

基于高光谱技术的穿心莲药材中穿心莲内酯类成分检测研究

肖丹¹, 王思曼¹, 张悦¹, 刘地发², 郝庆秀^{1,3}, 白瑞斌^{*1,3}, 杨健^{*1,3}

(1. 中国中医科学院 中药资源中心, 道地药材品质保障与资源持续利用全国重点实验室, 北京 100700;

2. 江西青峰药业有限公司, 江西 赣州 341000; 3. 江西省道地药材质量评价研究中心, 江西 赣江新区 330000)

摘要: 基于高光谱技术结合化学计量学, 建立不同种质穿心莲药材中穿心莲内酯类成分含量的检测方法。采集穿心莲样品的高光谱信息, 获得原始光谱数据 (Raw Data)。采用一阶导数 (D1)、二阶导数 (D2)、SG 平滑 (SG)、乘性散射校正 (MSC) 对 Raw Data 预处理, 结合偏最小二乘判别分析 (PLS-DA) 建立分类模型, 结合偏最小二乘回归 (PLSR)、反向传播神经网络 (BPNN)、随机森林回归 (RFR) 建立回归模型。应用连续投影算法 (SPA) 简化模型。不同种质的穿心莲最佳分类模型为 D1-PLS-DA。穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯、脱水穿心莲内酯 4 种穿心莲内酯类化合物总含量的最佳回归模型分别为 SG-PLSR、MSC-PLSR、Raw Data-SPA-BPNN、MSC-SPA-BPNN 和 Raw Data-PLSR。应用高光谱技术可实现穿心莲品质的快速准确检测。

关键词: 高光谱成像技术; 化学计量学; 穿心莲; 预测模型; BP 神经网络; 偏最小二乘法; 随机森林回归

中图分类号: R284.1 **文献标识码:** A **文章编号:** 0258-3283 (2024) 06-0089-10

DOI: 10.13822/j.cnki.hxsj.2024.0018

Research on the Detection of Andrographolide Components in *Andrographis paniculata* Based on Hyperspectral Technology XIAO Dan¹, WANG Si-man¹, ZHANG Yue¹, LIU Di-fa², HAO Qing-xiu^{1,3}, BAI Rui-bin^{*1,3}, YANG Jian^{*1,3} (1. State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China; 2. Jiangxi Qingfeng Pharmaceutical Co. Ltd., Ganzhou 341000, China; 3. Evaluation and Research Center of Daodi Herbs of Jiangxi Province, Ganjiang New Area 330000, China)

Abstract: Based on hyperspectral technology combined with chemometrics, to establish a method for detecting the content of andrographolide components in different germplasm of *Andrographis paniculata*. The hyperspectral information of *Andrographis paniculata* samples was collected to obtain the raw spectral data (Raw Data). The Raw Data was preprocessed using first derivative (D1), second derivative (D2), savitzky-golay (SG), multiplicative scatter correction (MSC), and a classification model was established using partial least squares discriminant analysis (PLS-DA). Regression models were established using partial least squares regression (PLSR), back propagation neural network (BPNN), and random forest regression (RFR). The application of successive projection algorithm (SPA) simplified the model. The optimal classification model for different germplasm of *Andrographis paniculata* is D1-PLS-DA. The optimal regression models for andrographolide, neoandrographolide, deoxyandrographolide, dehydroandrographolide, and total content of the four andrographolide components mentioned are SG-PLSR, MSC-PLSR, Raw Data-SPA-BPNN, MSC-SPA-BPNN, and Raw Data-PLSR, respectively. The application of hyperspectral technology can achieve rapid and accurate detection of the quality of *Andrographis paniculata*.

Key words: hyperspectral imaging technology; chemometric; *Andrographis paniculata*; prediction model; backpropagation neural network; partial least squares method; random forest regression

穿心莲为爵床科植物穿心莲 *Andrographis paniculata* (Burm.f.) Nees 的干燥地上部分^[1]。目

前, 商用穿心莲药材的栽培地区主要集中在广东、福建、广西, 年产量占全国 90% 以上^[2]。但由于

收稿日期: 2024-01-08; 网络首发日期: 2024-03-29

基金项目: 中药全产业链质量技术服务项目 (2022-230-221); 江西省主要学科科学技术带头人培养计划-领军人才项目 (20225BCJ22018); 江西省创新联合体协同攻关项目 (20224BBG72001); 中国中医科学院基本科研业务费优秀青年科技人才培养专项项目 (ZZ16-YQ-040); 中国中医科学院中药资源中心自主选题研究项目 (ZZXT202208)。

作者简介: 肖丹 (2002-), 女, 苗族, 湖南邵阳人, 硕士生, 主要研究方向为中药品质评价研究。

通讯作者: 杨健, E-mail: yangchem2012@163.com; 白瑞斌, E-mail: bairuibin2022@163.com。

引用本文: 肖丹, 王思曼, 张悦, 等. 基于高光谱技术的穿心莲药材中穿心莲内酯类成分检测研究[J]. 化学试剂, 2024, 46(6): 89-98。

自然留种繁育、种植环境限制以及种植技术落后等原因,出现成分含量不稳定、不同批次的产品质量差异较大等问题,因此,种质资源的分类十分必要。

二萜内酯类成分是穿心莲的主要功效成分,具有抗炎、抗感染、抗肿瘤等药理作用^[3,4]。穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯、脱水穿心莲内酯占总二萜内酯类的 75% 以上^[4],为《中国药典(2020 版)》规定的定量成分^[1]。目前针对以上 4 种活性物质的含量测定主要采用高效液相色谱法和超高效液相色谱法^[5-7]。这些检测方法虽然选择性好、灵敏度高,但存在检测成本高、时间长、试剂消耗量大等缺点。因此,建立一种快速、准确、高通量、无污染的检测方法十分必要。

光谱检测技术是一种快速、无损的检测方法,它根据样品的发射、吸收、散射光谱谱系特征来鉴别物质、测量物质含量信息。常见的光谱技术有拉曼光谱技术、近红外光谱技术和高光谱成像技术等,其中,近红外光谱技术是一种行之有效的农产品和食品质量分析技术,它利用产品中氢键(如 C—H、O—H 和 N—H)的伸缩和弯曲振动来获取信息。赖秀娣等^[8]运用近红外光谱技术和化学计量学方法,建立了穿心莲中穿心莲内酯的含量预测模型,但仅测量穿心莲内酯的含量不足以反映穿心莲的品质,而且传统的近红外方法只从样品的一个点收集光谱数据,同时检测多个目标仍具有挑战性。

