DOI:10.13193/j.issn.1673-7717.2025.06.006

基于深度学习的拉曼/太赫兹光谱对藏红花的分类识别

苏康慧[1],王超琪[1],张谷令[1,2],邹斌[1,2],杨玉平[1,2]
(1. 中央民族大学理学院,北京 100081;2. 光子系统工程软件教育部工程研究中心,北京 100081)

摘要:目的 采用光谱学结合机器学习方法实现菊科红花属植物草红花、鸢尾科番红花属植物西红花(产地伊朗)和藏红花(产地西藏)的种间与种内鉴别。方法 采用拉曼光谱(Raman spectroscopy,RS)与太赫兹(Terahertz,THz)光谱,以伊朗产西红花、甘肃产草红花、西藏产西红花(下文称藏红花)为研究对象,结合主成分分析(Principal component analysis,PCA)、支持向量机(Support - vector machine,SVM)、随机森林(Random forest,RF)等机器学习方法和多层感知机(Multilayer perceptron,MLP)、卷积神经网络(Convolutional neural network,CNN)、长短期记忆递归神经网络(Long short term memory,LSTM)、径向基函数神经网络(Radial Basis Function,RBF)、基于遗传算法的反向传播神经网络(Genetic Algorithm - Back Propagation,GA - BP)等深度学习方法,分析三类样本的THz吸收光谱与RS,实现中草药小样本的快速分类识别。结果 (1)SVM方法对三类样本的RS数据和经PCA降维后的RS数据进行分类精度分别为83.33%和94.44%;MLP对三类样本的原始THz光谱数据和经PCA降维后的THz光谱数据进行分类,准确率分别为94.00%和96.79%;表明PCA对机器学习分类精度提升较为明显,THz光谱用于深度学习结果较RS精度更高。(2)MLP、CNN、RBF、GA - BP、LSTM等神经网络对THz光谱数据的分类准确率高于对RS分类准确率,其中LSTM和RF对太赫兹原始吸收数据分类准确率高达100%,分类精度高、速度快、模型稳定性好。结论 光谱与机器学习方法结合可以实现草红花、西红花和藏红花的种间与种内鉴别,THz光谱鉴别精度较高。

关键词:太赫兹时域光谱;拉曼光谱;深度学习;草红花;藏红花
中图分类号:R284.1;R2 - 03 **文献标志码**:A **文章编号**:1673-7717(2025)06-0036-09

Deep Learning - Based Raman/THz Spectroscopy for Classification and Identification of Zanghonghua(Crocus sativus L.)

SU Kanghui[1],WANG Chaoqi[1],ZHANG Guling[1,2],ZOU Bin[1,2],YANG Yuping[1,2]
(1. School of Science,Minzu University of China,Beijing 100081,China;
2. Engineering Research Center of Photonic Design Soft Wave,Ministry of Education,Beijing 100081,China)

Abstract:*Objective* To achieve the inter and intraspecific identification of Caohonghua(Carthamus tinctorius L.)from Asteraceae,Xihonghua(Crocus sativus L.)from Iridaceae(originating from Iran)and Zanghonghua(Crocus sativus L.)(originating from Xizang)by using spectroscopic technologies coupled with machine learning methods. *Methods* Raman spectroscopy(RS)and terahertz(THz)spectral data was used to identify and characterize Xihonghua(Crocus sativus L.)(origin Iran),Caohonghua(Carthamus tinctorius L.)(origin Gansu)and Zanghonghua(Crocus sativus L.)(origin Xizang)by using machine learning methods including Principal Component Analysis(PCA),Support - Vector Machine(SVM),Random Forest(RF)and deep learning methods including Multilayer Perceptron(MLP),Convolutional Neural Network(CNN),Long Short Term Memory(LSTM)and Genetic Algorithm - Back Propagation(GA - BP)to analyze the THz and RS spectra of three types of samples. The rapid classification and recognition of Chinese herbal medicine with small sample size was realized. *Results* (1)The classification accuracies of the SVM method on the RS data of the three samples before and after dimensionality reduction by PCA were 83.33% and 94.44%,respectively. The MLP classified the THz data before and after dimensionality reduction by PCA with accuracies of 94.00% and 96.79%,respectively. The results indicated that the classification accuracy was further improved by PCA,and THz spectra for deep learning had a higher accuracy than RS spectra. (2)The classification accuracy of neural networks such as MLP,CNN,RBF,GA - BP and LSTM on THz spectral data was higher than that on RS spectra,among which the classification accuracy of LSTM and RF on THz raw data was as high as 100%. Also,these models had high classification accuracy,high speed and good stability. *Conclusion* The combination of spectroscopic and machine learning methods enables inter and intraspecific identification of Xihonghua(Crocus sativus L.),Caohonghua(Carthamus tinctorius L.)and Zanghonghua(Crocus sativus L.),with high accuracy of THz spectroscopic identification.

Keywords:Terahertz time - domain spectroscopy;Raman spectroscopy;deep learning;Caohonghua(Carthamus tinctorius L.);Zanghonghua(Crocus sativus L.)

基金项目:国家重点基础研究发展计划项目(2017YFB00405402,2020YFB2009303);国家自然科学基金项目(62075248)
作者简介:苏康慧(1997 -),女(藏族),黑龙江七台河人,博士在读,研究方向:传统医药的THz光谱检测与鉴定。
通讯作者:杨玉平(1976 -),女,山东德州人,教授,博士研究生导师,博士,研究方向:THz光谱学。E-mail:ypyang@muc.edu.cn。
邹斌(1980 -),男,山东威海人,副教授,硕士研究生导师,博士,研究方向:凝聚态物理理论。E-mail:zoubin@muc.edu.cn。

西红花，是鸢尾科番红花（*Crocus sativus* L.）属多年生草本植物，原产地为伊朗、西班牙、希腊等国，于唐朝期间经西藏传入国内，因此又名“藏红花”。市售的西红花产地多样，其中伊朗产的西红花和西藏产的西红花都属于番红花属；草红花是菊科红花属植物，多产于新疆、甘肃等地，虽然草红花与西红花外形相似，但它们不属于同一物种，其入药部位、活性成分、功效和产量也不同。西红花以其干燥柱头部分入药，而草红花的入药部位是其管状的小花，草红花产量比西红花高得多。根据 2020 版《中华人民共和国药典》的描述，西红花具有活血化瘀、保护心脏、抗抑郁、抗糖尿病、抗闭经痛经等功效，草红花具有淡斑祛印的功效[1]。由于西红花产量较低（1 朵红花仅有 3 根柱头），经过干燥后重量更轻，因而 150 朵红花仅能产生 1 g 左右的成品药材红花，这也导致优质红花在中药材市场上供不应求[2]。为了从中牟取暴利，不法商家以次充好、掺假售卖，或者以假乱真，将草红花染色后进行售卖，这些行为导致市场上红花真假难辨[3-4]。现代的检测方法如显微鉴别、薄层色谱、质谱法步骤繁琐、费时费力，且需要专业人员操作，无法实现快速、简便检测的需求，因而亟须一种新的方法实现对红花的快速检测[5-11]。

