



## 番茄早疫病高光谱图像早期判别及可视化研究\*

鲍浩<sup>1,2)</sup> 黄莉<sup>1)</sup> 张艳<sup>1,2)\*\*</sup> 庞浩<sup>1,2)</sup><sup>(1)</sup> 贵阳学院计算机科学学院, 农产品无损检测工程研究中心, 贵阳 550005; <sup>(2)</sup> 贵州大学大数据与信息工程学院, 贵阳 550025)

**摘要** **目的** 番茄是世界范围内产量高、种植面积广的经济作物之一, 具有较高的经济效益。番茄早疫病感染速度快、传播范围广、破坏性强, 导致番茄减产减收带来严重的经济损失。番茄早疫病早期病状的判别可以为病害的科学防治提供参考, 实现病害区域的可视化定位, 精准指导病害防治。本文使用高光谱图像结合机器学习算法建立了番茄早疫病早期病状的判别模型, 实现了番茄早疫病早期判别和病害区域的可视化定位。**方法** 针对高光谱图像的噪声干扰, 利用鲁棒主成分分析 (RPCA) 实现高光谱图像的去噪处理; 为了避免主观选择感兴趣区域造成信息表征不足的问题, 使用最大类间方差法 (OTSU) 算法从背景中有效提取了番茄叶片, 并以整个番茄叶片的平均光谱作为研究对象, 结合多元散射校正 (MSC) 算法以及标准化建立了光谱预处理的综合方法; 针对预处理后的光谱, 为了提高模型的判别能力, 建立了基于特征波长和线性核函数 SVM 的番茄早疫病早期判别模型。**结果** 较全光谱波长建模在测试集上的准确率提高 8.33%; 对 SVM 模型的参数进行优化, 当参数  $C=1.64$  时, 训练集和测试集上准确率分别达到了 91.67% 和 94.44%, 较参数优化前的模型在训练集上的准确率提高了 1.19%, 而测试集准确率保持不变, 有效缓解了模型欠拟合问题。**结论** 本文建立了番茄早疫病高光谱图像早期判别模型并实现了早期病状的可视化定位, 基于特征波长和线性核函数 SVM 的判别模型可以有效判别番茄早疫病病害的早期病状; 利用患病概率的形式对番茄早疫病早期病状进行可视化定位, 可视化分析更直观地发现早期病害并采取防治措施。本研究为植物病害早期病状的判别和可视化研究提供了参考, 对作物病害的监测识别和科学防治具有重要意义。

**关键词** 高光谱成像, 机器学习, 遗传算法, 番茄早疫病, 早期病状, 可视化

中图分类号 TP18, S432, O657.3

DOI: 10.16476/j.pibb.2024.0207

CSTR: 32369.14.pibb.20240207

针对农作物病害早期诊断的研究, 传统的病害检测方法主要包括人工感官评价和理化分析方法<sup>[1]</sup>。前者由于主观性强、耗时费力, 难以准确识别作物早期病状, 且不适用于大规模作物病害检测。后者虽然具有较高的准确率, 但由于时效性差、检测成本高且对样本具有破坏性, 难以满足现代农业发展的需求。而高光谱成像凭借其高分辨率、宽光谱范围以及“图谱合一”的优势<sup>[2]</sup>, 能够快速、准确且无损地实现植物病害检测识别。近年来, 高光谱成像技术在植物病害早期诊断中的应用越来越广泛。王建涛等<sup>[3]</sup>提出了一种基于卷积神经网络 (convolutional neural network, CNN) 和高光谱成像的柑橘病叶分类的方法, 取得了 98.75% 的准确率; Huang 等<sup>[4]</sup>利用高光谱成像结合偏最小二乘的判别模型检测了贮藏期间蓝莓的早期病害, 分别对健康蓝莓和早期病害蓝莓取得了 100% 和 99% 的准确率; 高荣华等<sup>[5]</sup>提出了一种基

于高光谱成像识别早期番茄灰霉病的方法, 模型的最高准确率为 99%, 平均准确率在 90% 以上, 表明该方法能有效识别番茄早期灰霉病且具有较好的鲁棒性。显然, 使用高光谱成像结合相关数据处理方法可以有效的检测识别作物病害。

当前作物病害的相关研究主要集中在复杂背景下作物病害病斑识别、病害严重程度定量分析等方面。林立杰等<sup>[6]</sup>针对作物生长的复杂环境为待识别的目标设定了统一标注标准, 保证模型能准确定位目标, 识别了数据集中的叶片及病斑, 结果显示, 训练准确率为 81.87%, 验证损失率为 15.22%, 能够达到一个良好的分割识别效果; 刘媛媛等<sup>[7]</sup>使用高光谱成像结合集成学习的算法实现了不同病

\* 国家自然科学基金 (62265003, 62141501) 资助项目。

\*\* 通讯联系人。

Tel: 0851-85402217, E-mail: Eileen\_zy001@sohu.com

收稿日期: 2024-05-14, 接受日期: 2024-09-29

害程度的库尔勒香梨黑斑病的检测识别,模型的总体准确率达到98.28%; Gao等<sup>[8]</sup>提出了一种基于SegNet的网络用于分割马铃薯晚疫病(PLB)病斑,以量化PLB的严重程度; Appeltans等<sup>[9]</sup>通过逐像素点的光谱建模识别了不同健康程度的韭菜白尖病,总体准确率为96.74%; 赵森等<sup>[10]</sup>使用高光谱成像检测了刺五加早期黑斑病; Adhikari等<sup>[11]</sup>通过主成分分析(principal component analysis, PCA)实现高光谱数据降维,并标记了不同的病害类别,轻微病害(绿色和黄色)和严重病害(红色),然后基于不同核函数的SVM对PCA降维后的高光谱数据建立了病害识别模型; Shi等<sup>[12]</sup>利用不同架构的CNN(经典CNN架构、改进CNN架构、基于CNN的分割网络)对植物病害程度评估方法进行了概述,并提出该方法在实际应用中可能面临的挑战以及解决方案。可见,现有研究缺乏对叶片患病区域的可视化定位,不仅无法了解患病区域的分布情况,也无法为后续的病状监测和治疗提供精确的信息。

早疫病是由茄链格孢菌导致的番茄易感病害之一,其发病率一般在10%左右,但在流行年份可能高达100%,造成的产量损失高达30%~40%<sup>[13]</sup>。本文旨在建立番茄早疫病早期病状的判别模型并进一步实现病状的可视化定位,解决目前对叶片患病区域缺乏可视化定位的问题。本文的主要研究内容包括:a.根据高光谱成像的特点,结合鲁棒主成分分析去除高光谱图像噪声;b.利用数字图像处理方法分离了样品高光谱图像的背景和叶片,选取合适的感兴趣区域;c.结合遗传算法(genetic algorithm, GA)选择了有利于早期病状判别的特征波长,并建立识别模型;d.对番茄早疫病早期病状判别模型的参数进行优化分析,从而增强模型性能;e.利用高光谱图像“图谱合一”的优势,实现早期病状的可视化定位。

## 1 实验数据采集

在实验室环境下使用无菌土种植番茄植株40盆,待番茄生长至结果期时采集无病虫害影响、形状大小相似、颜色相近且长势良好的番茄叶片共120片,其中70片接种茄链格孢菌作为接种组,50片不作任何处理作为健康对照组。病菌接种方式为悬浮液接种,即使用茄链格孢菌悬浮液喷洒至番茄叶片表面。在接种以后将120个番茄叶片样本分别放入温度为28~30℃,湿度为70%~90%的环境中

