油脂鉴别

DOI: 10.19902/j.cnki.zgyz.1003-7969.220460

分子光谱技术结合深度学习模型识别食用植物油种类

汤睿阳,王继芬

(中国人民公安大学 侦查学院,北京 102600)

摘要:为实现对食用植物油的快速无损识别,采用衰减全反射-傅里叶变换红外光谱获取10种食用植物油样本的340份谱图数据,经过预处理消除光谱数据中的噪声与背景干扰,通过主成分分析降维特征提取3个主成分,在此基础上构建KNN模型与基于SSA算法优化的BP神经网络模型,对植物油种类进行识别并对识别效果进行比较。结果表明:KNN模型的识别准确率可达97.7%;基于SSA算法优化的BP神经网络分类效果最佳,识别准确率达100%,而传统BP神经网络模型识别准确率仅为87.6%。综上,建立的分子光谱技术结合深度学习模型识别食用植物油种类的新方法,实现了对食用植物油种类的准确识别。

关键词:食用植物油;分子光谱;深度学习;种类识别

中图分类号:O657.33;TS227 **文献标识码**:A **文章编号**:1003-7969(2023)10-0116-06

Identifying types of edible vegetable oil by molecular spectroscopic technology combined with deep learning model

TANG Ruiyang, WANG Jifen

(School of Investigation, People's Public Security University of China, Beijing 102600, China)

Abstract: To achieve rapid and non-destructive identification of edible vegetable oil, attenuated total reflection-Fourier transform infrared spectroscopy was used to obtain 340 spectral data of 10 edible vegetable oil samples. After preprocessing, the noise and background interference in the spectral data were eliminated. Three principal components were extracted by principal component analysis, and base on which, the KNN model and the BP neural network model optimized based on the SSA algorithm were constructed for identification and their effects were compared. The results showed that the recognition rate of the KNN model could reach 97.7%. The BP neural network model optimized based on the SSA algorithm, with a recognition rate of 100%, had the best classification effect, while the recognition rate of traditional BP neural network model was only 87.6%. In summary, a new method for identifying edible vegetable oil types using molecular spectroscopy technology combined with deep learning models can realize the accurate identification of edible vegetable oil types.

Key words: edible vegetable oil; molecular spectroscopy; deep learning; type recognition

在当今公共安全视域下,随着社会资源的有效分配与重视程度的提高,人们对食品安全的呼吁越来越强烈。食用植物油作为生活中常用的烹饪原料,能够提供人体所必需的营养成分,完善膳食结构,保障人们的身体健康。然而近年来,一些不良商家为牟取暴利,会在食用植物油中掺假,违反了食品安全标准要求,导致食品安全隐患凸显[1]。全国公安机关曾联合开展"昆仑行动"[2],严厉打击食品安全相关犯罪行为,侦查人员通过对植物油样本的快速无损鉴别,得到种类、生产厂家等有效信息,为案件提供明确的侦查方向。此外在一些特殊案件中,现场提取到的植物油样本有助于判断嫌疑人的生活习惯以及职业特征,能够有效推动公安工作的开展进度。目前,对食用植物油进行的研究尚未形成判

收稿日期:2022-06-30;**修回日期**:2023-07-03

作者简介:汤睿阳(2000),男,在读本科,研究方向为刑事科学技术(E-mail)1455677580@qq.com。

通信作者:王继芬,教授(E-mail)wangjifen58@126.com。

别溯源方面的系统性成果，因而具有很高的研究价值。

目前，食用植物油的检验方法主要有指纹图谱法[3]、高效液相色谱法[4]、气相色谱－质谱法[5]、气相色谱－离子迁移谱法[6]等，这些方法检测性强、效果好，但存在仪器精密昂贵、步骤烦琐、样品消耗量大等缺陷，并不适合广泛运用于公安基层实战工作中。相比上述多种技术，红外光谱作为一种高效灵敏的检验方法，兼具样品用量少、分析快速、操作简单等优势[7]。各种食用植物油所具有的官能团差异，可以通过红外光谱准确定性，并由谱图直观反映出来，满足基层公安机关物证检验的快速无损检测要求，因而红外光谱法被广泛用于植物油检测的相关研究之中。He 等[8]将化学计量学融入山茶油的红外光谱数据中，采用偏最小二乘判别分析法构建对应的数据模型，完成了对成分相似的掺假山茶油的鉴别。孙一健等[9]借助衰减全反射－傅里叶变换红外光谱分析收集到的 5 种植物油样本数据，采用径向基函数神经网络和随机森林构建模型，实现了对植物油样本类别的有效鉴别，准确率达 100%。深度学习是在机器学习基础上，不断自我训练学习，丰富自身经验，从而能够有效解决新的问题[1]，随着其逐步应用于法庭科学[10]、材料分析[11]等多个领域，已成为分析测试领域的热点技术。在获取到植物油样本的红外光谱数据基础上，进一步构建合适的深度学习模型，利用深度学习的可视化、信息化特点对数据进行深度挖掘，可有效提高样本的判别准确率。

