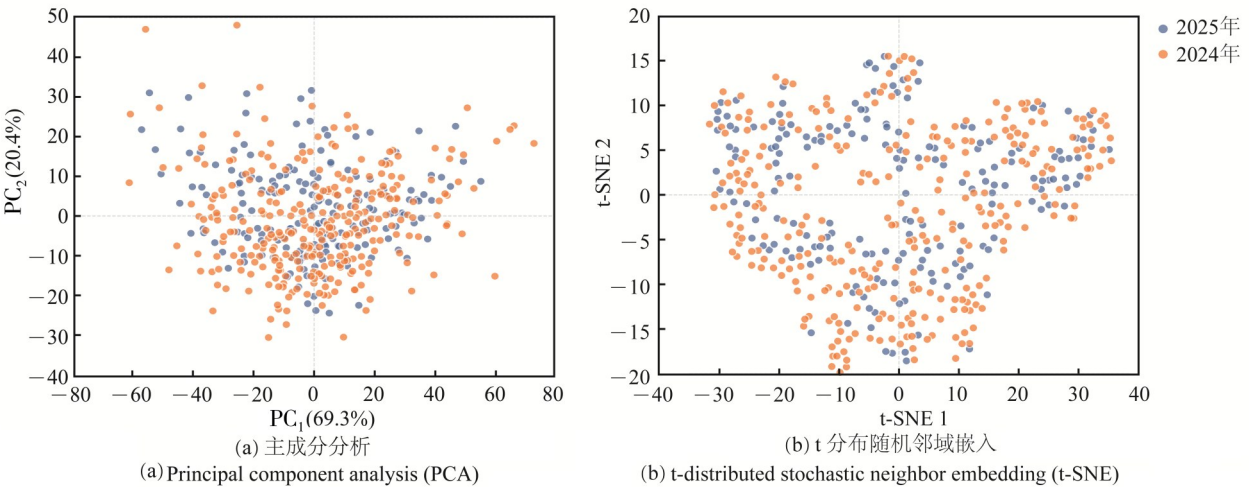


图 7 2024 和 2025 年万荣柴胡光谱主成分分析和 t-SNE 得分图

Fig. 7 The PCA and t-SNE scores of Wanrong *Bupleurum* in 2024 and 2025

表 6 基于校准数据 SNV-CARS-PLS-DA 建模准确率

Table 6 Modeling accuracy of SNV-CARS-PLS-DA based on calibration data %

校正算法 Calibration algorithm	训练集 Training sets	测试集 Testing set	万荣柴胡 Wanrong <i>Bupleurum</i>	
			2024 年	2025 年
TCA	85.0	85.0	78.7	96.1
CORAL	90.7	91.7	94.3	100.0
CORAL-TCA	85.7	86.1	94.3	96.7

