基于近红外高光谱成像鉴别不同产地的红参

沈国芳[1] 黄俊航[2] 许麦成[3] 金 强[3]

(1 杭州市食品药品检验研究院,杭州,310022; 2 浙江大学药学院,杭州,310058; 3 杭州胡庆余堂药业有限公司,杭州,311100)

摘要 **目的:** 建立一种基于近红外(NIR)高光谱成像技术融合近红外光谱和图像纹理特征,鉴别不同产地红参药材的方法。**方法:** 提取红参 ROI 近红外高光谱数据,采用多种预处理算法对光谱数据进行降噪处理。利用灰度共生矩阵(GLCM)和灰度游程矩阵(GLRLM)提取图像纹理特征,实现 NIR 光谱和图像纹理数据融合。利用偏最小二乘判别分析(PLS-DA)和支持向量机分类(SVC)建立产地分类模型。**结果:** 全波段光谱融合 GLRLM 所构建的模型性能最佳,准确率分别为 90.0%和 91.2%。进一步地使用混淆矩阵和 ROC 曲线对模型进行评估。混淆矩阵中 SVC 模型表现优异,对吉林、黑龙江和辽宁 3 个产地的分类准确率可达 100%、91%和 83%;经 ROC 特征曲线评估,2 个模型的最优曲线下面积值分别达到了 0.97 和 0.96。**结论:** 本研究为快速鉴别红参药材不同产地提供了一种新方法。

关键词 高光谱成像;红参;数据融合;产地鉴别

Discriminant Analysis of Red Ginseng from Different Origins Based on NIR-Hyperspectral Imaging Technology

SHEN Guofang[1], HUANG Junhang[2], XU Maicheng[3], JIN Qiang[3]

(1 *Hangzhou Institute for Food and Drug Control, Hangzhou* 310022, *China*; 2 *College of Pharmaceutical Science, Zhejiang University, Hangzhou* 310058, *China*; 3 *Hangzhou Huqingyutang Pharmaceutical Co., Ltd., Hangzhou* 311100, *China*)

Abstract **Objective:** To establish a method based on near-infrared (NIR) hyperspectral imaging technology which could identify red ginseng from different origins by fusing NIR spectra and image texture features. **Methods:** The near-infrared hyperspectral data of red ginseng ROI were extracted to be further de-noised by various pre-processing algorithm. Texture features are extracted from images using gray co-occurrence matrix (GLCM) and gray run matrix (GLRLM), and near-infrared spectroscopy and image data are fused. Partial least squares discriminant analysis (PLS-DA) and support vector machine classification (SVC) were used here to establish the origin classification model. **Results:** The model constructed by full-band spectral combined with GLRLM gained the best performance, with accuracy of 90.0% and 91.2%, respectively. The model was further evaluated using confusion matrix and ROC curves. The SVC model performed better in the confusion matrix, with the classification accuracy of 100%, 91% and 83% for red ginseng from Jilin, Heilongjiang and Liaoning. According to ROC characteristic curve evaluation, the areas under the optimal curve of the two models are 0.97 and 0.96, respectively. **Conclusion:** This research provides a new method for rapid identification of red ginseng from different origins.

Keywords Hyperspectral imaging; Red ginseng; Data fusion; Origin identification

中图分类号: R282.5 **文献标识码:** A **doi:** 10.3969/j.issn.1673-7202.2021.23.003

红参为五加科植物人参 *Panax ginseng* C. A. Mey. 的栽培品经蒸制后的干燥根和根茎。在我国主产于东北 3 省,具有大补元气,复脉固脱,益气摄血的功效[1]。红参作为一种常用中药材,在中药制剂中应用广泛,随着对人参需求量的提升,人参栽培受限于连作障碍问题,其产地由吉林为主向东北各地扩展,不同产地的人参及其加工品红参存在着较大的质量差异,为确保中药制剂质量稳定,加强红参原料质量控制,对红参产地进行鉴别区分具有较大意义[2-4]。

传统红参鉴别以经验判断真伪优劣和大致产地,受检验人员个人经验影响大、重复性差,随着化学分析手段的进步,薄层色谱、液相色谱等虽然能够准确检测出样本之间的差异,但前处理耗时费力且检测成本高,无法满足工业化生产对红参快速在线分选的要求[6-7]。高光谱成像(Hyper Spectral Imaging,HSI)技术能够同时采集对象品质属性的光谱信息和图像信息,是一种快速、无损、原位成像的检测技术[8]。近几年在食品、农产品等领域的应用较多,在中药材甄别掺假品、硫熏品、染色增重品和含量不合格等领域也逐步开始应用[9~10]。本研究以来源不同产地的红参样品为研究对象,利用高光谱成像技术、光谱预处理、数据融合方法和分类模型算法对不同产地的红参进行判别分析,使用混淆矩阵