高光谱成像技术是一种非接触式的光谱分析方法,可以在光谱范围内采集大量的连续光谱数据,以获取中药材表面感兴趣区域的光谱信息。化学计量学是借助统计学、数学等学科方法探索化学系统的测量值与系统状态之间的联系。高光谱成像技术结合化学计量学可实现对中药材的定性判别和定量分析。张璐等^[9]运用高光谱成像技术,结合化学计量学,建立了 D2-PLS-DA 模型,实现了对不同产地酸枣仁的鉴别;史磊^[10]基于高光谱成像技术,运用最小二乘支持向量回归模型和优化算法,实现了对三七皂苷含量的检测。本研究基于高光谱成像技术,结合化学计量学方法,建立不同种质穿心莲药材的分类模型和穿心莲中穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯、脱水穿心莲内酯以及上述 4 种穿心莲内酯类化合物总含量的回归模型,为穿心莲品质的快速检测提

供新方法。

1 实验部分

1.1 主要仪器与试剂

ACQUITY UPLC™ I-Class 型超高效液相色谱仪(美国 Waters 公司);HySpex 系列高光谱成像光谱仪(挪威 Norsk Elektro Optikk 公司)。

对照品穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯、脱水穿心莲内酯(98%,北京倍特仁康生物医药科技有限公司);乙腈(色谱纯,美国默克公司);蒸馏水(北京屈臣氏蒸馏水有限公司)。

1.2 穿心莲样品

本研究团队前期基于系统选育方法从不同种质来源的穿心莲中选育出 6 个品系,按表型性状差异分别命名为 A-1~A-6,全部品系移栽至广东湛江栽培基地。所有样品于 2022 年 9 月采自广东湛江,包括 A-1(30 批次)、A-2(50 批次)、A-3(31 批次)、A-4(68 批次)、A-5(19 批次)和 A-6(11 批次)共 209 批次。所有样品经中国科学院中药资源中心杨健副研究员鉴定为爵床科植物穿心莲 *Andrographis paniculata* (Burm. f.) Nees 的干燥茎叶,样品收集后常温保存并研磨成粉末,过 100 目筛,粉末密封在聚乙烯袋内并在 4℃ 下冷藏。各粉末样品用于高光谱数据采集,并按 7:3 的比例将样品随机分成训练集和预测集用于后续建模分析。

1.3 高光谱成像系统

高光谱成像设备,主要由 2 个高光谱相机、2 个 150 W/12 V 卤钨灯、1 个 CCD 探测仪、1 个移动平台和仪器自带计算机与内置软件组成。高光谱相机为光谱范围为 410~990 nm 的 SN0605 VNIR 可见-近红外相机和光谱范围为 950~2 500 nm 的 N3124 SWIR 短波红外相机,相机镜头的光谱分辨率为 6 nm,与样品的距离为 25 cm。平台移动速度为 1.5 nm/s。

1.4 高光谱图像采集、校正、兴趣区域数据提取

取约 2 g 穿心莲样品粉末于 35 mm×12 mm 一次性培养皿中,振荡,使粉末表面平整,均匀分布。依次取 5~10 个样品置于黑色水平移动平台上,同时放置 Teflon 白板,扫描,获得穿心莲粉末光谱数据。扫描过程中,SN0605 VNIR 可见-近红外相机镜头积分时间为 4 000 μs,帧时间为 19 000;N3124 SWIR 短波红外相机镜头积分时间

为 5 400 μs , 帧时间为 49 535。

为了消除图像采集过程中光源分布不均匀、照射不稳定和相机镜头暗电流等因素对样品数据的影响,需用仪器自带 RAD 校正软件对原始高光谱图像进行 RAD 校正。随后进行黑白校正,校正公式如(1)所示。

$$R = (R_{\text{raw}} - R_d) / (R_w - R_d) \quad (1)$$

式中, R 为校正后的光谱数据; R_{raw} 为原始光谱数据; R_w 为从反射率为 99% 的白板获得的白色参考数据; R_d 为关灯并遮挡相机镜头获得的暗参考数据。

校正后,采用 ENVI 5.3 软件进行兴趣区域提取,兴趣区域内的平均相对反射率即样品的 Raw Data。

1.5 含量测定

穿心莲样品中穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯、脱水穿心莲内酯含量参照《中国药典(2020 版)》^[1] 穿心莲项下含量测定方法并做适当改进。采用超高效液相色谱法,以 ACQUITY UPLC™ BEH C₁₈ (100 mm×2.1 mm, 1.8 μm) 为色谱柱;以水(A)-乙腈(B)为流动相,进行梯度洗脱(0~7.5 min, 20% B→25% B; 7.5~9 min, 25% B→28% B; 9~15 min, 28% B→32% B; 15~16 min, 32% B→35% B; 16~17 min, 35% B→90% B; 17~19 min, 90% B; 19~20 min, 90% B→20% B; 20~22 min, 20% B); 流速为 0.3 mL/min; 柱温为 35 $^{\circ}\text{C}$; 进样量为 1.0 μL , 检测波长为 205 nm。

1.6 光谱数据预处理

为了减少背景、噪声以及其他因素引起的误差,提高模型预测能力及稳定性,采用成性散射校正(Multiplicative scatter correction, MSC)、一阶导数(First derivative, D1)、二阶导数(Second derivative, D2)、SG 平滑(Savitzky-Golay, SG)4 种方法对光谱数据进行预处理。

D1、D2 属于基线校正,可以扣除仪器背景或漂移对信号的影响,D1 用来扣除斜线和曲线背景;D2 可以提高光谱分辨率;MSC 属于散射校正,用来消除颗粒分布不均匀及颗粒大小差异产生的散射现象和光程变换对漫反射的影响;SG 属于平滑处理,对光谱曲线进行平滑,消除光谱信号中的随机噪声,提高样本信号的信噪比^[11]。

1.7 模型的建立

采用偏最小二乘判别分析(Partial least squares discriminant analysis, PLS-DA)对不同种质

的穿心莲进行分类鉴别;采用偏最小二乘回归(Partial least squares regression, PLSR)、反向传播神经网络(Back propagation neural network, BPNN)、随机森林回归(Random forest regression, RFR)对穿心莲化学成分含量进行预测。

BPNN 是一种常用的人工神经网络模型,由输入层、隐藏层和输出层等 3 个或多个神经元组成。在分析过程中,每个神经元接收来自上一层神经元的输入,并通过激活函数将其转换为输出。通过多次迭代训练,调整神经元之间的权重,使得实际值与预测值尽可能接近^[12]。PLSR 是一种多元回归方法,PLS-DA 是一种分类方法,它们都是基于偏最小二乘法的统计分析技术。PLSR 模型能很好地解决自变量之间多重自相关性的问题,能提取出相对系统解释能力最强的变量,尽可能地排除无作用的信息^[13]。RFR 是一种基于 Bagging 技术和决策树的算法,对于某个实测样本,可构建含 K 棵决策树的 RFR 模型,每棵决策树分别输出预测结果,取 K 棵树的预测均值作为最终结果。RFR 模型并行建立 K 棵树,对于大样本数据集仍能快速处理^[14]。

