doi:10.6041/j.issn.1000-1298.2022.06.031

基于高光谱和集成学习的库尔勒香梨黑斑病潜育期诊断

刘媛媛[1]　张　凡[2]　师　琪[2]　马倩云[2]　王文秀[1,2]　孙剑锋[2]

(1.塔里木大学现代农业工程重点实验室,阿拉尔 843300;2.河北农业大学食品科技学院,保定 071001)

摘要:黑斑病是危害库尔勒香梨的真菌病害之一。若在黑斑病症状显证之前实现早期诊断,对于防止病害蔓延、减少经济损失具有重要的意义。结合高光谱成像技术和 Stacking 集成学习算法,构建了香梨黑斑病早期快速诊断模型。获取了健康、潜育期、轻度发病和重度发病的黑斑病库尔勒香梨的高光谱图像,提取感兴趣区域内的平均光谱,经标准正态变量变换、一阶导数、二阶导数及组合预处理后,利用主成分分析进行数据降维。然后,以 K 最近邻法(KNN)、最小二乘支持向量机(LS-SVM)和随机森林(RF)算法为基学习器,以 LS-SVM 为元学习器,构建了黑斑病病害程度的 Stacking 集成学习预测模型。结果表明,随着病害程度加深,光谱反射率整体呈下降趋势,且存在显著性差异,为分类模型的建立提供了理论依据。所建模型对健康和不同病害程度黑斑病库尔勒香梨的总体判别准确率为 98.28%,对潜育期香梨的判别准确率为 100%。与利用单一分类器建模结果相比,总体判别准确率和潜育期香梨判别准确率分别上升 5.18、23.08 个百分点。结果证明,Stacking 集成学习具有较强的特征学习能力,将其与高光谱成像技术结合,能实现库尔勒香梨黑斑病潜育期的识别。该结果为库尔勒香梨黑斑病的早期快速诊断和发病过程的实时监测提供了一种新的方法。

关键词:库尔勒香梨;黑斑病;潜育期;高光谱成像;Stacking 集成学习;早期诊断

中图分类号:S379.9　　**文献标识码**:A　　**文章编号**:1000-1298(2022)06-0295-09　　**OSID**:

Diagnosis of Korla Pear Black Spot Disease in Incubation Period Based on Hyperspectral Imaging and Ensemble Learning Algorithm

LIU Yuanyuan[1]　ZHANG Fan[2]　SHI Qi[2]　MA Qianyun[2]　WANG Wenxiu[1,2]　SUN Jianfeng[2]

(1. *Agricultural Engineering Key Laboratory*, *Tarim University*, *Alar* 843300, *China*

2. *College of Food Science and Technology*, *Hebei Agricultural University*, *Baoding* 071001, *China*)

Abstract: Black spot is one of the fungal diseases of Korla pear. It is of great significance to realize early diagnosis of black spot disease before the symptoms are evident, as it can prevent the spread of the disease and reduce the economic loss. Hyperspectral imaging technology was combined with Stacking ensemble learning algorithm to construct early and rapid diagnosis model of Korla pear black spot. Hyperspectral images of healthy, incubation period, mildly diseased and severely diseased Korla pear were obtained, and the average spectra in the region of interest were extracted. After pretreated by standard normal variable transformation, the first derivative, second derivative and their combinations, principal component analysis was implemented to reduce the data dimension. Then, the Stacking ensemble learning prediction model for black spot disease was constructed with K-nearest neighbor method (KNN), least squares-support vector machine (LS-SVM) and random forest (RF) algorithm as the base learner and LS-SVM as the meta-learner. The results showed that with the deepening of the disease degree, the reflectance spectra showed a downward trend, significant difference was observed, which provided a theoretical basis for the establishment of classification models. The total classification accuracy of healthy and different disease degrees of Korla pear was 98.28%, and the classification accuracy for incubation period pear was 100%. Compared with the results using single classifier, the classification

收稿日期:2021-11-23　修回日期:2022-03-24

基金项目:塔里木大学现代农业工程重点实验室开放项目(TDNG2020102)、河北省重点研发计划项目(20327111D)、河北省省属高等学校基本科研业务费研究项目(KY202002)和国家自然科学基金项目(31960498)

作者简介:刘媛媛(1982—),女,副教授,主要从事农产品品质安全快速无损检测技术及装备研究,E-mail:lyylvjin@163.com

通信作者:王文秀(1989—),女,副教授,主要从事食品无损检测技术与装置研究,E-mail:Godlovexiu@163.com

accuracy for all pear and incubation period pear was increased by 5.18 and 23.08 percentage points, respectively. The results showed that Stacking ensemble learning had strong feature learning ability, and its combination with hyperspectral imaging technology can realize the recognition of incubation period of black spot in Korla pear. The results can provide a method for the early diagnosis and real-time monitoring of black spot of Korla pear.

Key words: Korla pear; black spot disease; incubation period; hyperspectral imaging; Stacking ensemble learning; early diagnosis

0 引言

库尔勒香梨因营养丰富,香气馥郁、皮薄肉脆,果肉丰富多汁而深受消费者喜爱[1-2]。然而香梨易受病原菌的侵染,造成果实品质下降,腐烂率极高。黑斑病是梨的三大病害之一,是一种由链格孢属真菌 *Alternaria alternata* 引起的病害[3]。实现黑斑病胁迫下库尔勒香梨的快速诊断与检测,尤其是潜育期的早期识别具有重要意义。

目前库尔勒香梨黑斑病检测尚停留在抽样基础上的主观经验判断、分子生物学评价等阶段。主观经验法通过目测对有明显病症的样品进行识别,方便快捷,但对病害潜育期及症状不明显的样品辨识能力有限;分子生物学方法灵敏准确,但存在检测效率低、结果滞后、前处理繁琐、分析过程复杂等弊端,不适合于快速实时检测[4]。高光谱成像技术融合了图像和光谱信息,具有分辨率高、波数较多、“图谱合一”的特点,可同时获取被检测对象各波长的图像信息和各像元的光谱信息。目前,利用高光谱技术检测水果病害感染已有较多先例[5-10]。现有研究证明高光谱技术是水果病害等级区分的有效手段,但大多针对水果病害的显性病症进行识别,对更深层面的潜育期早期诊断仍需要深入研究。