光谱方法具有更高的准确性和灵敏度，拉曼光谱（Raman spectroscopy，RS）可以获得分子化学键（或官能团）振动与转动相关的信息，从而对化学成分不同的物质进行区分，有研究者将拉曼光谱法与其他光谱方法联用，可以实现藏红花的鉴别。余志引等[12]报道的西红花拉曼峰与密度泛函理论计算得到的西红花苷拉曼峰位基本一致，成功区分了伪品西红花（草红花）与真品西红花；VARLIKLIOZ ER S 等[13]用拉曼光谱联合激光诱导击穿光谱（Laser - induced breakdown spectroscopy，LIBS）设计出一种可以定量分析方法，可以用于西红花中掺杂万寿菊和姜黄的鉴别；DAI H C 等[14]将拉曼光谱与薄层色谱法（Thin - layer Chromatography，TLC）结合，在西红花掺伪鉴别中取得良好效果，验证了拉曼检测西红花的可行性。但是，对于分子结构相似的化合物，由于基团共振在红外波段引起的拉曼峰位几乎相同，因此在对其有效成分鉴别和种间（种内）评估方面产生了一定困难。此外，激发光源的高光子能量不可避免地会对样品产生一定的损伤，加上其测试的信噪比和稳定性不高，严重影响了分类精度。

太赫兹波频率范围介于 0.1 ~ 10 THz，具有穿透性、宽带性、安全性、瞬态性等特征，广泛应用于食品、药材等领域的无损检测。太赫兹时域光谱（Terahertz - time domain spectroscopy，THz - TDS）可以同时获取样品的振幅和相位信息，从而在化学分子和生物样品识别中具有较大潜力[15-18]。马品[18]等利用 THz - TDS 技术检测天麻等药物中的水分含量，并基于生物组织含水量差异实现生物组织健康情况的高精准判断[17]。近年来，为了进一步提高分类精度与检测效率，机器学习方法结合太赫兹光谱已被广泛用于药用植物和其中有效成分的分类鉴别，具有鉴别精度高、不损坏样本、快速高效、操作简便等优势。张活[19]用太赫兹光谱和机器学习方法相结合，成功地将青蒿素与吡哆素、苯甲酸、肉桂酸等几种化合物进行了区分，还实现了断肠草、忍冬草、五指毛桃、白英、寻骨风和龙葵草等六种外观相似中药材的分类，分类精度均在 94% 以上。徐振[20]结合 THz - TDS 和多种机器学习方法鉴别 4 种常见产地的贝母（川贝母、平贝母、伊贝母、浙贝母），其中，结合标准正态变换与 SVM 建立的多分类模型预测正确率高达 97.49%。2019 年，杨玉平等[21]采用 S - G 平滑（Savitzky - Golay，S - G）方法处理草红花和西红花的太赫兹吸收光谱，平滑后数据的 PCA 分类精度提高到 100%，较平滑前精度有较大改善；但草红花和西红花属于两个物种，目前的研究只实现了种间鉴别，并未提及不同产区西红花的种内鉴别。

本文综合利用太赫兹吸收光谱与拉曼光谱技术，并结合 PCA、SVM、LSTM、MLP、GA - BP 等分析方法，用于同时实现西红花、草红花（伪品）、藏红花的种内与种间鉴别，真伪鉴别与产地鉴别。基于 RS 和 THz 光谱数据用于深度学习的精度差异，高效地实现西藏西红花、甘肃草红花（西红花的伪品）和伊朗西红花的区分，为同时实现中药材的真伪区分与地域鉴别提供方法上的参考。

1 仪器与材料

1.1 样品制备 所需藏红花、草红花、西红花样本购置于同仁堂药业有限公司。为避免植物中水分对光谱的影响，同时模拟药用植物炮制过程，将样本和聚乙烯（Polyethylene，PE）均利用干燥箱在 40 ℃干燥 12 h，经玛瑙钵研磨后过 200 目筛，采用压片法将研磨均匀后的样本粉末和聚乙烯粉末按照 1∶3 的比例在 11 t 的压力下制成两面平行且光滑的圆片（总质量为 200 mg，每个样本 20 组）。压片的草红花平均厚度大约在 1.970 mm，藏红花和西红花的平均厚度约为 1.940 mm；压制同样重量的聚乙烯片作为对照，厚度为 1.556 mm。

测量草红花、西红花、藏红花粉末的 THz 光谱和拉曼光谱，每种样品各 20 个，共 20 × 3 = 60 个样本，每个样本测量 3 次取平均值。采集 3 组 PE 的 THz 透射光谱数据作为参考。

1.2 仪器 本实验室自主搭建的 THz - TDS 系统用于测试太赫兹光谱，其积分时间 300 ms，有效波段 0.2 ~ 3 THz。实验温度 19.1 ℃，测试环境充入氮气以去除水蒸气带来的影响。

法国 Jobin - Yvon 公司生产的 LabRam HR - 800 共聚焦激光拉曼光谱仪用于拉曼光谱测试。激发光源为 He - Ne 激光器，波长 632.8 nm，衍射光栅 600 gr/mm，针孔直径 100 μm；激光在样本表面的平均功率 10 mW，积分时间 5 s，积分次数 10，测量范围 200 ~ 2000 cm^{-1}，测试环境温度 20 ℃。

1.3 数据处理 测试纯 PE 和有待测物质的压片的太赫兹时域信号分别作为参考和样本信息，基于快速傅里叶变换，获取相应的频域振幅信息[参考 $A_R(\omega)$ 和样本 $A_S(\omega)$]，进而得到样本的吸光度，可表示为 A[22]，见公式（1）。

$$A = -\ln\left[\frac{A_S(\omega)}{A_R(\omega)}\right] \quad (1)$$

将 THz 吸收光谱数据与 S - G 滤波平滑并归一化后的样品粉末拉曼光谱数据作为做主成分分析的输入数据，获取降维后的光谱信息，并基于 SVM、CNN、LSTM、RF、RBF、MLP、GA - BP 机器学习方法，进而得到相应的分类结果，机器学习流程见插页Ⅳ图 1。

2 方法与结果

2.1 拉曼光谱分析 插页Ⅳ图 2（a）分别为 3 种样品在 900 ~ 1700 cm^{-1} 波段内的拉曼光谱，可以看到草红花没有特征峰，西红花和藏红花有明显的指纹特征且两者的峰位几乎相同，相对强弱和峰形稍有差异但不明显。其中 1537 cm^{-1} 左右的峰对应 C = C 的伸缩振动，1164 cm^{-1} 左右的峰对应 C—C 的伸缩振动，1209 cm^{-1} 左右的峰对应 C—C 和 C—H 的弯曲模式，1283 cm^{-1} 左右的峰对应 C—C 的伸缩振动和 C—H 面内弯曲振动，1019 cm^{-1} 左右的峰是附着在多烯链上的—CH_3 基团和 C—C 偶联的面内摇摆产生的，966 cm^{-1} 左右的峰对应 C—O—C 键[15,23-27]。为探究特征峰的物理来源，我们进一步测试了 3 种西红花有效物质，分别为西红花苷Ⅰ（Crocin I，分子式

$C_{44}H_{64}O_{24}$)、西红花苷Ⅱ(Crocin Ⅱ,分子式 $C_{38}H_{54}O_{19}$)和西红花苷Ⅲ(Crocin Ⅲ,分子式 $C_{32}H_{44}O_{14}$),如插页Ⅳ图 2(b)所示为 3 种有效物质的归一化拉曼光谱。西红花苷Ⅰ、Ⅱ、Ⅲ的拉曼光谱和西红花、藏红花的拉曼光谱相对应,但由于西红花苷Ⅰ、Ⅱ、Ⅲ的拉曼光几乎相同,难以确定西红花中具体是哪种物质在起作用,也无法据此实现西红花和藏红花的种内鉴别,需要采用聚类分析方法进一步分类。