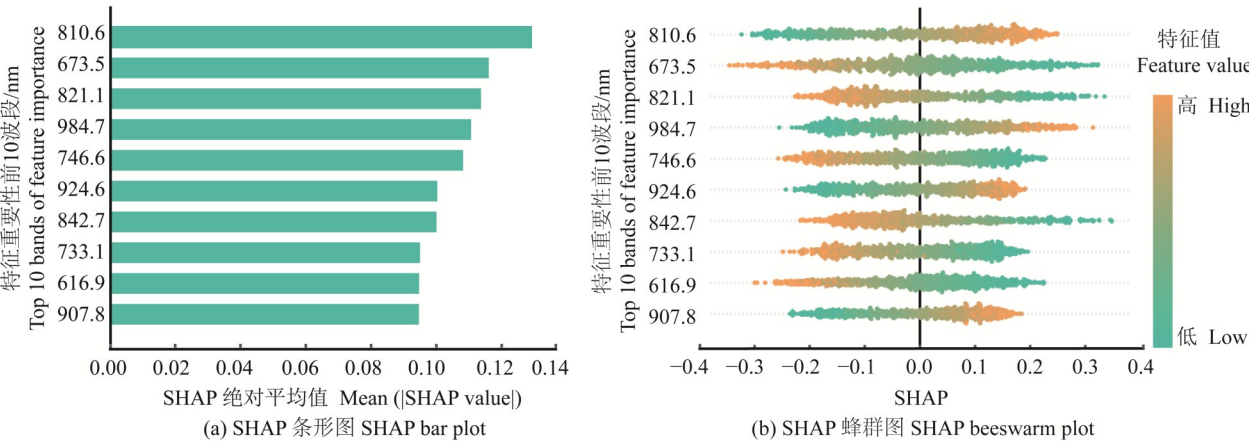
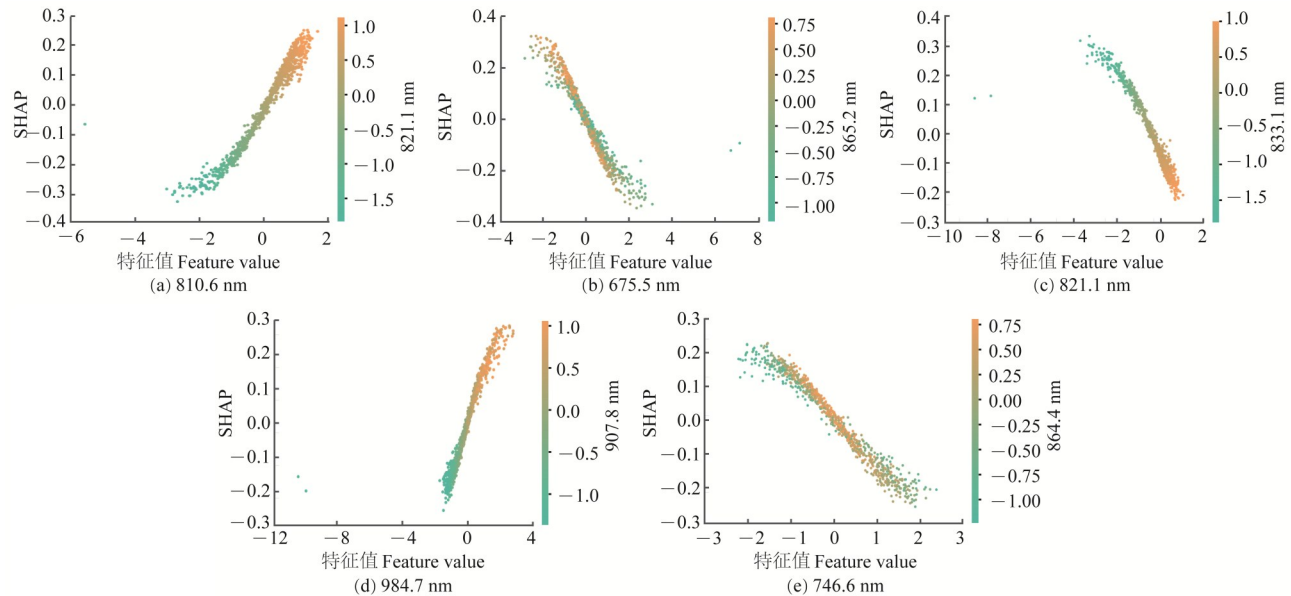


图 8 PLS-DA 模型 SHAP 条形图和蜂群图

Fig. 8 SHAP bar plot and beeswarm plot for the PLS-DA model

提取的均为线性特征。其中 810.6 nm 与模型预测呈近似正相关, 与其交互作用最大的波段为 821.1 nm, 同样为对模型预测有重要贡献的波段。除 984.7 nm 外, 其余 4 个波段交互作用最大的波段均集中在 820~860 nm。相较于其他 4 个波段,

821.1 nm 多数样本的特征值范围为 -4~2, SHAP 的变化范围为 -0.25~0.30, 变化率 0.09, 小于另 4 个波段。810.6、821.1 和 984.7 nm 与其交互作用最强的波段呈现协同变换, 即随着该波段反射率增加, 对应波段特征值亦同步增加。



颜色代表与其交互作用最大的波段的特征值。
Color represents the characteristic value of the band with the greatest interaction with it.