化学计量学方法是高光谱定性判断水果病害程度分析中的重要组成部分,常用的分类器算法包括 K 最近邻法(K-nearest neighbors, KNN)、最小二乘支持向量机(Least squares - support vector machine, LS-SVM)、随机森林(Random forest, RF)、线性判别分析(Linear discriminant analysis, LDA)等[11]。单一分类器往往具有泛化性能不佳的缺点,有学者提出将多个弱分类器集成为强分类器以完成高精度分类任务的集成学习方法[12]。其中,Stacking 集成学习策略是一种异构分类器集合的技术,能增强模型的准确性、鲁棒性和整体归纳能力,在夏玉米覆盖度估测[13]、黑枸杞快速分级[14]等研究上表现出较好的预测能力。鉴于 Stacking 集成学习策略较强的特征学习能力,将其与高光谱成像技术结合,有望实现潜育期病害样品的正确识别。

综合上述分析,本文基于高光谱成像特征,结合 Stacking 集成学习框架,解析潜育期不同病害程度黑斑病库尔勒香梨的图谱特征,提取表征样品性质的特征光谱信息,构建香梨黑斑病早期快速诊断模型,以期为香梨黑斑病潜育期快速诊断和发病过程实时监测提供一种新的方法。

1 材料与方法

1.1 实验样品制备

实验所用梨品种为库尔勒香梨,购于河北省保定市农贸市场。挑选 234 个成熟度、尺寸、质量相近且无明显缺陷、物理损伤、疾病感染的新鲜香梨,立即运输至河北农业大学食品科技学院实验室,用自来水冲洗后,再用 75% 酒精擦拭果实表面,最后用蒸馏水冲洗香梨 3 次,自然风干后备用。

选择自然发病的黑斑病香梨果实,切取染病组织后将其移入到马铃薯葡萄糖琼脂(PDA)培养基上,放置到 25℃恒温培养箱培养 7 d。然后,用无菌接种环挑取培养基中茁壮的菌落,接种到新的 PDA 培养基上继续培养 3 d,向其中加入 10 mL 含 0.5% Tween-80 的 0.85% 无菌生理盐水,并用平板涂布器反复刮擦 PDA 培养基上的病菌菌落,使病原菌的孢子脱落。随后,将制备好的菌悬液移入无菌三角瓶中,振荡 15 s,再用 4 层纱布进行过滤,加入适量的无菌生理盐水使孢子悬浮液菌体浓度调节至 1×10^6 个/mL 备用。

制备黑斑病香梨样品时,先用 75% 酒精擦拭健康香梨果实表面,再用无菌水淋洗 3 次后放入超净工作台中进行紫外线灭菌,待果皮表面干燥后,在梨的赤道线附近使用注射器注射 20 μL 孢子悬浮液,另注射 20 μL 无菌生理盐水作为对照样品。将接种好的样品放入恒温恒湿培养箱中,保证环境相对湿度为 80%,温度为(25 ± 2)℃。为获得不同病害程度的香梨样品,每天接种 17 个样品,整个实验持续 10 d,最终获得 170 个接菌的香梨样品和 64 个健康样品用于后续分析。接种 3 d 后,大部分果实表面开始出现显现的病斑,说明库尔勒香梨经历 3 d 的潜育期后开始发病。培养过程中,采用十字交叉法记录每个果实的病斑直径,参照文献[9]以病斑直径 40 mm 为重度发病的阈值,结合库尔勒香梨平均

表面积，制定表 1 的分级标准，将样品分为健康、潜育期、轻度发病、重度发病。

表 1　库尔勒香梨的分类标准

Tab. 1　Classification standards of Korla pear

类别	分级标准	数量/个
健康	未接种，无病斑	64
潜育期	接种，无病斑	54
轻度发病	病斑面积百分比小于等于 15%	83
重度发病	病斑面积百分比大于 15%	33

1.2　高光谱成像系统及图像采集

本研究采用的高光谱成像系统如图 1 所示，主要包括 SPECIM FX 10 型高光谱相机（Specim 公司，芬兰）、光源、镜头、位移平台、步进电机和计算机等。光谱范围为 400 ~ 1 000 nm，光谱分辨率为 5.5 nm，像素尺寸为 8 μm × 8 μm，焦距为 15 mm，光源为一对 12 V、100 W 的卤素灯。

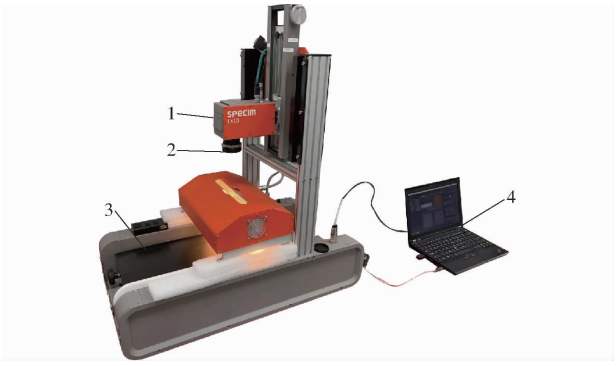


图 1　高光谱成像系统

Fig. 1　Hyperspectral imaging system

1. 相机　2. 镜头　3. 位移平台　4. 计算机

经过前期实验，将位移平台速度、相机曝光时间和采集距离分别设置为 7.5 mm/s、50 ms 和 300 mm。采集样品时，首先将仪器预热 30 min 以确保实验的稳定性，然后将 6 个样品按照 2 行 × 3 列病斑朝上的方式放置在位移平台上，高光谱相机连续线扫描，完成一幅高光谱图像的采集，在 ENVI 5.2 中进行批处理图像切割，得到单个样品的图像。为了减弱相机暗电流和光强变化对图像的影响，对原始高光谱图像进行黑白校正，即

$$I = \frac{I_0 - B}{W - B}$$

式中　I——校正后的高光谱图像

I_0——采集的原始高光谱图像

W——扫描标准白板得到的图像

B——旋上相机镜头盖采集得到的黑色图像

1.3　光谱信息提取

利用高光谱系统，得到 234 个样品的光谱图像立方体，每个像素点有 224 个波长处光谱信息。在图像中选择感兴趣区域（Region of interest，ROI）进行光谱信息的提取，ROI 的选择对后续模型的性能具有重要的影响。如图 2（图中 λ 表示波长）所示，以香梨损伤接种点为中心，横向像素个数设置为 150，选择方形感兴趣区域，通过 ENVI 5.2 获取 ROI 内 22 500 个像素点的平均反射光谱，即可得到样品的反射光谱曲线。由于相机有低和高光输出以及量子效率的存在，在整个光谱范围内，靠近两端光谱噪声比较大，因此最终选择的波长范围为 444 ~ 1 000 nm。依次提取所有样品的平均光谱，由此得到 234 × 206（样品数 × 变量数）的光谱矩阵用于后续分析。