2.2 拉曼光谱聚类分析 将 S-G 平滑后的拉曼光谱数据作 PCA 聚类分析,结果如插页Ⅳ图 3 所示,前两个主成分 PC1 和 PC2 的方差贡献率分别为 37.3% 和 19.3%,累计贡献率达 56.6%。从分散程度看,草红花(浅蓝色)的方差最小,表现出明显的聚类,能够与西红花和藏红花区分开;而西红花和藏红花的数据点较为离散,数据内部相似性一般,难以区分。

进一步采用 SVM 进行聚类分析。SVM 是典型的结构风险最小化模型,可以使间隔最大化,适用于小样本的精准分类。将草红花、西红花和藏红花的归一化拉曼光谱数据输入 Matlab R2019a,训练集 $12\times3=36$ 个,测试集 $6\times3=18$ 个,迭代次数 iter = 52,测试精度为 83.33%,如插页Ⅴ图 4(a)所示;进一步将 PCA 与 SVM 结合,通过交叉验证选择最优的主成分数量为 3,迭代次数 iter = 37 次,精度为 94.44%,PCA + SVM 比单用 SVM 精度提高了 11.11%,如插页Ⅴ图 4(b)所示。

2.3 太赫兹光谱分析 THz 时域波形和傅里叶变换光谱如插页Ⅴ图 5 所示。草红花、西红花和藏红花的时域、频域光谱均无明显差异,无法作为鉴别依据。

样品在 0.2 ~ 3.0 THz 的吸收曲线如插页Ⅵ图 6(a),曲线无明显吸收峰,在高频区,草红花和藏红花重叠部分较小,较易区分,西红花和藏红花重叠较多,难以区分,在低频区,三者几乎完全重叠。插页Ⅵ图 6(b)为插页Ⅵ图 6(a)平均后的结果,能更直观地展现 3 种样本吸收光谱的差别,草红花的平均吸收系数最大,西红花略大于藏红花。

2.4 太赫兹光谱的聚类分析 太赫兹吸收光谱数据的 PCA 聚类分析结果见插页Ⅵ图 7,主成分 PC1 和 PC2 的方差贡献率分别为 27.7% 和 13.1%。草红花的聚类效果较好,辨别准确率达 100%,与文献[21]中的结果一致,而同为鸢尾科红花属的西红花和藏红花重叠区域较大,依旧无法区分,这是由两者有效成分类似导致的,与拉曼光谱的 PCA 分类结果一致。

进一步,运用 SVM 聚类分析方法,选取 0.5 ~ 3.0 THz 范围内的光谱数据,其中训练集 $10\times3=30$ 组,测试集 $5\times3=15$ 组,迭代次数为 157 次,测试精度达 100%,如插页Ⅵ图 8 所示。使用 SVM 分类时,THz 作为输入数据的测试精度比 RS 数据作为输入数据时高 16.67%,比 PCA 降维后的 RS 数据作为输入数据时高 5.56%。

以上结果说明,THz 光谱用于深度学习时比拉曼光谱有更高的分类精度,这是由于 THz 时域光谱采用相干测试,信噪比更高,光谱中的冗余信息更少,且不易受测试环境干扰,在定量分析和聚类分析中具有更好的效果。

2.5 RS/THz-TDS 光谱数据的分类结果对比 为了进一步验证太赫兹光谱用于深度学习时的分类效果,将太赫兹原始吸收光谱数据、拉曼光谱数据作为输入数据,进行 RF 机器学习,决策树数目 100,最小叶子数 0.01,测试精度分别为 100%(THz + RF)和 94.44%(RS + RF),如插页Ⅵ图 9、插页Ⅶ图 10 所示,THz 的分类精度高于 RS。将 PCA 降维后的 RS 数据输入 RF 进行学习,得到的精度为 100%,分类效果得以改善,但操作步骤增加。可见,THz 光谱用于红花分类时可以省去 PCA 降维这一步骤,比拉曼光谱更快,准确率也更高。

类似地,将经 PCA 降维前后的 THz 吸收光谱和 RS 数据作为输入数据,进行 CNN 深度学习,THz 原始吸收数据的分类精度为 95%,RS 数据分类精度为 94.44%,经 PCA 降维后的 THz 分类精度为 100%,RS 为 94.44%。THz 数据降维后精度提高了 5%,RS 数据降维后分类精度未见提升。可见,PCA 降维前后,THz 作为输入数据时的分类精度都要优于 RS 的分类精度。见插页Ⅶ图 11、插页Ⅶ图 12。

同样地,分别将 PCA 降维前后的 THz 吸收数据和 RS 数据输入 Matlab 做 LSTM 分类,分类结果如插页Ⅷ图 13 所示,降维前后 THz 分类精度均为 100%,分类效果较好。RS 数据的训练集准确率只有 68.06%,测试集准确率仅为 63.89%,如插页Ⅷ图 14 所示;RS + PCA 的训练集分类准确率为 100%,测试集准确率为 100%,分类精度提升,但增加了处理步骤,使分类过程更复杂。

总结以上深度学习方法的分类效果列于表 1,其中包含 MLP、CNN、RBF、LSTM、RF、GA-BP 等神经网络分类方法的测试精度。原始 THz 吸收光谱的训练集和测试集的 MLP 分类精度分别为 98.41% 和 94%,经 PCA 降维后分类精度分别为 98.75% 和 96.79%,测试精度提高 2.79%。相比之下,PCA 降维前后的 RS 数据的 MLP 测试集分类精度分别为 91.67% 和 88.89%,均低于 THz 光谱数据的分类精度。类似的,经 PCA 降维前,THz 吸收光谱的 LSTM 分类精度为 100%,远高于 RS 数据的分类精度(63.89%);经 PCA 降维后,两种光谱分类精度相同,均为 100%,但拉曼光谱分类相较于 THz 光谱而言操作更繁琐。本文列出的几种机器学习方法,THz 吸收数据(未经 PCA 处理)的分类精度均高于 RS 数据,说明 THz 光谱数据联合一系列机器学习方法可以对药用植物这类混合物进行无损、快速鉴别,操作简便且灵敏度高,为中药这类复杂混合物的鉴别提供了参考。

表 1 THz-TDS 和 RS 深度学习分类精度结果

方法	THz-TDS		RS	
	训练集/(%)	测试集/(%)	训练集/(%)	测试集/(%)
RF	100.00	100.00	100.00	94.44
PCA + RF	100.00	90	100.00	100.00
MLP	98.41	94	98.61	91.67
PCA + MLP	98.75	96.79	100.00	88.89
CNN	100.00	95	100.00	94.44
PCA + CNN	100.00	100.00	100.00	94.44
LSTM	100.00	100.00	68.06	63.89
PCA + LSTM	100.00	100.00	100.00	100.00
RBF	100.00	93.75	97.25	89.99
PCA + RBF	100.00	95	100.00	89.81
GA-BP	98.61	95.56	95.63	84.84
PCA + GA-BP	99.06	96.25	97.22	94.44

由于神经网络具有复杂性高且可视性低的内部结构,深度学习往往比机器学习耗时更长;本研究采用小样本进行深度学习,避开了深度学习耗时长的短板。文中用到的几种机器学习方法中,GA-BP 耗时最长,准确率一般;SVM 最快,准确率较高、稳定性好;LSTM 准确率最高、稳定性最好。可见THz-TDS 对各种机器学习方式都很包容,在提高分类精度的同时,尽可能满足快速、简单、无损等要求。基于 THz 光谱和 LSTM 实现了草红花、西红花、藏红花的快速、无损鉴别,为中草药的无损鉴别提供了思路。

