

# 基于 PCA - SVM 结合共聚焦拉曼光谱的特级初榨橄榄油掺伪压榨菜籽油定量分析

彭楠<sup>1</sup>, 方俊<sup>1</sup>, 毛潭<sup>2</sup>

(1. 北京京北职业技术学院, 北京 101400; 2. 北方工业大学, 北京 100144)

**摘要:** 为了促进国内橄榄油市场的健康发展, 对掺伪同样存在天然类胡萝卜素的低温压榨菜籽油的特级初榨橄榄油进行了定量鉴别研究。采用共聚焦拉曼光谱技术对不同掺伪浓度油样进行测试, 基于密度泛函理论对油样的拉曼光谱峰的归属进行了理论分析, 并对拉曼光谱数据进行主成分分析(PCA), 然后利用支持向量机(SVM)构建 PCA - SVM 模型。另外, 对 PCA - SVM 模型的检出限进行了研究。结果表明: 特级初榨橄榄油与低温压榨菜籽油的拉曼光谱存在一定差异, 最明显的光谱差异主要集中在谱峰 1 008、1 161、1 528  $\text{cm}^{-1}$  和谱段 2 800 ~ 3 000  $\text{cm}^{-1}$  内, 与密度泛函理论对不同油样拉曼光谱峰的分析一致; 不考虑类胡萝卜素特征信号建立的 PCA - SVM 模型决定系数大于 0.989, 均方根误差小于 2.990%, 检出限为 2% (低温压榨菜籽油体积分数); 在特级初榨橄榄油掺伪定量分析中, 考虑类胡萝卜素的特征信号有助于提高模型预测精度, 但仅限于掺伪低价植物油中无类胡萝卜素存在的情况; PCA - SVM 模型在不考虑类胡萝卜素特征信号的情况下依然具有良好的定量预测效果。综上, 所建立的 PCA - SVM 模型可以用于掺伪 2% 以上低温压榨菜籽油的特级初榨橄榄油的定量鉴别。

**关键词:** 特级初榨橄榄油; 低温压榨菜籽油; PCA - SVM; 拉曼光谱; 密度泛函理论

**中图分类号:** TS201.6; O433 **文献标识码:** A **文章编号:** 1003 - 7969(2024)02 - 0070 - 05

## Quantitative analysis of extra virgin olive oil adulterated with pressed rapeseed oil based on PCA - SVM and confocal Raman spectroscopy

PENG Nan<sup>1</sup>, FANG Jun<sup>1</sup>, MAO Tan<sup>2</sup>

(1. Beijing Jingbei Vocational and Technical College, Beijing 101400, China;

2. North University of Technology, Beijing 100144, China)

**Abstract:** To promote the healthy development of the domestic olive oil market, a study was conducted on quantitatively identifying extra virgin olive oil adulterated with low - temperature pressed rapeseed oil containing natural carotenoids. The confocal Raman spectroscopy technology was used to test oil samples with different adulteration amounts. Theoretical analysis was conducted on attributing Raman spectral peaks of different oil samples based on density functional theory. Principal component analysis (PCA) was performed on the Raman spectral data, and then the support vector machine (SVM) was used to construct a PCA - SVM model. In addition, the detection limit of the PCA - SVM model was studied. The results showed that there was difference in the Raman spectra of extra virgin olive oil and low - temperature pressed rapeseed oil, and the most apparent spectral differences mainly concentrated in the

收稿日期: 2023 - 04 - 12; 修回日期: 2023 - 11 - 20

基金项目: 国家自然科学基金(62172006); 北京市职业教育改革项目(J201901, 2018 - 102); 北京市职业院校教师素质提高工程资助项目(028)

作者简介: 彭楠(1988), 女, 讲师, 主要从事计算机应用方面的研究(E-mail) pengnan\_beijing@163.com。

spectral peaks 1 008, 1 161, 1 528  $\text{cm}^{-1}$  and spectral bands 2 800 - 3 000  $\text{cm}^{-1}$ , which was consistent with the analysis of Raman spectral peaks of different oil samples using density functional theory. The PCA - SVM model established without considering the characteristic

signals of carotenoids had a coefficient of determination greater than 0.989, and the root mean square error was lower than 2.990%, and the detection limit was 2% (volume fraction of low-temperature pressed rapeseed oil). In the quantitative analysis of extra virgin olive oil adulteration, considering the characteristic signals of carotenoids could help to improve the prediction accuracy of the model, but it was only limited to the absence of carotenoids in adulterated low-price vegetable oils. The PCA-SVM model still had good quantitative prediction performance even when the characteristic signals of carotenoids were not considered. In summary, the established PCA-SVM model can be used to quantitatively identify extra virgin olive oil adulterated with over 2% low-temperature pressed rapeseed oil.