进行培养。以8 h为时间间隔,对高光谱图像进行动态监测,发现接种60 h以后的大部分番茄叶片表面出现了肉眼可见的病状。因此,将肉眼刚刚能区分健康和接种样本的时间定义番茄早疫病早期,本研究选择接种病菌60 h以后的番茄早疫病高光谱图像作为番茄早疫病早期病状判别模型的建模数据。

使用由四川双利合谱科技有限公司生产的高光谱成像系统(GaiaField-F-V10)采集番茄早疫病早期病状的实验数据。在采集图像前,为保证采集过程中系统光源稳定,先开启高光谱成像系统进行预热,预热时间约为30 min。在采集图像时,设置电荷耦合器件(charge coupled device, CCD)相机的曝光时间为15 ms,镜头与样本的距离为40 cm,电动平台的前进速度为1 cm/s,回退速度为2 cm/s。

## 2 高光谱图像去噪

光谱图像包含丰富的物理化学信息,在光谱分析时,提取关键信息有利于分析光谱波长与被测物质的相互作用。为了深入探讨染病番茄叶片在不同波长下的反射特性,通过细致的光谱信息可揭示病症的具体影响,然而高光谱图像在成像过程中易受到高斯噪声、脉冲噪声以及条纹噪声等污染,会严重破坏光谱和图像的信息结构,从而在研究中对采集的高光谱图像进行去噪处理至关重要。高光谱图像具有数百个光谱通道,传统的高光谱图像去噪往往采用逐通道的方式<sup>[13-14]</sup>,由于忽略了各通道间的强相关性,导致逐通道的去噪效果较差。鲁棒主成分分析(robust principal component analysis, RPCA)<sup>[15]</sup>是一种高光谱图像去噪方法,假设一幅受噪声污染的高光谱图像 $Y$ 由纯净的高光谱图像 $X$ 和稀疏噪声 $E$ 构成,如式(1)所示。

$$Y = X + E \quad (1)$$

RPCA认为纯净的高光谱图像具有低秩的特性,为了恢复原始的高光谱图像 $X$ ,RPCA的求解模型可以表示为:

$$\begin{aligned} \min_{X,E} \text{rank}(X) + \lambda \|E\|_0 \\ \text{s.t. } Y = X + E \end{aligned} \quad (2)$$

式(2)中, $\text{rank}(\cdot)$ 表示求秩函数, $\lambda > 0$ 为权重因子, $\|E\|_0$ 表示 $E$ 的0范数。

然而,式(2)是非凸优化问题,难以得到解析解,所以一般对式(2)进行松弛处理:

$$\begin{aligned} \min_{X,E} \|X\|_* + \lambda \|E\|_1 \\ \text{s.t. } Y = X + E \end{aligned} \quad (3)$$

式(3)中 $|X|_*$ 表示 $X$ 的核范数,  $|E|_1$ 为 $E$ 的1范数。

此时的式(3)一般可使用增广拉格朗日乘子法或交替方向乘子法等方法求解, 从而得到纯净的高光谱图像 $X$ 和稀疏噪声 $E$ 。

为了比较使用RPCA对番茄早疫病早期病状的高光谱图像去噪前后效果的差异, 分别提取患病区域和健康区域中任意一个像素的光谱曲线进行对比, 为了增加病状区域的可读性, 使用高光谱图像对应R (663.29 nm)、G (553.43 nm)、B (458.04 nm)

波段所构成的伪彩色图像进行说明(图1)。图1a为去噪前病状高光谱图像和点光谱曲线, RPCA去噪前的光谱曲线受到噪声的严重影响, 特别是首尾处的噪声较多, 导致有效信息大多被噪声淹没, 在建模时可能会严重降低模型性能; 图1b是RPCA去噪后高光谱图像和点光谱曲线, 经过RPCA处理以后的光谱曲线尽管在首尾处(392.53~450.97 nm以及951.03~1027.62 nm)仍然存在一些噪声, 但是大部分的有效信息已经从噪声中分离出来了, 可以进行后续的分析 and 建模处理。

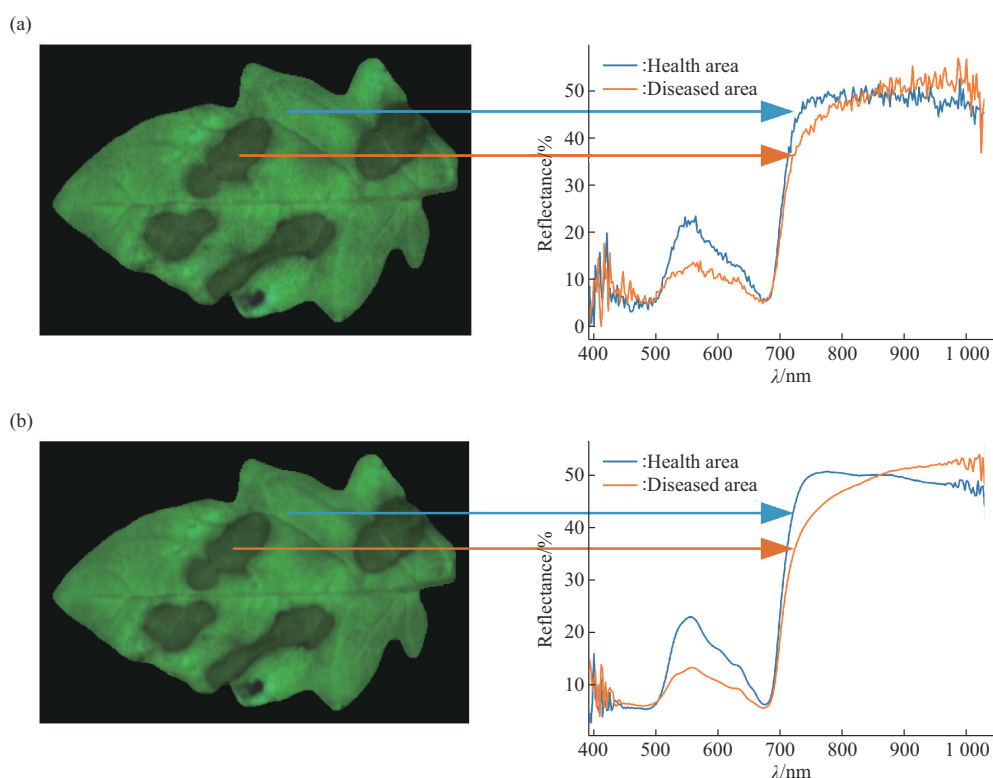


Fig. 1 Hyperspectral image denoising

(a) Spectral curve before RPCA denoising, (b) spectral curves after denoising with RPCA.

