基于高光谱成像技术和深度学习的烤后烟叶品种分类判别

孟令峰[1,2]，邢富康[1,2]，韦克苏[3]，王爱华[1]，程昌新[4]，曹　娜[5]，
林跃平[4]，王松峰[1*]，朱荣光[2*]
[1.中国农业科学院烟草研究所/农业农村部烟草生物学与加工重点实验室，青岛 266101；2.石河子大学机械电气工程学院，新疆 石河子 832003；3.贵州省烟草科学研究院，贵阳 550081；4.红云红河烟草(集团)有限责任公司，昆明 650202；5.云南省烟草公司保山市公司，云南 保山 678000]

摘　要：为了实现对不同品种烤后烟叶的快速判别，采集红花大金元、K326、云烟 105 三种烟叶的高光谱图像，利用阈值分割和边缘提取算法选择感兴趣区域，获得烟叶的光谱数据。分别使用卷积平滑结合一阶导数、标准正态变量变换、多元散射校正和中心化等 4 种常用方法对原始光谱数据进行预处理，在此基础上建立多层感知机网络(Multilayer Perceptron，MLP)、卷积神经网络(Convolutional Neural Network，CNN)和支持向量机(Support Vector Machine，SVM)模型，进行烤后烟叶品种分类的比较研究。结果表明，所建立的模型均可对烟叶的品种进行有效判别，其中一阶导数预处理在多个模型中表现均优于其他 3 种预处理方法，基于 CNN 建立的分类模型性能优于其他模型，两者结合的性能最优，其测试集正确率和准确率分别为 97.9% 和 98%，表明高光谱结合 CNN 可以实现烟叶品种的定性判别。上述研究为后续基于高光谱技术开发针对烤后烟叶的快速检测系统提供了理论依据和技术支撑。

关键词：烤后烟叶；品种分类；高光谱图像；深度学习；机器学习
中图分类号：S572；S126　　**文献标识码：**A　　**文章编号：**1007-5119（2024）04-0083-10

Classification and Identification of Tobacco Varieties after Curing Based on Hyperspectral Imaging and Deep Learning

MENG Lingfeng[1,2], XING Fukang[1,2], WEI Kesu[3], WANG Aihua[1], CHENG Changxin[4], CAO Na[5], LIN Yueping[4], WANG Songfeng[1*], ZHU Rongguang[2*]
(1. Institute of Tobacco Research of CAAS/Key Laboratory of Tobacco Biology and Processing, Ministry of Agriculture and Rural Affairs, Qingdao 266101, China; 2. College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832003, Xinjiang, China; 3. Guizhou Academy of Tobacco Science, Guiyang 550081, China; 4. Hongyun Honghe Tobacco Co., Ltd, Kunming 650202, China; 5. Baoshan Branch of Yunnan Province Tobacco Company, Banshan 678000, Yunnan, China)

Abstract: In order to realize rapid identification of different varieties of cured tobacco leaves, hyperspectral images of Honghua Dajinyuan, K326 and Yunyan 105 tobacco leaves were collected, and the region of interest was selected by threshold segmentation and edge extraction algorithm to obtain spectral data of tobacco leaves. The raw spectral data were preprocessed by four common methods: convolution smoothing combined with first-order derivative, standard normal variable transformation, multivariate scattering correction and centralization. On this basis, Multilayer Perceptron (MLP) network was established. MLP, Convolutional Neural Network (CNN) and Support Vector Machine (SVM) models were used to compare the classification of varieties using cured tobacco leaves. The results showed that all the established models could effectively distinguish the varieties of tobacco leaves, among which the first-order derivative pretreatment performed better than the other three pretreatment methods in several models, and the classification model based on CNN performed better than the other models. The combined performance of the two models was the best, and the correctness and accuracy of the test set were 97.9% and 98%, respectively. Hyperspectrum combined with CNN could realize the qualitative identification of tobacco varieties. The results of this study provide theoretical basis and technical support for the subsequent development of rapid detection system for cured tobacco based on hyperspectral technology.
Keywords: flue-cured tobacco; variety classification; hyperspectral image; deep learning; machine learning

---

基金项目：中国农业科学院科技创新工程(ASTIP-TRIC03)；中国烟草总公司重点项目(110202102007)；红云红河集团项目(HYHH2019YL02)
第一作者：孟令峰(1991—)，男，硕士，助理研究员，主要研究领域为烟叶采烤信息监测和智能烘烤装备研发。E-mail：menglingfeng@caas.cn
*通信作者：王松峰(1979—)，男，副研究员，主要研究领域为智能检测技术与装备研发。E-mail：wangsongfeng@caas.cn
朱荣光(1982—)，男，教授，主要研究领域为农畜产品品质快速智能检测。E-mail：rgzh_jd@163.com
收稿日期：2023-12-26　　**修回日期：**2024-03-07

烤烟作为重要的经济作物，在我国多个地区广泛种植[1]，主要品种包括云烟系列、K326、红花大金元、翠碧、秦烟等。不同品种烟叶的烟碱、总糖、还原糖等成分含量不同[2]，这些化学成分的含量及相对比例关系影响着烟叶的香型风格、香气质量、劲头、刺激程度和余味等风味特性，从而带来不同的评吸感受[3-4]。因此在卷烟的生产加工过程中，需将多品种的烟叶混合，进行科学调制配比，以满足不同消费者的风味需求。但目前烤后烟叶品种判断多依赖人工粗分，检测效率低，而且容易出现误分误判，影响后续的烟叶加工质量。因此对不同品种烟叶进行高效智能化判别具有重要的现实意义。

高光谱成像技术能够高效获取待测样品表面的图像和光谱信息，已广泛应用于无损检测领域[5]。目前已有针对茶叶[6]、红枣[7]、苹果[8]、大米[9]、玉米[10]等作物的品种检测研究，取得了较好的成果。在烟叶检测方面，主要集中于对鲜烟叶的成熟度[11-12]、病虫害[13]、叶绿素[14-15]和含水率[16]等指标进行检测，而针对品种判别的研究较少。王一丁等[17]基于高光谱成像构建了田间烟叶品种判别模型，取得了较好的效果。然而烟叶经过烘烤之后结构和素质发生较大变化，品种间区分难度更大。