**Key words:** extra virgin olive oil; low-temperature pressed rapeseed oil; PCA-SVM; Raman spectroscopy; density functional theory

特级初榨橄榄油是地中海饮食的主要食用油脂,含有多种对人体有益天然活性物质,如酚类物质、类胡萝卜素等<sup>[1]</sup>,其价格通常是普通植物油的7~9倍。在利益的驱动下,国内橄榄油市场出现以低价植物油掺假甚至冒充特级初榨橄榄油的现象,因此鉴别特级初榨橄榄油及其掺伪量是促进国内橄榄油市场健康发展的前提。

食品质量检测的重要需求包括非侵入和高通量,因而光谱技术具有巨大的应用潜力。2018年,曹晨鹏等<sup>[2]</sup>提出一种基于受激布里渊散射(SBS)及吸收光谱的橄榄油等级鉴别方法。2019年,Squeo等<sup>[3]</sup>采用FT-IR技术并结合化学计量学方法实现特级初榨橄榄油的分类鉴别。2019年,王泓鹏等<sup>[4]</sup>研究了特级初榨橄榄油的荧光特性和吸收光谱特性,并对比说明了两种技术的应用潜力。2020年,Li等<sup>[5]</sup>采用激光诱导荧光光谱技术和支持向量机(SVM)结合主成分分析(PCA)-线性判别分析算法研究了特级初榨橄榄油中低掺伪量大豆油的定量检测。2021年,马金鸽等<sup>[6]</sup>采用便携式拉曼光谱仪对进口散装橄榄油品质进行现场无损鉴别,基于偏小二乘法将拉曼强度转换为向量空间角度值,并建立了橄榄油品质指标分析预测模型。拉曼光谱技术是一种无损、快检的分子光谱技术,具有无样品前处理过程且对操作人员要求较低等优点,逐渐成为应用于特级初榨橄榄油鉴别和掺伪量化的强大分析工具之一。

根据特级初榨橄榄油中含有较多的类胡萝卜素,研究人员开展了将类胡萝卜素用于评估特级初榨橄榄油的地理来源、栽培品种、新鲜度以及掺伪量等研究<sup>[7-10]</sup>。但类胡萝卜素并非仅存在于油橄榄鲜果中,油菜籽中同样存在一定量的类胡萝卜素,采用低温压榨工艺制取的菜籽油中同样会保留天然类胡萝卜素,因而会影响特级初榨橄榄油掺伪低温压榨菜籽油定量分析结果。在上述掺伪检测的光谱技

术中,SBS及吸收光谱定量分析方法的缺陷在于设备过于昂贵,且仪器通常体积较为庞大;FT-IR技术光谱的稳定性较好,但仪器设备同样难以实现小型化,维护和保养也需要一定的专业知识;荧光和吸收光谱相结合的技术虽然易于推广,但面对同样具有相似光敏色素的低温压榨菜籽油掺入特级初榨橄榄油时,很难实现准确的量化分析;激光诱导荧光光谱技术同样是借助光敏色素的发光现象,研究特级初榨橄榄油的掺伪问题时低温压榨菜籽油会影响定量分析的精度;便携式拉曼光谱仪在手持式执法应用中具有潜在的应用价值,但其很难达到与台式拉曼光谱仪相同的检测灵敏度和检出限。综上所述,实现一种小型化的台式拉曼光谱仪并针对性开展特级初榨橄榄油掺伪低温压榨菜籽油的研究工作至关重要。本文提出一种基于共聚焦拉曼光谱结合化学计量学方法用于特级初榨橄榄油掺伪低温压榨菜籽油定量分析的研究,并基于密度泛函理论分析了油样的拉曼光谱特性。

## 1 材料与方法

### 1.1 实验材料

特级初榨橄榄油(产地,中国陇南;品牌,祥宇;产品执行标准,GB/T 23347—2021);低温压榨菜籽油(产地,中国;品牌,海天;产品执行标准,GB/T 1536—2021);类胡萝卜素(品牌,阿拉丁;PubChem编号,5280489)。