1.4　光谱预处理

所提取的光谱中包括一些与待测样品性质无关的因素带来的干扰，如光的散射、杂散光、仪器噪声等影响，导致光谱产生基线漂移和光谱的不重复[15]，因此需对提取的光谱进行预处理。对比研究采用标准正态变量变换（Standard normal variate transformation，SNV）、一阶导数（First derivative，FD）、二阶导数（Second derivative，SD）及组合使用（SNV + FD、SNV + SD）对光谱及建模结果的影响。其中，SNV 可用来校正样品间因散射引起的光谱误差，微分处理能消除基线漂移、强化谱带特征、克服谱峰重叠等。

1.5　光谱数据降维

由于获取的光谱存在冗余和多重共线性问题，因此采用主成分分析（Principal component analysis，

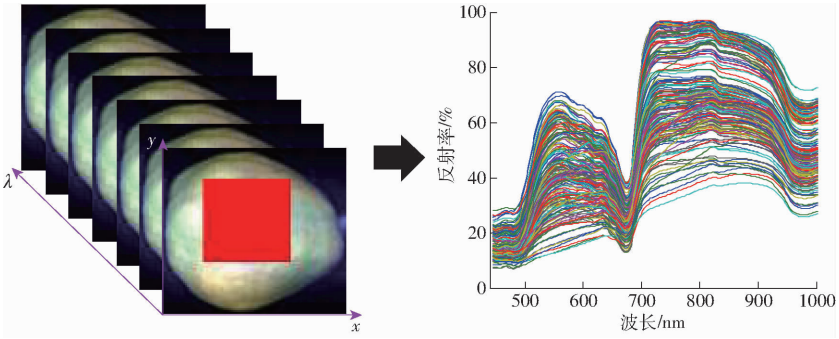


图 2　感兴趣区域选取示意图

Fig. 2　Schematic of ROI selection from hyperspectral image

PCA)对光谱进行降维。首先对光谱数据矩阵进行奇异值分解,得到一系列正交的变量(得分矩阵、对角矩阵和载荷矩阵),然后选取大特征值对应的特征得分矩阵替代原始光谱数据矩阵用于后续建模分析。主成分数的确定采用累计贡献率法,本研究取主成分的累计贡献率达到 99% 以上时主成分的数目。

1.6 建模结果分析

1.6.1 基于单一分类器的建模分析

首先,利用 Kennard - Stone 算法将 234 个健康和不同病害程度的库尔勒香梨按照比例 3∶1分为校正集 C(176 个样品)和验证集 P(58 个样品),其中校正集包括 48 个健康、41 个潜育期、62 个轻度发病和 25 个重度发病的香梨,验证集包括 16 个健康、13 个潜育期、21 个轻度发病和 8 个重度发病的香梨。然后,利用 KNN、LS - SVM 和 RF 算法,分别建立基于单一分类器的病害诊断模型。KNN 是以同类样品在模式空间内较为靠近为依据的分类方法,对待分析的样品,逐一计算其与校正集样品间的距离,找出最近的 K 个样品进行判别,本研究中 K 值设置为 4。LS - SVM 是适用于小样本的非线性机器学习算法,是支持向量机的一种扩展。RF 是一种包含多个决策树的分类器,对于待分析样品,由每一棵决策树进行分类,最后取所有结果中最多的分类。本研究决策树数量为 500,分裂属性集中的属性数量为 12。

1.6.2 基于 Stacking 集成学习的建模分析

Stacking 是一种通过元模型将多个单一模型进行组合的集成策略,其学习框架由两级分类器组成,其中低级别的分类器称为基学习器,而高级别的称为元学习器。基于 Stacking 集成学习的建模分析过程如图 3 所示,包括基于基学习器的基模型建立和基于元学习器的元模型建立两部分。具体步骤包括:

(1)利用上述基于单一分类器建模分析中的校正集 C 和验证集 P,按照五折交叉验证的方式先将 176 个校正集样品随机分为 5 个子集(C_1、C_2、C_3、C_4、C_5)。

(2)选择某一基学习器,以 C_2、C_3、C_4、C_5中样品为校正子集建立模型,对 C_1和 P 中样品进行预测,记为 x_1和 y_1。

(3)重复上述步骤,以 C_1、C_3、C_4、C_5中样品为校正子集,对 C_2和 P 中样品进行预测,记为 x_2和 y_2。以此类推,得到校正集样品 C 的预测值 X_1(由 x_1、x_2、x_3、x_4、x_5进行列合并所得)和验证集样品 P 的预测值 Y_1(由 y_1、y_2、y_3、y_4、y_5取众数所得)。

(4)选择其他基学习器,重复步骤(2)、(3),分别得到 X_2、X_3、Y_2、Y_3等。

(5)将第 1 层得到的结果合并,得到 $X=\{X_1, X_2, X_3\}$和 $Y=\{Y_1, Y_2, Y_3\}$。分别以 X 和 Y 作为校正集 C 和验证集 P 样品的新输入特征,以样品的实际类别 C_{label}和 P_{label}为输出特征,采用元学习器构建判别模型生成最终结果,实现对基模型能力的综合和特征的堆叠。