图9 PLS-DA 模型特征重要性前 5 波段 SHAP 交互作用部分依赖图

Fig. 9 SHAP interaction partial dependence plots for the top 5 bands based on feature importance in the PLS-DA model

3 讨论

为提高不同批次北柴胡种子跨年份真实性鉴定模型高光谱数据的鲁棒性,在种子高光谱数据因存在偏移而导致模型预测准确率下降的背景下,本研究通过结合CARS特征筛选与CORAL域适应算法构建鉴定模型,将SNV-PLS-DA模型预测不同批次跨年份采集的高光谱数据准确率提升了14.6%(表6)。CORAL算法可有效对齐种子2年高光谱数据中以线性偏移为主的混合偏移。

本研究发现,同一批次北柴胡种子经半年常温贮藏后,贮藏前后采集的光谱数据发生了以线性偏移为主的混合偏移,差异波段主要集中于近红外830~890 nm(图6)。由于近红外区域对水分含量高度敏感^[19],本试验中万荣柴胡种子采集于2024年10月,首组光谱数据采集于2024年11月,第二组采集于2025年5月,推测半年贮藏期内种子水分含量的改变导致近红外区域反射率显著变化,可能是该波段偏移的主要原因。此外,仪器老化也可能引发光谱数据偏移^[20];而种子贮藏过程中蛋白质、脂质等成分的变化,则可能导致光谱非线性偏移^[21]。这种偏移可使模型预测准确率从89.5%降至79.7%。对于同一批次万荣柴胡种子,在排除品种混杂因素后,时间因素成为导致模型性能下降的根本原因。

PLS-DA作为线性模型,通过线性变换以最小的误差将数据转换到低维空间^[22]。中药材种子个体间差异大,PLS-DA模型更适用于组内差异较大的情况^[23]。贮藏和仪器原因引起的偏移会改变特征的均值与方差^[21],而CORAL算法通过对齐2年数据的协方差矩阵,将模型预测准确率提升至94.3%,推测其可使校正后的数据在特征空间中类内分布更紧凑、类间分离度更高,进而优化了分类边界。

对比2种域适应方法可知,CORAL算法更适用于北柴胡光谱随时间产生的偏移(表6)。TCA作为子空间学习算法,对源域和目标域采用相同变换并降维至同一子空间^[17],但因不同批次种子偏移程度存在差异,维度选择难以统一,校准后需重新建立鉴定模型。由于PLS-DA的关键判别信息依赖特征间的线性协方差模式^[22],TCA侧重的全局分布对齐,因此TCA对模型性能提升有限。而CORAL算法不会改变目标域特征的数据结构,能够完整保留原始特征,无需通过降维压缩信息^[18]。同时,该算法直接对齐源域与目标域的协方差矩阵,可进行无监督处理跨域数据^[18],适用于数据分布差异主要由协方差偏移主导的场景。需注意的是,当特征数量大于样本数量时,样本协方差难以准确评估2批数据差异,易受异常值影响^[17],因此使用CORAL算法时需关注特征数量与样本量的匹配关系。

SHAP算法提取到重要性前十波段集中在红光和近红外的区域(图8(a))。其中673.5、733.1和746.6 nm 主要与叶绿素a的吸收有关,叶绿素a和叶绿素b在红光波段吸收光能用于光合作用,叶绿素在630~680 nm 内有强吸收峰,尤其是在670 nm 附近^[24];810.6、821.1和842.7 nm 与蛋白质等含氮化合物中N-H键的吸收有关^[25];984.7、924.6和907.8 nm 与水的O-H键震动有关^[26]。说明种子中叶绿素和其他相关色素以及蛋白质等物质是PLS-DA模型区分北柴胡和非北柴胡的关键信息。赵昕等^[27]基于1DCNN和PLS-DA鉴别酸枣仁真伪时,通过CARS和自定义选择层共同选择出了位于670、721和850 nm 附近的波段。Zhu等^[28]基于高光谱数据鉴定大豆种子时,利用CARS算法筛选出21个特征波段,其中也筛选到了680 nm 附近的波段。810.6和821.1 nm 属于830~900 nm 敏感区边缘,可能是由于这些波段反射率的变化导致模型预测准确率下降。