基金项目:浙江省市场监督管理局雏鹰计划培育项目(CY2022345)

通信作者:沈国芳(1978.12—),男,硕士,副主任药师,研究方向:药品质量控制与药物分析,E-mail:shengf@126.com

和 ROC 对不同模型预测性能进行评估，对比了不同模型的分类结果，为实现在线快速无损识别不同产地的红参提供参考。

1 材料与方法

1.1 药材 实验用红参药材选自东北 3 省，分别是辽宁（6 批次），吉林（11 批次）和黑龙江（5 批次）。经杭州市食品药品检验研究院郭怡飚主任中药师鉴定为五加科植物人参 *Panax ginseng* C. A. Mey. 的栽培品经蒸制后的干燥根。22 批次红参药材共收集到 304 个红参样品用于高光谱图像分析，其中辽宁产地红参 58 根，黑龙江产地红参 110 根，吉林产地红参 136 根。按照 Kennard-Stone 算法将样本分成训练集和测试集，其中训练集 203 样本，测试集 101 个样本。

1.2 仪器设备 本研究采用的高光谱成像系统由高光谱成像模块、移动平台、均匀光源、计算机与图像采集软件等部分组成。高光谱成像模块包含近红外光谱相机（OWL-640-mini，Raptor Photonics），可调节近红外聚焦透镜（OLE23，Specim），2 个 150 W 的卤素灯（3900ER，Illumination Technologies Inc.）和成像光谱仪（ImSpector-V10E，Specim）；移动平台（ETH14，TOYO）；图像采集软件为 Spectral Image 软件（Isuzu Optics）。

1.3 高光谱成像系统的参数设置 水平移动平台的移动速度为 2.4 mm/s，镜头与样品之间的距离为 25 cm，电荷耦合器件（CCD）相机的曝光时间为 26 ms。获得的高光谱数据立方体，其宽度为 640 像素，长为 1 000 像素，以 1.67 nm 的间隔从 898 ~ 1 751 nm的 512 个波长。

1.4 高光谱图像的黑白板校正 在采集得到高光谱图像后，为了减小光源不均匀、CCD 相机的暗电流以及仪器物理配置的差异等对所获得的高光谱反射率图像的影响，需要对采集的原始高光谱图像进行黑白板校正，校正公式为：$R_{cal}=\frac{R_{raw}-R_{dark}}{R_{white}-R_{dark}}$式中，$R_{cal}$为校正后的高光谱数据，$R_{raw}$为采集到的原始高光谱数据，$R_{dark}$为盖上相机镜头采集到的数据（反射率接近 0），$R_{white}$为对准 Teflon 白板采集到的数据（反射率接近 1）。校正步骤采用 HSI Analyzer 软件进行。

1.5 提取 ROI 平均光谱 为节约提取感兴趣区域（Region Of Interest，ROI）的时间，本实验采用 Python3.6 自动提取 ROI。高光谱图像经黑白板校正后，为最大程度区分开样品和背景，选择与背景差异最大波段的灰度图。将该灰度图进行二值化，得到二进制图像。最后提取二进制图像中红参的轮廓，得到 ROI。对 ROI 中每个波段灰度图所有像素点的光谱反射率求平均值，得到 ROI 平均光谱。

1.6 光谱预处理和特征波长提取 用合适的光谱预处理方法可以降低各种非目标因素对检测信号信息的影响，增强有用信息[7]。分别采用 Savitzky-Golay（SG）平滑算法，基于 SG 平滑的一阶导数，基于 SG 平滑的一阶导数二阶导数，标准正态变量变换（SNV）算法和多元散射矫正（MSC）算法等对光谱进行处理，并对这几种预处理算法进行比较。连续投影算法（Successive Projections Algorithm，SPA）是一种向前循环选择方法，可以最大限度地消除变量中存在的共线性信息[11]。使用 SPA 对波段范围内光谱数据进行特征波长提取，作为图像纹理特征提取时的特征波段。