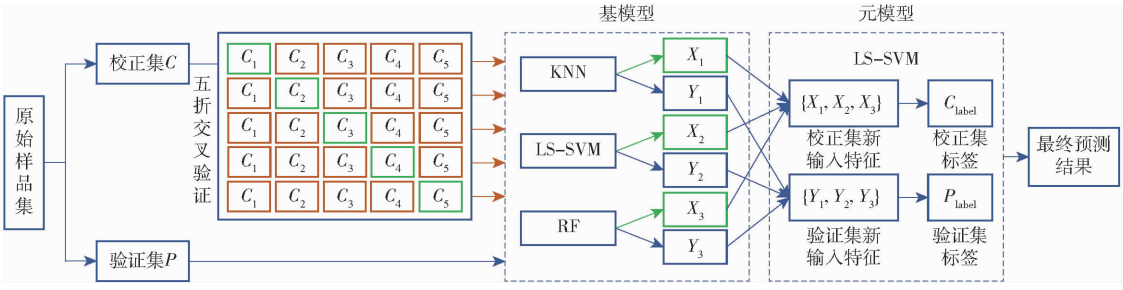


图 3 Stacking 集成学习模型

Fig. 3 Stacking ensemble learning model

上述分析过程中,基分类器和元分类器的设计至关重要。选择基分类器时应充分考虑学习器的充分性和多样性,使建立的基模型之间相互独立,实现模型间信息的有效互补。本研究以 KNN、LS - SVM 和 RF 算法为基分类器,以建模效果最好的基分类器作为元分类器进行第 2 层模型的训练。

1.7 模型评价指标

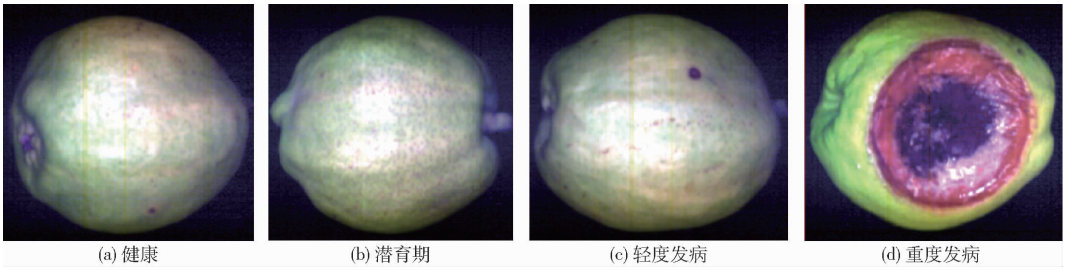
采用总体判别准确率和各类判别准确率对模型性能进行评价,分析过程在 ENVI 5.2 和 Matlab 2012b 中完成。

2 结果与讨论

2.1 黑斑病库尔勒香梨分析

健康和链格孢菌侵染不同时间的库尔勒香梨及病斑面积如图 4 和表 2 所示。对比图 4a、4b,在潜育期内,链格孢菌在香梨寄主体内吸收营养和水分并不断繁殖和扩展,形成局部侵染,使寄主部位新陈代谢发生改变,对寄主内部细胞及色素含量、水分和细胞间隙产生影响,但果实外表并未出现明显的可见症状。随着侵染时间的延长,链格孢菌从香梨果

实组织中获取生长所需的营养物质，致使梨果实的结构和生理机能发生变化，表面开始出现肉眼可见的黑褐色圆形小点（图 4c）。病原菌在香梨内部不断繁殖，产生溶解酶（纤维素酶、果胶酶等）和毒素（交链格孢酚等），使香梨果实细胞的超微结构（细胞膜、叶绿体等）发生改变，病斑区域逐渐扩大，发病样品的病斑面积范围为 0.19 ~ 36.30 cm^2，且稍有凹陷，部分病斑为同心轮纹斑（图 4d）。



(a) 健康　(b) 潜育期　(c) 轻度发病　(d) 重度发病

图 4　不同病害程度库尔勒香梨

Fig. 4　Korla pear with different disease degrees

表 2　不同病害程度库尔勒香梨病斑面积统计信息

Tab. 2　Statistical information of diseased spot area of Korla pear with different disease degrees

类别	病斑面积/cm^2	平均值/cm^2	标准偏差/cm^2	病斑面积百分比/%
轻度发病	0.19 ~ 12.56	4.14	3.72	0.23 ~ 14.79
重度发病	13.20 ~ 36.30	21.20	6.79	15.54 ~ 42.75

2.2　原始光谱信息及预处理

分别提取健康和不同病害程度黑斑病香梨高光谱图像 ROI 内的原始光谱，如图 5a 所示。可以发现，健康和潜育期样品光谱曲线较为相似，随着病害程度加深，光谱的反射率整体呈下降趋势。光谱反射率的降低表示梨果实中蔗糖、果糖、葡萄糖等成分含量在发生变化[10]。其中，波长 680 nm 处可见明显的吸收峰，与梨果皮和果肉中的叶绿素吸收有关；波长 740 nm 与 O—H 键伸缩的 3 级倍频有关；波长 980 nm 处可见明显的吸收峰，该吸收峰源于 O—H 键伸缩振动的 2 级倍频，与水分子结构密切相关[16]。潜育期香梨和健康香梨的光谱总体差异相对较小，与发病样品的差异较大。这可能是因为链格孢菌侵染初期，在表皮蜡质层和组织细胞壁的防御作用下，梨果实发生的变化相对较小。随着链格孢菌侵染时间的延长，病原菌含量增加，同时释放出寄主特异性毒素等代谢产物，香梨防御层遭到破坏，各种酶活性降低，梨果实中的果糖、葡萄糖、蔗糖、还原糖、总糖、总酸、水溶性蛋白质等化学成分减少。在病原菌和香梨自身变化的综合影响下，光谱反射率整体降低，在波长 550、680、730、980 nm 处较为明