1.8 模型的评估

采用模型准确率以及混淆矩阵对分类模型的性能进行评价。混淆矩阵是分类模型预测结果的评价方式,具体的评价指标包括准确率(Accuracy)、灵敏度(Sensitivity)和特异性(Specificity)^[15],各指标的计算公式如(2)所示:

$$\begin{aligned} \text{Accuracy} &= (\text{TP} + \text{TN}) / (\text{TP} + \text{FN} + \text{FP} + \text{TN}) \\ \text{Sensitivity} &= \text{TP} / (\text{TP} + \text{FN}) \\ \text{Specificity} &= \text{TN} / (\text{TN} + \text{FP}) \end{aligned} \quad (2)$$

式中,TP 为真阳性样本数;TN 为真阴性样本数;FP 为假阳性样本;FN 为假阴性样本数。

采用训练集和预测集的决定系数(Coefficient of determination, R^2)、均方根误差(Root mean squares error, RMSE)和预测集剩余预测残差(Residual predictive deviation, RPD)评估回归模型的性能。 R^2 越接近 1, RMSE 越小, RPD 越大,模型的性能越好^[16]。若 $0.60 < R^2 < 0.80$, $1.50 < \text{RPD} < 2.50$, 表明该模型可以用于预测;若 $0.81 < R^2 < 0.90$, $2.51 < \text{RPD} < 3.0$, 表明该模型具有良好的预测性能;若 $R^2 > 0.90$, $\text{RPD} > 3.0$, 表明该模型具有优异的预测能力。

1.9 数据处理及分析软件

图像校正软件为高光谱成像系统 HySpex

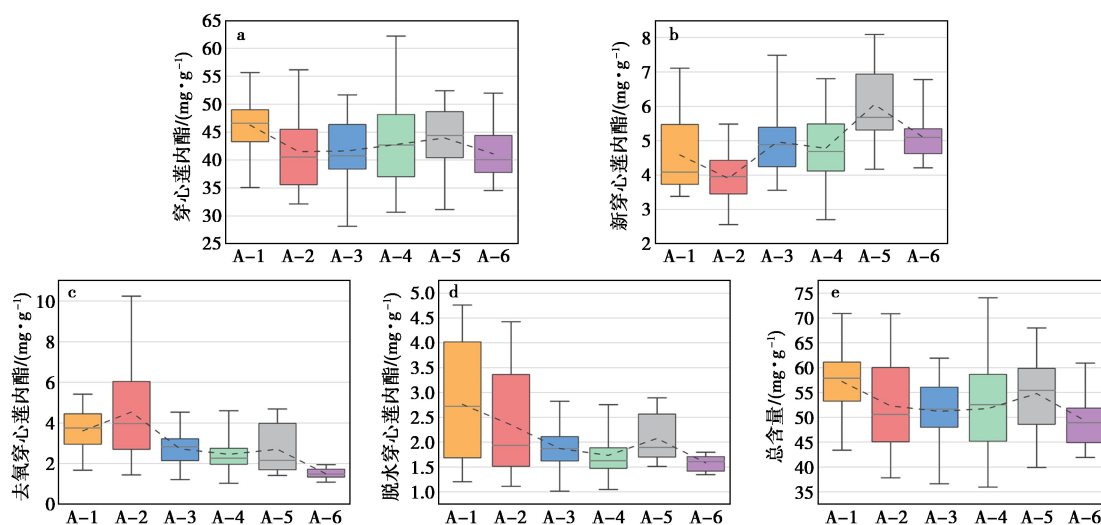
RAD 软件(挪威 Norsk Elektro Optikk 公司),黑白板校正和兴趣区域提取软件为 ENVI 5.3,光谱数据预处理及预测模型构建软件为 Matlab 2020a,脚本均为课题组自编。

2 结果与讨论

2.1 不同种质的穿心莲化学成分含量分析

不同种质的穿心莲样品含量如图 1 所示,穿心莲内酯的含量最高,范围为 28.16~62.19 mg/g,脱水穿心莲内酯的含量最低,范围为 1.01~4.76 mg/g,新穿心莲内酯和去氧穿心莲内酯的含量范

围分别为 2.56~8.09 mg/g、1.03~10.25 mg/g。不同种质的穿心莲样品穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯和脱水穿心莲内酯的含量有差异,穿心莲内酯平均含量最高的为 A-1,其余种质的平均含量相差不大;新穿心莲内酯平均含量最高的为 A-5,最低的为 A-2;去氧穿心莲内酯中 A-1、A-2 的平均含量较高,A-6 的最低;脱水穿心莲内酯中 A-1、A-2 的平均含量高于其余种质的。就穿心莲内酯类总含量而言,A-1 的平均含量最高,A-5 次之,A-2、A-3 和 A-4 的相差不大,A-6 最低。



图中折线为化学成分含量的平均值连接线

a. 穿心莲内酯; b. 新穿心莲内酯; c. 去氧穿心莲内酯; d. 脱水穿心莲内酯; e. 穿心莲内酯类总含量

图 1 不同种质穿心莲样品化学成分含量的箱式图

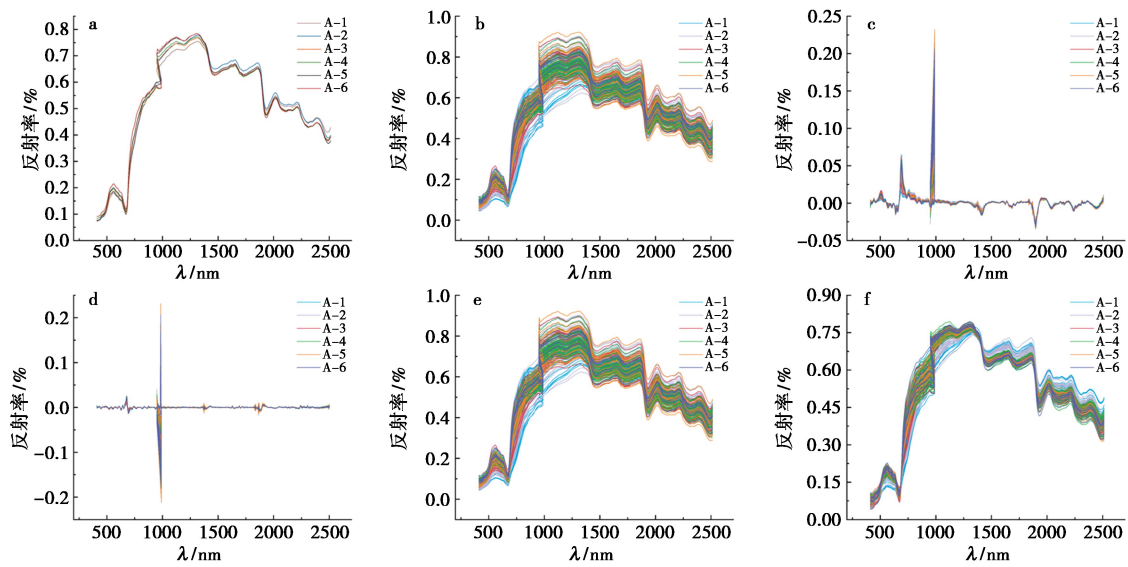
Fig.1 Box plot of chemical composition content in samples of different germplasms of *Andrographis paniculata*