K 近邻元素（K－nearest neighbor，KNN）算法是一种基于距离度量的非线性分类方法[12]，借助既定的训练集找到与新数据最接近的 K 条记录，并根据分析结果判定新数据的类别，完成对待检样本的归类。麻雀搜索算法（Sparrow－Search－Algorithm，SSA 算法）是 2020 年提出的一种新型群智能算法[13]，通过对麻雀群体寻觅食物和逃避捕猎过程的行为深入研究，将麻雀个体归为不同类别，并根据麻雀个体位置的实时更新来实现对模型数据的智能优化。采用 SSA 算法优化的 BP 神经网络（简称 SSA－BP 神经网络）进行建模分析，能够有效改进传统 BP 神经网络算法易陷入局部最优解的缺陷，进一步提高优化模型对食用植物油样本的识别准确率。现有研究中，采用分子光谱技术结合多种深度学习模型的方法进行植物油分类的应用研究尚显缺乏。

本文基于衰减全反射－傅里叶变换红外光谱分析技术，采集 340 份不同种类的食用植物油红外光谱数据，借助主成分分析（PCA）进行光谱数据的降维特征提取，构建 KNN 算法以及基于 SAA－BP 神经网络的对比分类模型，开展对不同植物油的可视化模型数据的判别分析，从而实现对食用植物油的快速无损识别分类，为公安工作中涉及植物油检材的相关案件提供明确的侦查方向与证据参考。

1　材料与方法

1.1　实验材料

为满足种类相对齐全的要求，并结合实际案件情况，共收集了 10 种常见的食用植物油，其中芝麻油 80 份，花生油 60 份，玉米油 40 份，亚麻籽油 40 份，橄榄油、菜籽油、调和油、藤椒油、花椒油以及椰子油各 20 份。340 份食用植物油样本信息统计见表 1。

表 1　340 份食用植物油样本信息

食用植物油	品牌	样本数量（份）
芝麻油	保定曲阳小磨香油	20
	永溢粮油小磨香油	20
	金起小磨香油	20
	六必居小磨香油	20
花生油	鲁花 5S 压榨一级花生油	20
	胡姬花古法花生油	20
	烟台胡姬花古法花生油	20
玉米油	鲁花压榨特香玉米胚芽油	20
	金龙鱼金滴玉米油	20
亚麻籽油	罗尔仕压榨亚麻籽油	20
	红井源压榨一级纯香亚麻籽油	20
橄榄油	贝蒂斯特级初榨橄榄油	20
菜籽油	成都红旗菜籽油	20
调和油	山茶橄榄食用植物调和油	20
藤椒油	么麻子藤椒油	20
花椒油	汉源鲜花椒油	20
椰子油	赫丽奇特初榨椰子油	20

Nicolet is10 型傅里叶变换红外光谱仪（金刚石 HATR 晶体），Thermo Scientific 公司。

1.2　实验方法

1.2.1　光谱采集

在采集红外光谱数据前，对这 10 种 340 份食用植物油样本进行标号并注明种类、品牌，置于样品池中进行测定。测定前排除环境及背景干扰，为尽量减小误差，每份食用植物油样本均取 2 mL，连续进行 3 次谱图采集，记录其平均值作为样本的光谱数据[9]。检测条件：KBr 分束器，采集范围 650～4 000 cm^{-1}，光谱分辨率 2 cm^{-1}，扫描次数 64 次。

1.2.2 光谱预处理

对采集到的食用植物油光谱数据采用峰面积归一化、基线校准以及 Z - score 方法进行平滑预处理[14]，消除谱图中的干扰与噪声，从而尽可能完整保留原光谱的真实信息，并清晰展现样本谱图特征。为规避表面散射、固体颗粒以及光程变化对光谱造成影响[1]，采用标准正态变换（Standard normal variate，SNV）方法处理光谱数据。

