浙 江 农 林 大 学 学 报, 2026, **43**(2): 283−292 https://zlxb.zafu.edu.cn
*Journal of Zhejiang A&F University* DOI: 10.11833/j.issn.2095-0756.20250418

**引用格式:** 潘威, 邓书文, 杨晓东, 等. 基于高光谱成像与集成模型的烟草种子活力分类研究[J]. 浙江农林大学学报, 2026, **43**(2): 283−292. PAN Wei, DENG Shuwen, YANG Xiaodong, *et al*. Tobacco seed vigor classification based on hyperspectral imaging and ensemble models[J]. *Journal of Zhejiang A&F University*, 2026, **43**(2): 283−292.

# 基于高光谱成像与集成模型的烟草种子活力分类研究

潘　威[1]，邓书文[2]，杨晓东[1]，乔　雨[1]，梅　媛[1]，张立猛[1]，关亚静[2]

（1. 玉溪中烟种子有限责任公司 科研中心，云南 玉溪 653100；2. 浙江大学 农业与生物技术学院 现代种业研究所，浙江 杭州 310058）

**摘要：【目的】**针对微小粒烟草 *Nicotiana tabacum* 种子，构建一种基于高光谱成像与集成模型的无损活力快速分类方法。**【方法】**以烟草品种‘MS 云烟 87’‘MS-Yunyan 87’、‘红花大金元’‘Honghua Dajinyuan’、‘云烟 99’‘Yunyan 99’为材料，设置多个劣变梯度，获取不同处理种子的群体高光谱数据。综合发芽势、根长、苗高等各项指标，按权重形成活力指数，并以阈值划分高低活力标签。所有光谱数据经卷积平滑 (SG) 降噪后，以无信息变量消除算法 (UVE) 筛选出判别波段，基于精简特征训练 CNN-LightGBM 分类器，并以独立品种数据进行外部验证。**【结果】**随劣变时长增加，活力指数整体显著下降，48、72 及 96 h 的劣变处理种子几乎全部归为低活力组，表明所定阈值具有良好的分界效应。与不同特征选择模型组合相比，UVE 所得精简特征配合 CNN-LightGBM 的综合表现最佳，测试集准确率为 88.90%、召回率为 97.40%、F1 分数为 91.40%。模型在烟草品种‘MS121’数据上的有效性验证中总体准确率达 85.58%，体现出模型良好的跨批次与跨品种泛化能力。**【结论】**建立了 SG 预处理与 UVE 特征波段筛选方法结合 CNN-LightGBM 模型的策略，可实现微小粒烟草种子活力的高效、无损分类，具备向其他微小粒作物种子迁移应用的潜力，为微小粒作物种子质量监测提供了新思路。图 6 表 2 参 25

**关键词：**烟草种子；高光谱成像；CNN-LightGBM；种子活力；无损检测；活力指数

**中图分类号：**S722.1+3　　**文献标志码：**A　　**文章编号：**2095-0756(2026)02-0283-10

# Tobacco seed vigor classification based on hyperspectral imaging and ensemble models

PAN Wei[1], DENG Shuwen[2], YANG Xiaodong[1], QIAO Yu[1], MEI Yuan[1], ZHANG Limeng[1], GUAN Yajing[2]

（1. Research Center, Yuxi Zhongyan Tobacco Seed Co., Ltd., Yuxi 653100, Yunnan, China; 2. Advanced Seed Institute, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, Zhejiang, China）

**Abstract:** [**Objective**] To develop a rapid, non-destructive vigor classification method for micro-sized tobacco seeds based on hyperspectral imaging and an ensemble model. [**Method**] Seeds of 3 cultivars (‘MS-Yunyan 87’ ‘Honghua Dajinyuan’ ‘Yunyan 99’) were subjected to multiple controlled-deterioration gradients, and population-level hyperspectral data were acquired across treatments. A seed vigor index (SVI) was constructed by weighting germination potential, primary root length, and seedling height, and a threshold was applied to assign high/low vigor labels. All spectra were denoised with Savitzky-Golay (SG) smoothing; discriminative wavelengths were selected via uninformative variable elimination (UVE); and a CNN-LightGBM classifier was trained on the compact features. External validation was performed using an independent cultivar ‘MS121’. [**Result**] SVI decreased markedly with longer deterioration; seeds treated for 48, 72, and 96 h were almost

收稿日期：2025-08-07；修回日期：2025-12-29

基金项目：中国烟草总公司云南省公司科技计划重大项目 (2024530000241004)

作者简介：潘威 (ORCID: 0009-0003-0085-8568)，高级农艺师，博士，从事烟草种子科学研究。E-mail: panwei_@126.com。通信作者：关亚静 (ORCID: 0000-0003-4763-2737)，教授，博士，从事种子科学研究。E-mail: vcguan@zju.edu.cn

entirely classified as low vigor, indicating that the predefined threshold provided a clear decision boundary. Among feature–model combinations, UVE-derived compact features coupled with CNN-LightGBM performed best, achieving 88.90% test accuracy, 97.40% recall, and an F1-score of 91.40%. On external validation with 'MS121', overall accuracy reached 85.58%, demonstrating robust cross-batch and cross-cultivar generalization. [**Conclusion**] Integrating SG preprocessing and UVE-based wavelength selection with a CNN-LightGBM classifier enables efficient, accurate, and non-destructive vigor classification for micro-sized tobacco seeds. The pipeline shows promising transferability to other small-seeded crops and offers a new avenue for quality monitoring in such crops. [Ch, 6 fig. 2 tab. 25 ref.]

Key words: tobacco seeds; hyperspectral imaging; CNN-LightGBM; seed vigor; non-destructive testing; seed vigor index

烟草 *Nicotiana tabacum* 是中国经济发展的重要产业支柱[1]，高质量的烟草种子是烟叶生产技术中最基本和最有效的增产措施[2]。其中，种子活力是衡量作物种子质量的关键指标，关系到作物的生长发育、产量和品质。然而，传统的种子活力评估方法主要依赖于胚根伸长计数、2,3,5-氯化三苯基四氮唑 (TTC) 染色、电导率测定和加速老化等方法。这些检测方法需要破坏种子，同时检测过程复杂、耗时耗力，已难以满足现代农业对高效、无损、精准检测种子质量的需求。高光谱图像技术[3]是融合了成像与光谱分析的先进检测技术，可在不破坏种子的情况下，短时间内对大量种子进行高精度检测，然后通过建立模型和算法，实现对种子活力的预测。基于高光谱图像技术开展高效精准的种子活力检测已成为当前的研究热点。AMBROSE 等[4]用 400~2 500 nm 波段的高光谱成像技术鉴别微波热处理后及未处理的 2 种活力水平玉米 *Zea mays* 种子批，所建立的偏最小二乘判别分析模型校准集和预测集的鉴别率分别高达 97.6% 和 95.6%。张婷婷等[5]以单粒小麦 *Triticum aestivum* 种子为研究对象，利用高光谱成像系统获取样品的光谱信息，利用多种预处理方法建立光谱与单粒小麦种子生活力的 PLS-DA 模型，其最优校正集和预测集的整体鉴别正确率分别为 86.7% 和 85.1%。石睿等[6]以不同活力的单粒小麦种子为研究对象，使用高光谱仪获得其 400~1 050 nm 的高光谱数据，使用 CARS 算法选取特征波段建立了支持向量机 (SVM)、K 近邻 (KNN)、一维卷积神经网络 (1DCNN) 和高效通道注意力卷积神经网络 (ECA-CNN) 共 4 种小麦种子活力检测模型，发现整体准确率为 86.67%，精确率为 92.31%，召回率为 80%。