### 3 光谱分析与处理

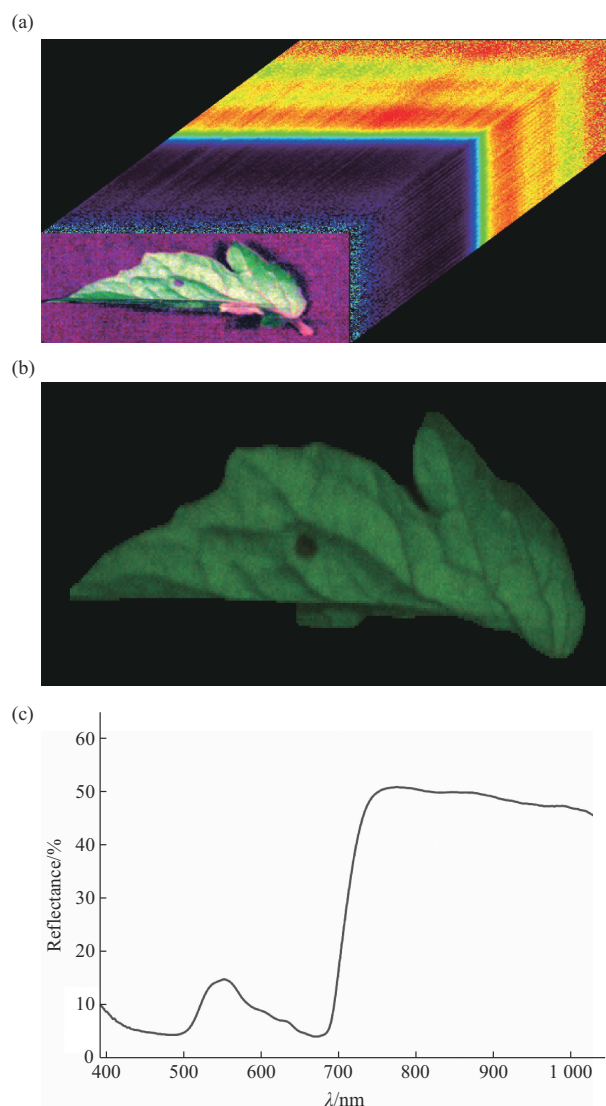
#### 3.1 感兴趣区域的光谱提取

人工选择染病叶片图像局部的感兴趣区域存在信息表征不足的问题, 可能导致建模存在偏差。为了获取染病叶片表征的完整信息, 利用了数字图像处理的方法提取整个叶片作为感兴趣区域。具体步骤为: a. 获取番茄叶片与背景差异最大的波长, 使用OTSU算法对该波长(本文为917.07 nm)对应的特征图像进行阈值分割得到掩膜图像; b. 对掩膜

图像进行腐蚀与膨胀处理, 消除掩膜图像噪声的同时连通区域; c. 使用掩膜图像与原始高光谱图像进行掩膜操作得到感兴趣区域, 然后提取感兴趣区域的平均光谱。番茄早疫病早期病状的高光谱图像、感兴趣区域及其平均光谱(图2)。

#### 3.2 番茄早疫病早期病状的光谱曲线

当植物受到病害胁迫时, 叶肉组织会遭到破坏、并且伴随着叶绿素和水分丢失, 从而发生叶片表面产生枯萎和变黄等现象, 这些现象可以通过光谱反射率的变化反映。为了分析番茄早疫病早期病



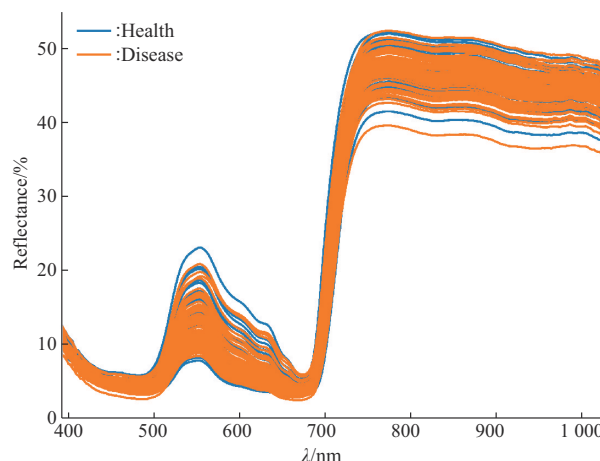
**Fig. 2 Region of interest selection**

(a) Hyperspectral image, (b) region of interest, (c) average spectrum.

状的光谱特性, 本研究分别从健康样本和感病样本的光谱曲线及其均值光谱曲线展开分析。

首先, 为了研究健康样本和感病样本的整体光谱反射率的差异, 分析了50个健康样本和70个感病样本的光谱曲线(图3)。各个番茄叶片样本光谱反射率的趋势大致相同, 相关研究表明<sup>[16-17]</sup>: 400 nm附近的反射率较高是由于叶绿素对紫外光的吸收较少; 由于叶绿素与类胡萝卜素对蓝光的吸收比例较高, 导致在500 nm附近存在反射低谷; 由于叶绿素对绿光的吸收率较低, 从而在550 nm附近形成叶绿素反射峰; 在680 nm附近出现反射低谷主要是由于叶绿素对橙光和红光具有强吸收性; 680~750 nm的光谱反射率迅速增大, 导致近红外

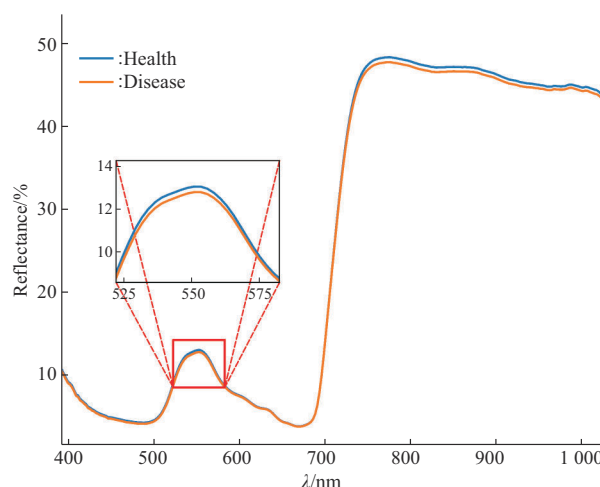
区域的光谱反射率大于可见光区域的光谱反射率, 主要是由于叶片细胞对近红外光谱有反射作用。



**Fig. 3 Full sample spectral curve**

In the spectral range from 400 to 1 000 nm, the full spectral reflectance of both healthy and disease samples.

为了研究在400~1 000 nm的光谱范围内哪些波段能实现病状的识别, 对50个健康样本与70个感病样本的平均光谱曲线进行分析(图4)。感病叶片的光谱走势与健康叶片的光谱走势基本一致, 可能因为是选择将整个样本作为感兴趣区域, 而病斑区域少且面积小, 弱化了病斑的光谱信息, 使整个样本的光谱信息由健康区域占主导。在可见光区域, 感病叶片的反射率略低于健康叶片的反射率, 可能是由于番茄叶片受病菌胁迫, 叶片的自我保护机制增加了细胞对光的吸收, 使得光反射率降低;



**Fig. 4 Health-disease spectrum mean curve**

In the spectral range from 400 to 1 000 nm, the differences in mean spectral reflectance of both healthy and disease samples.

在近红外区域, 感病叶片的反射率均低于健康叶片的反射率, 可能是病菌破坏了叶片细胞, 导致感病叶片对近红外光的反射率降低。

### 3.3 光谱预处理

在数据采集的过程中, 由于叶片表面的不均匀性、外界环境的干扰以及光源变化等因素的影响, 导致叶片表面反射的光谱出现偏移和畸变。为确保采集光谱的准确性和可靠性, 以便更准确地进行光谱分析和建模, 所以需要对光谱进行预处理。在高光谱成像过程中由于叶片表面的不均匀性、外界环境的干扰, 从而使叶片表面反射的光谱信息呈现出差异性。在光谱预处理中, 使用多元散射校正 (multiplicative scatter correction, MSC) 去除光谱散射, 以降低光散射带来的光谱偏移, 进行光谱标准化处理, 使得每个光谱维度的均值为0, 方差为1, 可以消除由于光谱畸变引入的数值偏差, 使模型可忽略各个光谱维度间光谱数值的分布, 能更好的学习光谱曲线中的内在规律, 光谱预处理结果如图5所示。

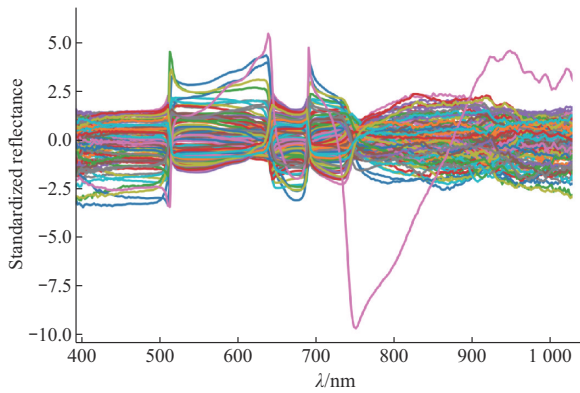


Fig. 5 Spectral pre-processing

The spectra were processed by multiplicative scatter correction and spectral normalization.