本研究基于10个批次北柴胡及5个近缘种的光谱数据,构建了北柴胡真实性SNV-CARS-PLS-DA模型。同一批次收获的万荣柴胡种子2024和2025年采集的光谱数据存在以线性偏移为主的混合偏移。为此采用CARS-CORAL联用策略,通过对齐协方差矩阵可有效校正2年份的数据偏移,将预测准确率提升14.3%。基于SHAP算法揭示模型通过水分(907.8~984.7 nm)、蛋白质(810.6~842.7 nm)及色素(673.5~746.6 nm)三大类生物标志物实现判别,其中关键波段与偏移敏感区存在重合。未来应该进一步扩大建模样本量,包含更多不同来源、不同贮藏时间的样本,以提高模型在复杂场景的泛化性。

4 结 论

通过高光谱成像技术结合域适应算法,对于北柴胡种源混杂的问题基于高光谱数据建立了北柴胡真实性鉴定模型,SNV-CARS-PLS-DA为最优模型,测试集准确率为91.4%。同一批次收获的万荣柴胡2024和2025年高光谱数据存在以线性偏移为主的混合偏移,其中400、830~900及1 000 nm 附近波段反射率差异显著,经过CORAL算法的校准,模型预测2024年万荣柴胡光谱数据准确率从79.7%提升到94.3%,CORAL算法能校准该偏移模式。可见,所构建的SNV-CARS-PLS-DA模型