高光谱成像技术在水稻 *Oryza sativa*、玉米、小麦等大粒种子活力检测中取得了显著进展，但针对小粒种子尤其微小粒种子的研究欠缺。例如烟草的种子，其平均千粒重仅为 0.080~0.092 g，单粒烟草种子直径约 315~630 μm，属于典型的微小粒种子。现有高光谱设备多为毫米级空间分辨率，难以有效实现单粒种子的精细成像，更难从单粒种子上获取足量的信息。且各类种子获取到的高光谱图像包含大量冗余的高维数据，如何从这些复杂数据中提取有用的特征，也是研究难点。

因此，本研究以烟草种子为研究对象，利用群体成像法采集种子光谱数据，通过数据预处理与特征波段选择不同算法，如无信息变量消除算法 (uninformative variables elimination algorithm，UVE)、竞争性自适应重加权算法 (competitive adaptive reweighted sampling algorithm，CARS) 和连续投影算法 (successive projections algorithm，SPA)，筛选与微小粒种子活力密切相关的特征变量，进而构建机器学习分类模型、深度学习分类模型卷积神经网络 (convolutional neural network，CNN) 以及基于 CNN-LightGBM (CNN-light gradient boosting machine) 集成网络的烟草种子活力水平分类检测模型，实现以烟草种子为例的典型微小粒种子群体活力等级的准确鉴别。该研究不仅有效弥补了烟草种子单粒检测的技术不足，也为其他作物微小粒种子质量的快速无损评价提供了新思路。

## 1 材料与方法

### 1.1 样品制备

1.1.1 *种子材料* 模型构建主要采用了国内主栽的 3 种烟草品种种子 (‘MS 云烟 87’ ‘MS-Yunyan 87’、‘红花大金元’‘Hongda Jinyuan’、‘云烟 99’‘Yunyan 99’)，在模型验证环节，增加了烟

草品种‘MS121’，均由玉溪中烟种子有限责任公司提供。收获种子材料后，对种子进行发芽率测定，均达 95% 以上，将种子储存在 4 ℃ 冰箱中备用。

1.1.2　控制劣变处理　对种子样品进行控制劣变处理使其加速老化。利用高温烘箱法测定烟草种子起始含水量，将种子含水量调至 20%，置于铝箔袋中密封，在 4 ℃ 冰箱中平衡水分 1 d，将铝箔袋置于老化箱中，(45±0.5) ℃ 精准控温下分别处理 0、12、24、36、48、72、96 h，每个烟草品种共获得 7 个不同处理时间的劣变种子，用于后续实验。

**1.2　数据收集**

1.2.1　种子的光谱图像获取　使用可见光/近红外高光谱成像系统 (ImSpector V10E，Spectral Imaghing Ltd.，芬兰) 获取种子的高光谱图像信息。此成像系统具有 414.6~1 017.5 nm 波段，包括光谱分辨率为 2.8 nm 的 CCD 相机 (ZYLA-4.2P-USB3.0，安道尔公国) 和像元尺寸为 6.5 μm 的镜头 (XENOPLAN 1.4/17" 400~1 000 nm)。样品由 2 个卤素灯照明，设置传送带与相机镜头间距 60 cm，每个时间点劣变处理后的种子设置 15 个重复，每重复 0.05 g 种子堆叠为椭圆形，均匀分布，种子间无空隙，每个处理种子以相同的排列方式 (5 行×3 列) 放置，然后以 8 mm·$s^{-1}$ 的传送带速度扫描，曝光时间为 20 ms，以获得高光谱图像。采集过程中使用的控制及图像校正软件为 Spectra VIEW。

1.2.2　标准发芽试验　每个品种的各劣变时间的种子均设置 15 个发芽重复，每个重复 50 粒种子。在 30 cm×40 cm×4 cm 的长方形塑料盒中放置湿润发芽纸，置种，发芽温度设置为 (25±1) ℃，以 12 h 光照/12 h 黑暗环境发芽 14 d，每天监测烟草种子的发芽进度，胚根长度≥2 mm 的种子视为发芽。第 6 天统计发芽势 (germination potential，$G_P$)，14 d 统计发芽率 (germination rate，$G_E$)。

$$G_P = \frac{N_e}{N_t} \times 100\%;$$

$$G_E = \frac{N_g}{N_t} \times 100\%。$$

其中，$N_e$ 为发芽第 6 天时正常发芽的种子数，$N_g$ 为发芽第 14 天时正常发芽的种子数，$N_t$ 为供试种子总数。

1.2.3　种子活力指数 (seed vigor index，$I_{SV}$) 计算　在发芽第 14 天，测量每个重复的幼苗根长 (primary root length，$L_{PR}$) 和苗高 (seedling height，$H_S$)。参考 ZHU 等[7] 的研究，为实现 $I_{SV}$ 中各项性状权重的客观确定，本研究采用了基于机器学习的特征重要性分析法。以 $G_P$、$L_{PR}$ 及 $H_S$ 指标为输入特征，第 14 天发芽率≥92% (GB/T 21 138—2019) 为高活力的基本标准，利用随机森林 (random forest，RF) 算法对活力高低进行分类建模，并提取模型训练过程中的特征重要性分数。各指标的重要性分数归一化后，作为最终用于 $I_{SV}$ 构建的性状权重，由此获得 $G_P$、$L_{PR}$ 及 $H_S$ 的权重，再根据公式计算 $I_{SV}$。

$$I_{SV} = W_{GP}G_P + W_{PRL}L_{PR} + W_{SW}H_S。$$

其中：$I_{SV}$ 为活力指数；$W_{GP}$、$W_{PRL}$、$W_{SW}$ 分别为 $G_P$、$L_{PR}$ 及 $H_S$ 的权重。

1.2.4　光谱数据提取　光谱仪得到的烟草种子光谱数据不仅包含有用的信息，还包含随机噪声。为了减少背景信息引起的噪声干扰，通过 ROI 工具划分种子和背景来创建感兴趣区域 (ROI)[8]，然后通过 ENVI5.1 (ITT Visual Information Solutions，美国) 提取 ROI 中的平均光谱反射率。