## 4 番茄早疫病早期病状的判别模型

将预处理后的数据集 (120 条光谱曲线) 以 7 : 3 的比例随机划分为训练集与测试集, 得到 84 个训练样本 (51 个感病样本, 33 个健康样本) 与 36 个测试样本 (19 个感病样本, 17 个健康样本)。虽然深度学习的识别准确率整体略高于机器学习的识别准确率, 但在识别准确率相差不大、样本数处

于相同量级的情况下, 深度学习的模型参数量会远大于机器学习, 增加了模型的复杂度, 从而也增加了计算资源占用和运算时间消耗。鉴于本实验样本量不大, 且机器学习识别模型已能提供较高的准确率、模型训练和调优也相对简单快速, 经过比较分析, 本文选择机器学习来建立识别模型。支持向量机 (support vector machine, SVM) 是一种适用于小样本建模的分类器, 具有泛化能力好的优点<sup>[18]</sup>, 同时具有能处理高维数据, 对噪声不敏感的特点。本文基于 SVM 建立番茄早疫病早期病状判别模型, 使用该判别模型对番茄早疫病病害高光谱图像进行逐像素点判别, 然后将判别结果转换为伪彩色图实现病状定位可视化。

### 4.1 基于全光谱波长的建模

为了验证全光谱波长对于早疫病早期病状判别的有效性, 使用 SVM 建立基于全光谱波长 (392.53~1 027.62 nm) 的番茄早疫病早期病状判别模型。调整 SVM 的核函数, 探究不同核函数对 SVM 性能的影响, SVM 常用核函数有:

a. 线性核函数, 求解速度快、可解释性好, 但仅能用于线性可分问题, 表达式为:

$$F(z_i, z_j) = z_i^T z_j \quad (4)$$

b. 高斯核函数, 参数少、映射维度高, 但速度慢、可解释性差且易过拟合, 表达式为:

$$F(z_i, z_j) = \exp\left(-\gamma \|z_i - z_j\|^2\right) \quad (5)$$

c. 多项式核函数, 可解决非线性可分问题, 但需要调整的参数过多, 表达式为:

$$F(z_i, z_j) = (\gamma z_i^T z_j + b)^c \quad (6)$$

d. Sigmoid 核函数是多层神经网络, 表达式为:

$$F(z_i, z_j) = \tanh(\gamma z_i^T z_j + b) \quad (7)$$

基于全光谱波长使用 4 种核函数的 SVM 建模的结果如表 1 所示。使用线性核函数建模的性能最好, 在训练集和测试集上的准确率分别达到 96.43% 和 86.11%, 说明使用线性核函数投影到高维空间能取得较好的性能; 而使用 Sigmoid 核函数建模的性能最差, 训练集和测试集上的准确率仅为 52.38% 和 58.33%; 使用高斯核函数建模的性能与使用多项式核函数建模的性能大致相同。因此, 本文选择使用带有线性核函数的 SVM 构建番茄早疫病早期病状的判别模型。

**Table 1 Comparison of SVM recognition performance with different kernel functions**

| Kernel Function   | Training set |            | Test set    |            |
|-------------------|--------------|------------|-------------|------------|
|                   | Sample size  | Accuracy/% | Sample size | Accuracy/% |
| Linear kernel     | 84           | 96.43      | 36          | 86.11      |
| Gaussian kernel   | 84           | 65.48      | 36          | 52.78      |
| Polynomial kernel | 84           | 66.67      | 36          | 52.78      |
| Sigmoid kernel    | 84           | 52.38      | 36          | 58.33      |

使用全光谱波长可以较好实现番茄早疫病早期病状的判别,其建模时输入的波长数为256个,较多的输入变量数需要耗费较多的计算资源,并且光谱波段间存在的冗余信息会降低模型的性能。所以,需要从全光谱波长中提取有利于判别番茄早疫病早期病状的特征波长进行研究。

#### 4.2 基于特征波长的建模

特征波长选择是光谱分析的关键步骤,从全光谱波段中选择具有代表性波长的同时不改变波段本身的物理信息<sup>[19]</sup>,使用选择特征波长建模时,不

仅大幅度减少了光谱维度,降低计算复杂度,提升分析速度,还可以有效避免光谱波长中的冗余信息对模型的影响,从而增强模型的泛化性能与稳定性。本文分别使用竞争性自适应重加权采样算法(competitive adaptive reweighted sampling, CARS)<sup>[20]</sup>、连续投影算法(successive projections algorithm, SPA)<sup>[21]</sup>以及遗传算法<sup>[22]</sup>用于番茄早疫病早期病状的高光谱图像特征波长选择。在选择番茄早疫病早期病状高光谱特征波长时设置CARS算法的蒙特卡洛采样次数为50次,交叉验证次数10次,SPA的特征数量搜索范围为2~50个,GA的种群数为50,最大迭代次数为500,交叉率为0.8,变异率为0.01。

CARS、SPA和GA选择的特征波长如表2所示。CARS、SPA以及GA分别得到了26、5以及14个特征波长。CARS算法选择的特征波长数最多,主要包含了蓝光、绿光、橙光、红光以及近红外波段;其次是GA选择的特征波长数,主要包含了红光和近红外波段,SPA算法选择的特征波长数最少,主要包含了近红外波段。