3 讨论

本研究对草红花、藏红花、西红花各 20 份样品在 THz 光谱和拉曼光谱测试的基础上，结合主成分分析和深度学习方法，全面、客观评价拉曼光谱数据和 THz 吸收光谱数据用于三类样品分类的分类效果。草红花的 THz 光谱和拉曼光谱数据的 PCA 均表现出较好的聚类，而西红花和藏红花 PCA 重叠区域大，可能由西红花和藏红花成分相似导致，这与拉曼光谱的测试结果一致。

进一步结合机器学习方法，将拉曼数据（训练集 36 组，测试集 18 组）和 THz 吸收光谱数据（训练集 30 组，测试集 15 组）以训练集∶测试集 = 2∶1 的比例输入 Matlab R2019a 中进行 SVM 分类，拉曼数据和原始 THz 吸收数据的分类精度分别为 83.33% 和 100%，THz 吸收光谱数据结合机器学习方法分类精度较高。

更进一步，使用更多机器学习方法进行测试，MLP 对原始 THz 吸收光谱数据和拉曼光谱数据的测试集分类精度分别为 94% 和 91.67%；RF 的精度分别为 100% 和 94.44%；CNN 的测试精度分别为 95% 和 94.44%；LSTM 精度分别为 100% 和 63.89%；RBF 为 93.75% 和 89.99%，GA - BP 分别为 95.56% 和 84.84%，整体看来，THz 吸收光谱数据用于机器学习的分类精度更高。

综上，本研究提出一种基于太赫兹时域光谱和深度学习实现小样本量的草红花、西红花、藏红花的快速、无损分类鉴别方法，不仅可以区分同属不同产地的西红花和藏红花，也可以区分西红花和其伪品——草红花，创新性地一次性实现真伪鉴别与产地区分，其中 THz 光谱的 RF、PCA + CNN、LSTM 分类精度可达 100%。相比拉曼光谱而言，THz 吸收光谱分类准确率更高、数据处理过程更简单、耗时更短，为不同产区的西红花药材的分类和真伪鉴别提供参考。

参考文献

[1] 国家药典委员会. 中华人民共和国药典：一部[S]. 北京：中国医药科技出版社，2020：153 - 1812.
[2] VIDA M, AKHTAR K, EBRAHIM A. Effect of different drying methods on saffron (*Crocus sativus* l) quality[J]. Iran J Chem & Chem Eng Int Engl Ed, 2009, 31: 85 - 89.
[3] 中国中药协会. 西红花质量等级：T/CATCM 002—2018[S]. 北京：中国标准出版社，2018.
[4] 周桂芬，留永咏，钱晓东，等. 基于中国药典与国际标准对西红花生产流通中质量评价方法的改良[J]. 药物分析杂志，2016，36(5)：835 - 841.
[5] UMBER F, ZAFAR M, ULLAH R, et al. Implication of light and scanning electron microscopy for pollen morphology of selected taxa of family Asteraceae and Brassicaceae[J]. Microsc Res Tech, 2022, 85(1): 373 - 384.
[6] MILOJKOVIĆ - OPSENICA D M, TRIFKOVIĆ J D, RISTIVOJEVIĆ P M, et al. Thin - layer chromatography in the authenticity testing of bee - products[J]. J Chromatogr B Analyt Technol Biomed Life Sci, 2022, 1188: 123068.
[7] REGUIGUI A, HEIL J, GORAI M, et al. Profile comparison and valorization of Tunisian *Salvia aegyptiaca* and S. verbenaca aerial part extracts *via* hyphenated high - performance thin - layer chromatography[J]. J Chromatogr A, 2022, 1673: 463057.
[8] GUPTA I, ADIN S N, AQIL M, et al. High - performance thin - layer chromatography method development and validation for quantification of piperine in different extracts of *Piper longum* L. and its antioxidant activity[J]. JPC, 2023, 36(2): 147 - 155.
[9] KHAIRNAR A, LOHIDASAN S, DUBEY R, et al. Analytical method development and validation for the simultaneous estimation of quercetin, berberine, rutin and curcumin in a polyherbal formulation using high - performance thin - layer chromatography[J]. JPC, 2023, 36(1): 63 - 70.
[10] KIM S, KIM D, JUNG M J, et al. Analysis of environmental organic matters by Ultrahigh - Resolution mass spectrometry - a review on the development of analytical methods[J]. Mass Spectrom Rev, 2022, 41(2): 352 - 369.
[11] TONG Q, ZHANG C, TU Y, et al. Biosynthesis - based spatial metabolome of *Salvia miltiorrhiza* Bunge by combining metabolomics approaches with mass spectrometry - imaging[J]. Talanta, 2022, 238: 123045.
[12] 余志引，朱奇彪. 拉曼光谱技术检测真假藏红花[J]. 现代农业科技，2021(15)：223 - 224，230.
[13] VARLIKLIOZ ER S, EKSI - KOCAK H, YETIM H, et al. Novel spectroscopic method for determination and quantification of saffron adulteration[J]. Food Anal Meth, 2017, 10(5): 1547 - 1555.
[14] DAI H C, GAO Q X, HE L L. Rapid determination of saffron grade and adulteration by thin - layer chromatography coupled with Raman spectroscopy[J]. Food Anal Meth, 2020, 13(11): 2128 - 2137.
[15] 王馨达，宋正勋，李岩，等. 基于太赫兹光谱的 DNA 检测及应用综述[J]. 激光与光电子学进展，2020，57(17)：170003.
[16] 张卫方，杜潮海，刘濮鲲. 太赫兹技术在医学成像方面的应用和进展[J]. 标记免疫分析与临床，2019，26(9)：1603 - 1608.
[17] 马品，杨玉平，范丽洁，等. 肝脏和肺组织的太赫兹光谱成像[J]. 太赫兹科学与电子信息学报，2016，14(6)：838.
[18] 马品，杨玉平. 运用太赫兹光谱技术检测天麻中的水分含量[J]. 太赫兹科学与电子信息学报，2017，15(1)：26.
[19] 张活. 基于太赫兹时域光谱技术的中药检测方法研究[D]. 西安：西安电子科技大学，2018.
[20] 徐振. 基于太赫兹时域光谱技术的川贝母品质检测方法研究[D]. 南昌：华东交通大学，2022.
[21] 杨玉平，张成，刘海顺，等. 两类红花和牛黄的太赫兹光谱法真伪鉴别分析[J]. 光谱学与光谱分析，2019，39(1)：45.
[22] BAI C Y, NIU Q, YANG G, et al. Surface - enhanced Raman spectroscopy substrate based on silver/titanium dioxide nanocomposites and demonstration on rhodamine 6G and severe acute respiratory syndrome coronavirus 2 spike - 1 protein[J]. Spectrosc Lett, 2022, 55(6): 401 - 413.
[23] TARANTILIS P A, BELJEBBAR A, MANFAIT M, et al. FT - IR, FT - Raman spectroscopic study of carotenoids from saffron (*Crocus sativus* L.) and some derivatives[J]. Spectrochim Acta Part A Mol Biomol Spectrosc, 1998, 54(4): 651 - 657.
[24] SCHULZ H, BARANSKA M, BARANSKI R. Potential of NIR - FT - Raman spectroscopy in natural carotenoid analysis[J]. Biopolymers, 2005, 77(4): 212 - 221.
[25] PRATA J C, DA COSTA J P, FERNANDES A J S, et al. Selection of microplastics by Nile Red staining increases environmental sample throughput by micro - Raman spectroscopy[J]. Sci Total Environ, 2021, 783: 146979.
[26] MARY S J J, SIDDIQUE M U M, PRADHAN S, et al. Quantum chemical insight into molecular structure, NBO analysis of the hydrogen - bonded interactions, spectroscopic (FT - IR, FT - Raman), drug likeness and molecular docking of the novel anti COVID - 19 molecule 2 - [(4, 6 - diaminopyrimidin - 2 - yl) sulfanyl] - N - (4 - fluorophenyl) acetamide - dimer[J]. Spectrochim Acta A Mol Biomol Spectrosc, 2021, 244: 118825.
[27] QU F, FU H Y, LI Y Q, et al. Temperature effect on electronic and vibrational properties of β - carotene aggregates in aqueous ethanol solution[J]. Dyes Pigm, 2019, 166: 323 - 329.