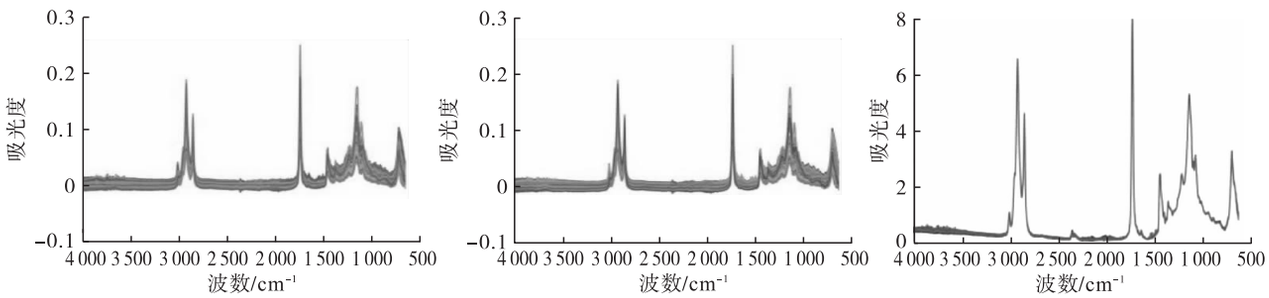
1.2.3 食用植物油识别模型的建立

采用 PCA，对经预处理后的 340 份食用植物油样本的红外光谱数据进行压缩降维。以各样本的 3 个主成分为特征变量，采用 KNN 算法以训练集即为测试集的方法进行交互验证[15]，建立 KNN 模型对未知样本进行判别分类。通过 MATLAB 软件进行多元建模分析，采用基于 SSA - BP 神经网络模型对 10 种食用植物油样本的光谱数据按照训练集和测试集 7∶3 的比例进行训练判别[1]。

2 结果与讨论

2.1 光谱预处理

10 种食用植物油样本的原始光谱、平滑处理后及 SNV 处理后的红外光谱图如图 1 所示。



a) 原始光谱 b) 平滑处理后光谱 c) SNV处理后光谱

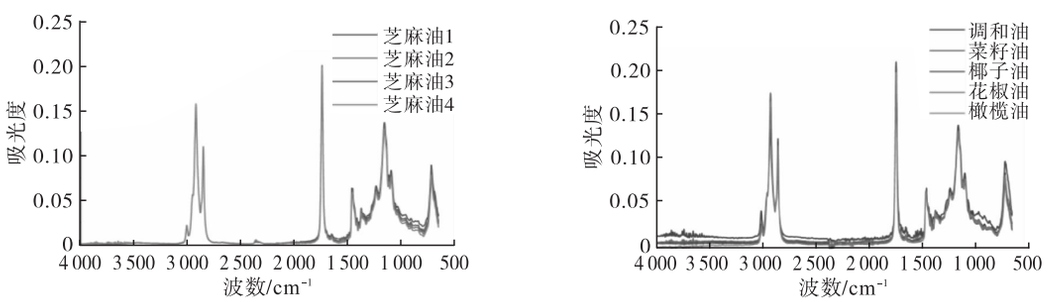
图 1 食用植物油样本红外光谱图

由图 1 可知，经平滑处理和 SNV 处理后，消除了植物油红外光谱图中的干扰与噪声，有效提高了谱图的灵敏度和分辨率。

2.2 样本光谱图分析

采集食用植物油样本在 650 ~ 4 000 cm^{-1} 波段范围内经预处理后的红外光谱数据，4 个不同品牌的芝麻油样本的红外光谱图和 5 种不同种类植物油样本的红外光谱图如图 2 所示。