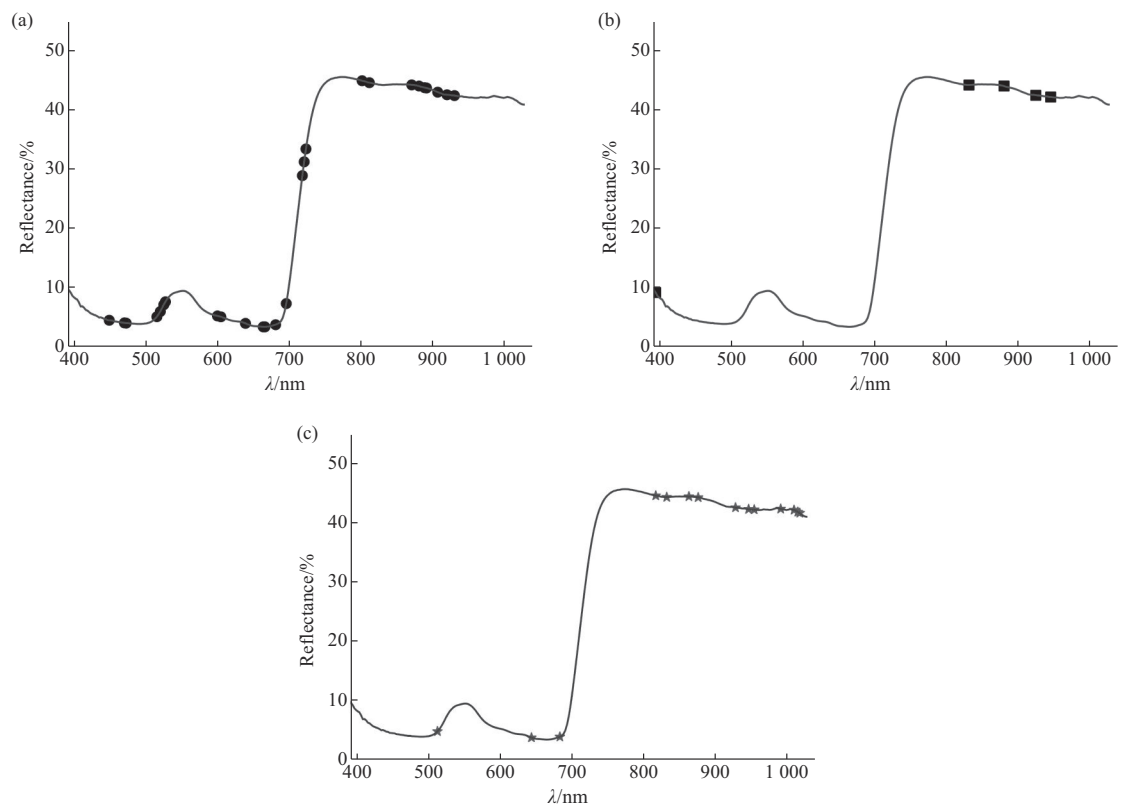
**Table 2 Selected characteristic wavelength**

| Feature wavelength selection algorithm | Feature wavelength/nm                                                                                                                                                                                          |
|----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| CARS                                   | 448.62, 469.85, 472.21, 515.02, 519.80, 524.59, 526.99, 599.49, 604.36, 638.65, 663.29, 665.76, 680.62, 695.53, 717.98, 720.48, 722.98, 801.22, 811.41, 870.41, 880.74, 888.50, 891.09, 906.66, 919.67, 930.10 |
| SPA                                    | 394.85, 831.85, 880.74, 924.88, 945.79                                                                                                                                                                         |
| GA                                     | 512.63, 643.56, 683.10, 816.51, 831.85, 862.67, 875.57, 927.49, 945.79, 953.66, 990.51, 1 009.03, 1 014.33, 1 016.99                                                                                           |

CARS、SPA以及GA选择的特征波长在光谱曲线上的位置如图6所示,在图6a中,CARS算法选择的特征波长主要集中在500~550 nm、600~700 nm以及900 nm附近;在图6b中,SPA选择的特征波长主要集中在900 nm附近;在图6c中,GA选择的特征波长主要集中在800~850 nm、900~1 000 nm区域。其中,500 nm附近的光谱属于蓝光和绿光的交界,叶绿素和类胡萝卜素对该部分的光具有很强的吸收率,主要帮助植物进行光合作用以及合成蛋白质和氨基酸<sup>[23]</sup>,600~700 nm附近的光谱是橙光与红光的交界,可以促进植物的光合作用,从而增加叶绿素、类胡萝卜素、可溶性糖物质积累的作用<sup>[24]</sup>,800~850 nm和900~1 000 nm属于近红外区域,易被植物细胞反射,可以反映叶片的细胞结构。此外,CAPS、SPA以及GA算法选择的特征波长主要集中900~1 000 nm(近红外光)附近,

说明近红外区域的光谱信息对番茄早疫病的判别具有指导性作用。

使用特征波长建模的性能比较如表3所示。基于GA选择的14个特征波长建立的模型在测试集上的准确率达到94.44%,其准确率较全光谱波长提高了8.33%,说明GA选择的特征波长深度揭示了番茄早疫病与特征波长的联系。使用CARS算法选择了26个特征波长,建立的模型在训练集与测试集上的准确率为67.86%和66.67%,分别较全光谱波长的准确率降低了28.57%和19.44%,说明CARS算法不适用于番茄早疫病早期病状的高光谱特征波长选择;使用SPA取得了5个特征波长,虽然取得了最少的特征波长数,但是建立的模型在训练集与测试集上的准确率分别为70.24%和52.78%,分别较全光谱波长降低了26.19%和33.33%,说明SPA不适用于番茄早疫病早期病状的



**Fig. 6 Selected characteristic wavelengths**  
Characteristic wavelength selected by CARS (a), SPA (b), and GA (c).

特征波长选择。GA 在番茄早疫病早期病状高光谱特征波长选择中的性能较好，选择使用GA结合线性核函数的SVM建立番茄早疫病早期病状的判别模型。

**Table 3 Performance comparison of feature wavelength-based SVM**

| Feature wavelength selection algorithm | Number of feature wavelengths | Training set |            | Test set    |            |
|----------------------------------------|-------------------------------|--------------|------------|-------------|------------|
|                                        |                               | Sample size  | Accuracy/% | Sample size | Accuracy/% |
| —                                      | 256                           | 84           | 96.43      | 36          | 86.11      |
| CARS                                   | 26                            | 84           | 67.86      | 36          | 66.67      |
| SPA                                    | 5                             | 84           | 70.24      | 36          | 52.78      |
| GA                                     | 14                            | 84           | 90.48      | 36          | 94.44      |

可以发现，使用全光谱波长以及GA选择的特征波长建立的模型均可以用于实现番茄早疫病早期病状的判别。然而，使用全光谱波长建立的模型出现了过拟合现象，即训练集上的性能高于测试集上的性能，说明全光谱波长（392.53~1 027.62 nm）中存在信息冗余。经过GA算法选择特征波长以后，虽然挑选了具有代表性的特征波长，但是建立的模型却出现了欠拟合问题，即测试集上的性能优于训练集上的性能。因此，为了缓解模型欠拟合，需要在使用特征波长建模的基础上对SVM模型的

参数进行优化调整。

**4.3 模型参数优化**

模型参数优化的目的是寻找一组最佳的参数，使模型不仅在训练集上表现良好，并且在测试集上也有优秀的泛化能力。影响SVM模型性能的参数主要是惩罚参数C和核函数参数<sup>[25-26]</sup>。其中，惩罚系数C表示模型对于错误分类样本的容忍度，核函数参数则影响了投影到高维空间的特征分布。由于选择了带有线性核函数的SVM，所以不需要考虑核函数参数。GA常用于机器学习模型的参数优

化<sup>[27]</sup>，在优化SVM参数时具有收敛性且基本是全局收敛<sup>[28]</sup>。因此使用GA对SVM的参数进行优化调整，得到了参数寻优曲线（图7）。

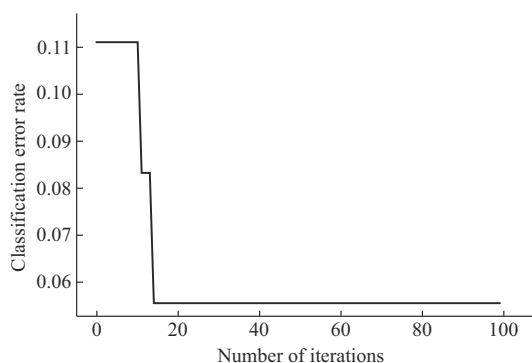


Fig. 7 Optimization of model parameters

Table 4 Performance comparison of GA optimized SVM model parameters

| Model  | Number of input wavelengths | Training set |            | Test set    |            |
|--------|-----------------------------|--------------|------------|-------------|------------|
|        |                             | Sample size  | Accuracy/% | Sample size | Accuracy/% |
| SVM    | 14                          | 84           | 90.48      | 36          | 94.44      |
| GA-SVM | 14                          | 84           | 91.67      | 36          | 94.44      |