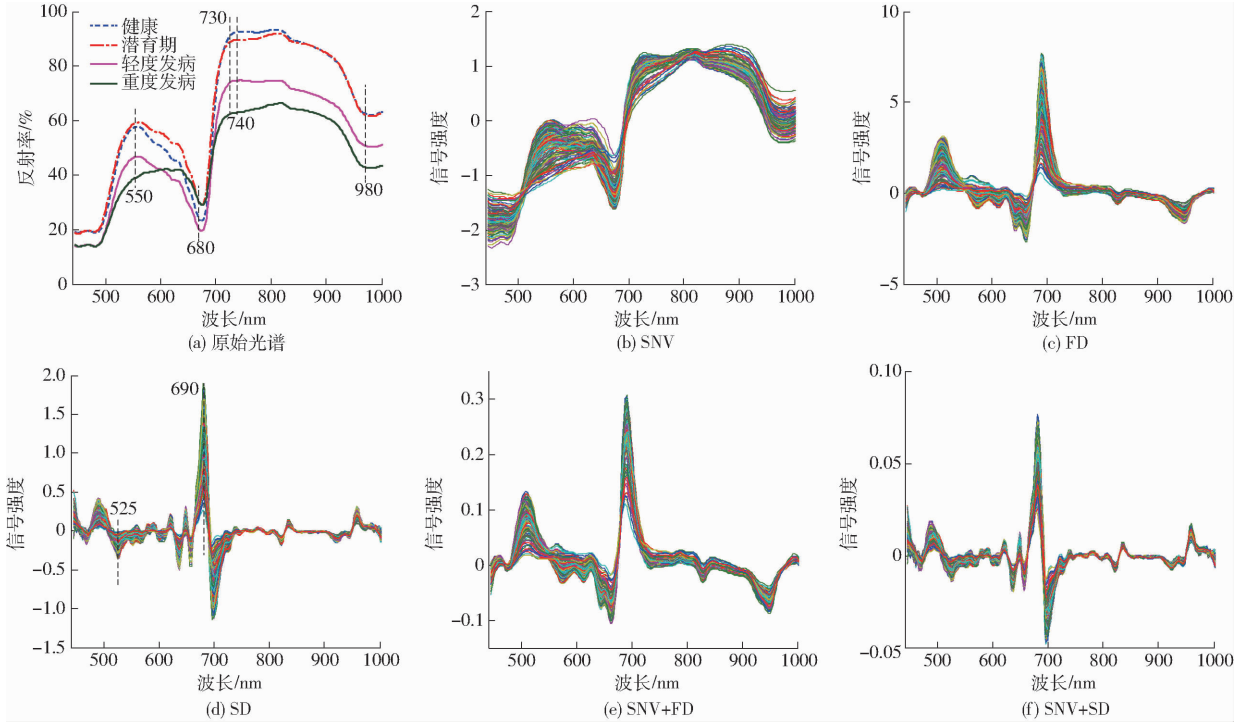


图 5　不同病害程度库尔勒香梨原始及预处理光谱

Fig. 5　Original and pretreated spectra of Korla pear with different disease degrees

显。对健康和不同病害程度黑斑病香梨光谱进行 Kruskal－Wallis 显著性差异分析，其渐近显著性 $P<0.0001$，表明不同等级样品的光谱信息存在差异，为后续分类模型的建立提供了理论依据。

经过 5 种不同方法预处理后的光谱如图 5b～5f 所示。SNV 预处理有效减小了样品间因散射引起的光谱误差（图 5b）；微分是消除基线漂移、强化谱带特征的常用方法，经 FD 处理可有效去除与波长无关的漂移（图 5c），波长 525、690 nm 处的吸收峰更加明显，而经 SD 处理能有效去除同波长线性相关的漂移（图 5d）。将 SNV 处理与微分处理结合，能在散射校正的同时有效提高谱图信息。如图 5e、5f 所示，SNV＋FD 和 SNV＋SD 处理后的光谱与 FD 和 SD 处理后的光谱总体趋势保持一致，数值上有较大差异。

2.3 光谱数据降维

由于光谱数据量较大，因此对其进行 PCA 处理，只取其中的主成分进行投影分析，在最大程度保留有效信息的基础上，通过减少特征维度来提高计算速度。所选取的各主成分之间彼此相互正交，从而使数据从高维空间向低维空间投影时能尽可能多地保留有用信息。以原始光谱为例，前 10 个主成分的贡献率和累计贡献率如图 6a 所示，前 3 个主成分的累计贡献率超过 85%，前 6 个主成分的累计贡献率超过 99%。为了尽可能多地利用光谱特征信息，选取前 6 个主成分用于后续分类建模。前 3 个主成分的载荷如图 6b 所示，载荷反映了各主成分在光谱范围内不同波长处的相关系数，波峰或波谷表示局部最大绝对权重系数。由图 6 可知，波长 480、530、550、650、680、720、745、805、925、970、980 nm 处具有较高的权重系数，其中 480 nm 与类胡萝卜素有关；波长 530 nm 和 550 nm 反映梨果实的绿色信息；波长 650～680 nm 由梨果表皮叶绿素吸收差异引起，反映了香梨从健康到发病后表皮颜色的变化；波长 720 nm 与 O—H 键 3 级倍频有关；波长 745 nm 与 CH_2 伸缩振动有关；波长 805 nm 与 N—H 键 3 级倍频有关；波长 925 nm 与 C—H 键 3 级倍频有关；波长 970 nm 反映了 C—H 键 2 级倍频的吸收，与果实中糖分的吸收密切相关[17-18]；波长 980 nm 则与果实中含水率密切相关。不同病害程度梨果实的颜色、水分、化学成分之间具有差异，所选主成分能够反映待测样品特征信息。以第 1 和第 2 主成分进行投影分析，健康和不同病害程度黑斑病梨果的空间分布如图 6c 所示。从图中可知，直接采用主成分分析投影，为 4 类样品的正确区分提供了有效信息，但健康和潜育期样品区分效果较差，且轻度发病和重度发病样品边缘区域的区分效果不好。因此，基于所选前 6 个主成分，进一步采用分类器算法进行不同病害程度梨果的区分和潜育期梨果的识别。