现有研究针对高光谱数据的建模方法主要为机器学习中的 SVM 模型[18]，SVM 模型的性能依赖超参数的选择，超参数优化算法寻优能够在一些分类任务中获得较好的结果，而在特征复杂度过高时的性能较差。多层感知机(MLP)作为最经典的深度学习模型，结合非线性激活函数在处理高光谱数据中也有较多应用[19-20]。卷积神经网络(CNN)利用卷积运算提取数据中的潜在特征，在多个研究中表现出最好的性能[21-22]。然而由于深度学习参数复杂，运算量大，其结合高光谱成像技术在烤后烟叶检测中的研究较少。

因此，本试验采集多品种的烤后烟叶，获取高光谱图像并开发程序完成图像处理和光谱提取，选择不同的预处理方式和建模方法，建立品种判别模型并优选最佳模型，探讨利用高光谱成像技术和深度学习算法建立烤后烟叶品种判别模型的可行性。

1 材料与方法

1.1 供试材料

供试烤烟(由云南省烟草公司保山市公司提供)品种为红花大金元、K326 和云烟 105，烟叶全部采集自保山市下辖的 4 个区县(隆阳县、昌宁县、腾冲县、施甸县)，表 1 为各品种样本数。

表 1 样本分布

Table 1 Quantity of each sample

品种 Varieties	类别 Label	类别编码 Category coding	样本数量 Sample size
红花大金元 Honghuadajinyuan	Type1	100	64
K326	Type2	010	64
云烟105 Yunyan 105	Type3	001	64
总计 Total	3	3	192

按照 3∶1 的比例从每个品种的烟叶样本中随机划分出校正集和测试集，其中校正集 144 片，测试集 48 片。将烤后烟叶样品逐片展平，以主脉为中线将叶片分成两半，沿主脉方向等长度划分叶尖、叶中、叶基 3 个部位，共计 6 个采样区域。在每个区域中随机选择大小为 3 cm×3 cm 的叶块裁剪并编号，将所有样品置密封袋内保存，待测。

1.2 高光谱图像数据获取与光谱信息提取

1.2.1 高光谱图像采集 在本研究中，采用的高光谱图像采集系统主要包括成像光谱仪(ImSpector V10E-QE，ImSpector 公司，芬兰)、线阵电荷耦合器件(Charge Coupled Device, CCD)相机(Infinity 3-1，Lumenera，加拿大)、光源(Fiber-Lter DC950 型，DolanJenner 公司，美国)、脉冲输送装置(SC300-1A，Zolix 公司，中国)、暗箱和计算机等，高光谱图像数据采集软件为 SpecView。自主设计并用黑色尼龙材料制作多区域样本数据采集托盘，辅助系统获取两片烟叶的图像。采集系统及烟叶摆放方式如图 1 所示。

在高光谱图像数据采集前，先打开光源和相机预热 30 min。高光谱图像采集系统参数设置如下：光源角度与水平面为 60°，曝光时间为 0.6 ms，调整镜头焦距，对焦完成保持样品与镜头间的距离为 33.5 cm，图像采集速度 69.53 mm/s。

为了消除光照不均匀和外部环境因素的影响，需对高光谱数据进行黑白校正。保持采集条件一致，通过遮盖镜头并关闭光源获得全黑标定图像 I_B(反

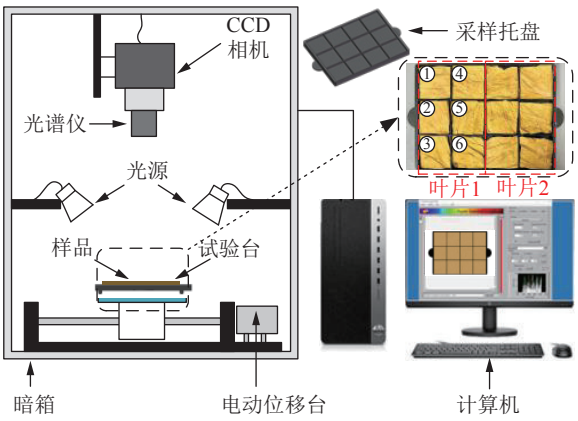

图 1　高光谱图像数据采集系统组成示意图

Fig. 1　Schematic diagram of hyperspectral image data acquisition system

射率约为 0)，开启镜头和光源扫描聚四氟乙烯标准白板获得全白图像 I_W(反射率约为 99%)，对原始图像 I_R 按照公式(1)进行黑白校正，得到校正后的图像数据 I。

$$I = \frac{I_R - I_B}{I_W - I_B} \tag{1}$$

1.2.2　叶片光谱数据获取　将样品校正后的高光谱图像数据归一化到 0~255 范围内，得到灰度图。由于黑色尼龙材料与烟叶在高光谱相机中的反射率具有较大差异，在灰度图中较易区分出烟叶与背景。在图 2 中看出，当背景分离阈值过小时背景去除不干净，阈值过大时造成叶片中较多光谱信息的丢失，因此需要选择合适的分离阈值，在本文中通过对比探讨将分离阈值设置为(80，255)。由于烟叶存在折痕和部分褶皱，对烟叶的光谱数据采集具有较大影响，本研究采用自适应阈值分割算法和轮廓提取算法，识别出烟叶中存在的折痕和褶皱，建立去折痕-褶皱掩膜，得到烟叶样品的高光谱图像。随后按相对位置的比例进行分割，得到 12 个局部区域(图 2)。由于单个叶块区域在图像中的像素面积约为 120×120 pixels，为了避免相邻叶块间的干扰，保证区域间的独立性，以叶块的区域中心为矩形中心划分 100×100 pixels 的区域作为感兴趣区域(Region of interest，ROI)。从叶块 ROI 内有效像素点中随机抽取 4 000 个像素点的平均光谱作为该叶块的光谱数据，来自同一片烟叶 6 个叶块的光谱数据再次平均，得到该烟叶样本的代表性光谱数据，重复 3 次随机抽取的步骤，每个烟叶样本获得 3 条不完全相同的代表性光谱数据。