1.7 图像纹理特征提取 分别通过灰度共生矩阵（Gray-Level Co-occurrence Matrix，GLCM）和灰度游程矩阵（Gray-Level Run-Length Matrix，GLRLM）方法提取图像纹理特征。GLCM 是一种经典的图像纹理提取方法，其主要描述局部空间域的强度变化。本研究从 4 个方向（0°，45°，90°，135°）提取 GLCM 的对比度、能量、逆差矩和熵，共获得 16 个参数。GLRLM 描述相同像素值在特定方向的分布，游走的长度为游走方向的像素值[12]。长游走长度提取粗糙纹理，短游走长度提取细腻纹理。本研究从 GLRLM 提取 10 个特征，分别为短游程因子、长游程因子、灰度不均匀度、游程比、游程长不均匀度、低灰度游程因子、高灰度游程因子、低灰度短游程因子、高灰度短游程因子、高灰度长游程因子。高光谱图像有数百个波段，若对所有波段相对应的灰度图像计算 GLCM 和 GLRLM 纹理，存在大量的冗余信息，且增加计算复杂性。因此，本研究中仅提取通过 SPA 算法得到的特征波段灰度图像中的纹理信息。

1.8 NIR 光谱和图像纹理融合 数据融合可以有效提高多分类模型的性能，不同种类的数据可以实现信息的互补，从而提高模型的准确率和鲁棒性。近红外高光谱包含丰富的近红外光谱信息和图像纹理信息，通过对 NIR 光谱与图像纹理特征进行融合，可以提升鉴别模型的性能[13]。

1.9 红参产地分类模型的构建和模型评价 本研究分别采用偏最小二乘判别分析（Partial Least Squares Discriminate Analysis，PLS-DA）和支持向量机分类（Support Vector Classification，SVC）进行产地鉴别。准确率常常用来评估模型的性能，但评估基

于类不平衡数据集建立的分类模型时存在明显缺陷[14]。因此，本研究进一步采用混淆矩阵和受试者工作特征曲线（Receiver Operating Characteristic Curve，ROC）对模型性能进行评价。混淆矩阵是数据分析中对分类模型预测结果的一种评价方式，ROC 是反映模型敏感性和特异性连续变量的综合指标，一般采用 ROC 曲线下的面积（Area Under ROC Curve，AUC）作为模型评价指标，其值最大为 1，值越大代表其模型的探测效果越好[7-8]。

1.10　数据分析　本研究使用的图像校正工具为五铃光学公司高光谱成像系统 HSI Analyzer 分析软件，后续感兴趣区域提取、预处理、特征波长提取、等数据处理分析操作用到的软件为 Spyder（Python 3.6）。

2　结果

2.1　样品的原始光谱曲线　高光谱成像代表性图像见图 1（A），不同产地红参的平均近红外光谱见图 1（B）。

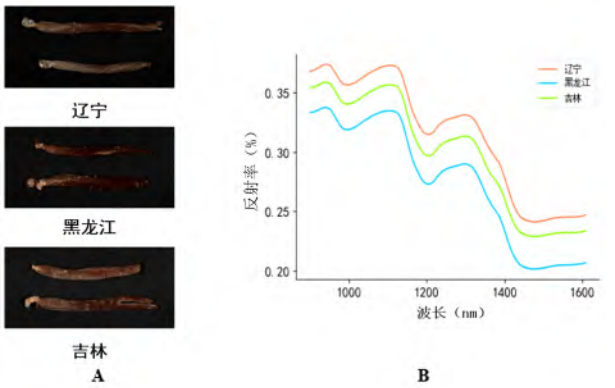

图 1　不同产地红参高光谱数据图谱

注：A. 不同产地红参高光谱伪彩色图像；B. 不同产地红参的平均光谱

2.2　预处理方法的选择　为实现对不同产地红参的鉴别分析，使用主成分分析（Principal Component Analysis，PCA）对原始光谱进行处理，由于第一主成分（PC1）和第二主成分（PC2）的累计方差贡献度在 99% 以上，因此使用前 2 个主成分绘制 PCA 图。图 2 为原始光谱的 PCA 图。由图可见，不同产地红参的样本重叠交织在一起，不易区分。因此通过原始光谱无法实现对不同产地红参的准确鉴别分析。为进一步提高鉴别准确率，采用 SG 平滑、一阶导数、二阶导数、SNV、MSC 对原始光谱进行预处理，并使用 Kennard-Stone 算法进行样本划分，计算出 PLS-DA、SVC 2 种分类方法的准确率见表 1。根据结果，在本研究中使用基于 SG 平滑的二阶导数作为光谱预处理的方式。见图 3。