2.2 光谱曲线分析

对不同种质的穿心莲样品 Raw Data 求平均值,绘制平均光谱曲线图(图 2a)。从总体上看,不同种质的穿心莲平均光谱曲线变化趋势较为相似。光谱曲线在 550 nm 处形成吸收峰,在 670~1 300 nm 内呈上升趋势,尤其是 670~960 nm,反射率急剧增大,在 1 300~2 500 nm 范围内呈波动下降趋势,有明显的吸收峰和吸收谷。对不同峰值进行归因,550 nm 附近的吸收峰对应 O—H 的第五泛音;840 nm 附近的吸收与 C=C—H 中 C—H 的第三泛音有关;970 nm 附近的吸收由 O—H 的第二泛音引起;1 120 nm 处的吸收峰和 1 300~1 400 nm 区域的吸收与 C—H 的第二拉伸泛音有关;1 470 nm 处的吸收对应 O—H 的第一泛音;1 600~1 700 nm 处的吸收与 C—H 的第一

泛音有关;1 950~2 000 nm 范围的吸收与酯键中 C=O 的第二拉伸泛音有关;在 2 000 nm 处的吸收峰与 O—H 伸缩振动和弯曲振动有关,2 200 nm 处是 C—H 和 C—O 的联合吸收峰;2 370 nm 附近的吸收对应 C—H 的第二泛音^[17-20]。不同种质的穿心莲吸收峰的位置具有较为相似的特征,但吸收强度存在差异,说明不同种质的穿心莲化学成分种类差异不大,但含量存在差异。

对 Raw Data 和预处理后的光谱数据绘图,结果见图 2b~2f。原始光谱图(图 2b)谱线重叠,基线漂移,难以直接观察吸收峰。经 D1、D2 预处理后,光谱曲线基线基本水平,吸收峰特征明显,尤其是 D2 预处理后,但也引入了噪声。与图 2b 相比,图 2e 光谱曲线更平滑,图 2f 光谱曲线间距明显缩小。



a.平均光谱;b.原始光谱;c.D1 预处理;d.D2 预处理;e.SG 预处理;f.MSC 预处理

图 2 不同种质穿心莲的平均光谱曲线、原始光谱曲线和预处理光谱曲线

Fig.2 Average spectral curve,original spectral curve,and preprocessed spectral curve of different germplasms of *Andrographis paniculata*

2.3 不同种质的穿心莲分类性能评估

将穿心莲样品 Raw Data 经 D1、D2、SG、MSC 预处理后,建立 PLS-DA 模型,结果见表 1。Raw Data 和不同预处理后的光谱数据建立的模型训练集和预测集准确率均大于 90%,经 D1、D2、MSC 预处理后,模型的训练集和预测集准确率分别提高 4.1%和 4.6%、4.1%和 2.7%、0.2%和 2.7%。D1-PLS-DA 模型的性能最好,预测结果的混淆矩阵如图 3 所示,不同种质的灵敏度均大于 90%,精确率均大于 75%。因此,对于不同种质的穿心莲样品,利用高光谱技术,结合有效的预处理方法,可建立 PLS-DA 模型加以鉴别。其原因可能是不同种质的穿心莲药材穿心莲内酯类成分含量差异较大。

表 1 不同种质的穿心莲在各种预处理方法下建立的 PLS-DA 模型的分​​类准确率

Tab.1 Classification accuracy of PLS-DA models established under various preprocessing methods for different germplasms of *Andrographis paniculata*

预处理	训练集	预测集
Raw Data	0.953	0.917
D1	0.994	0.963
D2	0.994	0.944
SG	0.953	0.900
MSC	0.955	0.944

预测值	A-1	A-2	A-3	A-4	A-5	A-6
	7 13.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	0 0.0%	11 20.4%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	0 0.0%	0 0.0%	7 13.0%	1 1.9%	0 0.0%	87.5% 12.5%
	0 0.0%	0 0.0%	0 0.0%	19 35.2%	0 0.0%	100% 0.0%
	0 0.0%	0 0.0%	0 0.0%	0 0.0%	5 9.3%	100% 0.0%
	0 0.0%	0 0.0%	0 0.0%	1 1.9%	0 0.0%	3 5.6%
真实值						
	100% 0.0%	100% 0.0%	100% 0.0%	90.5% 9.5%	100% 0.0%	100% 0.0%
						96.3% 3.7%

图 3 不同种质的穿心莲 D1-PLS-DA 分类模型预测集预测结果的混淆矩阵

Fig.3 Confusion matrix of prediction results for D1-PLS-DA classification model prediction sets of different germplasms of *Andrographis paniculata*

2.4 穿心莲中 4 种穿心莲内酯类化合物的含量预测评估

2.4.1 基于全波长数据预测穿心莲中 4 种穿心莲内酯类化合物的含量

利用穿心莲样品 Raw Data,结合含量测定结果,分别建立穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯、脱水穿心莲内酯和穿心莲内酯类总含量的 PLSR、BPNN、RFR 模型,结果见表 2。穿心莲内酯、新穿心莲内酯、穿心莲内酯类总含量的 PLSR 模型性能最好,BPNN 模型次之,去氧穿心莲内酯、脱水穿心莲内酯的 BPNN 模型性能最好,PLSR 次之,除脱水穿心莲内酯外,其余 3 种穿心莲内酯类化合物及穿心莲内酯类总含量的 RFR

模型预测集 R^2 均小于 0.6, RPD 值均小于 1.5, 预测效果不理想。

表 2 穿心莲全波段 Raw Data 不同模型的预测结果

Tab.2 Prediction results of different models for the full band Raw Data of *Andrographis paniculata*

化学成分	模型	训练集		预测集		
		R^2	RMSE	R^2	RMSE	RPD
穿心莲内酯	PLSR	0.835	2.67	0.823	2.80	2.32
	BPNN	0.804	2.90	0.735	3.42	1.67
	RFR	0.761	3.21	0.195	5.97	0.54
新穿心莲内酯	PLSR	0.901	0.36	0.758	0.51	2.10
	BPNN	0.906	0.36	0.706	0.56	1.99
	RFR	0.681	0.65	0.004	1.03	0.55
去氧穿心莲内酯	PLSR	0.954	0.34	0.909	0.52	3.70
	BPNN	0.967	0.29	0.946	0.40	4.53
	RFR	0.832	0.65	0.546	1.17	1.14
脱水穿心莲内酯	PLSR	0.975	0.14	0.833	0.25	2.91
	BPNN	0.978	0.13	0.849	0.23	3.00
	RFR	0.897	0.28	0.625	0.37	1.92
穿心莲内酯类总含量	PLSR	0.903	2.54	0.853	3.06	2.74
	BPNN	0.800	3.65	0.818	3.39	2.08
	RFR	0.779	3.83	0.283	6.74	0.75