2.4 基于单一分类器的香梨分类结果

将健康、潜育期和不同病害程度黑斑病库尔勒香梨分别赋予类别标签 1、2、3、4，以所选主成分特征变量为输入，分别采用 KNN、LS－SVM 和 RF 算法建立分类模型，结果如表 3 所示。可以发现，基于预处理光谱的总体和各类判别准确率均优于原始光谱建模，表明采用合适的预处理方法能有效减少光谱采集过程中无关变量和噪声的干扰，对于有效信息的保留、模型稳健度的提高有重要的意义。对比不同方法预处理后的建模结果可知，导数处理后的建模结果优于 SNV 预处理后的建模结果，尤其是对提高潜育期样品的识别准确率较为明显，表明微分处理可以放大局部谱峰信息，分辨重叠峰，提高分辨率和灵敏度，对易混淆为健康和轻度发病的潜育期样品具有较好的识别效果。在 SNV 预处理的基础上增加 FD 或 SD 处理能有效提高模型的判别准确率，且 SNV＋SD 处理后的总体和各类判别准确率相对更高。

对比不同分类器算法的建模结果发现，LS－SVM 算法的分类效果较好，验证集和潜育期的判别准确率分别为 93.10% 和 76.92%。利用 KNN 和 RF 算法建立的模型，验证集总体判别准确率分别为 91.38% 和 89.83%。LS－SVM 算法致力于将香梨

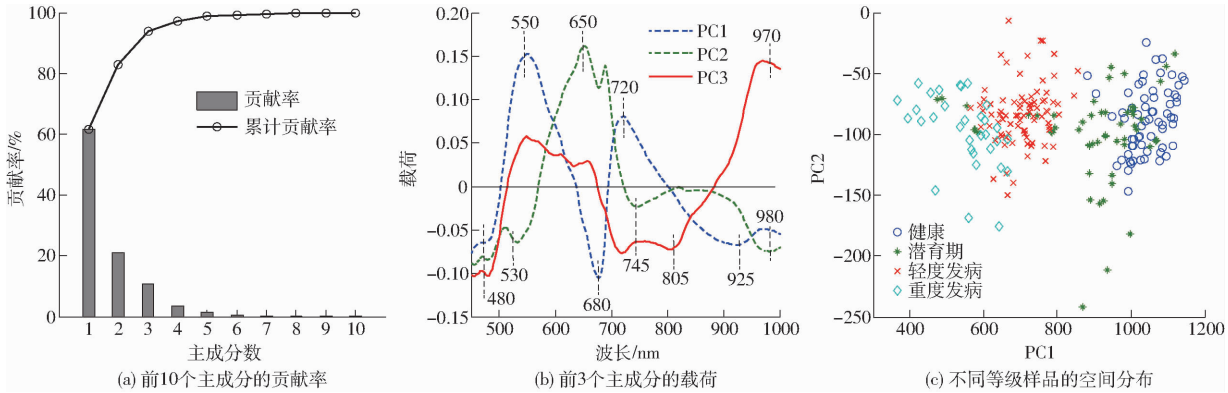


图 6 主成分分析结果

Fig. 6 Principal component analysis results

主成分光谱特征映射到高维特征空间中，通过寻找一个超平面将不同病害程度的香梨样品分开。模型建立过程中，通过对径向基核函数边界参数和核宽度的优化，较好地解决了本研究中的小样本、高维度的分类问题。KNN 算法通过计算待分析香梨与训练集香梨主成分特征之间的空间距离，由最近邻的 K 个样品“表决”确定其归属情况。RF 算法通过在香梨主成分特征空间的多次采样和变量空间的随机选择构建多个决策树模型，投票确定其所属的类别。后两种算法在对待测样品归类时，受各类别样品数量不一致的影响，容易使信息增益的结果偏向于具有更多样品数量的特征。而本研究中不同类别香梨样品的数量有所不同，从而使 KNN 算法和 RF 算法较 LS－SVM 算法稍差。

表 3　不同分类模型的总体判别准确率和各类判别准确率比较

Tab. 3　Comparison of classification accuracy for total samples and each category samples by using different classification models　%

预处理方式	建模方法	校正集					验证集				
		健康	潜育期	轻度发病	重度发病	总体	健康	潜育期	轻度发病	重度发病	总体
无	KNN	95.83	41.46	95.16	64.00	78.41	93.75	38.46	100	62.50	79.31
	LS－SVM	97.92	90.24	100	92.00	96.02	93.75	84.62	95.24	62.50	87.93
	RF	100	100	100	100	98.86	81.25	78.57	90.48	87.50	84.75
SNV	KNN	100	51.22	100	84.00	86.36	87.50	46.15	90.48	87.50	79.31
	LS－SVM	100	90.24	98.39	100	97.16	93.75	84.62	90.48	75.00	87.93
	RF	100	100	100	100	98.86	87.50	76.92	95.24	75.00	86.21
FD	KNN	100	82.93	100	88.00	94.32	100	61.54	90.48	100	87.93
	LS－SVM	100	92.68	98.39	92.00	96.59	100	76.92	95.24	75.00	89.66
	RF	100	100	100	100	98.86	100	92.86	80.95	87.50	89.83
SD	KNN	95.83	92.68	95.16	84.00	93.18	100	84.62	90.48	75.00	89.66
	LS－SVM	100	95.12	100	100	98.88	100	76.92	90.48	87.50	89.66
	RF	100	100	100	100	98.86	93.75	92.31	90.48	62.50	87.93
SNV＋FD	KNN	93.75	58.54	95.16	88.00	85.23	100	53.85	85.71	100	84.48
	LS－SVM	100	92.68	98.39	100	97.73	93.75	76.92	95.24	100	91.38
	RF	100	100	100	100	99.43	81.25	78.57	100	87.50	88.14
SNV＋SD	KNN	100	82.93	95.16	92.00	93.18	100	84.62	85.71	100	91.38
	LS－SVM	97.92	95.72	98.39	100	97.73	100	76.92	95.24	100	93.10
	RF	100	100	100	100	99.43	100	71.43	100	75.00	89.83