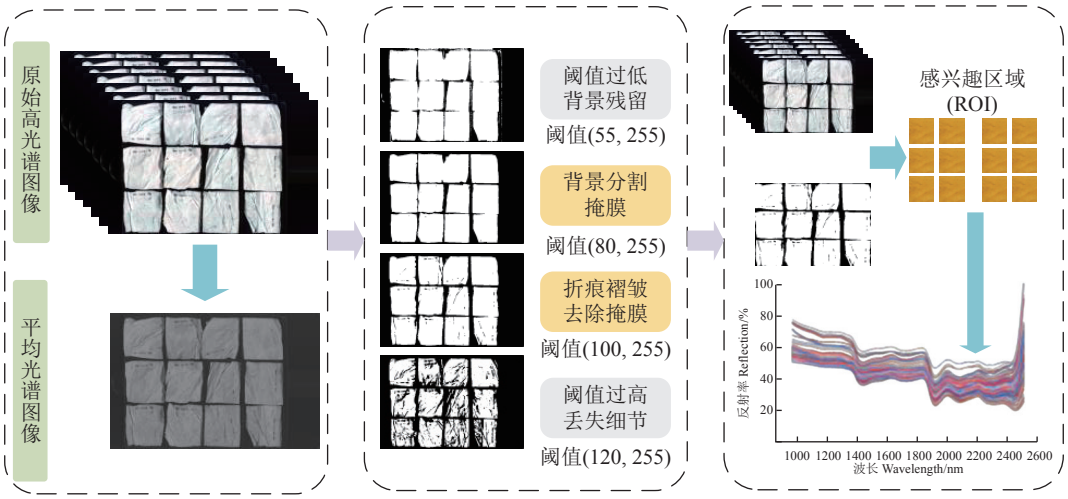

图 2　图像处理流程

Fig. 2　Image processing

1.3　光谱数据预处理

叶片的光谱包含 900 ~ 2 700 nm 范围内的 1 800 个波长，由于<1 000 nm 和>2 500 nm 波长的光谱噪声较大，故本研究选择 1 000 ~ 2 500 nm 范围内的 1 500 个波长进行预处理和后续分析。

在光谱数据采集时，光谱曲线易受到外界因素的影响，导致光谱中包含大量噪声和干扰信息，因此需对光谱进行预处理，去除无关信息的干扰。在本研究中，首先对光谱数据进行 SG 平滑，然后分别利用一阶导数(First Derivative，1D)、中心化

(Mean Center，MC)、多元散射校正(Multiplicative Scattering Correction，MSC)和标准正态变量变换(Standard Normalized Variate，SNV)这 4 种常用的预处理方法进行光谱预处理。

1.4 分类模型建立

1.4.1 试验环境 试验的硬件如下：CPU 为 Intel i7-12700H，GPU 为 NVIDIA GeForce RTX 3060 Laptap，内存为 16 GB。软件环境如下：操作系统为 Windows 11，编程语言为 Python 3.7，图像基于 OpenCV 处理，深度学习基于 Tensorflow 2.4 框架搭建。CUDA 版本为 11.3，CUDNN 版本为 11.3。

1.4.2 数据集划分 将样本按照 3∶1 的比例划分校正集和测试集(表 2)，校正集用于训练验证、模型参数寻优，测试集用于评价模型在未知样本上的性能表现。在模型训练过程中，主要对校正集数据进行划分，在 SVM 模型中采用 *K* 折交叉验证法对其进行内部交叉验证，目标函数选择交叉验证集正确率(Cross-validation set accuracy)；在 MLP 和 CNN 中采用固定随机数种子、随机划分 30% 为验证集的方法，选择交叉熵损失函数(Cross-entropy Loss)作为模型训练的损失函数。

表 2 光谱数据集划分

Table 2 Spectral quantities

品种 Varieties	类别 Label	校正集 Calibration set 训练集 Train set	校正集 Calibration set 验证集 Validation set	测试集 test set	总计 Total
红花大金元 Honghuadajinyuan	Type1	101	43	48	192
K326	Type2	101	43	48	192
云烟105 Yunyan 105	Type3	101	43	48	192

1.4.3 模型参数确定 采用多层感知机网络(MLP)和卷积神经网络(CNN)建立烤后烟叶品种检测模型，将批大小设为 32，使用 Adam 优化器，初始学习率 0.001。并与机器学习模型 SVM 进行比较。

(1)多层感知机网络(MLP)。本研究中搭建的 MLP 网络结构如图 3 所示，主要由 5 个全连接层、2 个批归一化层和 4 个随机失活层构成。全连接层的神经元个数分别为 512，512，256，64。随机失活的比例分别为 0.5，0.5，0.5，0.6。从而缓解深层全连接神经网络的过拟合问题。最后输出 3 个神经元节点利用 Softmax 激活函数输出烟叶品种的 3 分类概率分布。

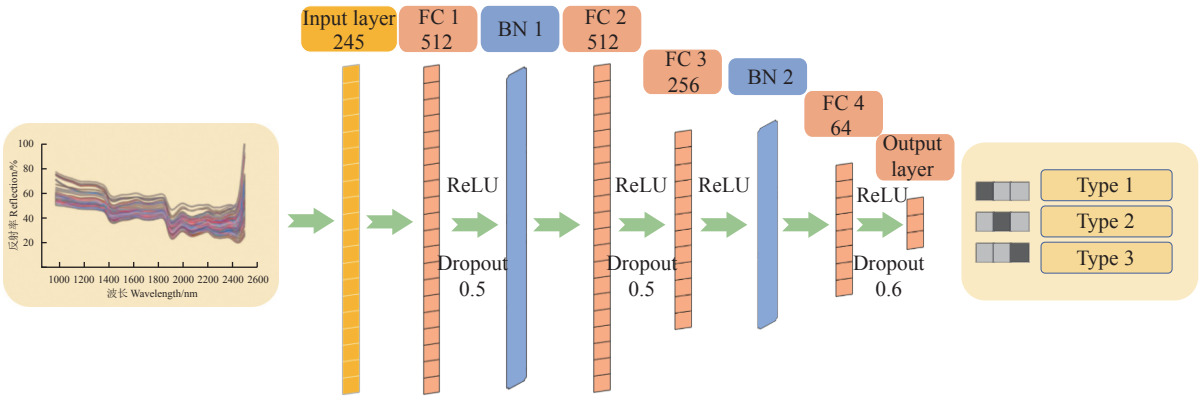

图 3 MLP 模型结构图

Fig. 3 MLP model structure diagram

(2)卷积神经网络(CNN)是利用卷积核的滑移进行特征提取的模型。本研究搭建的 CNN 网络结构如图 4 所示，其主要包括 3 个一维卷积层、2 个批归一化层、2 个全连接层、2 个随机失活层和 1 个最大值池化层。卷积层对特征提取后，通过最大值池化(MaxPooling1D)降低特征映射维度，减小计算复杂度并保留关键信息；同时添加批量归一化层(Batch-Normalization)将特征聚拢归一化，提高模型的稳定性；引入随机失活层(Dropout)随机将部分神经元失活，从而缓解过拟合。