共聚焦拉曼光谱仪,自制,其光路示意图如图1所示。该装置主要有激光器、光谱仪和显微物镜等组成。其中:激光光源为532 nm蓝绿激光器,线宽为0.000 1 nm,作用于焦点处的激光能量达10 mW;如海XS7031光谱仪波数范围150~3 200 cm<sup>-1</sup>,面阵制冷探测器,狭缝宽度25 μm;20倍奥林巴斯MXPlanFLN显微物镜;Semrock滤光片(LP03-532RE-25、LPD02-532RU-25)。

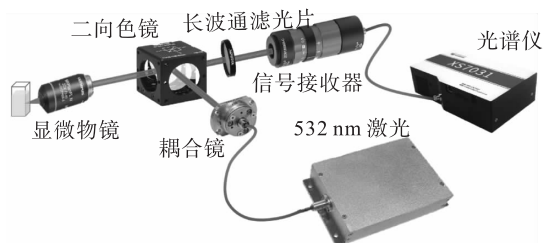


图1 共聚焦拉曼光谱仪光路示意图

## 1.2 实验方法

### 1.2.1 油样配制

取一定体积的特级初榨橄榄油,加入一定量的低温压榨菜籽油,充分搅拌并静置 12 h,配制特级初榨橄榄油体积分数分别为 100.0%、50.0%、35.0%、25.0%、20.0%、16.0%、15.0%、0.0% 的油样。每种油样各制备 10 个样本。

### 1.2.2 光谱数据采集

使用一次性巴氏滴管将油样滴入石英比色皿内,再将石英比色皿置于激光聚焦点处(激光聚焦在油品 2 mm 内),利用共聚焦拉曼光谱仪采集油样的光谱数据。实验过程中,每条光谱的积分时间为 1 s,每个样本各随机采集 50 条光谱。

### 1.2.3 光谱数据预处理

光谱数据预处理包括去基线和降噪两部分。在拉曼光谱实验过程中,荧光、背景噪声等会产生较强的基底信号,为此需要对原始拉曼光谱进行数据预处理,以降低拉曼光谱的噪声并去除线性和非线性基底信号。本文使用 Matlab R2020a 采用小波变换算法,通过多尺度分解产生的低频系数,求解原始信号的基线趋势,从而实现荧光及背景噪声的去除。

### 1.2.4 掺伪识别模型的建立及评价标准

PCA 算法是一种线性降维算法,可实现较少的数据维度表达出较多的原始数据特征<sup>[11]</sup>。在机器学习中,SVM 是一种具有相关学习算法的监督学习模型,其可应用于分析分类和回归的数据问题<sup>[12]</sup>。本文将 PCA 与 SVM 相结合,即首先对经预处理的数据进行 PCA 降维,再将降维数据输入 SVM 模型中,从而实现回归问题的分析。本文采用 Matlab R2020a 进行数据处理和分析。

本文选用决定系数( $R^2$ )和均方根误差(RMSE)评价模型预测效果。其中: $R^2$ 是一个介于 0 和 1 之间的数字,用于衡量统计模型预测结果的程度,其值越趋近于 1,说明模型在预测方面越好;RMSE 是预测值与真实值偏差的平方与观测次数比值的平方根,其衡量的是预测值与真实值之间的偏差,其值越小越好。

## 2 结果与讨论

### 2.1 油样共聚焦拉曼光谱

图 2 为每种掺伪油样的平均共聚焦拉曼光谱。

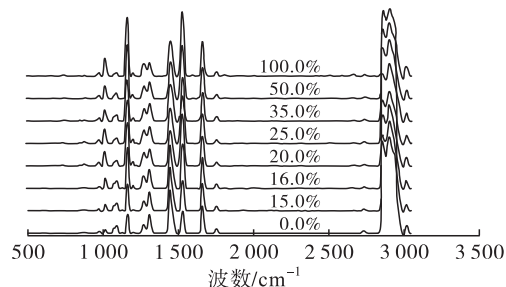


图2 不同特级初榨橄榄油掺伪油样的平均拉曼光谱

由图 2 可知,特级初榨橄榄油与低温压榨菜籽油的拉曼光谱存在一定差异,最明显的光谱差异主要集中在谱峰 1 008、1 161、1 528  $\text{cm}^{-1}$  和波段 2 800 ~ 3 000  $\text{cm}^{-1}$  内,其中 1 008、1 161、1 528  $\text{cm}^{-1}$  的 3 个拉曼特征峰主要由类胡萝卜素的甲基、碳碳单键和碳碳双键产生,2 800 ~ 3 000  $\text{cm}^{-1}$  内的拉曼特征峰主要由脂肪酸分子的甲基和亚甲基振动产生<sup>[10]</sup>。