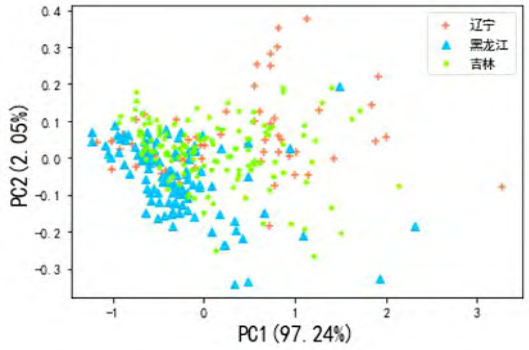

图 2　不同产地红参药材原始光谱的 PCA 得分

表 1　NIR 光谱预处理方法和分类模型的成对组合分类准确率

预处理方法	分类模型	校正集（%）	测试集（%）
原始光谱	PLS-DA	75.12	70.40
	SVC	75.81	64.55
SG 平滑（11 点）	PLS-DA	74.63	69.70
	SVC	73.00	64.35
一阶导数	PLS-DA	91.63	89.50
	SVC	93.79	82.87
二阶导数	PLS-DA	93.79	89.60
	SVC	95.67	84.75
SNV	PLS-DA	84.09	78.42
	SVC	90.64	69.80
MSC	PLS-DA	45.12	42.77
	SVC	58.57	45.94

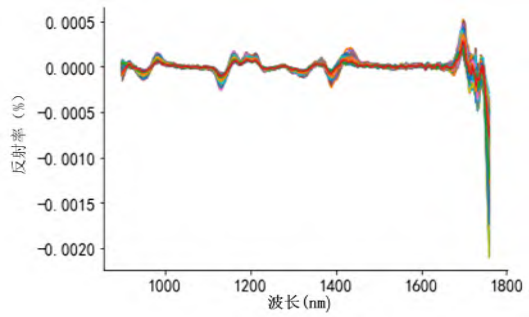

图 3　二阶导数处理后的光谱曲线

2.3　特征波段的提取　采用 SPA 法挑选出 10 个近红外特征波长（961 nm、1 069 nm、1 157 nm、1 323 nm、1 332 nm、1 377 nm、1 401 nm、1 457 nm、1 500 nm、1 526 nm）。见图 4。

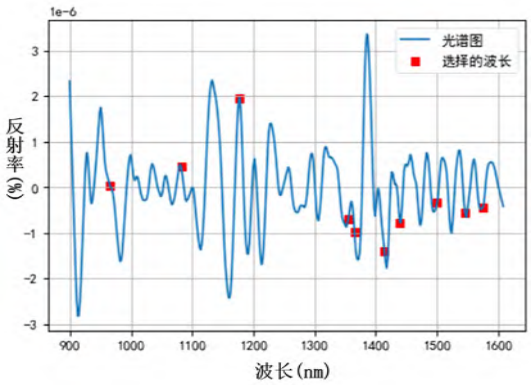

图 4　SPA 在训练集中提取的 10 个特征波长

表 2 NIR 光谱与图像纹理融合后分类准确率(%)

分类模型	NIR 光谱		图像纹理				光谱与纹理融合			
	全波段		GLCM		GLRLM		全波段 + GLCM		全波段 + GLRLM	
	校正集	测试集	校正集	测试集	校正集	测试集	校正集	测试集	校正集	测试集
PLS-DA	93.8	89.6	76.8	69.1	81.5	75.6	95.7	90.0	96.3	91.2
SVC	95.7	84.8	86.6	70.2	91.6	78.5	95.0	86.0	99.0	89.6

2.4 NIR 光谱与图像纹理融合 通过 PLS-DA 和 SVC 模型对 NIR 光谱和图像纹理信息融合前后的分类性能进行比较,基于全光谱、图像纹理和光谱纹理融合后模型分类准确率见表 2。

2.5 分类模型的性能评估 采用混淆矩阵评估红参产地分类模型的性能。图 5 展示了基于全光谱和 GLRLM 融合数据的 PLS-DA 和 SVC 模型预测结果的混淆矩阵: 吉林产地的红参表现最好,在 PLS-DA 和 SVC 模型中都达到了 100% 准确率; 黑龙江产地的红参次之,其准确率分别为 86% 和 91%; 辽宁产地的红参结果最差,准确率都为 83%。