(3)支持向量机(SVM)是一种有监督的机器学习方法，广泛应用于统计分类以及回归分析。其原理是将向量映射到一个更高维的空间里，在高维空间里建立一个最大间隔超平面。在分隔数据的超平面的两边建有两个互相平行的超平面，分隔超平面使两个平行超平面的距离最大化。本研究利用

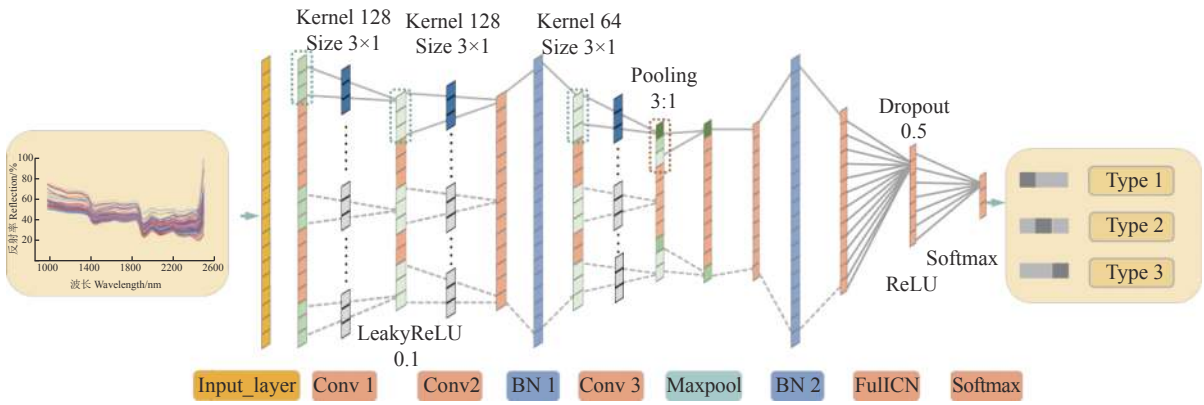

图 4　CNN 模型结构图

Fig. 4　CNN model structure diagram

SVM 模型进行烟叶品种类型判别，主要超参数有惩罚系数（C）、核函数（Kernel）、核函数参数（Gamma），在本研究中采用贝叶斯优化算法对其中的超参数进行组合寻优。

1.4.4　分类模型评估　借助混淆矩阵，采用正确率、精确率、召回率和 F1 分数等[12] 对分类模型的判别性能进行评价。

2　结　果

2.1　烟叶样品的光谱分析

不同品种的烟叶光谱反射率曲线如图 5 所示，对比发现其光谱曲线整体趋势一致。光谱曲线的吸收峰主要出现在 1 380、1 650、1 820、1 920 nm 波长处。结合烟草不同品种的化学成分和光学特性，说明品种是影响烟叶结构及化学成分的重要因素。可见光谱信息能够在一定程度上反映烟草品种之间的差异。如图 6 所示，对光谱数据预处理之后数据发生了较大变化，光谱曲线的特征峰数量增多，潜在特征被进一步挖掘，这将有利于烟叶品种分类模型的构建。

2.2　检测模型构建

2.2.1　MLP 模型结构的参数优化　MLP 模型具有较好的非线性拟合能力，其中神经元的个数和层数对模型性能具有较大的影响。本研究分别将神经元的个数设置为 32、64、128、256、512；神经网络的层数设置为 3、4、5、6。探讨神经元个数和层数对 MLP 模型的影响。由表 3、表 4 可见，在多层感知机模型中，随着神经元个数的增加，模型在校正集和验证集上的判别效果都在不断增加，这表明模型的特征提取能力不断增强。然而这将导致模型参数量较大，训练和推理时间变慢，同时也存在过拟合风险；而随着隐藏层个数增加，模型在校正集和验证集上的判别效果存在波动现象，当神经元层数为 5 时，模型的表现变差，说明过深的网络结构不适合光谱数据特征的提取，对比看来，当神经元层数为 4 时，校正集和验证集的判别效果均较好。