由图 2 可见，植物油的红外光谱图中较明显峰主要出现在 1 200 cm^{-1} 处碳氧伸缩振动、1 450 cm^{-1} 处和 2 900 cm^{-1} 处碳氢伸缩振动及 1 650 ~ 1 750 cm^{-1} 处的碳氧双键伸缩振动[9]。由图 2a 可知，不同品牌同种植物油的红外光谱在峰的整体走向与形状上几乎没有区别。由图 2b 可见，不同种类食用植物油的红外光谱具有相似的走势和吸收峰，且存在相互交织重叠的现象，差异并不明显，难以直观区分出食用植物油的种类，考虑引入深度学习模型来实现对食用植物油红外光谱图数据的识别。



a) 不同品牌同种食用植物油 b) 不同种类食用植物油

图 2 样本光谱图

2.3 PCA 降维

采用 PCA 对食用植物油样本的红外光谱数据进行降维，可用相对少的数据变量概括原有谱图数据的特征，尽可能减少后续建模分析中的计算成本，有效提高实验效率与准确率。食用植物油的红外光谱数据的 PCA 结果见表 2。

表 2 PCA 结果

主成分	特征值	方差贡献率/%	累积方差贡献率/%
PC1	61.508	85.427	85.427
PC2	4.575	11.775	97.202
PC3	1.437	1.584	98.786

由表 2 可见，第一主成分即 PC1 的特征值高达 61.508，显著高于其他主成分，可以较大程度地概括大部分原始数据。前 3 个主成分的累积方差贡献率达到了 98.786%，表明可以概括 98.786% 的样本谱图数据信息，几乎能够解释样本数据的所有内容。根据特征值大于 1 以及累积方差贡献率高于 85% 的要求[16]，选择前 3 个成分（PC1、PC2、PC3）作进一步建模分析。

2.4　KNN 模型分析

采用 KNN 模型进一步对光谱数据进行分析。为弥补 KNN 模型面对冗多数据，计算量大的缺点，采用 PCA 得到的降维数据进行建模。以各样本的 3 个主成分为特征变量，采用以训练集即为测试集的方法进行交互验证[15]，建立 KNN 模型对未知样本进行判别分类，特征变量重要性图如图 3 所示，K 选择错误统计图如图 4 所示。

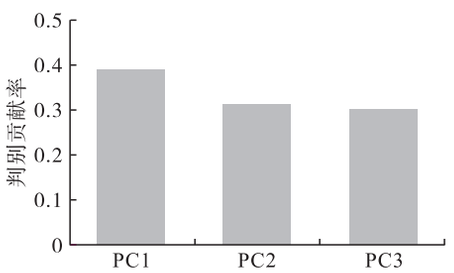


图 3　特征变量重要性图

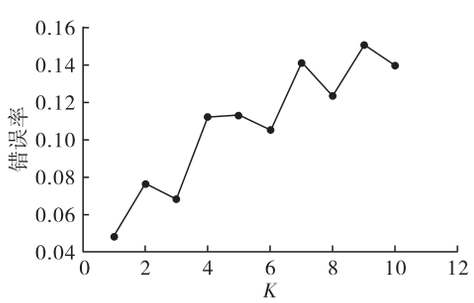


图 4　K 选择错误统计图

由图 3 可知，3 个特征变量中 PC1（即特征变量 1）进行样本判别时贡献最大，判别贡献率为 0.39，而 PC3 是区分贡献最小的特征变量，其值为 0.30，3 个特征变量的重要程度之和为 1[13]。由图 4 可知，各样本在进行判别预测的，分类错误率呈波动递增趋势，当 K 值为 1 时，错误率最低，为 0.048 2，即当 K 值为 1 时进行模型分类的准确率较高，能较好地实现不同植物油样本的准确区分。因此，选取 K 值为 1，重要程度最高的 PC1 为主要特征自变量，PC2 和 PC3 为协变量参数，运用以训练集即为测试集的交互验证方法构建分类模型，10 种食用植物油样本的空间分布如图 5 所示。