基于深度学习的拉曼/太赫兹光谱对藏红花的分类识别

（正文见 36－39 页）

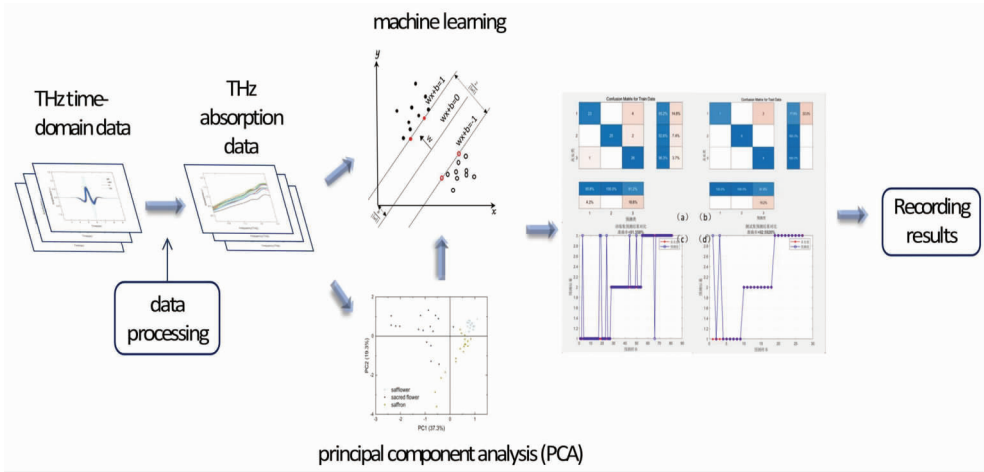
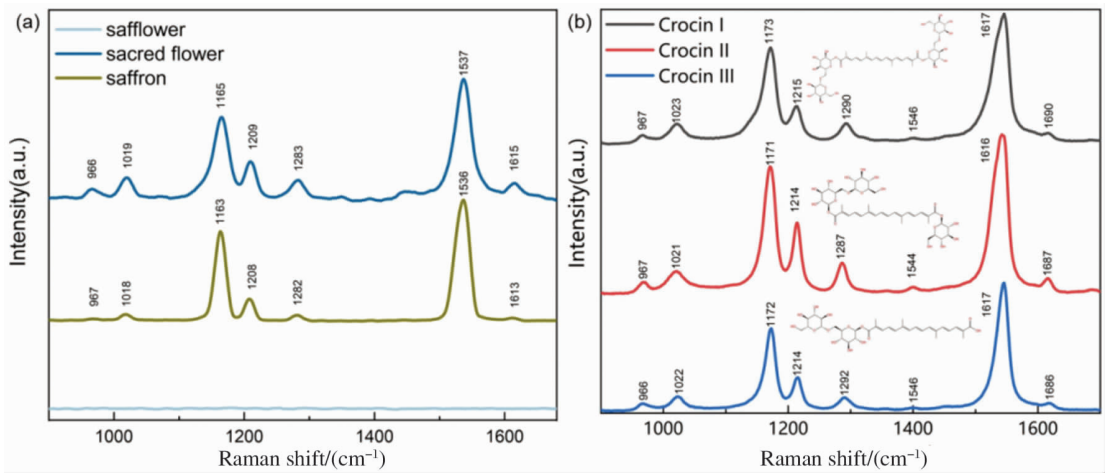