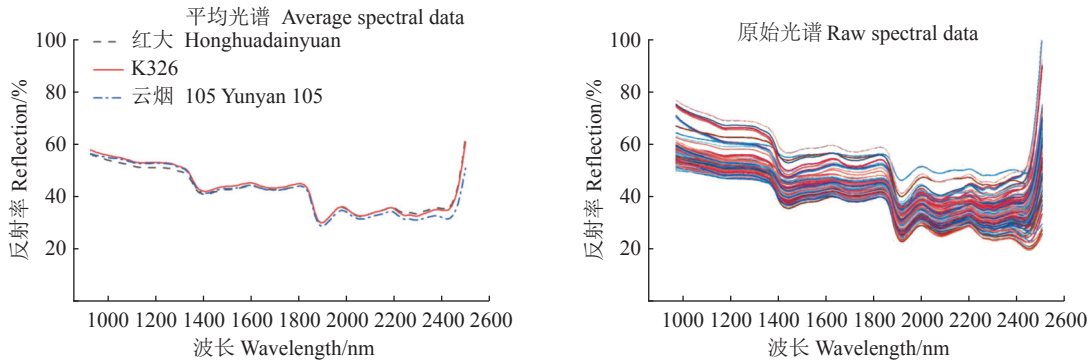

图 5　不同品种的原始光谱和平均光谱数据

Fig. 5　Raw and average spectral data of different varieties of tobacco leaves

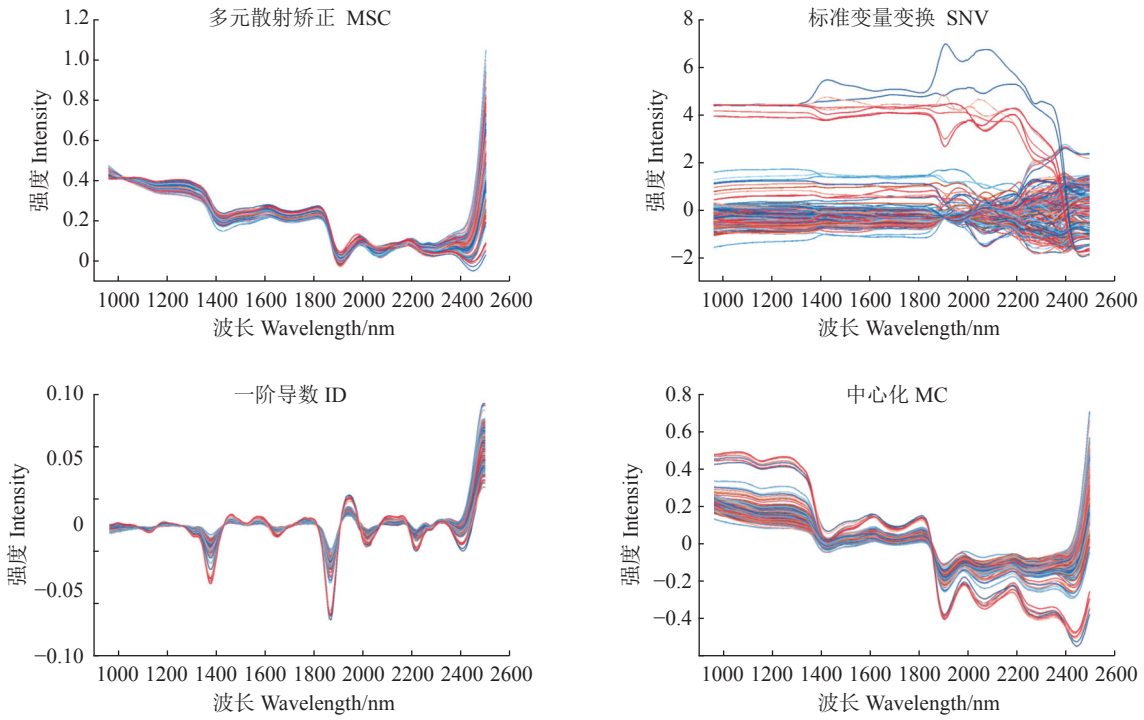

图 6 不同预处理算法下的光谱数据

Fig. 6 Spectral data of different preprocessing algorithm

表 3 不同神经元个数的 MLP 模型性能对比

Table 3 Performance comparison of MLP models with different number of neurons

神经元个数 Number of neurons	校正集正确率 Correct rate of correction set	验证集正确率 Correct rate of verification set	测试集正确率 Correct rate of test set	平均精确率 Average precision	平均召回率 Average recall	平均F1 Average F1
32	0.761	0.735	0.722	0.736	0.722	0.716
64	0.8	0.807	0.708	0.73	0.708	0.694
128	0.823	0.821	0.799	0.804	0.799	0.795
256	0.907	0.887	0.854	0.863	0.854	0.851
512	0.946	0.934	0.91	0.913	0.91	0.911

表 4 不同隐藏层数的 MLP 模型性能对比

Table 4 Performance comparison of MLP models with different neural network layers

神经元层数 Number of layers	校正集正确率 Correct rate of correction set	验证集正确率 Correct rate of verification set	测试集正确率 Correct rate of test set	平均精确率 Average precision	平均召回率 Average recall	平均F1 Average F1
2	0.942	0.936	0.889	0.889	0.889	0.889
3	0.933	0.923	0.854	0.872	0.854	0.856
4	0.954	0.937	0.896	0.898	0.896	0.896
5	0.923	0.887	0.854	0.854	0.854	0.854

综上，本文的 MLP 模型的最优结果为神经元个数设置为 512，网络层数设置为 4，并以此作为后续预处理方法对比的模型结构。

2.2.2 CNN 模型结构的参数优化 适当的卷积核数量和尺寸能够使 CNN 模型在训练过程中，充分利用局部和全局特征。本研究首先将卷积核大小设置为 3×1，卷积核的滑移步长固定设置为 1。将 3 个卷积层中卷积核的数量分别设置为 128、64、32、16 的不同组合，根据模型评价指标的优化结果确定每个卷积层中最优卷积核个数。由表 5 可以看出，增加卷积核个数有利于特征提取，训练过程中损失函数收敛速度过快，校正集验证集正确率都较高，但测试集正确率较差，容易过拟合；反之，减少卷积核个数不利于特征提取，训练过程不稳定，损失函数波动下降，收敛速度减缓，最终校正集、验证集、测试集的正确率均较低。因此，本研究选择(128，64，64)的组合，通过多层卷积堆叠，提取局部特征和广域特征，在满足特征提取效率的前提下尽可能缓解模型的过拟合问题。

表 5　不同卷积核数量组合的 CNN 模型性能对比

Table 5　Performance comparison of CNN models with different combinations of the number of convolution kernels

卷积核个数组合 Combination of number of convolution kernels	校正集正确率 Correct rate of correction set	验证集正确率 Correct rate of verification set	测试集正确率 Correct rate of test set	平均精确率 Average precision	平均召回率 Average recall	平均F1 Average F1
(128，128，128)	0.984	0.980	0.854	0.854	0.854	0.85
(128，128，64)	0.963	0.946	0.896	0.908	0.896	0.895
(128，64，64)	0.957	0.953	0.875	0.88	0.875	0.87
(128，64，32)	0.938	0.927	0.833	0.855	0.833	0.833
(64，64，32)	0.931	0.931	0.875	0.88	0.875	0.873
(64，32，32)	0.924	0.917	0.875	0.878	0.873	0.873
(64，32，16)	0.915	0.884	0.833	0.849	0.833	0.831
(32，32，16)	0.915	0.890	0.854	0.86	0.854	0.853

为了进一步探讨卷积核尺寸对 CNN 模型的性能影响，将三个卷积层的卷积核尺寸设置为 3、5 的组合，不同卷积尺寸组合下基于全光谱数据的烟叶品种检测结果于表 6 中。由表 6 可以看出，不同的卷积核尺寸影响着训练的速度和感受野，较大尺寸的卷积核可以增大感受野，但引入了更多参数，极大地增加运算量，较小的卷积核通过多层堆叠也有较好的性能。因此，本研究选择的卷积核尺寸组合为(3，3，3)。

激活函数是影响 CNN 模型性能的重要因素，合适的激活函数有利于神经网络对光谱数据中非线性特征的提取。为了选用合适的激活函数建立烟叶种类判别模型，分别采用 LeakyReLU、ReLU、Tanh 和 Sigmoid 四种激活函数进行建模，根据模型性能选择最优的激活函数，不同激活函数基于全光谱数据的烟叶品种检测结果列于表 7。从表 7 可以看出，LeakyReLU 激活函数在校正集和验证集的正确率均为最高，表明该激活函数对于非线性拟合效果最好。