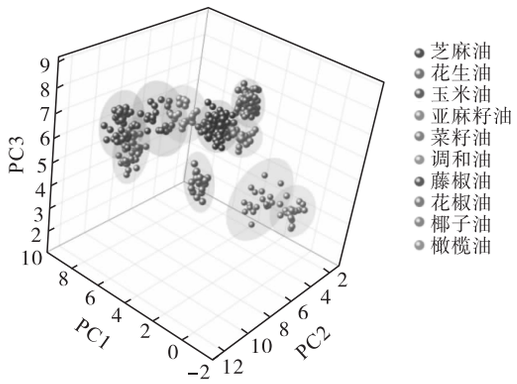


图 5　10 种食用植物油样本的空间分布

由图 5 可知，10 种食用植物油样本被完全分离开来，不同种类的食用植物油样本在空间中分散距离较远，如花椒油和椰子油之间，表明不同种类间的差异明显，而同类型食用植物油则分布较为集中，芝麻油、花生油、玉米油和亚麻籽油整体内部分布很密集，表明这几种植物油的不同品牌之间差异不大，借助 KNN 模型构建空间模型进一步实现了对待测植物油样本的准确分类。采用此算法模型对所有植物油样本进行预测判别，大部分样本都得到了准确归类，识别准确率达到 97.7%。分析认为，KNN 模型是一种基于样本距离的算法模型，主要通过分析样本与周围有限的邻近样本间距来判断其所属类别，植物油样本红外谱图的降维数据类域内交叉重叠明显，而不同类数据的域间距较为显著，故判别效果较为理想。

2.5　SSA－BP 神经网络模型分析

SSA－BP 神经网络模型的训练曲线如图 6 所示，基于 SSA－BP 神经网络的识别结果如图 7 所示。

由图 6 可知，模型中的训练集在 45 次后开始收敛，测试集于 47 次达到收敛，当训练次数达到 53 次后，各指标均趋于平稳收敛，预测效果达到最佳。

由图 7 可知，模型中训练集与测试集数据结果高度拟合。其中数据总体的拟合程度达到 99.790%，训练集和测试集分别与原始数据的拟合程度达 99.917% 和 99.599%，说明 SSA－BP 神经网络模型对食用植物油种类的识别率较高，效果理想。