图 1　机器学习流程图（以 THz 光谱为例）



注：（a）草红花、西红花、藏红花的，（b）有效成分西红花苷Ⅰ、Ⅱ、Ⅲ。

图 2　归一化拉曼光谱

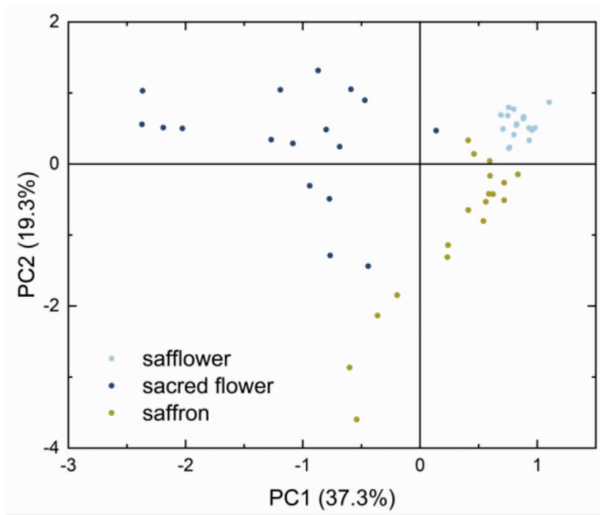
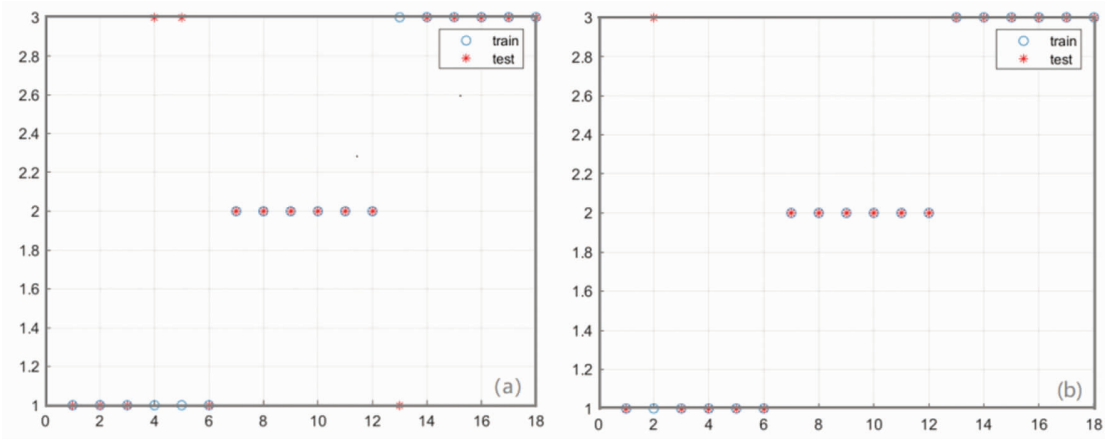
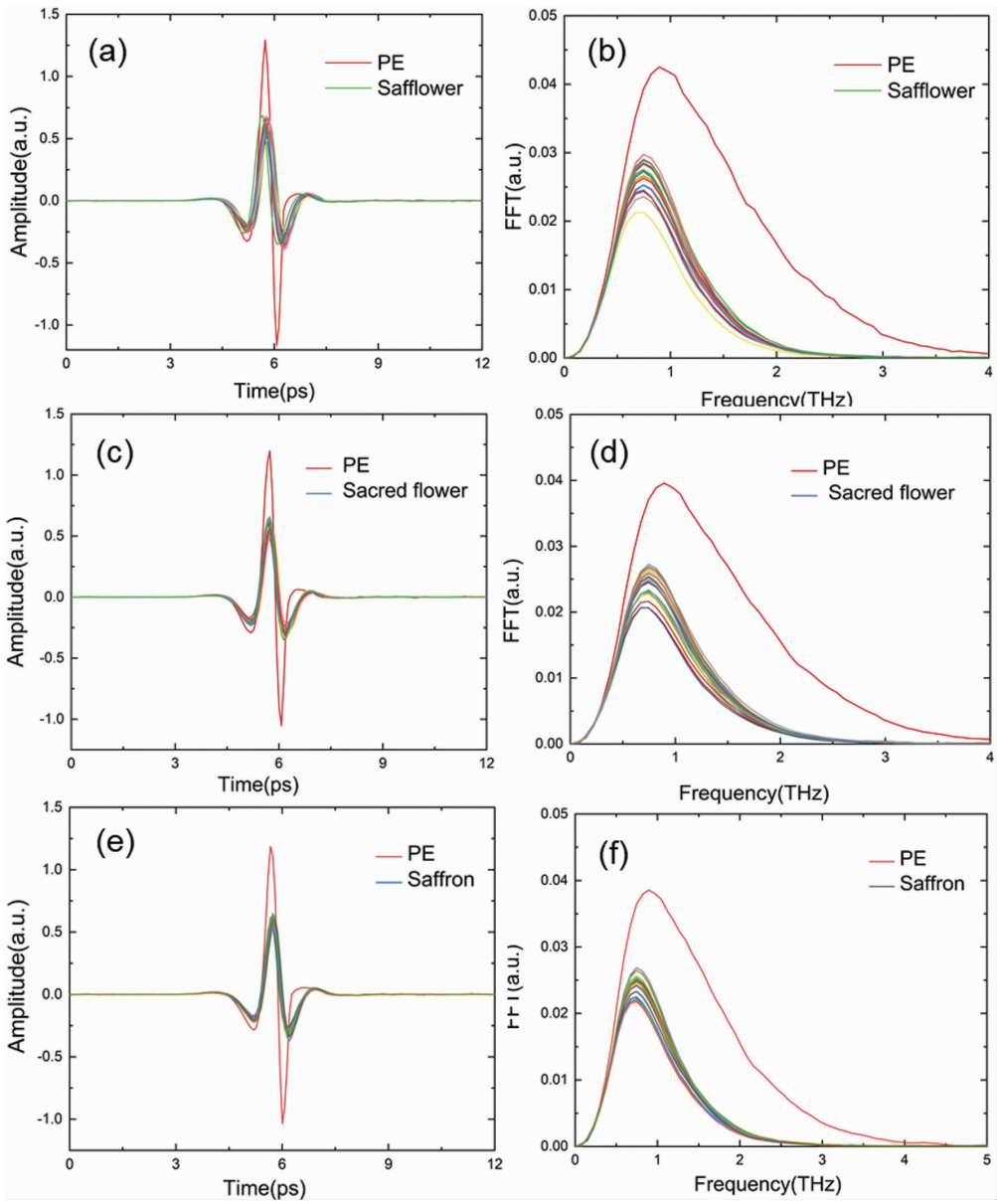