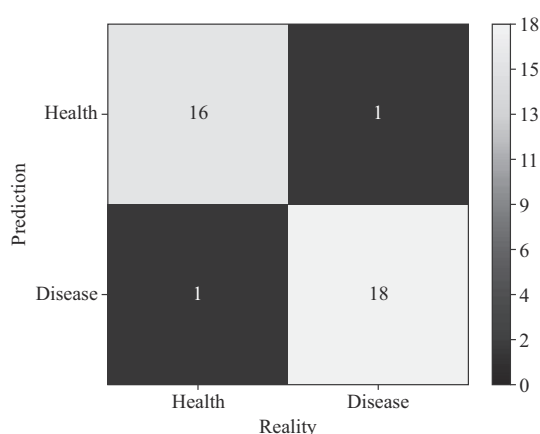


Fig. 8 GA-SVM confusion matrix on the test set

在本轮实验的36个样本中，能够进行准确识别的样本是34个，其中实际为健康样本且模型预测也为健康样本的样本数量是16个，实际为感病样本且模型预测也为感病样本的样本数量是18个，有两个样本被误判，将实际为感病样本预测为健康样本的样本数量是1个，将实际为健康样本预测为感病样本的样本数量是1个。利用该模型进行识别的准确率为94.4%，说明了模型对于番茄早疫病早期病害的判别具有较好的性能。

在GA算法迭代初期，模型的错误率不断下降，说明GA算法正在选择有利于提高模型性能的参数，迭代20次左右时算法就已经收敛，在达到最大迭代次数时得到了最佳参数 $C=1.64$ 。将最佳参数带入SVM番茄早疫病早期病状判别模型，得到的性能比较如表4所示。最优参数的SVM模型（GA-SVM）在训练集和测试集上分别得到了91.67%和94.44%的准确率，在保证测试集上的性能不变的情况下提高了训练集上1.19%的准确率，从而缓解了欠拟合问题，增强了模型的泛化性能。

通过模型在测试集上的混淆矩阵，客观评价了番茄早疫病早期病状判别模型（GA-SVM）的性能（图8）。

#### 4.4 病状的可视化定位

实现病状的可视化定位可以在农业生产中精细化指导农药的使用，由于高光谱图像同时包含样品表面的光谱信息和图像信息，即高光谱图像的任意像素点均包含光谱信息。因此本文将每一个像素点的光谱作为输入，由模型输出对应的判别结果。为了提高判别结果的可读性，使用伪彩色图处理技术将判别结果转换为伪彩色图，从而实现病状的可视化定位。

使用GA-SVM模型对早期病状的高光谱图像进行病状的可视化定位（图9）。图9a为随机从接种组中挑选样本并提取其对应R（663.29 nm）、G（553.43 nm）、B（458.04 nm）波段所构成的伪彩色图像，可以看到番茄叶片上已经存在4个较大的病斑，且病害具有向外扩散的趋势。使用GA-SVM模型对图9a的高光谱图像中叶片样本的像素点进行逐一的判别，利用伪彩色图像处理技术将每个像素点的判别结果转换为伪彩色图（图9b），图中的数值越大则颜色越深，表明患病的概率越高。红色区域是患病概率比较高的位置，蓝色区域的患病概率比较小（可以认为是健康区域），而黄色区域则是病斑向外扩张的区域。从可视化结果可以看出患病区域已经较准确地表现出来，并且还可以观察到在病斑周围还呈现出向外扩张的病变趋势。

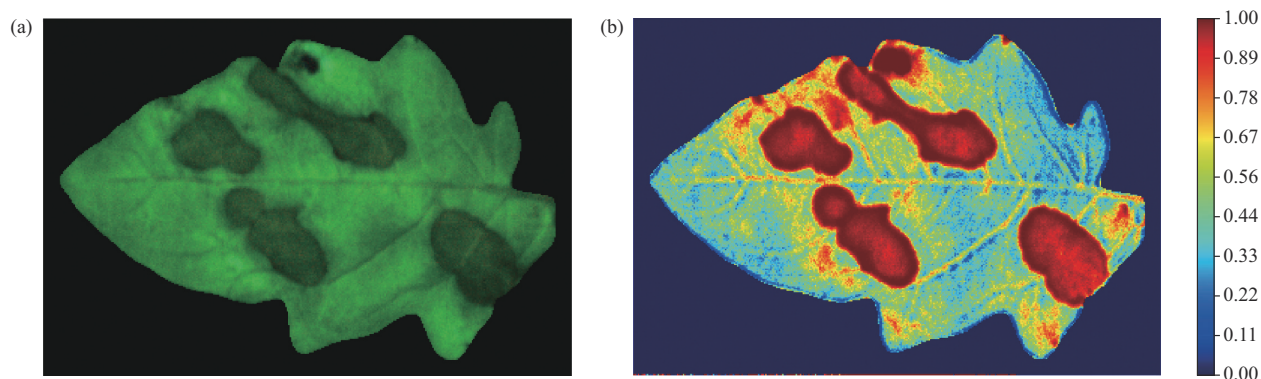


Fig. 9 Symptom visualization results

(a) RGB image, (b) visual positioning of disease evolution trends.

## 5 结果与讨论

本文将机器学习算法应用于高光谱图像建模,建立了基于特征波长的SVM番茄早疫病早期病状判别模型,实现了病状定位可视化,研究结果主要有:

a. 为了避免高光谱图像受到噪声干扰,使用RPCA从受噪声污染的高光谱图像中恢复了具有较低噪声的高光谱图像。

b. 使用数字图像处理方法分离了高光谱图像中的叶片和背景区域,以叶片作为研究的感兴趣区域。针对MSC结合标准化处理后的光谱信息建模,使用GA得到了14个具有代表性的特征波长,分别为512.63、643.56、683.1、816.51、831.85、862.67、875.57、927.49、945.79、953.66、990.51、1 009.03、1 014.33以及1 016.99 nm。此外,使用GA优化了SVM的参数,得到最优参数 $C=1.64$ ,缓解了模型欠拟合的问题,此时的模型在训练集和测试集分别达到了91.67%和94.44%的准确率。

c. 根据高光谱图像“图谱合一”的特点对图像信息进行研究,使用建立的番茄早疫病早期病害判别模型对每个像素点进行判别,以患病概率的形式实现了番茄早疫病早期病状的可视化定位。

尽管本研究弱化了病斑区域的光谱信息,但这种全面的方法有效避免了局部选择带来的信息不足问题,使得结果更具整体性和多样性。这为未来植物病害的早期检测与可视化分析提供了有价值的参考。

## 6 结 论

本文通过RPCA去除噪声,结合数字图像处理技术,研究了番茄叶片的关键光谱信息,筛选出14个特征波长为识别模型建立提供了重要依据。优化后的遗传算法与SVM结合的GA-SVM模型在训练集和测试集上分别达到了91.67%和94.44%的准确率,验证了机器学习方法在高光谱图像分析中的有效性。利用高光谱图像“图谱合一”的特性,本文在建立番茄早疫病早期病状的判别模型基础上进一步实现病状的可视化定位,针对每个像素的病状判别,提升了可视化定位的准确性,解决目前对叶片患病区域缺乏可视化定位的问题。结果表明,利用高光谱图像的特征波段选择对光谱特性变化进行分析,对番茄早疫病病状识别具有重要意义,本文建立的基于线性核函数的SVM判别模型能有效揭示番茄早疫病早期病状的光谱特性。本文研究成果增强了对番茄早疫病早期诊断的能力,为智慧农业的发展提供了新的研究思路和技术支撑。后续研究可进一步探讨不同植物病害的检测方法及其适用性,促进农业生产的精准化管理。