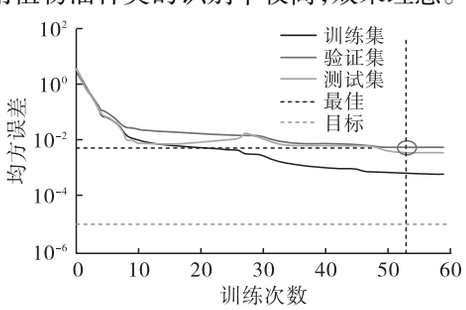


图 6　SSA－BP 神经网络模型的训练曲线

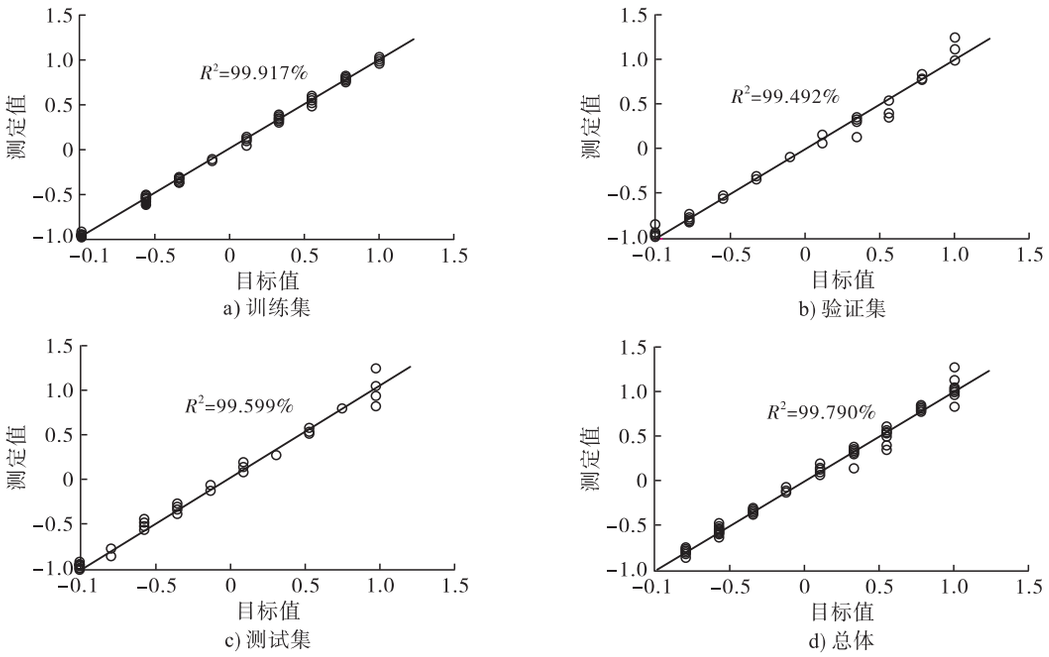


图 7　基于 SSA－BP 神经网络的识别结果

采用传统 BP 神经网络模型和 SSA－BP 神经网络模型对 10 种食用植物油样本进行识别，结果如图 8 所示。

经计算，传统 BP 神经网络模型对这 10 种食用植物油样本的识别准确率仅为 87.6%。由图 8 可知：根据误差绝对值需小于 1 且尽可能接近 0 的要求，传统 BP 神经网络多组数据的判别误差在－2 上下波动，识别效果较差；SSA－BP 神经网络模型对所有样本的分类识别准确率高达 100%，各组数据的误差均稳定趋于 0，效果较为理想。传统 BP 神经网络模型的优势在于训练海量数据过程中，逐步调整训练参数与下一步幅度与方向，最终得到优化目标结果。但囿于 BP 自身一定的局限性[7]，且该研究中样本数量相对较少，不足以建立和完善模型，导致预测结果准确率较低，识别效果尚不及 KNN 模型。采用 SSA－BP 神经网络，显著提高了模型的识别准确率，相较于 KNN 模型，SSA－BP 神经网络模型更具有优势，所有食用植物油样本数据都得到了准确的预测判别。

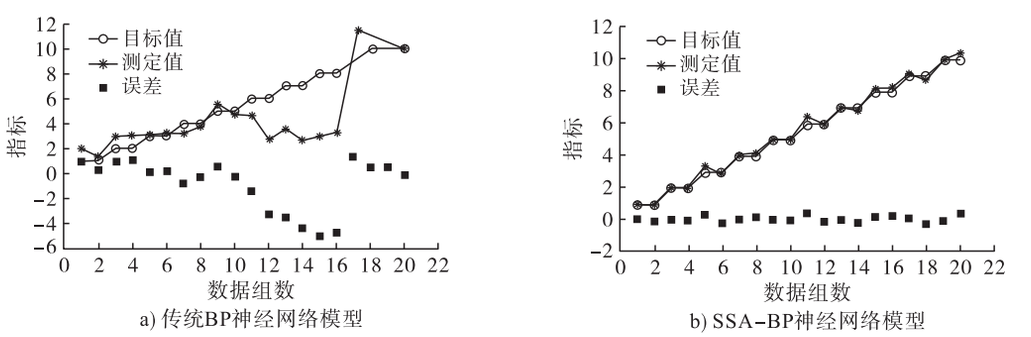


图 8　植物油种类预测识别

3　结　论

采用衰减全反射－傅里叶变换红外光谱技术结合深度学习多元分析方法，通过 KNN 算法和 SSA－BP 神经网络算法分别构建了食用植物油样本光谱数据的识别模型，实现对 10 种植物油样本的准确识别归类。结果显示：植物油样本的红外光谱数据之间的非线性关系显著，采用主成分分析对光谱数据进行特征提取及降维处理，规避了 KNN 难以处理海量数据的缺陷。KNN 模型的预测结果较为准确，分类识别准确率达到了 97.7%，能基本完成对样本数据的有效判别。相比之下，SSA－BP 神经网络模型在食用植物油样本归类方面更具优势，克服了传统 BP 神经网络中易局限于局部最优解的问题，能更高效地处理参数冗杂带来的不便，实现了对植物油样本的准确预测鉴别，识别准确率达 100%，实验效果理想。研究发现，采用分子光谱技术可以现场快速

获取案件中植物油检材的光谱数据，结合深度学习进行建模分析，为其在公安工作中植物油乃至其他物证分析检验的应用提供了指导与方向，具有广泛应用的潜力。

参考文献：

[1] 接昭玮，刘卓，王继芬，等. 植物油的红外光谱结合神经网络快速识别[J]. 中国油脂：2023，48(1)：79－83，93.

[2] 孙一健，王继芬. 太赫兹时域光谱技术在食品、药品和环境领域中的应用研究进展[J]. 激光与光电子学进展，2022，59(16)：22－31.

[3] 姚云平，李昌模，刘慧琳，等. 指纹图谱技术在植物油鉴定和掺假中的应用[J]. 中国油脂，2012，37(7)：51－54.

[4] 周子焱，邢家溧，应璐，等. 食用植物油中黄曲霉毒素 B_1 调查分析[J]. 中国油脂，2017，42(12)：66－69.