接着将经 D1、D2、SG、MSC 预处理的光谱数据结合含量测定结果分别建立 PLSR、BPNN 模型, 结果见表 3。在穿心莲内酯的含量预测中, 与 BPNN 模型相比, PLSR 模型性能更好, 具有更高的 R^2 和 RPD 值。其中, SG-PLSR 模型的训练集和预测集的 R^2 分别为 0.841 和 0.829, RPD 值为

2.36, 预测结果散点图如图 4a 所示, 斜率和 R^2 均大于 0.8, 表明 SG-PLSR 模型可用于穿心莲内酯的含量预测。在新穿心莲内酯的含量预测方面, 与 BPNN 模型相比, PLSR 模型也展现出更佳的性能, 其中 MSC-PLSR 模型的训练集和预测集的 R^2 分别为 0.895 和 0.783, RPD 值为 2.25, 预测结果散点图如图 4b 所示, 斜率为 0.933, R^2 为 0.806, 说明 MSC-PLSR 模型可用于新穿心莲内酯的含量预测。在去氧穿心莲内酯的含量预测中, 除 D2-PLSR 和 D2-BPNN 模型外, 其余模型的训练集和预测集 R^2 均大于 0.9, RMSE 值均趋近于 0, RPD 值均大于 3.0。其中, Raw Data-BPNN 模型性能最好, 预测结果散点图如图 4c 所示, 斜率接近 1, R^2 大于 0.9, 模型的预测值与参考值相差较小。在脱水穿心莲内酯的含量预测中, 除 D2-PLSR 和 D2-BPNN 模型外, 其余模型均具有良好的预测能力。其中, MSC-BPNN 模型的训练集和预测集 R^2 最大, RMSE 值最小, RPD 值最大。该模型的预测结果散点图如图 4d 所示, 斜率接近 1, R^2 大于 0.9, 模型的预测值与参考值的差值较小。在穿心莲内酯类总含量的预测方面, Raw Data-PLSR、SG-PLSR、D1-PLSR 模型的训练集和预测集 R^2 均大于 0.8, RPD 值均大于 2.5, 其中 Raw Data-PLSR 模型性能最好, 预测结果散点图如图 4e 所示, 斜率为 0.969, R^2 为 0.867。

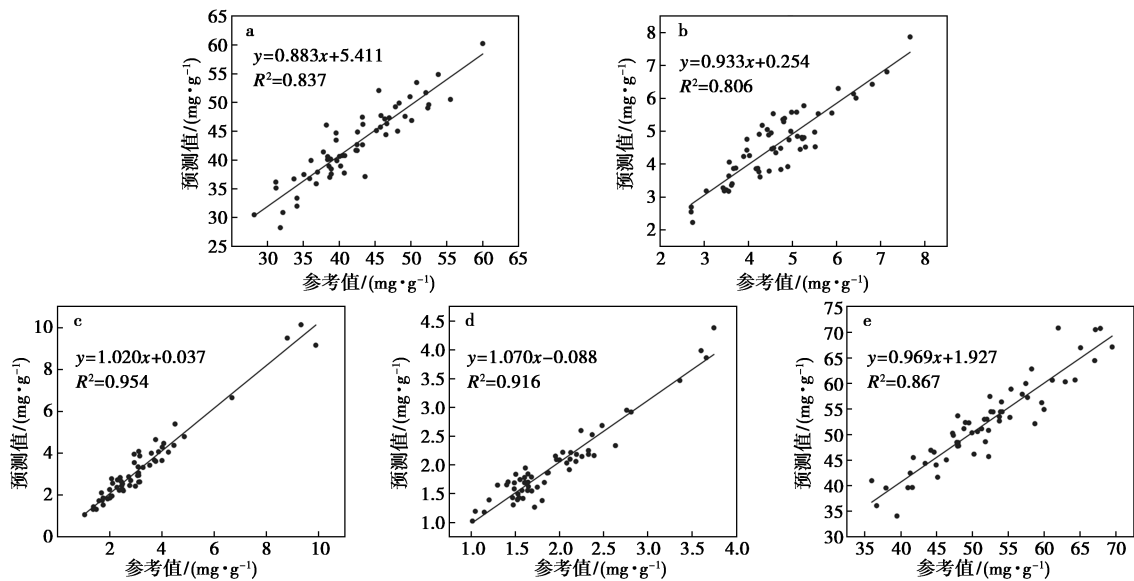
表 3 穿心莲全波段光谱数据经不同的方法预处理后建立的 PLSR 和 BPNN 模型的预测结果

Tab.3 Prediction results of PLSR and BPNN models established after preprocessing the full band spectral data of *Andrographis paniculata* using different methods

化学成分	预处理	PLSR					BPNN				
		训练集		预测集			训练集		预测集		
		R^2	RMSE	R^2	RMSE	RPD	R^2	RMSE	R^2	RMSE	RPD
穿心莲内酯	Raw Data	0.835	2.67	0.823	2.80	2.32	0.804	2.90	0.735	3.42	1.67
	D1	0.850	2.54	0.842	2.65	2.26	0.797	2.96	0.753	3.30	1.55
	D2	0.672	3.76	0.658	3.89	1.67	0.764	3.19	0.617	4.11	1.52
	SG	0.841	2.62	0.829	2.75	2.36	0.777	3.10	0.761	3.25	2.05
	MSC	0.833	2.69	0.854	2.55	2.25	0.835	2.67	0.709	3.59	1.78
新穿心莲内酯	Raw Data	0.901	0.36	0.758	0.51	2.10	0.906	0.36	0.706	0.56	1.99
	D1	0.865	0.42	0.604	0.65	1.58	0.888	0.39	0.417	0.79	1.42
	D2	0.830	0.48	0.541	0.70	1.38	0.705	0.63	0.538	0.70	1.18
	SG	0.900	0.36	0.770	0.50	2.20	0.823	0.49	0.622	0.64	1.64
	MSC	0.895	0.38	0.783	0.48	2.25	0.905	0.36	0.622	0.64	1.60
去氧穿心莲内酯	Raw Data	0.954	0.34	0.909	0.52	3.70	0.967	0.29	0.946	0.40	4.53
	D1	0.936	0.40	0.935	0.44	4.11	0.963	0.31	0.921	0.49	3.63
	D2	0.902	0.50	0.844	0.69	2.16	0.873	0.56	0.830	0.72	1.87
	SG	0.951	0.35	0.909	0.52	3.67	0.952	0.35	0.938	0.43	4.14
	MSC	0.948	0.36	0.905	0.54	3.48	0.965	0.30	0.905	0.54	3.66