由表 3 可知，基于 SNV＋SD 预处理后的光谱数据，利用 LS－SVM 的建模结果最优。为进一步明确该模型对不同类别库尔勒香梨的识别效果，利用混淆矩阵对结果进行分析，如图 7 所示。在混淆矩阵中，用颜色代表样品数量，颜色越深，表示数量越大。对角线上数字表示被正确识别的样品数量，对角线以外的数字表示被错误识别的样品数量。如图 7a 所示，校正集 176 个样品有 4 个识别错误，其中 1 个健康香梨识别为潜育期，2 个潜育期香梨分别识别为健康和轻度发病，1 个轻度发病香梨识别为潜育期。如图 7b 所示，验证集 58 个样品中有 4 个识别错误，其中，3 个潜育期样品分别被识别为健康（1 个）和轻度发病（2 个），1 个轻度发病样品被识别为重度发病样品。这说明潜育期香梨和健康、轻度发病尤其是发病初期的样品光谱较为相似，判别时会相互影响。同时，未出现潜育期判别为重度发病的样品，表明此二类样品之间差别较为明显，更适合于选择最优信息点，建立分类模型时误差降低，从而提高分类准确度。总体而言，验证集中潜育期样品的判别准确率最低，仅为 76.92%，因此进一步采用 Stacking 集成学习的方法，提高模型的预测准确率。

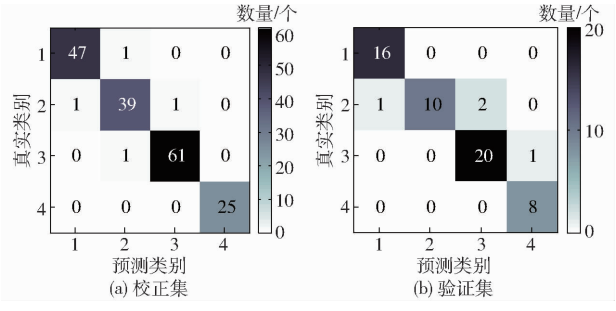


图 7　LS－SVM 算法的混淆矩阵

Fig. 7　Confusion matrix of LS－SVM model

2.5　基于 Stacking 集成学习的香梨分类结果

上述分析可知，利用 SNV＋SD 预处理光谱的建模结果整体上优于其他预处理方法，LS－SVM 算法的识别效果优于 KNN 和 RF。因此，基于 SNV＋SD

预处理光谱,以 KNN、LS－SVM 和 RF 作为第 1 层分类器算法,以 LS－SVM 作为第 2 层分类器算法,进行 Stacking 集成学习的模型构建,预测结果的混淆矩阵如图 8 所示。由图可知,Stacking 集成学习模型表现出良好的分类性能,校正集中各类别样品的判别准确率均为 100%,验证集中健康、潜育期和重度发病样品的判别准确率均为 100%,轻度发病样品中有 1 个判别为重度发病,准确率为 95.24%,总体准确率为 98.28%,比基于单一分类器建模的准确率高 5.18 个百分点。该判别错误与临界样品有关,处于轻度发病和重度发病阈值点附近的样品被错误分类。与单一分类器模型相比,Stacking 集成学习的多模型组合策略可以减少总误差,提高单个模型的性能,在潜育期样品的识别上具有显著的优势,判别准确率高 23.08 个百分点。因此,基于高光谱结合集成学习算法建立的库尔勒香梨黑斑病模型能较好地实现潜育期样品的正确识别。

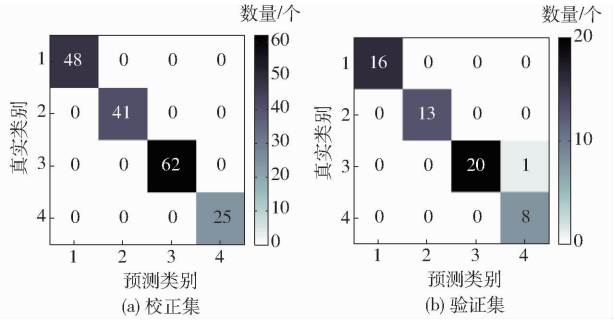


图 8 Stacking 集成学习模型的混淆矩阵

Fig. 8 Confusion matrix of Stacking ensemble learning model

本研究 Stacking 集成学习的分类器组合模型第 1 层选择了学习能力较强、差异度较大的 KNN、LS－SVM和 RF 作为基分类器,3 种算法的组合使用有助于优势互补地提升模型的预测效果。第 2 层学习器采用泛化能力较强、建模效果最佳的 LS－SVM 模型,用于纠正多个分类器算法对于校正集的偏置情况,并通过集合方式防止过拟合效应的出现[19]。元模型构建过程中采用五折交叉验证进行训练,缓解了过拟合风险。目前已有较多研究证明集成模型的性能优于单一模型[20－21],本研究的实验结果得到相同的结论,即与单一分类器模型相比,Stacking 集成学习模型能提高对不同病害程度黑斑病库尔勒香梨的识别效果,尤其是针对潜育期样品,模型的预测准确率有较大幅度的提升。这可能是因为,单一分类器在训练过程中极易陷入局部最优点,导致模型泛化性能不佳,而 Stacking 集成学习集合多个基分类器,能够在保持高可信度和稳定性的前提下兼顾一定的泛化能力,有效减少了陷入局部最优点的风险[22]。

3 结论

(1)基于高光谱成像技术和 Stacking 集成学习算法,构建了库尔勒香梨黑斑病的早期快速诊断模型,实现黑斑病潜育期和发病过程的实时检测。

(2)对比分析了健康、潜育期、轻度发病和重度发病的黑斑病库尔勒香梨的光谱曲线,健康和潜育期样品光谱较为相似,随着病害程度加深,光谱的反射率整体呈下降趋势。对不同病害程度黑斑病香梨光谱进行显著性差异分析,光谱信息存在显著性差异,为分类模型的建立提供了理论依据。

(3)以 KNN、LS－SVM 和 RF 为基学习器,以 LS－SVM为元学习器,构建了黑斑病病害程度的 Stacking 集成学习预测模型。该模型对健康和不同病害程度黑斑病库尔勒香梨的总体判别准确率为 98.28%,对潜育期香梨的判别准确率为 100%。

(4)Stacking 集成学习算法充分结合了多个学习器的能力,能实现模型间信息的有效互补,具有较强的特征学习能力。该方法集成多个基分类器,有效减少了陷入局部最优点的风险,结果相比单一分类器具有明显的提高。

参 考 文 献

[1] 张金阁,周婷,王鹏,等. 香梨脆度的力声同步检测[J]. 农业工程学报,2021,37(1):290－298.
ZHANG Jin'ge, ZHOU Ting, WANG Peng, et al. Simultaneous mechanical-acoustic measurement of the crispness of Korla pears [J]. Transactions of the CSAE, 2021, 37(1): 290－298. (in Chinese)

[2] SUN T R, OUYANG H, SUN P C, et al. Postharvest UV－C irradiation inhibits blackhead disease by inducing disease resistance and reducing mycotoxin production in 'Korla' fragrant pear (*Pyrus sinkiangensis*) [J]. International Journal of Food Microbiology, 2022, 362: 109485.