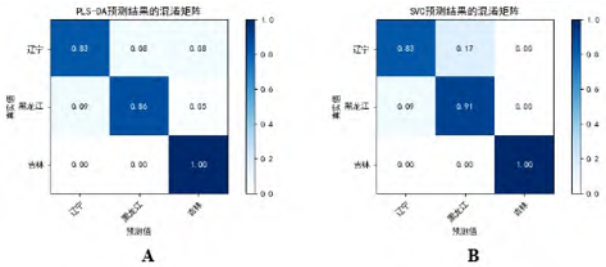

图 5 模型预测结果的混淆矩阵

注: A. PLS-DA 模型; B. SVC 模型

ROC 描述的是各种不同阈值下真正率(Ture Positive Rate, TPR) 相对于假正率(False Positive Rate, FPR) 取值变化情况,本研究使用 ROC 曲线来评估融合方法的性能。图 6 展示了光谱和图像纹理信息融合前后模型的 ROC,其中融合方法的 ROC 曲线均优于未融合的方法,全光谱信息融合 GLRLM 特征提取图谱纹理信息及全光谱信息融合 GLCM 特征提取图谱纹理信息均达到了最优曲线下面积(Area Under Curve, AUC) 值 0.97。

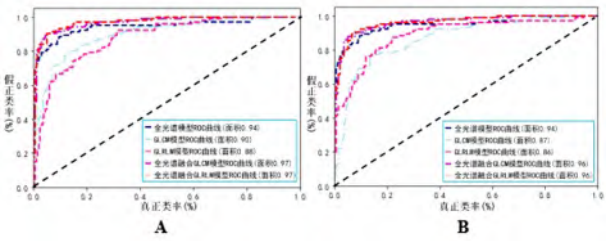

图 6 模型的 ROC

注: A. PLS-DA 模型; B. SVC 模型

3 讨论

红参作为一种常用中药材,在中药制剂中应用广泛,由于不同产地红参存在较大的质量差异,因此其产地鉴别对于保障制剂稳定性至关重要。目前红参的产地鉴别主要依靠外观性状和一些化学分析方法,外观形状受限于技术人员个人经验且工作量大,化学分析方法无法做到对每个药材的无损鉴别,难于在实际中药制剂生产中运用。近红外高光谱技术可采集红参近红外光谱信息和图像纹理信息,经信息预处理并建立合适的模型,能做到无损准确鉴别。

近红外光谱所主要反映的信息主要是分子内含氢基团(如 C-H、O-H、N-H 等) 振动时倍频和合频的吸收,其分析领域基本包括了全部的有机化合物和混合物。从不同产地红参的平均近红外光谱看,光谱有一定的差异,其中黑龙江红参的反射强度最低,与其他 2 个产地的红参区别较大,吉林和辽宁红参的光谱曲线较为接近。这一现象可能由于吉林和辽宁产地的土壤环境、纬度、光照条件更接近所致。但仅仅依靠原始光谱,鉴别准确率不高,需去除光谱的干扰信息。从不同预处理方法看,简单的 SG 平滑无法提高鉴别效果,MSC 算法反而降低了鉴别准确率,可能采用 MSC 算法丢失了部分有用信息,基于 SG 平滑的一阶导数和二阶导数以及 SNV 算法处理后光谱数据的模型准确率有不同程度地提高,其中使用二阶导数预处理后的光谱数据构建的模型效果最佳,在测试集中准确率接近 90%。从二阶导数预处理后的光谱曲线可以看出预处理后能够减轻光谱基线移位、漂移等干扰。

高光谱采集得到的近红外平均光谱只是对高光谱信息的简单信息运用,高光谱是对样品整个面的光谱扫描,进一步提取样品图像信息是增加鉴别准确率的关键。通过采用 SPA 法挑选出 10 个近红外特征波长,其中,961 nm 的波长对应多糖和树脂类的 O-H 的伸缩振动(第二泛频),1 069 nm 和 1 157 nm处的波长对应 C-H 伸缩振动(第三泛频),1 323 nm、1 332 nm、1 377 nm 处的波长对应 C-H 的伸缩振动(第一泛频),1 401 nm 处的波长对应 O-H 的伸缩振动(第一泛频),1 457 nm、1 500 nm、1 526 nm处的波长对应 N-H 的伸缩振动,反映出不同产地红参在糖类、蛋白质和其他有机化合物有一