## 参 考 文 献

- [1] 张德荣,方慧,何勇. 可见/近红外光谱图像在作物病害检测中的应用. 光谱学与光谱分析, 2019, 39(6): 1748  
Zhang D R, Fang H, He Y. Spectrosc Spectr Anal, 2019, 39(6): 1748
- [2] Wang C, Liu B, Liu L, *et al.* A review of deep learning used in the hyperspectral image analysis for agriculture. Artif Intell Rev, 2021, 54(7): 5205-5253

- [3] 王建涛, 吴叶兰, 廖禹, 等. 基于卷积神经网络的柑橘病叶高光谱分类. 信息技术与信息化, 2020(3): 84-87  
Wang J T, Wu Y L, Liao Y, *et al.* Inf Technol Informatization, 2020(3): 84-87
- [4] Huang Y, Wang D, Liu Y, *et al.* Measurement of early disease blueberries based on vis/NIR hyperspectral imaging system. Sensors (Basel), 2020, **20**(20): 5783
- [5] 高荣华, 冯璐, 张月, 等. 基于多维随机森林的番茄灰霉病高光谱图像早期检测. 光谱学与光谱分析, 2022, **42**(10): 3226-3234  
Gao R H, Feng L, Zhang Y, *et al.* Spectrosc Spectr Anal, 2022, **42**(10): 3226-3234
- [6] 林立杰, 代佳依, 杨予熙, 等. 基于深度学习水稻病害及其严重程度识别模型研究进展//中国植物病理学会. 中国植物病理学会2023年学术年会论文集. 北京: 中国农业科学技术出版社, 2023: 1  
Lin L J, Dai J Y, Yang Y X, *et al.* Research progress on deep learning based models for identifying rice diseases and their severity//Chinese Society for Plant Pathology. Proceedings of the Annual Meeting of Chinese Society for Plant Pathology (2023). Beijing: China Agricultural Science and Technology Press, 2023: 1
- [7] 刘媛媛, 张凡, 师琪, 等. 基于高光谱和集成学习的库尔勒香梨黑斑病潜育期诊断. 农业机械学报, 2022, **53**(6): 295-303  
Liu Y Y, Zhang F, Shi Q, *et al.* Trans Chin Soc Agric Mach, 2022, **53**(6): 295-303
- [8] Gao J, Westergaard J C, Sundmark E H R, *et al.* Automatic late blight lesion recognition and severity quantification based on field imagery of diverse potato genotypes by deep learning. Knowledge-Based Systems, 2021, **214**: 106723
- [9] Appeltans S, Pieters J G, Mouazen A M. Detection of leek white tip disease under field conditions using hyperspectral proximal sensing and supervised machine learning. Comput Electron Agric, 2021, **190**: 106453
- [10] 赵森, 付芸, 崔江南, 等. 高光谱的刺五加黑斑病的早期检测研究. 光谱学与光谱分析, 2021, **41**(6): 1898-1904  
Zhao S, Fu Y, Cui J N, *et al.* Spectrosc Spectr Anal, 2021, **41**(6): 1898-1904
- [11] Adhikari P, Oh Y, Panthee D R. Current status of early blight resistance in tomato: an update. Int J Mol Sci, 2017, **18**(10): 2019
- [12] Shi T, Liu Y, Zheng X, *et al.* Recent advances in plant disease severity assessment using convolutional neural networks. Sci Rep, 2023, **13**(1): 2336
- [13] 胡政, 张艳. 降维降噪处理对番茄早疫病潜育期高光谱识别效果的影响. 光谱学与光谱分析, 2023, **43**(3): 744-752  
Hu Z, Zhang Y. Spectrosc Spectr Anal, 2023, **43**(3): 744-752
- [14] Dabov K, Foi A, Katkovnik V, *et al.* Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans Image Process, 2007, **16**(8): 2080-2095
- [15] Wright J, Ganesh A, Rao S, *et al.* Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization//Bengio Y, Schuurmans D, Lafferty J, *et al.* Advances in Neural Information Processing Systems 22. Vancouver: NIPS, 2009. [https://papers.nips.cc/paper\\_files/paper/2009/hash/c45147dee729311ef5b5c3003946c48f-Abstract.html](https://papers.nips.cc/paper_files/paper/2009/hash/c45147dee729311ef5b5c3003946c48f-Abstract.html)
- [16] Li X, Feng J, Wang X. Hyperspectral imaging techniques for diagnosis and monitoring of potato diseases//Zhang C, Zhang X, Tani M. Infrared, Millimeter-Wave, and Terahertz Technologies V. Beijing: SPIE, 2018, **10826**: 326-332
- [17] 伍观娣, 赖敏婷, 汪迎利, 等. 近红外光谱分析技术应用于植物叶片研究综述. 林业与环境科学, 2020, **36**(1): 118-127  
Wu G D, Lai M T, Wang Y L, *et al.* For Environ Sci, 2020, **36**(1): 118-127
- [18] 陈永义, 俞小鼎, 高学浩, 等. 处理非线性分类和回归问题的一种新方法(I)——支持向量机方法简介. 应用气象学报, 2004, **15**(3): 345-354  
Chen Y Y, Yu X D, Gao X H, *et al.* Q J Appl Meteor, 2004, **15**(3): 345-354
- [19] Wang Y, Wang C, Dong F, *et al.* Integrated spectral and textural features of hyperspectral imaging for prediction and visualization of stearic acid content in lamb meat. Anal Methods, 2021, **13**(36): 4157-4168
- [20] 王海龙, 杨国国, 张瑜, 等. 竞争性自适应重加权算法和相关系数法提取特征波长检测番茄叶片真菌病害. 光谱学与光谱分析, 2017, **37**(7): 2115-2119  
Wang H L, Yang G G, Zhang Y, *et al.* Spectrosc Spectr Anal, 2017, **37**(7): 2115-2119
- [21] 吴迪, 吴洪喜, 蔡景波, 等. 基于无信息变量消除法和连续投影算法的可见-近红外光谱技术白虾种分类方法研究. 红外与毫米波学报, 2009, **28**(6): 423-427  
Wu D, Wu H X, Cai J B, *et al.* J Infrared Millim Waves, 2009, **28**(6): 423-427
- [22] 黄清, 薛河儒, 刘江平, 等. 基于蚁群-遗传算法的光谱选择方法与应用. 光谱学与光谱分析, 2022, **42**(7): 2262-2268  
Huang Q, Xue H R, Liu J P, *et al.* Spectrosc Spectr Anal, 2022, **42**(7): 2262-2268
- [23] 傅明华, 汪羞德, 顾仲兰, 等. 多功能转光塑料薄膜应用效应研究. 农业工程学报, 2000, **16**(6): 81-84  
Fu M H, Wang X D, Gu Z L, *et al.* Trans Chin Soc Agric Eng, 2000, **16**(6): 81-84
- [24] Brazaitytė A, Vaštakaitė V, Viršilė A, *et al.* Changes in mineral element content of microgreens cultivated under different lighting conditions in a greenhouse. Acta Hort, **2018**(1227): 507-516
- [25] 潘曦, 李冉, 魏敏, 等. 萤火虫优化支持向量机参数的近红外光谱技术鉴别卷烟牌号. 食品与机械, 2022, **38**(7): 85-90, 98  
Pan X, Li R, Wei M, *et al.* Food Mach, 2022, **38**(7): 85-90, 98
- [26] 高虹雷, 门昌骞, 王文剑. 多核贝叶斯优化的模型决策树算法. 国防科技大学学报, 2022, **44**(3): 67-76  
Gao H L, Men C Q, Wang W J. J Natl Univ Def Technol, 2022, **44**(3): 67-76
- [27] Hemeida A M, Bakry O M, Mohamed A A A, *et al.* Genetic algorithms and satin bowerbird optimization for optimal allocation of distributed generators in radial system. Appl Soft Comput, 2021, **111**: 107727
- [28] Huang W, Liu H, Zhang Y, *et al.* Railway dangerous goods transportation system risk identification: comparisons among SVM, PSO-SVM, GA-SVM and GS-SVM. Appl Soft Comput, 2021, **109**: 107541