高光谱图像 $\boldsymbol{I}_0$ 上的黑白校正是为了减少光源波动和暗电流的负面影响。将反射率为 99.99% 的标准白色校准板放置在与样品相同的高度，扫描收集标准白光校准数据 $\boldsymbol{W}$。用黑色镜头盖盖住镜头，暗场标准黑板图像 $\boldsymbol{B}$ 被收集。根据公式，计算黑白校正后的光谱数据 $\boldsymbol{I}$。

$$I = \frac{I_0 - B}{W - B} \times 100\%。$$

1.2.5　数据分析和处理方法　在烟草种子高光谱图像采集和提取过程中存在很多干扰因素。为了减少干扰因素，对采集的高光谱数据分别进行多元散射校正 (multiplicative scatter correction，MSC)、标准正态变换 (standard normal variate，SNV)、一阶导数 (first derivative，FD) 以及卷积平滑 (Savitzky-Golay，SG) 处理。为诊断数据质量，基于马氏距离法对不同预处理后的光谱数据进行了离群值检测，对比选择

效果最优的预处理方法[9]。由于收集了大量的种子光谱数据变量，许多光谱变量之间存在共线性、冗余性，影响模型判别的准确率，因此对预处理后的光谱信息进行特征波段筛选。采用连续投影算法 (successive projections algorithm，SPA)、遗传算法 (genetic algorithm，GA)、无信息变量消除 (uniformative variable elimination，UVE) 等方法进行特征变量筛选[10]。

1.2.6 *模型构建与验证* 采用 Kennard-Stone 方法以 4∶1 的比例将光谱数据分为 252 个训练样本和 63 个测试样本。传统机器学习方法通常高度依赖于手动特征选择和特征工程，容易导致信息丢失，而深度学习模型虽然能够自动提取深层次特征，但需要大量高质量样本进行有效训练，在样本规模受限或数据不均衡情况下易发生过拟合，泛化能力受到一定制约。

为此，本研究所构建的 CNN-LightGBM 集成模型首先利用 CNN 的特征提取优势对高光谱数据进行深度特征学习，通过一维卷积层捕捉光谱波段之间的空间特征关联性，有效地提取出高层次的隐含特征；随后再以 CNN 提取的深度特征为输入，通过 LightGBM 分类器对特征进行高效学习与决策判别。模型充分融合了深度学习的特征表达能力和 LightGBM 的高效学习能力，有效提升了烟草种子活力检测任务的准确性和稳定性，为烟草种子活力无损检测提供了更准确的方法。

此外，研究采用网格搜索-交叉验证 (GridSearchCV) 方法[11]，对关键超参数进行系统搜索与优化，进而获取最优参数组合以提升模型的泛化能力与运算效率。

1.2.7 *结果评估* 通过结果对比、混淆矩阵以及常见的分类评估指标对模型进行评估。

①训练集和测试集预测结果对比。通过对比图可以直观地看到模型在两者上的预测效果。为进一步评估模型性能，计算了准确率 (accuracy，$A$)，即正确预测的样本数与总样本数之比：

$$A = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}。$$

其中，$T_P$ 为真阳性样本数；$T_N$ 为真阴性样本数；$F_P$ 为假阳性样本数；$F_N$ 为假阴性样本数。

②混淆矩阵。混淆矩阵可以直观地展示分类模型在不同类别之间的误分类情况。通过混淆矩阵，可以进一步计算模型的精确率 (precision，$P$)、召回率 (recall，$R$) 和 F1 分数 ($F_1$) 等指标，这些指标能综合反映分类模型的性能。

③精确率、召回率和 F1 分数。精确率衡量的是被分类器预测为正类的样本中，实际为正类的比例。

$$P = \frac{T_P}{T_P + F_P} \times 100\%;$$

召回率衡量的是实际为正类的样本中，被正确预测为正类的比例。

$$R = \frac{T_P}{T_P + F_N} \times 100\%;$$

$F_1$ 是精确率和召回率的调和平均，用于综合考虑二者的表现。

$$F_1 = 2 \times \frac{P \times R}{P + R}。$$

所有预处理、特征波段选择以及模型构建在 MATLAB R2023a (MathWorks，美国)、Python3.12 中进行。

1.2.8 *模型有效性验证* 取烟草品种‘MS121’种子进行相同劣变处理，对新数据集进行 SVI 权重分析，并用分类效果最优模型进行模型有效性验证。

## 2 结果与讨论

### 2.1 劣变处理对烟草种子发芽的影响

由表 1 可见：3 个品种的烟草种子发芽势和发芽率等随劣变时间延长而普遍下降，但下降速率与耐劣变能力存在明显差异。具体而言，未劣变的‘MS 云烟 87’种子的发芽势达 93.33%，经 48 h 劣变后，发芽势快速下降，72 h 时发芽势降为 0，发芽率仅为 43.20%；‘红花大金元’的原始发芽势为 91.07%，与‘MS 云烟 87’无显著差异。其在 48 h 劣变后，发芽势也快速下降，72 h 时发芽势降为 0，发芽率显著降低至 56.13%；‘云烟 99’在短时间 (36 h) 劣变后仍保持良好发芽势，72 h 时发芽势仍达

65.20%，96 h 时发芽势降为 0，耐高温劣变能力明显优于前两者。3 个品种根长和苗高的变化均随劣变时间的延长显著降低。综合分析表明，不同品种由于其遗传特性的差异等，其耐劣变能力有明显差异，表现为发芽势、根长及苗高等性状的显著变化。