结合CORAL算法,可对不同批次北柴胡种子跨年份的真实性进行快速无损鉴定。

参考文献 References

- [1] 黄涵筌,王潇晗,付航,王妍,杨世海. 柴胡属药用植物资源研究进展[J]. 中草药, 2017, 48(14): 2989-2996
Huang H Q, Wang X H, Fu H, Wang Y, Yang S H. Research progress on medicinal plant resources of *Bupleurum* L [J]. *Chinese Traditional and Herbal Drugs*, 2017, 48(14): 2989-2996(in Chinese)
- [2] 丁自勉,张旭. 柴胡生物学研究进展[J]. 中国野生植物资源, 2005, 24(2): 11-13
Ding Z M, Zhang X. Research advance on *Bupleurum* biology[J]. *Chinese Wild Plant Resources*, 2005, 24(2): 11-13(in Chinese)
- [3] DB 15/T 1296—2017. 北柴胡种子质量分级[S]. 呼和浩特:内蒙古自治区质量技术监督局, 2018
DB 15/T 1296—2017. Seed quality grading of *Bupleurum chinense* DC[S]. Hohhot: Inner Mongolia Autonomous Region Quality and Technical Supervision Bureau, 2018(in Chinese)
- [4] DB 14/T 1491—2017. 柴胡种子质量标准[S]. 太原:山西省质量技术监督局, 2017
DB 14/T 1491—2017. Quality standard of *Bupleurum* seed[S]. Taiyuan: Shanxi Provincial Bureau of Quality and Technical Supervision, 2017(in Chinese)
- [5] 刘志浩,查良平,罗丽,赵玉洋,李晓琳,袁媛. 中药材种子鉴别技术研究进展[J]. 中国实验方剂学杂志, 2024, 30(4): 1-11
Liu Z H, Zha L P, Luo L, Zhao Y Y, Li X L, Yuan Y. Methods for identifying seeds of Chinese medicinal materials: A review[J]. *Chinese Journal of Experimental Traditional Medical Formulae*, 2024, 30(4): 1-11(in Chinese)
- [6] GB/T 3543.5-1995. 农作物种子检验规程 真实性和品种纯度鉴定[S]. 北京: 中国标准出版社, 1996
GB/T 3543.5-1995. Rules for agricultural seed testing: Verification of genuineness and cultivar[S]. Beijing: Standards Press of China, 1996(in Chinese)
- [7] 赵莎,郑司浩,曾燕,戴玮,李鹏英,王继永. 中药材种子种苗真实性鉴定技术研究进展[J]. 中国现代中药, 2021, 23(10): 1831-1836
Zhao S, Zheng S H, Zeng Y, Dai W, Li P Y, Wang J Y. Technology for authenticity identification of seeds and seedlings of Chinese medicinal plants: A review [J]. *Modern Chinese Medicine*, 2021, 23(10): 1831-1836(in Chinese)
- [8] Lu Y Z, Saeys W, Kim M, Peng Y K, Lu R F. Hyperspectral imaging technology for quality and safety evaluation of horticultural products: A review and celebration of the past 20-year progress[J/OL]. *Postharvest Biology and Technology*, (2020-08-01). DOI: doi. org/10.1016/j. postharvbio.2020.111318
- [9] Wu J X, Zhang Y, Hu P F, Wu Y Y. A review of the application of hyperspectral imaging technology in agricultural crop economics[J/OL]. *Coatings*, (2024-10-09). DOI: doi. org/10.3390/coatings
- [10] Zhao Y Y, Zhu S S, Zhang C, Feng X P, Feng L, He Y. Application of hyperspectral imaging and chemometrics for variety classification of maize seeds[J]. *Royal Society of Chemistry Advances*, 2018, 8(3): 1337-1345
- [11] Zhang W D, Li Z X, Li G H, Zhuang P X, Hou G J, Zhang Q, Li C Y. GACNet: Generate adversarial-driven cross-aware network for hyperspectral wheat variety identification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 1-14
- [12] Liu Y, Wu T, Yang J J, Tan K Z, Wang S W. Hyperspectral band selection for soybean classification based on information measure in FRS theory[J]. *Biosystems Engineering*, 2019, 178: 219-232