综上，本研究搭建的 CNN 模型主要由 3 层卷积层构成，卷积核个数分别为 128、128、64，卷积核尺寸为 3×1，激活函数为 LeakyReLU。

2.2.3　SVM 模型的超参数优化　利用贝叶斯优化算法对 SVM 的超参数进行寻优，训练过程均采用 5 折交叉验证，优化方向为使交叉验证集正确率最大化，优化迭代轮数为 100 轮。由于 SVM 模型的结果与数据预处理算法联系较大，故模型优化算法需要结合不同的预处理算法进行探讨，基于贝叶斯优化 SVM 模型的超参数空间设置与参数优化结果如表 8 所示。从表 8 中可以看出，在相同的参数优化空间里，不同的预处理算法会带来不同的优化结果；这与图 6 中原始光谱数据经过预处理之后的数据发生了较大变化的结果一致。

2.3　模型的优选比较

将预处理算法与上文建立的烟叶品种检测 SVM、MLP 和 CNN 模型相结合，探讨光谱预处理对不同模型判别性能的影响，并优选出最佳判别

表 6　不同卷积核尺寸组合的 CNN 模型性能对比

Table 6　Performance comparison of CNN models with different kernel size combinations

卷积核尺寸 Kernel size	校正集正确率 Correct rate of correction set	验证集正确率 Correct rate of verification set	测试集正确率 Correct rate of test set	平均精确率 Average precision	平均召回率 Average recall	平均F1 Average F1
(5，5，5)	0.977	0.963	0.875	0.88	0.875	0.87
(5，5，3)	0.966	0.961	0.903	0.902	0.903	0.902
(3，3，3)	0.976	0.966	0.924	0.94	0.938	0.938
(3，5，3)	0.956	0.930	0.833	0.857	0.833	0.832
(3，5，5)	0.963	0.969	0.875	0.88	0.875	0.87

表 7　不同激活函数的 CNN 模型性能对比

Table 7　Performance comparison of CNN models with different activation functions

激活函数 Activation function	校正集正确率 Correct rate of correction set	验证集正确率 Correct rate of verification set	测试集正确率 Correct rate of test set	平均精确率 Average precision	平均召回率 Average recall	平均F1 Average F1
ReLU	0.976	0.966	0.924	0.94	0.938	0.938
LeakyReLU	0.99	0.996	0.91	0.847	0.847	0.842
Sigmod	0.98	0.975	0.917	0.918	0.917	0.938
Tanh	0.98	0.96	0.875	0.924	0.924	0.922

表 8 贝叶斯优化算法的参数空间设置与寻优结果

Table 8 Parameter space settings and optimization results of Bayesian optimization algorithm

优化空间设置			预处理算法 Preprocessing algorithm	惩罚系数（C）	核函数（Kernel）	核函数参数（Gamma）
惩罚系数(C)	核函数(Kernel)	核函数参数(Gamma)				
			None	2.66	poly	3.57
	‘linear’		MSC	64.2	linear	9.45
[1,100]	‘rbf’	[0.01,10]	SNV	54.4	poly	7.32
	‘poly’		1D	6.06	rbf	0.01
			MC	5.5	poly	0.136

模型。由模型检测准确率(表 9)和测试集混淆矩阵(图 7)可见，基于原始光谱数据和预处理之后的光谱数据建立的最优模型均为 CNN 模型，基于原始光谱的 CNN 模型，在训练集、验证集和测试集的正确率分别为 99.3%、96.9% 和 88.9%。数据预处理方式也会影响模型的判别性能，1D 预处理方法在多个模型中相较于其他预处理方法均表现出了较好的性能，其在测试集上的表现排序为 1D-CNN>1D-MLP>1D-SVM，最优模型为 1D-CNN 模型，其相较于原始光谱的最优模型 None-CNN 在训练集、验证集、测试集上分别提高 1.4%、2.4%、9%。

3 讨 论

现有研究表明利用光谱信息能够对鲜烟叶的品种进行鉴别[17]。烟叶经过烘烤之后结构素质发生了较大变化，品种判别难度进一步加大。本研究发现不同品种烤后烟叶的组分差异能够在光谱信息中得到体现，具体表现在光谱的反射率强度、反射峰数量及部分波段下的反射率等发生变化。光谱数据进行合适的预处理之后，能够进一步消除外界干扰，增强光谱数据与烤后烟叶品种之间的相关关系，从

表 9 不同数据预处理方法处理的烟叶品种检测模型准确率

Table 9 Accuracy of tobacco varieties detection model treated by different data pretreatment methods

模型 Model	预处理算法 Preprocessing algorithm	校正集正确率 Correct rate of correction set	验证集正确率 Correct rate of verification set	测试集正确率 Correct rate of test set	平均精确率 Average precision	平均召回率 Average recall	平均F1 Average F1	正确率差值 Correct rate difference
多层感知机 MLP	None	0.904	0.877	0.847	0.846	0.847	0.844	0.057
	MSC	0.966	0.961	0.792	0.796	0.792	0.791	0.174
	SNV	0.96	0.915	0.882	0.881	0.882	0.881	0.078
	1D	0.984	0.973	0.958	0.96	0.958	0.958	0.026
	MC	0.937	0.915	0.833	0.837	0.833	0.834	0.104
卷积神经网络 CNN	None	0.993	0.969	0.889	0.892	0.889	0.89	0.104
	MSC	0.984	0.97	0.875	0.876	0.875	0.873	0.109
	SNV	0.907	0.9	0.833	0.87	0.833	0.824	0.074
	1D	0.998	0.977	0.979	0.98	0.979	0.979	0.019
	MC	0.924	0.917	0.917	0.919	0.917	0.917	0.007
支持向量机 SVM	None	0.986	0.969	0.833	0.842	0.833	0.834	0.153
	MSC	0.993	0.964	0.889	0.891	0.889	0.889	0.104
	SNV	0.983	0.983	0.799	0.805	0.799	0.799	0.184
	1D	0.993	0.953	0.931	0.931	0.931	0.931	0.062
	MC	0.984	0.983	0.822	0.885	0.852	0.88	0.162