**表 1　不同劣变时间的烟草种子萌发情况**

Table 1　Germination performance of tobacco seeds under different deterioration times

| 品种 | 处理时间/h | 第6天发芽势/% | 第14天发芽率/% | 第14天根长/cm | 第14天苗高/cm | 品种 | 处理时间/h | 第6天发芽势/% | 第14天发芽率/% | 第14天根长/cm | 第14天苗高/cm |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ‘MS云烟87’ | 0 | 93.33±0.02 a | 97.87±0.02 a | 1.49±0.12 a | 0.58±0.06 a | ‘红花大金元’ | 48 | 41.33±0.11 d | 88.53±0.07 b | 1.03±0.06 c | 0.28±0.03 d |
| | 12 | 90.13±0.00 a | 96.27±0.02 a | 1.13±0.11 b | 0.51±0.07 b | | 72 | 0 e | 56.13±0.10 c | 0.86±0.05 d | 0.21±0.02 e |
| | 24 | 82.93±0.07 b | 92.80±0.05 ab | 0.86±0.10 c | 0.47±0.04 bc | | 96 | 0 e | 30.00±0.13 d | 0.33±0.04 e | 0.14±0.02 f |
| | 36 | 75.07±0.06 c | 90.80±0.04 b | 0.65±0.04 de | 0.43±0.05 c | ‘云烟99’ | 0 | 95.73±0.03 a | 98.93±0.01 a | 1.63±0.06 a | 0.38±0.03 a |
| | 48 | 37.20±0.10 d | 78.53±0.07 c | 0.65±0.03 d | 0.43±0.04 c | | 12 | 92.93±0.05 ab | 96.67±0.03 a | 1.42±0.07 b | 0.40±0.03 a |
| | 72 | 0 e | 43.20±0.07 d | 0.58±0.07 e | 0.23±0.05 d | | 24 | 93.73±0.04 ab | 95.60±0.03 ab | 1.35±0.06 c | 0.37±0.04 a |
| | 96 | 0 e | 17.60±0.08 e | 0.21±0.06 f | 0.053±0.04 e | | 36 | 93.87±0.03 a | 95.33±0.03 ab | 1.22±0.13 d | 0.33±0.04 b |
| ‘红花大金元’ | 0 | 91.07±0.03 a | 98.53±0.01 a | 1.52±0.05 a | 0.37±0.05 a | | 48 | 87.07±0.11 b | 92.13±0.04 b | 1.07±0.05 e | 0.29±0.02 c |
| | 12 | 83.33±0.06 b | 97.73±0.02 a | 1.24±0.01 b | 0.33±0.04 bc | | 72 | 65.20±0.14 c | 81.47±0.05 c | 0.81±0.03 f | 0.21±0.02 d |
| | 24 | 78.13±0.069 b | 95.20±0.03 a | 1.22±0.08 b | 0.35±0.06 ab | | 96 | 0 d | 36.13±0.09 d | 0.36±0.04 g | 0.13±0.01 e |
| | 36 | 69.87±0.1 c | 92.93±0.03 ab | 1.05±0.05 c | 0.32±0.02 cd | | | | | | |

说明：数据为平均值±标准差，同列同品种数字后的不同小写字母表示差异显著 ($P$<0.05)。

## 2.2　种子活力指数及活力分类

由图 1 可见：$G_P$、$L_{PR}$ 及 $H_S$ 等 3 项性状指标的权重分别为 0.48、0.31 和 0.21。根据公式计算获得 $I_{SV}$。鉴于劣变处理对于 3 个品种的影响不同，制定烟草种子活力分类策略时，等级划分过多会使各类别样本量不足且差异不显著，同时，由于实际生产用种中需快速掌握种子活力高低，因此，根据 $I_{SV}$ 结果，将样本分为高活力与低活力 2 类种子。将 $I_{SV}$>0.859 6 (>35%) 的种子归为高活力种子，其余为低活力种子。随着处理时间的延长，种子 $I_{SV}$ 整体呈下降趋势，高活力样本比例逐渐减少。在 48、72 及 96 h 时段，样本几乎全部归为低活力组。这表明劣变处理能有效区分种子活力水平，且 $I_{SV}$ 阈值 (0.859 6) 能够较好地反映活力高低分界。

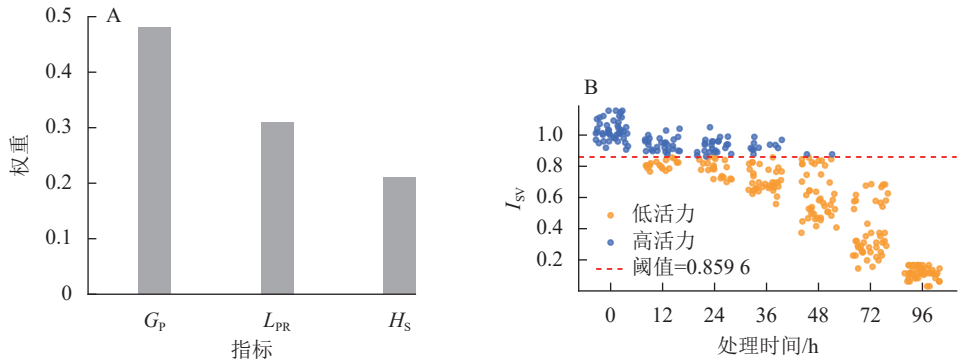
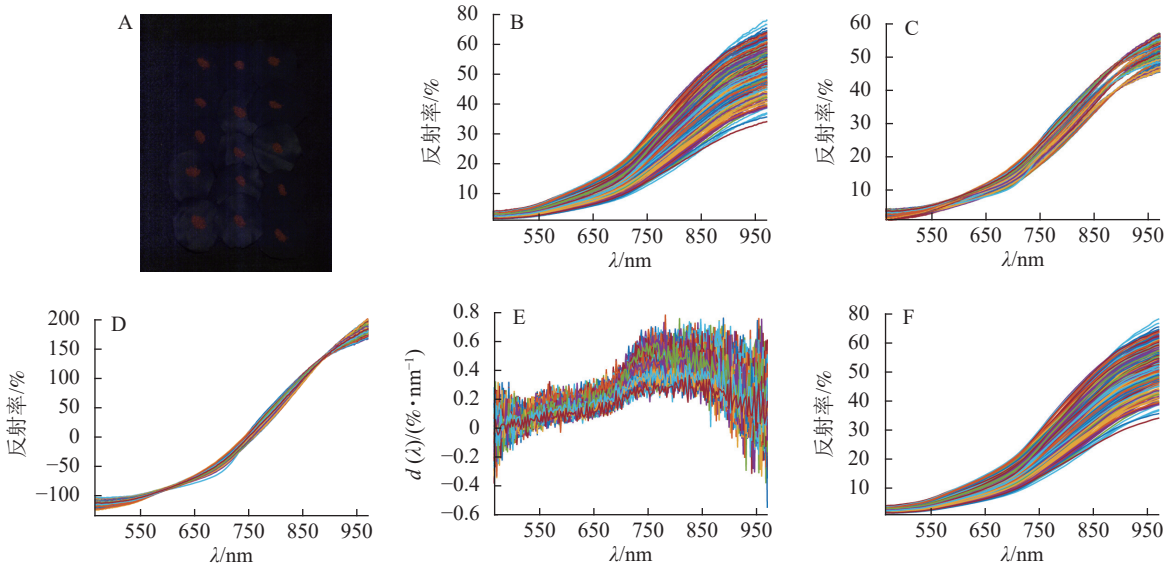



图 1　3 项性状指标特征权重 (A) 及不同处理时间下种子活力指数分组散点图 (B)