- [13] Huang M, Tang J Y, Yang B, Zhu Q B. Classification of maize seeds of different years based on hyperspectral imaging and model updating[J]. *Computers and Electronics in Agriculture*, 2016, 122:139-145
- [14] Wang Y, Song S R. Variety identification of sweet maize seeds based on hyperspectral imaging combined with deep learning[J/OL]. *Infrared Physics & Technology*, (2023-04-19). DOI: doi. org/10.1016/j.infrared.2023.104611
- [15] 李灵巧, 王卓健, 陈江海, 卢丰, 黄殿贵, 杨辉华, 李泉. 迁移成分分析结合直接校正的模型传递方法[J]. 光谱学与光谱分析, 2024, 44(12):3399-3405
Li L Q, Wang Z J, Chen J H, Lu F, Huang D G, Yang H H, Li Q. A model transfer method based on transfer component analysis and direct correction [J]. *Spectroscopy and Spectral Analysis*, 2024, 44(12): 3399-3405 (in Chinese)
- [16] 王雪, 王子文, 张广月, 马铁民, 陈争光, 衣淑娟, 王长远. 基于迁移成分分析的近红外光谱定量分析通用模型[J]. 光谱学与光谱分析, 2024, 44(11): 3213-3221
Wang X, Wang Z W, Zhang G Y, Ma T M, Ceng Z G, Yi S J, Wang C Y. A universal model for quantitative analysis of near-infrared spectroscopy based on transfer component analysis [J]. *Spectroscopy and Spectral Analysis*, 2024, 44(11):3213-3221(in Chinese)
- [17] Cheng Z W, Chen C, Chen Z H, Fang K, Jin X Y. Robust and high-order correlation alignment for unsupervised domain adaptation [J]. *Neural Computing and Applications*, 2021, 33(12):6891-6903
- [18] Sun B C, Feng J S, Saenko K. Return of frustratingly easy domain adaptation [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, 30(1): 2058-2065
- [19] 王冬, 栾云霞, 王欣然, 贾文坤. 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, Jia W S. Progress in non-destructive analysis of meat quality by near-infrared spectroscopy[J]. *Meat Research*, 2024, 38(5):61-70(in Chinese)
- [20] Zhang Y Q, Chen Z C, Zhang W J, Zhang H. Quantitative effects of the spectral calibration accuracy of the imaging spectrometer on the vegetation red edge[C]. in: 2013 35th International Symposium on Remote Sensing of Environment. Beijing: Institute Remote Sensing & Digital Earth, 2013
- [21] Du W, Chen Z P, Zhong L J, Wang S X, Yu R Q, Nordon A, Littlejohn D, Holden M. Maintaining the predictive abilities of multivariate calibration models by spectral space transformation[J]. *Analytica Chimica Acta*, 2011, 690(1):64-70
- [22] Ruiz-Perez D, Guan H B, Madhivanan P, Mathee K, Narasimhan G. So you think you can PLS-DA?[C]. in: 2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCBMS). Las Vegas: Institute of Electrical and Electronics Engineers (IEEE), 2018
- [23] Barker M, Rayens W. Partial least squares for discrimination[J]. *Journal of Chemometrics*, 2003, 17(3):166-173
- [24] Eccher Zerbini P, Vanoli M, Rizzolo A, Grassi M, de Azevedo Pimentel R M, Spinelli L, Torricelli A. Optical properties, ethylene production and softening in mango fruit[J]. *Postharvest Biology and Technology*, 2015, 101: 58-65
- [25] Caporaso N, Whitworth M B, Grebby S, Fisk I D. Non-destructive analysis of sucrose, caffeine and trigonelline on single green coffee beans by hyperspectral imaging[J]. *Food Research International*, 2018, 106:193-203
- [26] Wang L, Sun D W, Pu H B, Zhu Z W. Application of hyperspectral imaging to discriminate the variety of maize seeds[J]. *Food Analytical Methods*, 2016, 9(1):225-234
- [27] 赵昕, 石玉娜, 刘怡彤, 姜洪喆, 褚璇, 赵志磊, 王宝军, 陈晗. 基于IDCNN和PLSDA酸枣仁真伪高光谱图像鉴别中的关键特征分析[J]. 光谱学与光谱分析, 2025, 45(3):869-877
Zhao X, Shi Y N, Liu Y T, Jiang H Z, Chu X, Zhao Z L, Wang B J, Chen H. Key Feature analysis in identification and authenticity of ziziphi spinosae semen by using hyperspectral images based on IDCNN and PLSDA[J]. *Spectroscopy and Spectral Analysis*, 2025, 45(3):869-877(in Chinese)
- [28] Zhu S L, Chao M N, Zhang J Y, Xu X J, Song P W, Zhang J L, Huang Z W. Identification of soybean seed varieties based on hyperspectral imaging technology [J/OL]. *Sensors*, (2019-11-28). DOI: doi. org/10.3390/s19235225

责任编辑:吕晓梅



第一作者简介: 冷君娇,中国农业大学农学院种子科学与技术专业硕士研究生。参与新疆维吾尔自治区重点研发计划项目(2024B040039),获得中国农业大学2022—2023学年硕士研究生二等学业奖学金和2023—2024学年硕士研究生二等学业奖学金。



通讯作者简介: 孙群,农学博士,中国农业大学农学院副教授,博士生导师。教育部高等学校种子科学与工程专业教学指导分委员会秘书长,中国作物学会作物种子专业委员会副秘书长(2011—2022年)。主要从事种子活力无损检测研究,主持及参与完成多项国家及省部级科研项目,以通讯作者或第一作者在*Food Chemistry*、*Computers and Electronics in Agriculture*、*Journal of Integrative Agriculture*、《中国农业大学学报》等期刊发表论文80余篇。授权发明专利3项,授权实用新型专利6项,软著权3项,制定种子活力检测及种子加工的团体标准10项。