续表

化学成分	预处理	PLSR					BPNN				
		训练集		预测集			训练集		预测集		
		R^2	RMSE	R^2	RMSE	RPD	R^2	RMSE	R^2	RMSE	RPD
脱水穿心莲内酯	Raw Data	0.975	0.14	0.833	0.25	2.91	0.978	0.13	0.849	0.23	3.00
	D1	0.970	0.15	0.820	0.25	2.76	0.974	0.14	0.858	0.23	2.70
	D2	0.936	0.22	0.790	0.28	2.34	0.873	0.31	0.694	0.33	1.99
	SG	0.984	0.11	0.854	0.23	3.16	0.979	0.13	0.872	0.22	3.20
	MSC	0.980	0.12	0.855	0.23	3.01	0.985	0.11	0.884	0.20	3.31
穿心莲内酯类总含量	Raw Data	0.903	2.54	0.853	3.06	2.74	0.800	3.65	0.818	3.39	2.08
	D1	0.886	2.76	0.853	3.06	2.51	0.898	2.60	0.712	4.28	1.68
	D2	0.852	3.13	0.754	3.95	2.05	0.680	4.61	0.671	4.57	1.14
	SG	0.886	2.75	0.849	3.10	2.59	0.839	3.27	0.810	3.47	2.21
	MSC	0.898	2.61	0.842	3.16	2.27	0.856	3.10	0.787	3.68	1.89



a.穿心莲内酯的 SG-PLSR 模型;b.新穿心莲内酯的 MSC-PLSR 模型;c.去氧穿心莲内酯的 Raw Data-BPNN 模型;
d.脱水穿心莲内酯的 MSC-BPNN 模型;e.穿心莲内酯类总含量的 Raw Data-PLSR 模型

图 4 基于全波长建立的穿心莲化合物含量的最佳预测模型的参考值和预测值的散点图

Fig.4 Scatter plots of reference and predicted values for the optimal prediction model of compound contents in *Andrographis paniculata* based on full wavelengths

综上所述,本研究分别建立了穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯和脱水穿心莲内酯的含量预测模型,实现了对上述 4 种穿心莲内酯类化合物的含量预测。《中国药典(2020 版)》^[1]规定,穿心莲药材中含穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯和脱水穿心莲内酯的总量不得少于 1.5%,鉴于此,本研究建立了穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯和脱水穿心莲内酯的总含量的预测模型,为穿心莲品质快速检测提供新方法。

2.4.2 基于特征波长数据预测穿心莲中 4 种穿心莲内酯类化合物的含量

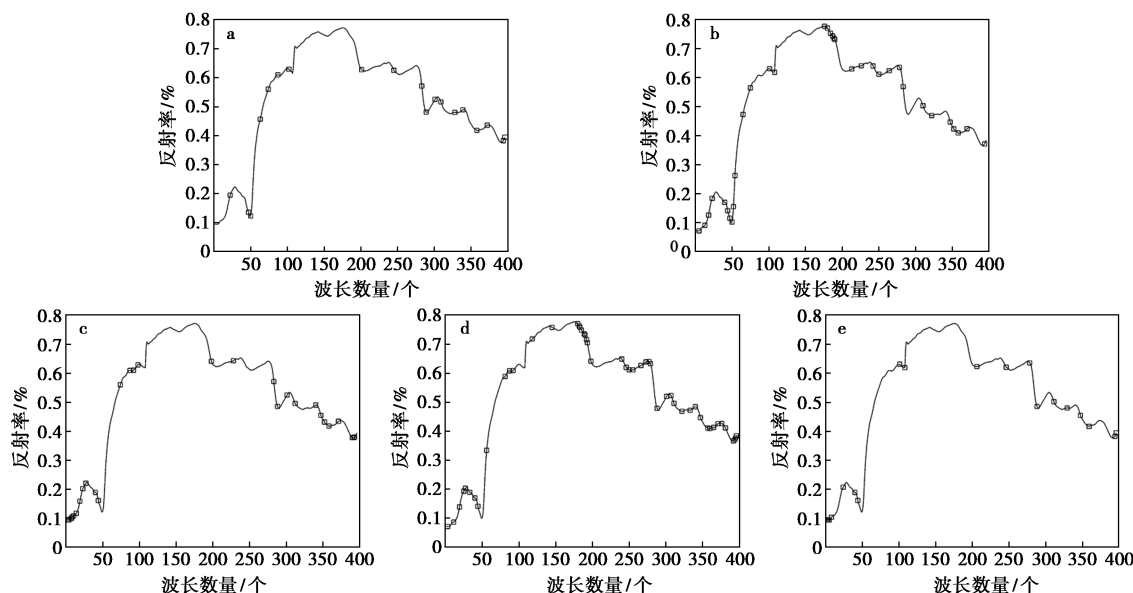
根据全波长的预测结果,进一步选取一些预测效果较好的模型结合 SPA 来选择特征波长。

具体选择的模型为:用于预测穿心莲内酯的 SG-PLSR 模型、预测新穿心莲内酯的 MSC-PLSR 模型、预测去氧穿心莲内酯的 Raw Data-BPNN 模型、预测脱水穿心莲内酯的 MSC-BPNN 模型和预测穿心莲内酯类总含量的 Raw Data-PLSR 模型。

SPA 筛选的特征波长如图 5 所示,这些特征波长大部分分布在吸收峰和吸收谷附近。穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯、脱水穿心莲内酯和穿心莲内酯类总含量筛选出的特征波长数目分别为 19、34、29、49、19,波长数目分别减少为全波长的 4.80%、8.59%、7.32%、12.37%、4.80%。穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯和脱水穿心莲内酯的结构相似,均含有—C=C—、—COOR 和—OH 官能团。结合图 2

可知,SPA 筛选的特征波长与 4 种穿心莲内酯类化合物的结构存在密切关系。在 410~670、900~1 000、1 470 nm 附近选择的特征波长可能与 4 种穿心莲内酯类化合物的—OH 有关,在 800~900 nm 附近选择的特征波长可能与 4 种穿心莲内酯类化合物的—C=C—有关,在 1 950~2 000、2 200 nm

区域选择的特征波长可能与 4 种穿心莲内酯类化合物的一COOR 有关,在 1 120、1 300~1 400、1 600~1 700、2 370 nm 附近选择的特征波长可能与 4 种穿心莲内酯类化合物的一CH、—CH₂、—CH₃ 有关。因此,可通过穿心莲的光谱信息分析其官能团信息,进而对穿心莲品质进行分析。