Figure 1　Feature importance of GP, PRL and SH (A), and grouped scatter plot of SVI across different treatment times (B)

## 2.3　光谱数据分析

由图 2 可见：微小粒种子在高光谱图像下每粒种子所获得的像素极少，无法准确获取单粒种子的图像信息。图 2B 中 3 个品种 7 个不同劣变梯度的共计 315 个种子样本反射光谱。每个品种设置了 105 个样本，由于首尾波段区域噪声干扰较大，为了减少光谱数据引起的噪声干扰，去除了前后共 39 个波段，选取了 465.67~970.77 nm 的 198 个高光谱波段作为原始光谱用于后续分析。由图 2 看出原始光谱存在高频噪声，需要进行预处理。本研究使用 MSC、SNV、FD、SG 等 4 种预处理方法来优化烟草种子光谱数据，以提高分类模型的性能。图 3 结果表明：基于 99% 置信度区间的数据分析经 SG 平滑预处理后的数据没有超出阈值线，即异常值完全消除，而 MSC、FD 预处理有少量异常值，SNV 预处理效果较差，大




A. 群体法检测排列；B. 原始光谱图；C. MSC预处理；D. SNV预处理；E. FD预处理；F. SG预处理

图 2 各预处理后的光谱曲线图

Figure 2 Spectral curves after different preprocessing methods

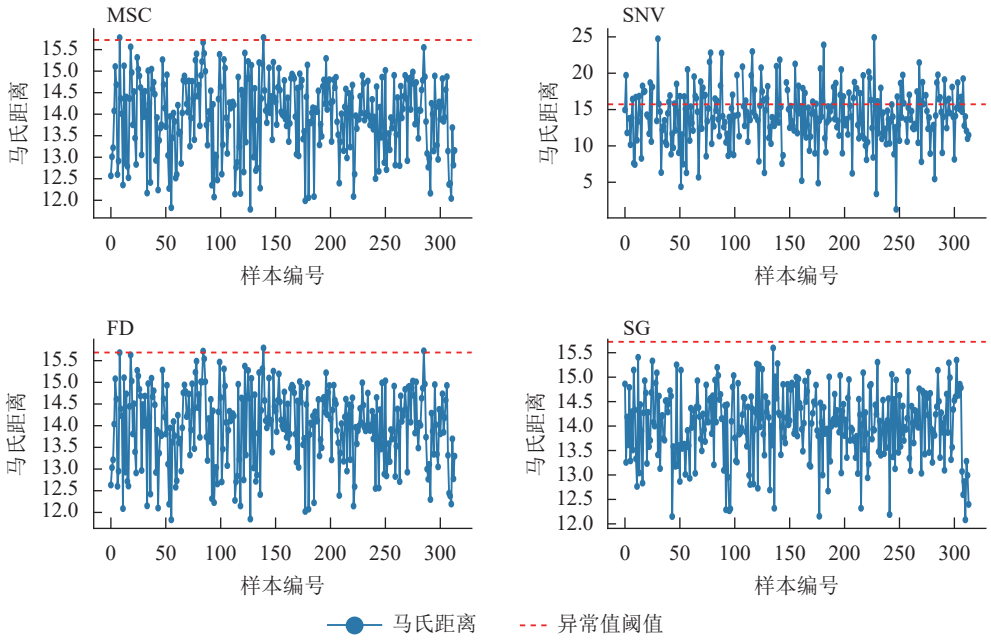



图 3 马氏距离离群值检测

Figure 3 Outlier detection using Mahalanobis distance

部分结果异常。因此，本研究最终选用 SG 平滑预处理方法为后续特征波长的提取提供最佳光谱数据。

### 2.4 平均光谱分析

由图 4 可见：经 SG 预处理后，噪声被有效平滑，各组样本的平均光谱曲线在关键波段上呈现出明显差异，尤其在 650 nm 以上的区域，低活力样本与高活力样本之间的反射率差异更为显著。具体来看，高活力种子的平均光谱曲线整体反射率较低，且曲线呈现平滑上升趋势，而低活力种子则在高波段区域显示出较高且波动较大的反射率。这一现象可能与种子在不同劣变状态下的水分含量、细胞完整性以及化学成分变化有关[12]。

### 2.5 基于特征波段筛选方法的优选波段建模分析

SPA 算法通过逐步投影选择最具差异性的波长变量子集。如图 5A 所示：随着波段数增加，交叉验证均方根误差 (root mean square error of cross validation, RMSECV) 逐渐降低；波段数达 15 时，均方根误

差 (root mean square error， RMSE) 最小 (0.228 9)，随后略有升高。图 5B 显示经 SG 预处理后，SPA 从 198 个全光谱波长中筛选出 15 个特征波段，其中 450~650 nm 区间被选择的波段较多。

UVE 算法通过消除信息量较低的变量提高模型的准确性与泛化能力。从图 5C 可见：多数真实变量稳定性高于随机噪声，尤其在前部变量索引区域明显超出阈值，表明这些波段对烟草种子活力分类贡献较大。UVE 共保留 66 个波段，尤其集中在 415~600 和 800~970 nm 区域 (图 5D)。

CARS 方法结合加权自适应抽样与竞争选择以确定最优变量子集。如图 5E 所示：在 EDP 作用下，波段数初期快速降低，随后降低速度减缓。RMSECV 随着迭代次数的增加先降低后升高，在第 43 次采样达到最低值，此时变量子集包含了预测烟草种子活力的 28 个关键波段。图 5F 显示 CARS 方法尤其在 600~800 nm 区域内波段采集比例最大。