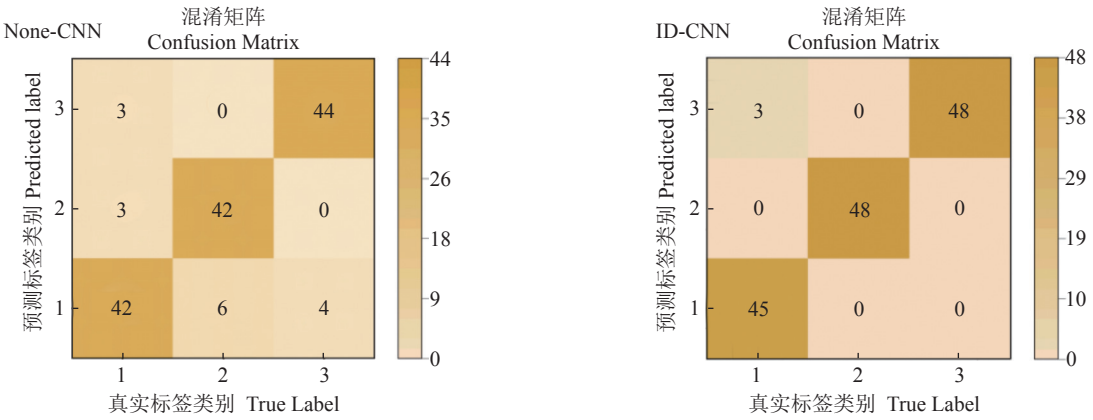

图 7 最优模型测试集混淆矩阵

Fig. 7 Optimal model test set confusion matrix

而提高模型的判别精度。本文对比了利用原始光谱和经过 4 种预处理算法后光谱的建模性能，发现预处理对模型性能有不同程度的提高，这与殷彩云等[23] 的研究结果相似。在这 4 种预处理算法中，一阶导数预处理在多个模型中均有较好表现，与张恒等[24] 研究结果相近。这可能是由于一阶导数预处理能够更敏感地提取到反射率在不同波段上的变化特征，增大了光谱曲线中的有效信息，从而明显提高了品种模型的判别精度。

本研究将 4 种预处理算法与 3 种建模方法结合建立品种判别模型，发现使用相同预处理算法建立不同模型时，MLP 和 CNN 模型的整体表现优于 SVM 模型。这可能由于 SVM 模型的性能受超参数影响较大，模型优劣更多取决于优化算法的结果，并且对于非线性关系的拟合效果较差。MLP 和 CNN 通过堆叠神经层和使用激活函数从而具有更好的非线性拟合性能。在深度学习的 MLP 和 CNN 模型参数寻优过程中，本文发现由于烟叶的光谱数据特征数量较少，模型结构较浅难以学习到数据中的深层特征，较深则会导致模型退化，因此模型不宜过深或过浅；随着神经元个数的增加，模型的判别效果逐步增加，然而随着个数进一步增加，模型效果不再提高，同时计算量倍增，因此应当在保证模型效果时尽可能控制神经元个数。研究结果表明模型层数在 3~4 层、MLP 和 CNN 模型的神经元个数设置为 512 和 128 较为适宜，既可保证判别效果，又能控制计算复杂度，推理时间较短，判别效率远超人工，在实际应用中更具可行性。此外研究还发现 CNN 对于光谱数据的特征提取优于 MLP，这可能是由于 CNN 是利用卷积核滑移计算，能够有效地关注到光谱数据中不同波段之间的联系。基于一阶导数预处理和 CNN 建立的模型是烤后烟叶品种判别的最佳模型，其中对未知样本的品种判别正确率达到 97.9%。

相比于其他检测技术手段如近红外[25-26] 和电子鼻技术[27]，高光谱成像技术可以更加准确快速地判别烤后烟叶的品种。与采用同类检测技术的相关研究对比，本研究采用深度学习的建模方法，建立的 CNN 模型对烤后烟叶的品种判别性能更好。

4　结　论

本研究结合近红外高光谱技术与传统机器、深度学习方法对烤后烟叶的品种进行判别。结果表明：(1)基于高光谱成像技术采集的烟叶光谱数据能够在一定程度上反映出品种信息，一阶导数预处理有利于提取相关特征，说明高光谱成像技术结合一阶导数预处理方法可以实现烟叶的品种判别；(2)在基于一阶导数预处理后的光谱数据建立的模型中，CNN 模型在烟叶品种判别任务中表现出色，对未知样本的判别正确率能够达到 97.9%，为烟叶品种的快速判别提供了解决方案。研究结果为实现烤后烟叶的快速无损检测提供了理论依据和技术支持，后续将考虑涵盖更多烟叶品种，以进一步提升模型的泛化能力和实用性，并开展基于高光谱技术的烟叶精细化的分析研究。

参考文献

[1] WANG Y C, QIN L. Research on state prediction method of tobacco curing process based on model fusion[J]. Journal of Ambient Intelligence and Humanized Computing, 2022, 13(6): 2951-2961.

[2] 王鹏，孙兰茜，彭忠，等. 国内主要烟区烟叶常规化学成分含量的差异[J]. 贵州农业科学，2020，48(5)：50-53.
WANG P, SUN L X, PENG Z, et al. Differences in conventional chemical composition of tobacco in major tobacco-growing areas in China[J]. Guizhou Agricultural Sciences, 2020, 48(5): 50-53.

[3] 敖金成，赵剑华，张晓龙，等. 云南不同生态区有机种植方式下烟叶常规化学成分及安全性分析[J]. 河南农业科学，2015，44(4)：66-71.
AO J C, ZHAO J H, ZHANG X L, et al. Analysis of routine chemical components and safety of flue-cured tobacco leaves under organic cultivation in different ecological regions of Yunnan Province[J]. Journal of Henan Agricultural Sciences, 2015, 44(4): 66-71.

[4] 韩富根. 烟草化学[M]. 2 版. 北京：中国农业出版社，2010.
HAN F G. Tobacco chemistry[M]. 2nd ed. Beijing: China Agriculture Press, 2010.

[5] BAI Z X, DU D D, ZHU R G, et al. Establishment and comparison of *in situ* detection models for foodborne pathogen contamination on mutton based on SWIR-HSI[J]. Frontiers in Nutrition, 2024, 11: 1325934.

[6] 孙俊，靳海涛，武小红，等. 基于低秩自动编码器及高光谱图像的茶叶品种鉴别[J]. 农业机械学报，2018，49(8)：316-323.
SUN J, JIN H T, WU X H, et al. Tea variety identification based on low-rank stacked auto-encoder and hyperspectral image[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(8): 316-323.

[7] 刘立新，何迪，李梦珠，等. 基于高光谱技术与机器学习的新疆红枣品种鉴别[J]. 中国激光，2020，47(11)：1111002.
LIU L X, HE D, LI M Z, et al. Identification of Xinjiang jujube varieties based on hyperspectral technique and machine learning[J]. Chinese Journal of Lasers, 2020, 47(11): 1111002.

[8] 马惠玲，王若琳，蔡骋，等. 基于高光谱成像的苹果品种快速鉴别[J]. 农业机械学报，2017，48(4)：305-312.
MA H L, WANG R L, CAI C, et al. Rapid identification of apple varieties based on hyperspectral imaging[J]. Transactions of the

Chinese Society for Agricultural Machinery, 2017, 48(4): 305-312.