定差异。以这 10 个特征波长采集样品高光谱数据的纹理特征，分别比较全光谱、GLCM 和 GLRLM 提取的纹理特征、光谱与纹理融合这几种提取信息建立的模型，结果可见，仅仅基于全光谱或者图像纹理特征提取建立的模型性能均不理想，全光谱信息和图像纹理信息融合能够有效提高模型的预测性能。全光谱和 GLRLM 提取的纹理特征融合后模型的性能最佳，在测试集中达到了 91.2% 的正确率。

采用混淆矩阵评估分类模型的性能，结果表明，辽宁和黑龙江产地分类准确率高于辽宁产地，可能的原因是辽宁产地的红参与其他 2 个产地差异较小。采用 ROC 进行评估，结果表明全光谱融合图像纹理信息融合的结果与正确率的结果保持一致，说明光谱与图像纹理信息融合方法能有效提高模型分类的准确率。

本研究基于近红外高光谱成像系统采集 3 个产地红参的 NIR 光谱和图像信息，对原始光谱进行二阶导数预处理后，融合 GLRLM 提取图像纹理特征后，可有效提取不同产地红参特征信息，采用 PLS-DA 模型，对吉林、黑龙江和辽宁产地样本进行准确分类。实验结果表明基于近红外高光谱成像技术的红参产地鉴别技术有望为建立稳健、切实可行的红参产地溯源模型提供思路和方法参考。

参考文献

[1] 国家药典委员会. 中华人民共和国药典(一部)[S]. 北京: 中国医药科技出版社，2020: 160.

[2] 董鹏凯，赵上勇，郑柯鑫，等. 激光诱导击穿光谱技术结合神经网络和支持向量机算法的人参产地快速识别研究[J]. 物理学报，2021，70(4)：67-75.

[3] 吴雪松，叶正良，郭巧生，等. 东北不同产地人参及其加工品人参皂苷类成分的比较分析[J]. 中草药，2013，44(24)：3551-3556.

[4] 沈亮，李西文，徐江，等. 人参无公害农田栽培技术体系及发展策略[J]. 中国中药杂志，2017，42(17)：3267-3274.

[5] 沈亮，徐江，董林林，等. 人参栽培种植体系及研究策略[J]. 中国中药杂志，2015，40(17)：3367-3373.

[6] 赵幻希，王秋颖，孙秀丽，等. HPLC-MS 结合多元统计分析区分人参产地及筛选皂苷类标志物[J]. 高等学校化学学报，2019，40(2)：246-253.

[7] 殷文俊，茹晨雷，郑洁，等. 基于高光谱成像技术融合光谱和图像特征鉴别不同产地的甘草[J]. 中国中药杂志，2021，46(4)：923-930.

[8] 吉海彦，任占奇，饶震红. 基于高光谱成像技术的不同产地小米判别分析[J]. 光谱学与光谱分析，2019，39(7)：2271-2277.

[9] 陶益，陈林，江恩赐，等. 人工智能和工业 4.0 视域下高光谱成像技术融合深度学习方法在中药领域中的应用与展望[J]. 中国中药杂志，2020，45(22)：5438-5442.

[10] Xia Z，Zhang C，Weng H，et al. Sensitive Wavelengths Selection in Identification of Ophiopogon japonicus Based on Near-Infrared Hyperspectral Imaging Technology[J]. Int J Anal Chem，2017，2017：6018769.

[11] 李江波，郭志明，黄文倩，等. 应用 CARS 和 SPA 算法对草莓 SSC 含量 NIR 光谱预测模型中变量及样本筛选[J]. 光谱学与光谱分析，2015，35(2)：372-378.

[12] Hu W，Huang Y，Wei L，et al. Deep Convolutional Neural Networks for Hyperspectral Image Classification[J]. J Sensors，2015，2015：1.

[13] Ru C，Li Z，Tang R. A Hyperspectral Imaging Approach for Classifying Geographical Origins of Rhizoma Atractylodis Macrocephalae Using the Fusion of Spectrum-Image in VNIR and SWIR Ranges (VNIR-SWIR-FuSI)[J]. Sensors(Basel)，2019，19(9)：2045.

[14] Fawcett T. An introduction to ROC analysis[J]. Pattern Recogn Lett，2006，27(8)：861-874.

(2021-10-25 收稿　责任编辑：王明)