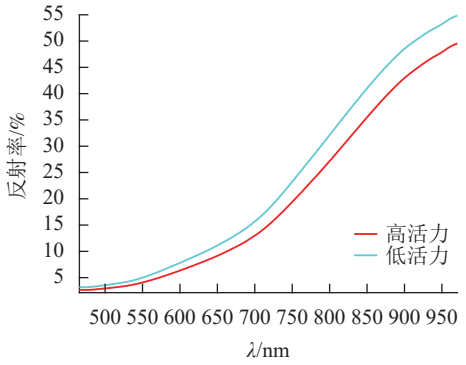
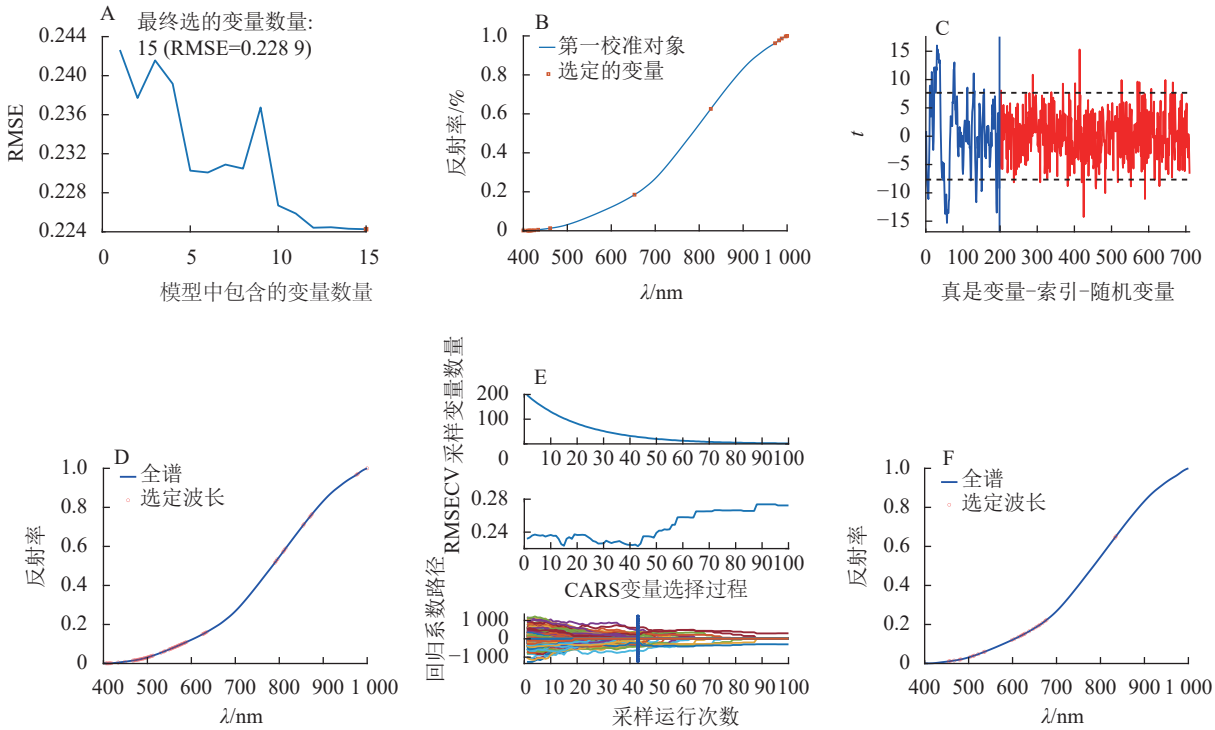



图 4　SG 预处理后依据 SVI 分类的不同活力等级烟草种子平均光谱曲线图

Figure 4　Mean spectrum after of tobacco seeds with different vigor levels classified by SVI after SG pretreatment




A. SPA的PLS-RMSECV随变量数的变化曲线；B. 样本的反射率曲线，橙色标记为SPA选中的特征波段；C. UVE变量稳定性统计；D. 全谱与UVE保留的稳定变量(红色空心圆)；E. CARS迭代过程；F. 全谱与CARS最终保留的关键波段(红色空心圆)。

图 5　基于 SPA、UVE 与 CARS 的特征波段筛选结果

Figure 5　Feature-wavelength selection results using SPA, UVE, and CARS

经过特征选择方法处理后的模型表现如表 2 所示。CARS 特征选择后的模型在训练集和测试集上的分类准确率普遍较低，仅 RF 模型的测试集准确率达到了 80.95%。相比之下，UVE 特征选择下的 CNN-LightGBM 集成模型在训练集和测试集上的表现均最为优异，训练集各项指标均超过 90%，测试集准确率达 88.90%，召回率高达 97.40%。SPA 方法虽能提升部分模型的分类效果，但仍不及 UVE。总体来看，UVE 方法筛选出的特征变量更能反映样本类别差异，配合 CNN-LightGBM 模型能达到较高的精度，是实现烟草种子活力无损分类的最优策略。

### 2.6　模型有效性验证

烟草品种‘MS121’数据集 GP、PRL 及 SW 的权重分别为 0.43，0.32 和 0.25。使用 SG-UVE-CNN-

**表 2 特征选择后模型效果**

Table 2 Model performance after feature selection

| 模型输入 | 模型 | 变量数 | 训练集 | | | | 测试集 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 精确率/% | 召回率/% | F1分数 | 准确率/% | 精确率/% | 召回率/% | F1分数 | 准确率/% |
| SPA | LightGBM | 15 | 100.00 | 100.00 | 100.00 | 100.00 | 76.70 | 89.20 | 82.50 | 77.80 |
| | RF | 15 | 100.00 | 100.00 | 100.00 | 100.00 | 88.64 | 88.64 | 88.64 | 84.13 |
| | CNN | 15 | 76.30 | 76.84 | 76.55 | 78.57 | 76.61 | 74.18 | 75.00 | 77.78 |
| | CNN-LightGBM | 15 | 97.60 | 98.80 | 98.20 | 97.60 | 80.50 | 89.20 | 84.60 | 81.00 |
| UVE | LightGBM | 66 | 83.40 | 94.00 | 88.40 | 83.70 | 76.10 | 92.10 | 83.30 | 77.80 |
| | RF | 66 | 93.49 | 95.18 | 94.33 | 92.46 | 81.40 | 89.74 | 85.37 | 80.95 |
| | CNN | 66 | 78.52 | 78.36 | 78.44 | 80.56 | 78.53 | 78.11 | 78.29 | 79.37 |
| | CNN-LightGBM | 66 | 95.70 | 94.00 | 94.80 | 93.20 | 86.00 | 97.40 | 91.40 | 88.90 |
| CARS | LightGBM | 28 | 75.70 | 97.60 | 85.30 | 77.70 | 66.10 | 97.40 | 78.70 | 68.30 |
| | RF | 28 | 92.68 | 92.68 | 92.68 | 90.48 | 85.37 | 85.37 | 85.37 | 80.95 |
| | CNN | 28 | 75.12 | 74.03 | 74.44 | 76.16 | 77.78 | 75.55 | 75.55 | 75.55 |
| | CNN-LightGBM | 28 | 87.30 | 82.50 | 84.80 | 80.50 | 73.50 | 94.70 | 82.80 | 76.20 |

LightGBM 模型进行模型有效性验证，分类结果以混淆矩阵展示。考虑到不同批次中高/低活力种子的基线比例 (流行率) 不同，研究未使用固定 0.5 的决策阈值，而采用流行率匹配阈值，将预测为“高活力”的比例对齐训练集的标注规则 (>35%)，从而在不利用验证集标签信息的前提下自适应缓解批次效应。如图 6 所示：模型整体分类准确率达到 85.58%，表现出较强的泛化性能。结果表明 SG-UVE-CNN-LightGBM 模型能准确区分不同类别的烟草种子活力，在实际应用中具备较强的分类能力，具有较好的潜在推广价值。