图 3　草红花、西红花、藏红花的拉曼光谱 PCA 分析结果



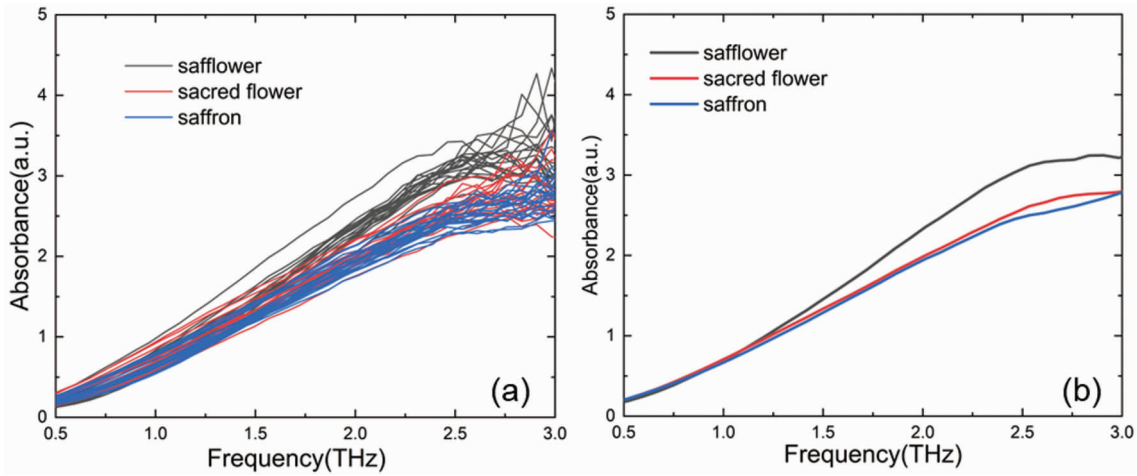
注:(a)SVM,(b)PCA + SVM。

图 4 草红花、西红花、藏红花的拉曼光谱分类结果



注:(a)草红花,(c)西红花,(e)藏红花;与对应的频域光谱(b)草红花,(d)西红花,(f)藏红花。

图 5 草红花、西红花、藏红花的 THz 时域波形



注:(a)原始吸收光谱,(b)平均吸收光谱。

图 6　草红花、西红花、藏红花的 THz 吸收光谱

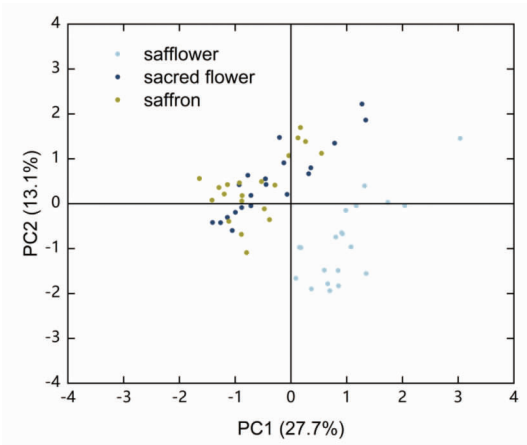


图 7　草红花、西红花、藏红花的 THz 数据 PCA 分析结果

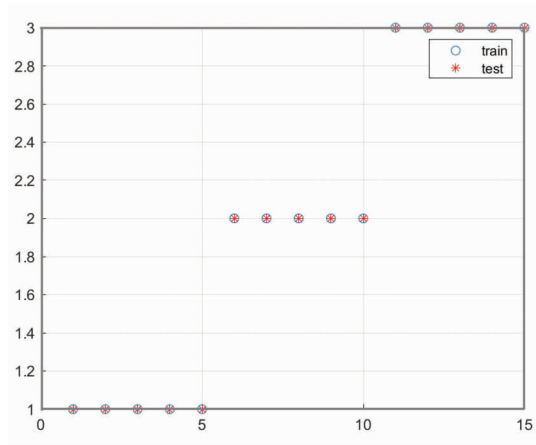
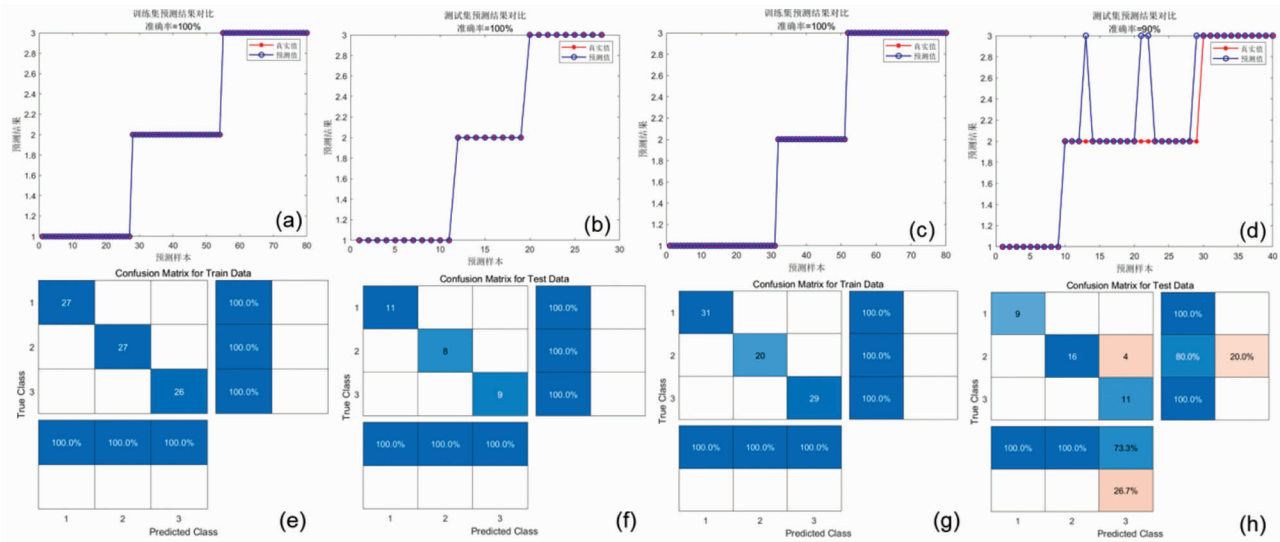
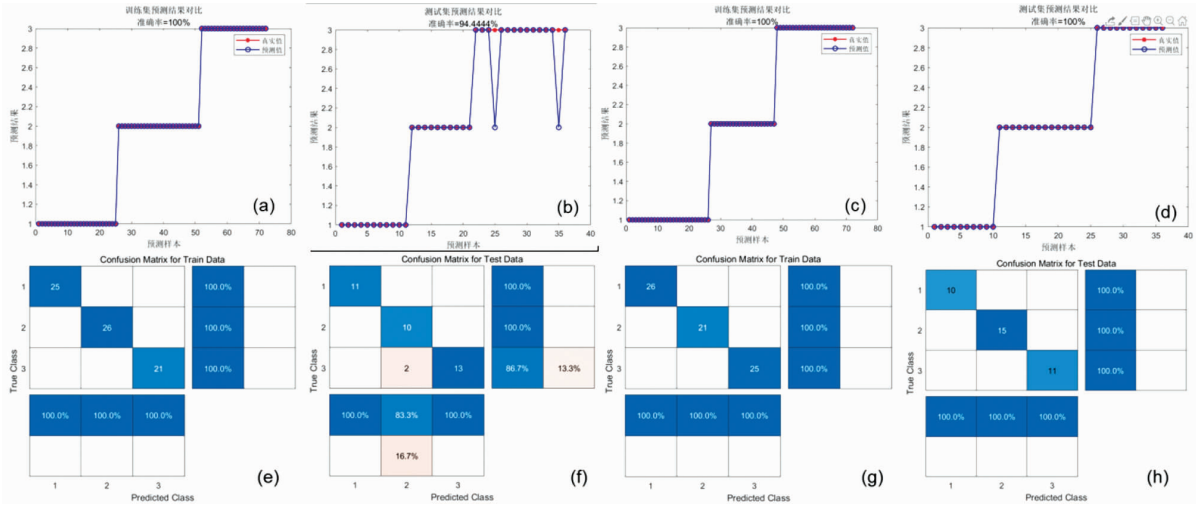


图 8　THz 数据测试集的 SVM 分类结果



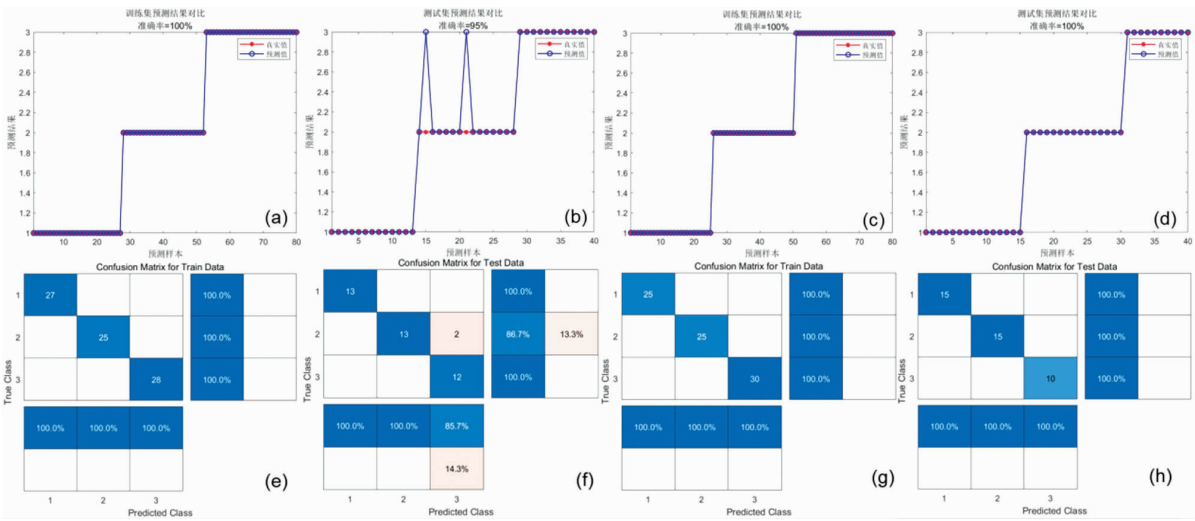
注:(a)、(e)RF 训练集,(b)、(f)RF 测试集,(c)、(g)PCA + RF 训练集,(d)、(h)PCA + RF 测试集。