“□”即提取的特征波长

a.穿心莲内酯;b.新穿心莲内酯;c.去氧穿心莲内酯;d.脱水穿心莲内酯;e.穿心莲内酯类总含量

图 5 SPA 筛选特征波长

Fig.5 SPA filtering feature bands

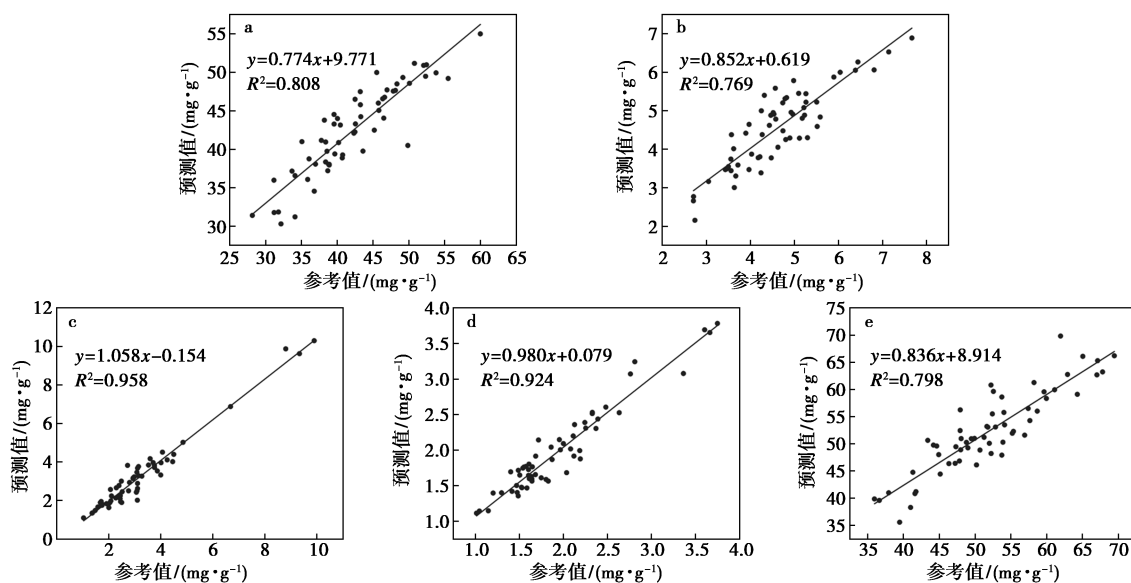
经 SPA 筛选后建模,预测结果见表 4。穿心莲内酯的 SG-SPA-PLSR 模型和新穿心莲内酯的 MSC-SPA-PLSR 模型的训练集和预测集 R^2 均大于 0.6,RPD 值均大于 1.5,预测结果散点图如图 6a 和 6b 所示,斜率和 R^2 均大于 0.75,说明经 SPA 提取特征波长后,降低了穿心莲内酯和新穿心莲内酯含量预测模型的复杂程度,在波长数量有限的情况下实现了含量预测。在去氧穿心莲内酯的含量预测中,与全波长 Raw Data-BPNN 模型相比,提取特征波长后建立的 Raw Data-SPA-BPNN 模型,RPD 值提高 4.42%,预测结果散点图如图 6c 所示,斜率接近 1, R^2 大于 0.9,模型的预测性能好。在脱水穿心莲内酯的含量预测中,与全波长 SG-BPNN 模型相比,筛选特征波长后建立的 SG-SPA-BPNN 模型,RPD 值提高 7.25%,预测结果散点图如图 6d 所示,斜率和 R^2 均大于 0.9,模型的预测值与参考值相接近。因此,在去氧穿心莲内酯和脱水穿心莲内酯的含量预测中,SPA 提取特征波长后,剔除了不相关变量,简化了建模

过程,提高了模型的精度。在穿心莲内酯类总含量的预测方面,Raw Data-SPA-PLSR 模型的训练集和预测集 R^2 在 0.61~0.80 范围内,RPD 值大于 2.0,预测结果散点图如图 6e 所示,斜率大于 0.8,说明该模型可用于穿心莲内酯类总含量的预测。因此,SPA 筛选的特征波长可有效表征 4 种穿心莲内酯类化合物的光谱信息,简化模型,为后

表 4 穿心莲样品光谱数据经 SPA 筛选后不同组合模型的预测结果

Tab.4 Prediction results of different combination models for spectral data of *Andrographis paniculata* samples after SPA screening

化学成分	模型	特征波长个数	训练集		预测集		
			R^2	RMSE	R^2	RMSE	RPD
穿心莲内酯	SG-PLSR	19	0.784	3.05	0.805	2.93	1.97
新穿心莲内酯	MSC-PLSR	34	0.826	0.48	0.756	0.51	1.98
去氧穿心莲内酯	Raw Data-BPNN	29	0.961	0.31	0.947	0.40	4.73
脱水穿心莲内酯	MSC-BPNN	49	0.981	0.12	0.916	0.17	3.55
穿心莲内酯类总含量	Raw Data-PLSR	19	0.796	3.69	0.794	3.62	2.08



a. 穿心莲内酯的 SG-SPA-PLSR 模型; b. 新穿心莲内酯的 MSC-SPA-PLSR 模型; c. 去氧穿心莲内酯的 Raw Data-SPA-BPNN 模型; d. 脱水穿心莲内酯的 MSC-SPA-BPNN 模型; e. 穿心莲内酯类总含量的 Raw Data-SPA-PLSR 模型

图 6 基于特征波长建立的穿心莲化合物含量的最佳预测模型的参考值和预测值的散点图

Fig.6 Scatter plots of reference and predicted values for the optimal prediction model of compound contents in *Andrographis paniculata* based on the selected wavelengths

续穿心莲品质检测专属小型化高光谱设备的开发提供了方法参考。

2.5 分析与讨论

2.5.1 不同预处理方法对穿心莲中 4 种穿心莲内酯类化合物含量预测模型的影响

含量预测模型构建前,对 Raw Data 进行合适的预处理,可有效减少杂散光、噪声、基线漂移等因素的干扰,提高模型精度。在本研究中,采用了 D1、D2、MSC、SG 这 4 种预处理方法。其中,D1 是一种基线校正方法,通过对重叠光谱进行拆分,增大光谱曲线中波峰和波谷的特征,从而实现基线校正^[21]。在去氧穿心莲内酯的含量预测中,与 Raw Data-PLSR 模型相比,D1-PLSR 模型的 RPD 值提高了 11.08%,说明经 D1 预处理后,有效扣除基线漂移对信号的影响,提高模型精度。D2 也是一种基线校正方法,但在本研究中,4 种穿心莲内酯类化合物的 Raw Data 经 D2 预处理后建立的模型,精度均未提高,可能是在预处理过程中,对基线进行校正的同时,也放大了噪声。MSC 属于散射校正,在脱水穿心莲内酯的含量预测中,全波段 Raw Data 经 MSC 预处理后建模,无论是 PLSR 模型还是 BPNN 模型,RPD 值均提高,这可能是 MSC 消除了由于颗粒分布不均匀及颗粒大小不同所产生的散射对光谱的影响^[22]。SG 是一种平滑处理方法,在穿心莲内酯的含量预测中,最好的