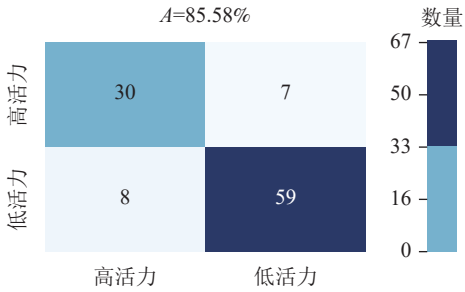



图 6 SG-UVE-CNN-LightGBM 模型分类结果的混淆矩阵

Figure 6 Confusion matrix of the SG-UVE-CNN-LightGBM model for classification results

## 3 讨论

光谱曲线上各关键波段的变化通常反映的是种子在劣变过程中的微观结构与内在生理和化学成分的多重动态演变[13]。烟草种子在劣变过程中，其种子细胞壁可能出现结构松弛或降解，从而改变细胞壁对光的散射能力[14]，导致种子的光谱反射率发生变化[15]。同时，种子中过氧化物酶、超氧化物歧化酶及过氧化氢酶等关键酶的活性降低，促使活性氧水平上升并诱发脂质过氧化反应，其产物的积累会进一步促使种子的光谱特性发生改变[16]，该机制在近年的种子老化与氧化胁迫研究中亦得到支持[17−18]。此外，高温高湿促使种子呼吸加剧，导致可溶性糖和蛋白质含量下降，从而改变种子对特定波段尤其是近红外区域的吸收情况[19]。值得注意的是，780~970 nm 波段的光谱信息被认为与种子内部脂肪含量密切相关，不仅反映种子能量物质储备的变化，也可能关联种子整体活力水平[20]。种子劣变过程中内部结构和化学成分的变化对高光谱数据中特征波段的选择影响显著，这为利用不同特征选择方法提取与种子活力相关的有效波段提供了理论支撑。

近年来，高光谱成像结合机器学习模型在多种作物种子活力评价中得到广泛应用，并在甜玉米、向日葵 *Helianthus annuus*、水稻等作物上实现了高准确率识别[21−23]，但不同作物因其种子结构、成分以及活力劣变机制的差异，模型性能表现存在明显差异。例如，在玉米种子活力识别研究中，SVM、RF 和卷积神经网络 (CNN-DC) 等模型均能有效区分不同活力水平的种子，CNN-DC 模型准确率高达 92.06%，显著优于 SVM 与 RF(准确率均在 70% 以上)[24]。而对于如烟草等微小粒种子，由于其体积小、

信号弱，传统模型对高维光谱数据的表达能力有限。相较之下，集成深度学习模型 (如 CNN-LightGBM) 能够更好地捕捉微小种子在空间与光谱维度的复杂表征特征，实现对关键生理变化的敏感识别，从而有效提升模型在烟草种子活力分类中的表现。本研究结果显示：UVE 特征选择结合 CNN-LightGBM 模型在烟草种子上取得了优于其他组合的分类效果，验证了其对于小粒种子复杂信息挖掘与表征的优势。因此，依托高光谱成像的高维数据获取能力以及深度模型的强特征学习能力，有望突破传统检测手段的灵敏度瓶颈，实现对烟草等微小粒作物种子活力的高精度、无损检测。这为后续多作物种子质量评价体系的建立与完善提供了可借鉴的技术路径和理论基础。

## 4　结论

本研究表明：高光谱成像结合机器学习可在不破坏样品的前提下实现烟草微小粒种子活力的快速分类，通过 SG 预处理方法能够显著提高模型的准确性，而 UVE 则有效提升了模型的分类效果，基于 SG-UVE-CNN-LightGBM 组合模型的准确率达 88.90%，能够准确区分不同活力等级的烟草种子，并在测试集与独立验证集上取得稳定、均衡的识别表现。但本研究也存在一定局限性，研究的样本主要来自人工老化处理，与自然老化在损伤机制上的差异可能影响模型的可迁移性；批次效应、设备误差造成的 UVE 波段对齐误差，也可能影响模型稳定性，降低模型准确率。后续工作将面向自然老化与多来源数据开展更大规模验证，同时尝试更高分辨率的单粒采集与多模态融合，以获取更细粒度的表征并提升模型在真实生产场景中的可部署性。此外，本研究构建的活力指数判别体系有助于跨研究对比，建议在进行种子活力等数据测定时，严格遵守中国以及国际种子检验协会 (International Seed Testing Association，ISTA) 的最新规程[25]。

## 5　参考文献


[1] 王国平, 索文龙, 周东洁, 等. 烟草种子技术研究进展[J]. 种子, 2017, **36**(10): 50−58. WANG Guoping, SUO Wenlong, ZHOU Dongjie, *et al*. Research progress of tobacco seed technology[J]. *Seed*, 2017, **36**(10): 50−58. DOI: 10.16590/j.cnki.1001-4705.2017.10.050.

[2] 利站. 烟草种子发育、贮藏和引发过程中的质量变化和机理研究[D]. 杭州: 浙江大学, 2018. LI Zhan. *Study on Quality Change and Mechanism of Tobacco Seeds during Development, Storage and Priming*[D]. Hangzhou: Zhejiang University, 2018.

[3] 王冬, 王坤, 吴静珠, 等. 基于光谱及成像技术的种子品质无损速测研究进展[J]. 光谱学与光谱分析, 2021, **41**(1): 52−59. WANG Dong, WANG Kun, WU Jingzhu, *et al*. Progress in research on rapid and non-destructive detection of seed quality based on spectroscopy and imaging technology[J]. *Spectroscopy and Spectral Analysis*, 2021, **41**(1): 52−59. DOI: 10.3964/j.issn.1000-0593(2021)01-0052-08.

[4] AMBROSE A, KANDPAL L M, KIM M S, *et al*. High speed measurement of corn seed viability using hyperspectral imaging[J]. *Infrared Physics & Technology*, 2016, **75**: 173−179. DOI: 10.1016/j.infrared.2015.12.008.

[5] 张婷婷, 向莹莹, 杨丽明, 等. 高光谱技术无损检测单粒小麦种子生活力的特征波段筛选方法研究[J]. 光谱学与光谱分析, 2019, **39**(5): 1556−1562. ZHANG Tingting, XIANG Yingying, YANG Liming, *et al*. Wavelength variable selection methods for non-destructive detection of the viability of single wheat kernel based on hyperspectral imaging[J]. *Spectroscopy and Spectral Analysis*, 2019, **39**(5): 1556−1562. DOI: 10.3964/j.issn.1000-0593(2019)05-1556-07.