图 9　THz 数据的 RF 分类结果



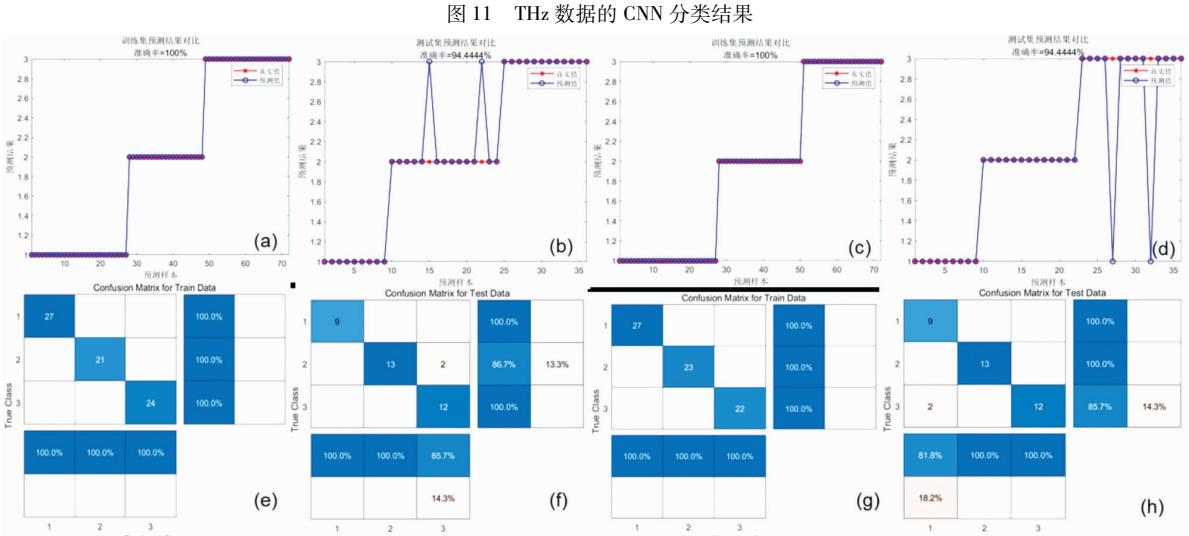
注：(a)、(e)RF 训练集，(b)、(f)RF 测试集，(c)、(g)PCA + RF 训练集，(d)、(h)PCA + RF 测试集。

图 10　RS 数据的 RF 分类结果



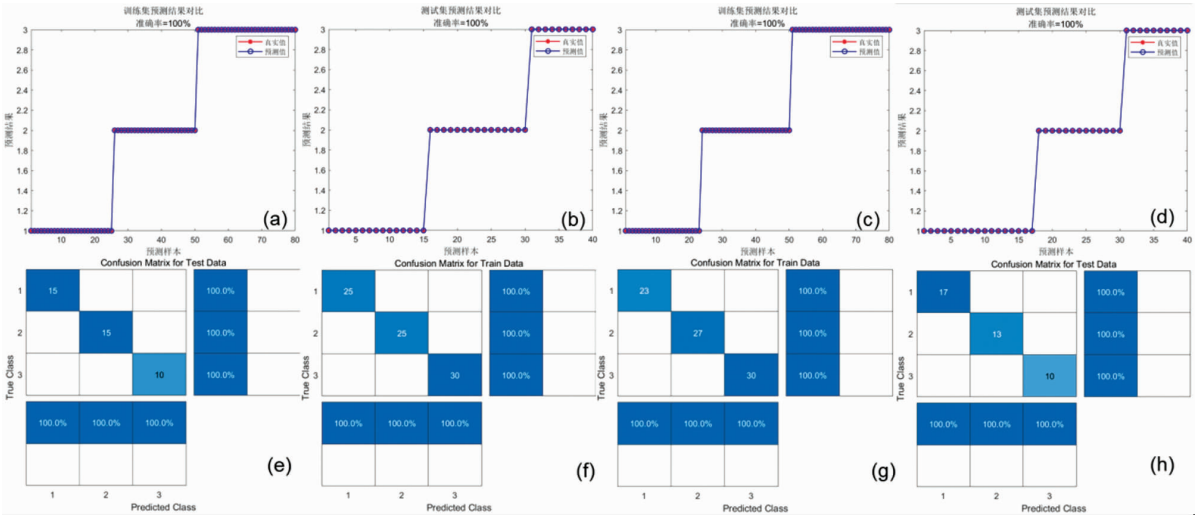
注：(a)、(e)RF 训练集，(b)、(f)RF 测试集，(c)、(g)PCA + RF 训练集，(d)、(h)PCA + RF 测试集。

图 11　THz 数据的 CNN 分类结果



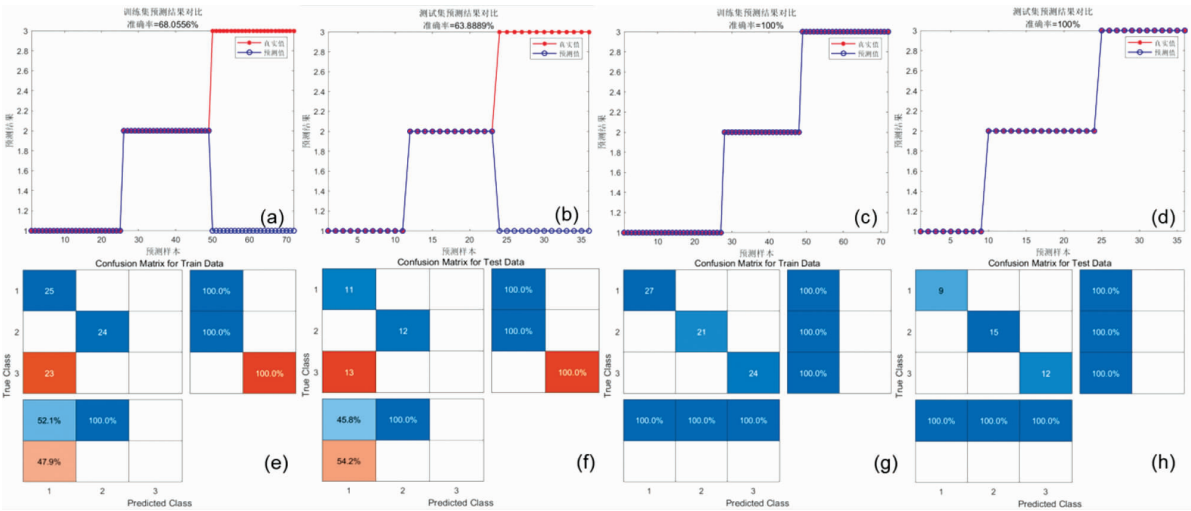
注：(a)、(e)RF 训练集，(b)、(f)RF 测试集，(c)、(g)PCA + RF 训练集，(d)、(h)PCA + RF 测试集。

图 12　RS 数据的 CNN 分类结果



注：(a)、(e) RF 训练集，(b)、(f) RF 测试集，(c)、(g) PCA + RF 训练集，(d)、(h) PCA + RF 测试集。

图 13　THz 数据的 LSTM 分类结果



注：(a)、(e) RF 训练集，(b)、(f) RF 测试集，(c)、(g) PCA + RF 训练集，(d)、(h) PCA + RF 测试集。

图 14　RS 数据的 LSTM 分类结果

基于机器学习方法预测胃癌前状态的比较研究

(正文见 31 – 35 页)

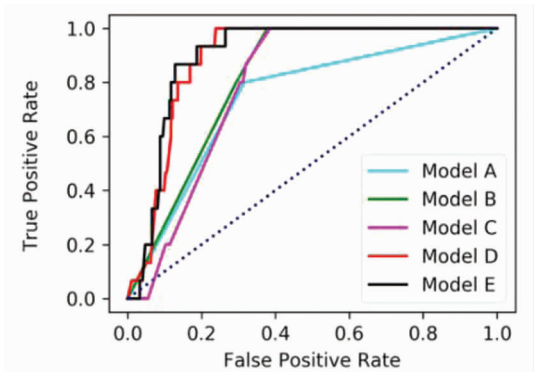


图 1　逐步添加影响因素后 XGBoost 模型表现出的 ROC 曲线