[3] LIU Y X, LI Y C, BI Y, et al. Induction of defense response against *Alternaria* rot in Zaosu pear fruit by exogenous L-lysine through regulating ROS metabolism and activating defense-related proteins[J]. Postharvest Biology and Technology, 2021, 179: 111567.

[4] 任美佳,张华平,钟聪慧,等. 库尔勒香梨黑斑病病原鉴定[J]. 石河子大学学报(自然科学版),2020,38(5):574－579.
REN Meijia, ZHANG Huaping, ZHONG Conghui, et al. Identification of the pathogens causing Korla pear black spot in Alea[J]. Journal of Shihezi University (Natural Science), 2020, 38(5): 574－579. (in Chinese)

[5] QIAO S C, TIAN Y W, WANG Q H, et al. Nondestructive detection of decayed blueberry based on information fusion of

hyperspectral imaging (HSI) and low-field nuclear magnetic resonance (LF - NMR)[J]. Computers and Electronics in Agriculture,2021,184:106100.

[6] FOLCH-FORTUNY A,PRATS-MONTALBAN J M,CUBERO S,et al. Vis/NIR hyperspectral imaging and N-way PLS - DA models for detection of decay lesions in citrus fruits[J]. Chemometrics and Intelligent Laboratory Systems,2016,156:241 - 248.

[7] LOPEZ-MAESTRESALASA A,KERESZTESB J C,GOODARZIB M,et al. Non-destructive detection of blackspot in potatoes by Vis - NIR and SWIR hyperspectral imaging[J]. Food Control,2016,70(6):229 - 241.

[8] JAROLMASJED S,CARLOS Z E,SANKARAN S. Near infrared spectroscopy to predict bitter pit development in different varieties of apples[J]. Journal of Food Measurement and Characterization,2017,11(3):987 - 993.

[9] SUN Y,XIAO H,TU S C,et al. Detecting decayed peach using a rotating hyperspectral imaging testbed[J]. LWT—Food Science and Technology,2018,87(1):326 - 332.

[10] PAN T T,ERKINBAEV C,SUN D W,et al. Pathogenetic process monitoring and early detection of pear black spot disease caused by *Alternaria alternata* using hyperspectral imaging[J]. Postharvest Biology and Technology,2019,154(2):96 - 104.

[11] YUN Y H,LI H D,DENG B C,et al. An overview of variable selection methods in multivariate analysis of near-infrared spectra[J]. Trends in Analytical Chemistry,2019, 113(11):102 - 115.

[12] DONG X,YU Z,CAO W,et al. A survey on ensemble learning[J]. Frontiers of Computer Science,2020,14(2):241 - 258.

[13] 张宏鸣,陈丽君,刘雯,等. 基于 Stacking 集成学习的夏玉米覆盖度估测模型研究[J]. 农业机械学报,2021,52(7):195 - 202.
ZHANG Hongming,CHEN Lijun,LIU Wen,et al. Estimation of summer corn fractional vegetation coverage based on Stacking ensemble learning[J]. Transactions of the Chinese Society for Agricultural Machinery,2021,52(7):195 - 202. (in Chinese)

[14] 卢伟,蔡苗苗,张强,等. 高光谱和集成学习的黑枸杞快速分级方法[J]. 光谱学与光谱分析,2021,41(7):2196 - 2204.
LU Wei,CAI Miaomiao,ZHANG Qiang,et al. Fast classification method of black goji berry (*Lycium ruthenicum* Murr.) based on hyperspectral and ensemble learning[J]. Spectroscopy and Spectral Analysis,2021,41(7):2196 - 2204. (in Chinese)

[15] ZHAO N,WU Z S,CHENG Y Q,et al. MDL and RMSEP assessment of spectral pretreatments by adding different noises in calibration/validation datasets[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,2016,163:20 - 27.

[16] LI B C,HOU B L,ZHANG D W,et al. Pears characteristics (soluble solids content and firmness prediction, varieties) testing methods based on visible-near infrared hyperspectral imaging[J]. Optik,2016,127(5):2624 - 2630.

[17] YUAN L M,MAO F,CHEN X J,et al. Non-invasive measurements of 'Yunhe' pears by Vis - NIRS technology coupled with deviation fusion modeling approach[J]. Postharvest Biology and Technology,2020,160:111067.

[18] FAN S X,WANG Q Y,TIAN X,et al. Non-destructive evaluation of soluble solids content of apples using a developed portable Vis/NIR device[J]. Biosystems Engineering,2020,193:138 - 148.

[19] 袁培森,杨承林,宋玉红,等. 基于 Stacking 集成学习的水稻表型组学实体分类研究[J]. 农业机械学报,2019,50(11):144 - 152.
YUAN Peisen,YANG Chenglin,SONG Yuhong,et al. Classification of rice phenomics entities based on Stacking ensemble learning[J]. Transactions of the Chinese Society for Agricultural Machinery,2019,50(11):144 - 152. (in Chinese)

[20] 冯建英,苏允汇,龚劭齐,等. 基于集成学习的农业生产技术效率评价方法[J]. 农业机械学报,2021,52(增刊):148 - 155.
FENG Jianying,SU Yunhui,GONG Shaoqi,et al. Evaluation method of agricultural production technical efficiency based on ensemble learning[J]. Transactions of the Chinese Society for Agricultural Machinery,2021,52(Supp.):148 - 155. (in Chinese)

[21] HU Z,QIU H,SU Z,et al. A Stacking ensemble model to predict daily number of hospital admissions for cardiovascular diseases[J]. IEEE Access,2020,8:138719 - 138729.

[22] JIA W,CHENG J,HU H. A cluster-stacking-based approach to forecasting seasonal chlorophyll-a concentration in coastal waters[J]. IEEE Access,2020,8:99934 - 99947.