[5] 鲍晓瑾，倪炜华，沈锡贤. GC－MS 法识别二元混合植物油掺混量的方法研究[J]. 中国油脂，2016，41(12)：81－84.

[6] 陈通，陆道礼，陈斌. GC－IMS 技术结合化学计量学方法在食用植物油分类中的应用[J]. 分析测试学报，2017，36(10)：1235－1239.

[7] 汤睿阳，王之宇，王继芬，等. 基于分子光谱模式识别的个体指甲无损鉴别及性别刻画[J/OL]. 激光与光电子学进展，2022：1－16[2022－06－27]. http://kns.cnki.net/kcms/detail/31.1690.TN.20220527.1251.004.html.

[8] HE W X，LEI T X. Identification of camellia oil using FT－IR spectroscopy and chemometrics based on both isolated unsaponifiables and vegetable oils[J/OL]. Spectrochim Acta A，2020，228：117839[2022－06－27]. https://doi.org/10.1016/j.saa.2019.117839.

[9] 孙一健，王继芬，张震. 基于红外光谱的食用植物油种类鉴别[J]. 中国油脂，2023，48(1)：120－124.

[10] HOFFMAN B L，HACKMAN L，LINDENFELD L A. Training for communication in forensic science[J]. Emerg Top Life Sci，2021，5(3)：359－365.

[11] 付严宇，杨桄，李德军，等. 基于高光谱遥感技术的伪装材料的光谱特性分析[J]. 激光与光电子学进展，2021，58(20)：518－524.

[12] 何欣龙，陈利波，王继芬，等. 基于 K 近邻算法的塑钢窗拉曼光谱分析[J]. 激光与光电子学进展，2018，55(5)：409－413.

[13] 赵婧宇，池越，周亚同. 基于 SSA－LSTM 模型的短期电力负荷预测[J]. 电工电能新技术，2022，41(6)：71－79.

[14] 宋丽梅，罗菁. 模式识别[M]. 北京：机械工业出版社，2015.

[15] 桑国通，廖晓曦，何欣龙，等. K 近邻算法结合红外光谱对轮胎橡胶颗粒的鉴别研究[J]. 化学通报，2019，82(1)：87－91.

[16] 卫辰洁，王继芬，范琳媛，等. 基于光谱数据融合和人工神经网络的汽车灯罩鉴别[J]. 中国塑料，2020，34(12)：59－64.

（上接第 108 页）

[2] 刘林奇. 基于粮食安全视角的我国主要粮食品种进口依赖性风险分析[J]. 农业技术经济，2015(11)：37－46.

[3] 朱再清，袁圣弘，涂涛涛. 我国油菜籽及菜子油进口依赖性与进口安全研究[J]. 中国农业大学学报，2014，19(4)：253－264.

[4] 杨艳涛，丁琪，王国刚. 全球疫情下我国玉米供应链体系的风险问题与对策[J]. 经济纵横，2020(5)：58－65.

[5] 崔连标，翁世梅，宋马林. 贸易冲突、“一带一路”与中国农产品进口多元化策略研究[J]. 科学决策，2021(1)：31－53.

[6] 张洋，严茂林，葛玮玮，等. 我国食用植物油供给现状分析及未来发展战略研究[J]. 中国油脂，2022，47(4)：1－8.

[7] 曹景武. 供需态势、风险摆脱与食用植物油料的安全保障[J]. 改革，2015(9)：130－141.

[8] 周静，谷强平，杜吉到. 中国大豆进口依赖性及其对大豆进口安全的影响[J]. 大豆科学，2015，34(3)：503－506，511.

[9] 傅龙波，钟甫宁，徐志刚. 中国粮食进口的依赖性及其对粮食安全的影响[J]. 管理世界，2001(3)：135－140.

[10] 贾兴梅，李平. 中国大豆产业安全度初步评估[J]. 华南农业大学学报(社会科学版)，2012，11(3)：25－32.

[11] 祝孔超，牛叔文，赵媛，等. 中国原油进口来源国供应安全的定量评估[J]. 自然资源学报，2020，35(11)：2629－2644.

[12] 谷强平，周静，杜吉到. 基于贸易视角的中国大豆产业安全分析[J]. 大豆科学，2015，34(2)：314－319.

[13] 龚谨，孙致陆，李先德. 中国大麦进口的替代弹性及可依赖性研究[J]. 中国流通经济，2019，33(10)：85－93.