[9] 杨森，张新寡，王振民，等. 基于近红外光谱和深度学习数据增强的大米品种检测[J]. 农业工程学报，2023，39(19)：250-257.
YANG S, ZHANG X A, WANG Z M, et al. Rice variety detection based on near-infrared spectroscopy and deep learning data augmentation[J]. Transactions of the Chinese Society of Agricultural Engineering, 2023, 39(19): 250-257.

[10] 张伏，王新月，崔夏华，等. 高光谱结合极限学习机的玉米品种鉴别[J]. 光谱学与光谱分析，2023，43(9)：2928-2934.
ZHANG F, WANG X Y, CUI X H, et al. Identification of maize varieties by hyperspectral combined with extreme learning machine[J]. Spectroscopy and Spectral Analysis, 2023, 43(9): 2928-2934.

[11] 李鑫，汤卫荣，张永辉，等. 基于高光谱成像技术的烟叶田间成熟度判别模型[J]. 烟草科技，2022，55(7)：17-24.
LI X, TANG W R, ZHANG Y H, et al. Discriminant model for field maturity of tobacco leaves based on hyperspectral imaging technology[J]. Tobacco Science & Technology, 2022, 55(7): 17-24.

[12] 王大彬，孟贵星，王雪松，等. 不同成熟度茄衣烟叶高光谱特征分析及分类模型构建[J]. 中国烟草科学，2023，44(3)：85-91.
WANG D B, MENG G X, WANG X S, et al. Analysis of hyperspectral characteristics of coat-oriented cigar tobacco with different maturity and construction of a classification model[J]. Chinese Tobacco Science, 2023, 44(3): 85-91.

[13] GU Q, SHENG L, ZHANG T H, et al. Early detection of tomato spotted wilt virus infection in tobacco using the hyperspectral imaging technique and machine learning algorithms[J]. Computers and Electronics in Agriculture, 2019, 167: 105066.

[14] GUO T, TAN C W, LI Q, et al. Estimating leaf chlorophyll content in tobacco based on various canopy hyperspectral parameters[J]. Journal of Ambient Intelligence and Humanized Computing, 2019, 10(8): 3239-3247.

[15] 李梦竹，刘国顺，贾方方. 干旱胁迫下旺长期烤烟冠层叶绿素密度的高光谱估测[J]. 中国烟草科学，2017，38(1)：23-28.
LI M Z, LIU G S, JIA F F. Hyperspectral estimation of canopy chlorophyll density in flue-cured tobacco under different drought stress at the vigorous growth stage[J]. Chinese Tobacco Science, 2017, 38(1): 23-28.

[16] 魏雷，徐迎波，曹海兵，等. 基于 AE-BiRNN 的片烟含水率高光谱检测[J]. 烟草科技，2024，57(2)：70-78.
WEI L, XU Y B, CAO H B, et al. AE-BiRNN-based hyperspectral detection of moisture content in tobacco strips[J]. Tobacco Science & Technology, 2024, 57(2): 70-78.

[17] 王一丁，赵铭钦，付博，等. 基于可见光-近红外光谱技术的烤烟品种鉴别研究[J]. 山东农业科学，2016，48(2)：119-124.
WANG Y D, ZHAO M Q, FU B, et al. Identification of flue-cured tobacco varieties based on visible and near-infrared spectroscopy technique[J]. Shandong Agricultural Sciences, 2016, 48(2): 119-124.

[18] YOON H I, LEE S H, RYU D, et al. Non-destructive assessment of *Cannabis* quality during drying process using hyperspectral imaging and machine learning[J]. Frontiers in Plant Science, 2024, 15: 1365298.

[19] GUTIÉRREZ S, FERNÁNDEZ-NOVALES J, DIAGO M P, et al. On-the-go hyperspectral imaging under field conditions and machine learning for the classification of grapevine varieties[J]. Frontiers in Plant Science, 2018, 9: 1102.

[20] LI Y J, AL-SARAYREH M, IRIE K, et al. Identification of weeds based on hyperspectral imaging and machine learning[J]. Frontiers in Plant Science, 2021, 11: 611622.

[21] LI Y J, MA B X, LI C, et al. Accurate prediction of soluble solid content in dried Hami jujube using SWIR hyperspectral imaging with comparative analysis of models[J]. Computers and Electronics in Agriculture, 2022, 193: 106655.

[22] YU Z Y, FANG H, ZHANGJIN Q N, et al. Hyperspectral imaging technology combined with deep learning for hybrid okra seed identification[J]. Biosystems Engineering, 2021, 212: 46-61.

[23] 殷彩云，白子金，罗德芳，等. 基于高光谱数据的土壤全氮含量估测模型对比研究[J]. 中国土壤与肥料，2022(1)：9-15.
YIN C Y, BAI Z J, LUO D F, et al. Comparative study on estimation models of soil total nitrogen content based on hyperspectral data[J]. Soil and Fertilizer Sciences in China, 2022(1): 9-15.

[24] 张恒，梁太波，冯文强，等. 基于连续投影算法的土壤全氮和碱解氮含量高光谱估测[J]. 中国烟草科学，2023，44(5)：103-113.
ZHANG H, LIANG T B, FENG W Q, et al. Hyperspectral estimation of total nitrogen and alkali hydrolysable nitrogen contents in tobacco growing soil based on successive projection algorithm[J]. Chinese Tobacco Science, 2023, 44(5): 103-113.

[25] 杜文，易建华，谭新良，等. 基于近红外光谱的烟叶 SIMCA 模式识别[J]. 中国烟草学报，2009，15(5)：1-5.
DU W, YI J H, TAN X L, et al. SIMCA pattern recognition of tobacco leaves based on near infrared spectra[J]. Acta Tabacaria Sinica, 2009, 15(5): 1-5.

[26] 冯莉，常爱霞，郭丛涛，等. 电子鼻检测烤后烟叶挥发性组分的方法研究[J]. 中国烟草科学，2014，35(4)：92-98.
FENG L, CHANG A X, GUO C T, et al. Study on the method of using electronic nose to detect the volatile components in cured leaves[J]. Chinese Tobacco Science, 2014, 35(4): 92-98.

[27] MARCELO M C A, SOARES F F, ARDILA J A, et al. Fast inline tobacco classification by near-infrared hyperspectral imaging and support vector machine-discriminant analysis[J]. Analytical Methods, 2019, 11(14): 1966-1975.