模型即 SG-PLSR,这可能与 SG 可以减小随机误差,降低噪声,提高谱图信噪比有关^[22]。因此,合适的预处理方法,可以有效消除曲线漂移和散射效应,增大光谱数据的信噪比,从而提高模型的准确度和稳定性。

2.5.2 SPA 对穿心莲中 4 种穿心莲内酯类化合物含量预测模型的影响

高光谱数据庞大,其中包含了大量的冗余信息,数据间存在共线性和重叠问题,而且还可能存在与待测样品性质无关的变量,因此,本研究运用 SPA 筛选特征波长,减少数据的维度,简化模型。在去氧穿心莲内酯和脱水穿心莲内酯的含量预测中,经 SPA 筛选后建立的模型,RPD 值分别增加 4.42% 和 7.25%,模型精度提高,这与 SPA 所选择的波长具有最小的冗余和包含最多的特征相关信息有关^[23]。在穿心莲内酯、新穿心莲内酯和 4 种穿心莲内酯类化合物总含量的预测模型构建中,经 SPA 筛选后建立的模型预测效果没有全波段的好,这可能是在降维过程中,损失了部分重要的光谱信息,导致模型精度有差距,但经 SPA 筛选后,减少了数据量,提高了运算效率,节省了运算时间。

3 结论

本实验以穿心莲为对象,研究了基于高光谱

技术的不同种质穿心莲药材的判别方法和穿心莲化学成分含量的检测方法。研究结果表明高光谱技术结合化学计量学方法可实现穿心莲品质的准确、快速检测。对于不同种质穿心莲的判别,建立 D1-PLS-DA 模型可获得较好的判别准确率。对于穿心莲内酯、新穿心莲内酯、去氧穿心莲内酯、脱水穿心莲内酯和穿心莲内酯类总含量的检测,最佳性能的模型分别为 SG-PLSR、MSC-PLSR、Raw Data-SPA-BPNN、MSC-SPA-BPNN 和 Raw Data-PLSR。该研究可为不同种质穿心莲品质的快速准确检测提供方法和技术支持,但不足之处在于未能使用优化算法对分类模型和回归模型进行参数优化,后续将结合优化算法建立不同种质穿心莲品质检测的分类模型和回归模型,提高模型精度。

参考文献:

- [1] 国家药典委员会. 中华人民共和国药典[M]. 一部. 北京: 中国医药科技出版社, 2020.
- [2] 陈东亮, 钟楚, 林阳. 药用植物穿心莲种质资源、育种及栽培研究进展[J]. 江苏农业科学, 2020, **48**(21): 34-40.
- [3] 周芳, 孙铭阳, 梅瑜, 等. 药用植物穿心莲研究进展[J]. 广东农业科学, 2021, **48**(1): 9-16.
- [4] 张晓, 唐力英, 吴宏伟, 等. 穿心莲现代研究进展[J]. 中国实验方剂学杂志, 2018, **24**(18): 222-234.
- [5] 肖传学, 孙玉侠, 耿晓梅, 等. 不同药用部位及产地穿心莲药材中 4 个二萜内酯类成分的研究[J]. 中国现代中药, 2017, **19**(5): 675-678.
- [6] 滕云霞, 张飞宇, 毛鑫, 等. 复方穿心莲片 HPLC 含量测定方法研究[J]. 广东化工, 2023, **50**(13): 201-204.
- [7] 李振雨, 赵书运, 刘晓霞, 等. 基于 UPLC 特征图谱和一测多评法的穿心莲药材质量研究[J]. 中国中医药信息杂志, 2023, **30**(6): 141-147.
- [8] 赖秀娣, 林晓菁, 龚雪, 等. 近红外光谱法快速测定穿心莲中穿心莲内酯的含量[J]. 中国医药工业杂志, 2018, **49**(9): 1 300-1 305.
- [9] 张璐, 茹晨雷, 殷文俊, 等. 基于近红外高光谱成像结合分水岭算法鉴别酸枣仁药材的产地[J]. 药物分析杂志, 2021, **41**(4): 726-734.
- [10] 史磊. 基于高光谱成像技术的三七皂苷无损检测研究[D]. 昆明: 昆明理工大学, 2023.
- [11] 李欣庭, 张峰, 冯洁. 卷积神经网络结合改进光谱处理方法用于马铃薯病害检测[J]. 光谱学与光谱分析, 2024, **44**(1): 215-224.
- [12] 谢潮勇. 基于可见/近红外光谱技术结合化学计量学测量土壤养分含量[D]. 上海: 华东交通大学, 2023.
- [13] 易翔. 基于无人机滴灌棉田氮营养监测与分区管理研究[D]. 石河子: 石河子大学, 2023.
- [14] 韩宣宣. 多源无人机遥感影像的冬小麦长势监测及产量估算[D]. 合肥: 安徽大学, 2022.
- [15] 张悦, 周骏辉, 王思曼, 等. 基于可见光-近红外高光谱成像技术的陈皮陈化年份鉴别研究[J]. 光谱学与光谱分析, 2023, **43**(10): 3 286-3 292.
- [16] 刘玲玲, 王游游, 杨健, 等. 基于高光谱技术的枸杞子化学成分含量快速检测技术研究[J]. 中国中药杂志, 2023, **48**(16): 4328-4336.
- [17] 张悦, 王游游, 张婷, 等. 高光谱结合图分割算法快速鉴别不同尺度产地陈皮[J]. 化学试剂, 2023, **45**(1): 136-143.
- [18] YOUYOU W, FENG X, YUE Z, et al. Application of hyperspectral imaging assisted with integrated deep learning approaches in identifying geographical origins and predicting nutrient contents of Coix seeds [J]. *Food Chem.*, 2023, **404**: 134 503-134 512.
- [19] QIN O, LI W, BOSOON P, et al. Simultaneous quantification of chemical constituents in matcha with visible-near infrared hyperspectral imaging technology[J]. *Food Chem.*, 2021, **350**: 129 141-129 148.
- [20] ZHANG H, JIA B, LU Y, et al. Detection of aflatoxin B₁ in single peanut kernels by combining hyperspectral and microscopic imaging technologies [J]. *Sensors (Basel)*, 2022, **22**(13): 4 864-4 878.
- [21] 张恒, 梁太波, 宋效东, 等. 基于高光谱成像的烟田土壤 pH 估测研究[J]. 西南农业学报, 2023, **36**(12): 2 771-2 779.
- [22] 第五鹏瑶, 卞希慧, 王姿方, 等. 光谱预处理方法选择研究[J]. 光谱学与光谱分析, 2019, **39**(9): 2 800-2 806.
- [23] 岳明慧, 张姗姗, 张静, 等. 基于高光谱成像技术的薯麦混合粉淀粉含量检测[J]. 中国粮油学报, 2023, **38**(12): 197-202.