[6] 石睿, 张晗, 王成, 等. 高光谱图谱结合策略检测小麦单粒种子活力[J]. 光谱学与光谱分析, 2024, **44**(11): 3206−3212. SHI Rui, ZHANG Han, WANG Cheng, *et al*. Detection of wheat single seed vigor using hyperspectral imaging and spectrum fusion strategy[J]. *Spectroscopy and Spectral Analysis*, 2024, **44**(11): 3206−3212. DOI: 10.3964/j.issn.1000-0593(2024)11-3206-07.

[7] ZHU Hongfei, YANG Ranbing, LU Miaomiao, *et al*. Identification of maize seed vigor under different accelerated aging times using hyperspectral imaging and spectral deep features[J]. *Computers and Electronics in Agriculture*, 2025, **231**: 109980. DOI: 10.1016/j.compag.2025.109980.

[8] CUI Huawei, BING Yang, ZHANG Xiaodi, *et al*. Prediction of maize seed vigor based on first-order difference characteristics of hyperspectral data[J]. *Agronomy*, 2022, **12**(8): 1899. DOI: 10.3390/agronomy12081899.

[9] 王新忠, 卢青, 张晓东, 等. 基于高光谱图像的黄瓜种子活力无损检测[J]. 江苏农业学报, 2019, **35**(5): 1197−1202. WANG Xinzhong, LU Qing, ZHANG Xiaodong, *et al*. Non-destructive detection of cucumber seeds vigor based on hyperspectral imaging[J]. *Jiangsu Journal of Agricultural Sciences*, 2019, **35**(5): 1197−1202. DOI: 10.3969/j.issn.1000-4440.2019.05.028.

[10] 杨波, 段明磊, 杨童. 基于高光谱成像技术的西瓜种子活力等级分类方法研究[J]. 河南农业科学, 2022, **51**(9): 151−158. YANG Bo, DUAN Minglei, YANG Tong. Research on the classification method of watermelon seed vigor level based on hyperspectral imaging technology[J]. *Journal of Henan Agricultural Sciences*, 2022, **51**(9): 151−158. DOI: 10.15933/j.cnki.1004-3268.2022.09.016.

[11] 王昭栋, 王自法, 李兆焱, 等. 基于机器学习-网格搜索优化的砂土液化预测[J]. 振动与冲击, 2024, **43**(5): 82−93. WANG Zhaodong, WANG Zifa, LI Zhaoyan, *et al*. Prediction of sandy soil liquefaction based on machine learning-GridSearchCV[J]. *Journal of Vibration and Shock*, 2024, **43**(5): 82−93. DOI: 10.13465/j.cnki.jvs.2024.05.009.

[12] 刘贵华, 袁龙义, 苏睿丽, 等. 储藏条件和时间对 6 种多年生湿地植物种子萌发的影响[J]. 生态学报, 2005, **25**(2): 371−374. LIU Guihua, YUAN Longyi, SU Ruili, *et al*. Effects of storage condition and duration on seed germination of six wetland perennials[J]. *Acta Ecologica Sinica*, 2005, **25**(2): 371−374. DOI:1000-0933(2005)02-0371-04

[13] REDDY P, GUTHRIDGE K M, PANOZZO J, *et al*. Near-infrared hyperspectral imaging pipelines for pasture seed quality evaluation: an overview[J]. *Sensors*, 2022, **22**(5): 1981. DOI: 10.3390/s22051981.

[14] KUREK K, PLITTA-MICHALAK B, RATAJCZAK E, *et al*. Reactive oxygen species as potential drivers of the seed aging process[J]. *Plants*, 2019, **8**(6): 174. DOI: 10.3390/plants8060174.

[15] 禹晓梅, 马文广, 郑昀晔. 烟草种子的老化规律研究[J]. 种子, 2016, **35**(3): 21−24. YU Xiaomei, MA Wenguang, ZHENG Yunye. Study on aging law of tobacco seeds[J]. *Seed*, 2016, **35**(3): 21−24. DOI: 10.16590/j.cnki.1001-4705.2016.03.021.

[16] 孙俊, 张林, 周鑫, 等. 采用高光谱图像深度特征检测水稻种子活力等级[J]. 农业工程学报, 2021, **37**(14): 171−178. SUN Jun, ZHANG Lin, ZHOU Xin, *et al*. Detection of rice seed vigor level by using deep feature of hyperspectral images[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2021, **37**(14): 171−178. DOI: 10.11975/j.issn.1002-6819.2021.14.019.

[17] YE Tiantian, MA Tianxiao, CHEN Yang, *et al*. The role of redox-active small molecules and oxidative protein post-translational modifications in seed aging[J]. *Plant Physiology and Biochemistry*, 2024, **213**: 108810. DOI: 10.1016/j.plaphy.2024.108810.

[18] RAUF A, KHALIL A A, AWADALLAH S, *et al*. Reactive oxygen species in biological systems: pathways, associated diseases, and potential inhibitors: a review[J]. *Food Science & Nutrition*, 2024, **12**(2): 675−693. DOI: 10.1002/fsn3.3784.

[19] BARNES W J, ANDERSON C T. Release, recycle, rebuild: cell-wall remodeling, autodegradation, and sugar salvage for new wall biosynthesis during plant development[J]. *Molecular Plant*, 2018, **11**(1): 31−46. DOI: 10.1016/j.molp.2017.08.011.

[20] QIAO Juxiang, LIAO Yun, YIN Changsheng, *et al*. Vigour testing for the rice seed with computer vision-based techniques[J]. *Frontiers in Plant Science*, 2023, **14**: 1194701. DOI: 10.3389/fpls.2023.1194701.

[21] WANG Yi, SONG Shuran. Detection of sweet corn seed viability based on hyperspectral imaging combined with firefly algorithm optimized deep learning[J]. *Frontiers in Plant Science*, 2024, **15**: 1361309. DOI: 10.3389/fpls.2024.1361309.

[22] HUANG Peng, YUAN Jinfu, YANG Pan, *et al*. Nondestructive detection of sunflower seed vigor and moisture content based on hyperspectral imaging and chemometrics[J]. *Foods*, 2024, **13**(9): 1320. DOI: 10.3390/foods13091320.

[23] QI Hengnian, HUANG Zihong, SUN Zeyu, *et al*. Rice seed vigor detection based on near-infrared hyperspectral imaging and deep transfer learning[J]. *Frontiers in Plant Science*, 2023, **14**: 1283921. DOI: 10.3389/fpls.2023.1283921.

[24] XU Peng, FU Lixia, PAN Yongfei, *et al*. Identification of maize seed vigor based on hyperspectral imaging and deep learning[J]. *Bulletin of the National Research Centre*, 2024, **48**(1): 84. DOI: 10.1186/s42269-024-01239-6.

[25] International Seed Testing Association (ISTA). *International Rules for Seed Testing*[S]. 2024 ed. Bassersdorf: ISTA, 2024.