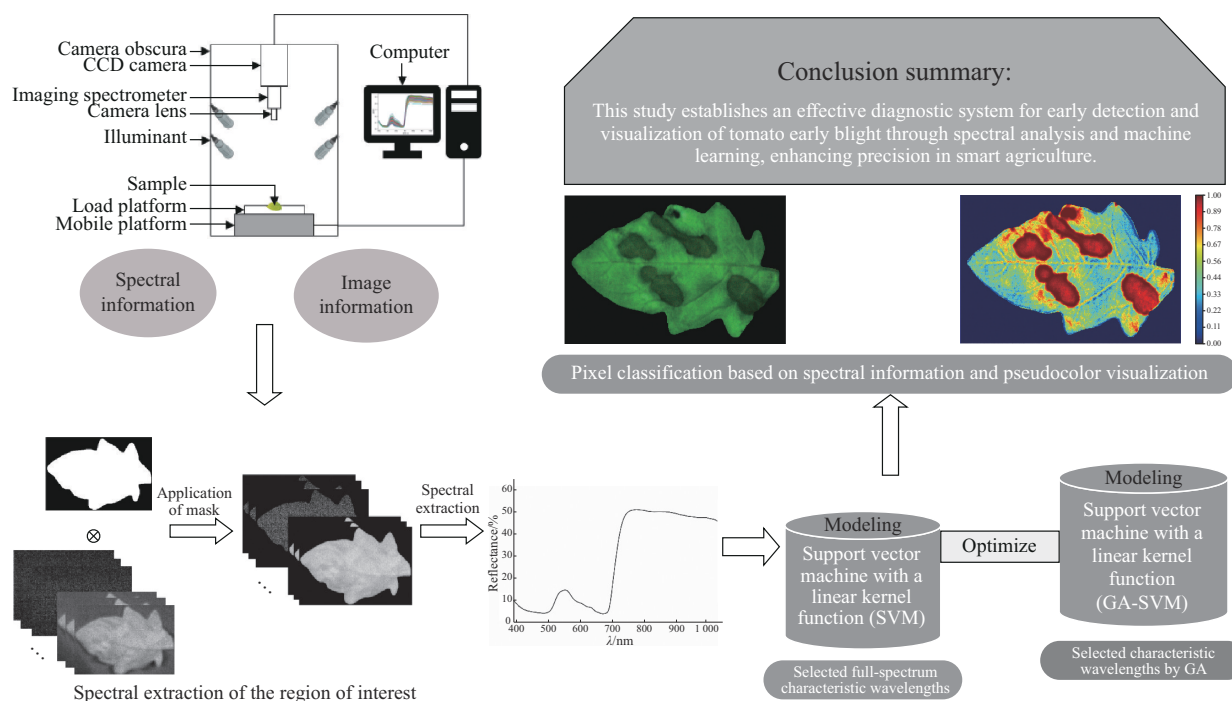
## Early Identification and Visualization of Tomato Early Blight Using Hyperspectral Imagery\*

BAO Hao<sup>1,2)</sup>, HUANG Li<sup>1)</sup>, ZHANG Yan<sup>1,2)\*\*</sup>, PANG Hao<sup>1,2)</sup>

(<sup>1)</sup>Engineering Research Centre for Non-Destructive Testing of Agricultural Products, College of Computer Science, Guiyang University, Guiyang 550005, China;

<sup>2)</sup>School of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

### Graphical abstract



**Abstract Objective** Tomatoes are one of the highest-yielding and most widely cultivated economic crops globally, playing a crucial role in agricultural production and providing significant economic benefits to farmers and related industries. However, early blight in tomatoes is known for its rapid infection, widespread transmission, and severe destructiveness, which significantly impacts both the yield and quality of tomatoes, leading to substantial economic losses for farmers. Therefore, accurately identifying early symptoms of tomato early blight is essential for the scientific prevention and control of this disease. Additionally, visualizing affected areas can provide precise guidance for farmers, effectively reducing economic losses. This study combines hyperspectral imaging technology with machine learning algorithms to develop a model for the early

\* This work was supported by grants from The National Natural Science Foundation of China (62265003, 62141501).

\*\* Corresponding author.

Tel: 86-851-85402217, E-mail: Eileen\_zy001@sohu.com

Received: May 14, 2024 Accepted: September 29, 2024

identification of symptoms of tomato early blight, facilitating early detection of the disease and visual localization of affected areas. **Methods** To address noise interference present in hyperspectral images, robust principal component analysis (RPCA) is employed for effective denoising, enhancing the accuracy of subsequent analyses. To avoid insufficient information representation caused by the subjective selection of regions of interest, the Otsu's thresholding method is utilized to extract tomato leaves effectively from the background, with the average spectrum of the entire leaf taken as the primary object of study. Furthermore, a comprehensive spectral preprocessing workflow is established by integrating multivariate scatter correction (MSC) and standardization methods, ensuring the reliability and effectiveness of the data. Based on the processed spectral data, a discriminant model utilizing a linear kernel function support vector machine (SVM) is constructed, focusing on characteristic wavelengths to improve the model's discriminative capability. **Results** Compared to full-spectrum modeling, this approach results in an 8.33% increase in accuracy on the test set. After optimizing the parameters of the SVM model, when  $C=1.64$ , the accuracies of the training set and test set reach 91.67% and 94.44%, respectively, demonstrating a 1.19% increase in training set accuracy compared to the unoptimized model, while maintaining the same accuracy on the test set, effectively alleviating issues of underfitting. **Conclusion** This study successfully establishes an early discriminant model for tomato early blight using hyperspectral imaging and achieves visualization of early symptoms. Experimental results indicate that the SVM discriminant model based on characteristic wavelengths and a linear kernel function can effectively identify early symptoms of tomato early blight. Visualization of these symptoms in terms of disease probability allows for a more intuitive detection of early diseases and timely implementation of corresponding control measures. This visual analysis not only enhances the efficiency of disease identification but also provides farmers with more straightforward and practical information, aiding them in formulating more reasonable prevention strategies. These research findings provide valuable references for the early identification and visualization of plant diseases, holding significant practical implications for monitoring, identifying, and scientifically preventing crop diseases. Future research could further explore how to apply this model to disease detection in other crops and how to integrate IoT technology to create intelligent disease monitoring systems, enhancing the scientific and efficient management of crops.

**Key words** hyperspectral imaging, machine learning, genetic algorithm, tomato early blight, early disease, visual analysis

**DOI:** DOI: 10.16476/j.pibb.2024.0207

**CSTR:** 32369.14.pibb.20240207