### 2.2 基于密度泛函理论的拉曼光谱归属分析

为了对拉曼谱峰分子振动归属判别提供理论依据,本文利用 Gaussian 09 软件预测了油酸、亚油酸(特级初榨橄榄油的主要脂肪酸)以及  $\beta$ -胡萝卜素(特级初榨橄榄油类胡萝卜素的主要成分)的拉曼光谱。3 种分子的构型通过 Gaussian view 5.0 实现,采用密度泛函理论(DFT)分别优化 3 种分子模型的空间结构,使用 B3LYP/6-31+G(d,p) 基组对 3 种分子结构进行精优化并预测分子振动光谱,结果如图 3 所示。

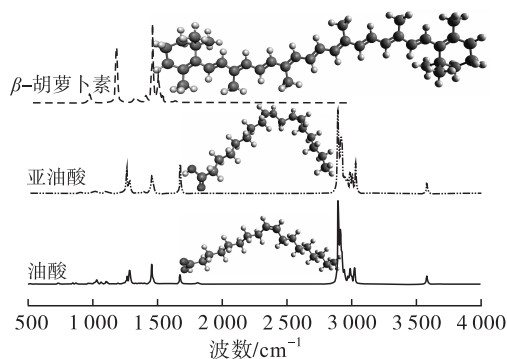


图3 3种分子的分子构型及其理论拉曼光谱

由图 3 可知,在 1 000 ~ 3 200  $\text{cm}^{-1}$  的区域内,油酸和亚油酸的分子振动光谱相似度较高,其原因主要为 2 种分子有相似度较高的官能团和化学键,通过振动谱峰归属判别发现 2 种分子的拉曼光谱峰主要由碳氢键和碳碳双键产生。 $\beta$ -胡萝卜素分子理论计算的拉曼光谱峰主要由甲基的平面摇摆振动(1 008  $\text{cm}^{-1}$ )、碳碳单键的伸缩振动(1 190  $\text{cm}^{-1}$ )

和碳碳双键的伸缩振动( $1\ 510\ \text{cm}^{-1}$ )产生。可见,理论分析与实测数据(图2)基本吻合。

### 2.3 基于PCA - SVM 掺伪定量分析模型

在构建PCA - SVM模型中,将所有数据的每种浓度的混合油样按3:1:1划分为训练集、验证集和测试集,因此可获得2 400条训练集数据、800条验证集数据以及800条测试集数据。

首先利用训练集数据构建PCA - SVM模型,图4是训练集的3个主成分的累积方差直方图。

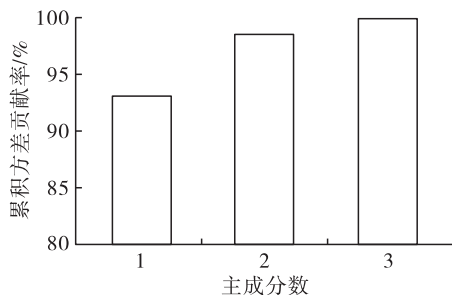


图4 累积方差直方图

由图4可以看出,PC1对拉曼光谱矩阵的方差贡献率为93.15%,PC2对拉曼光谱矩阵的方差贡献率为5.37%,PC3对拉曼光谱矩阵的方差贡献率为1.35%,前3个主成分对拉曼光谱矩阵的累积方差贡献率达99.87%。因此,混合油样训练集在PCA得分图中的分布(如图5所示)基本反映了数据特征。

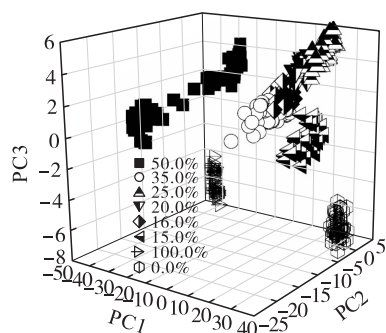
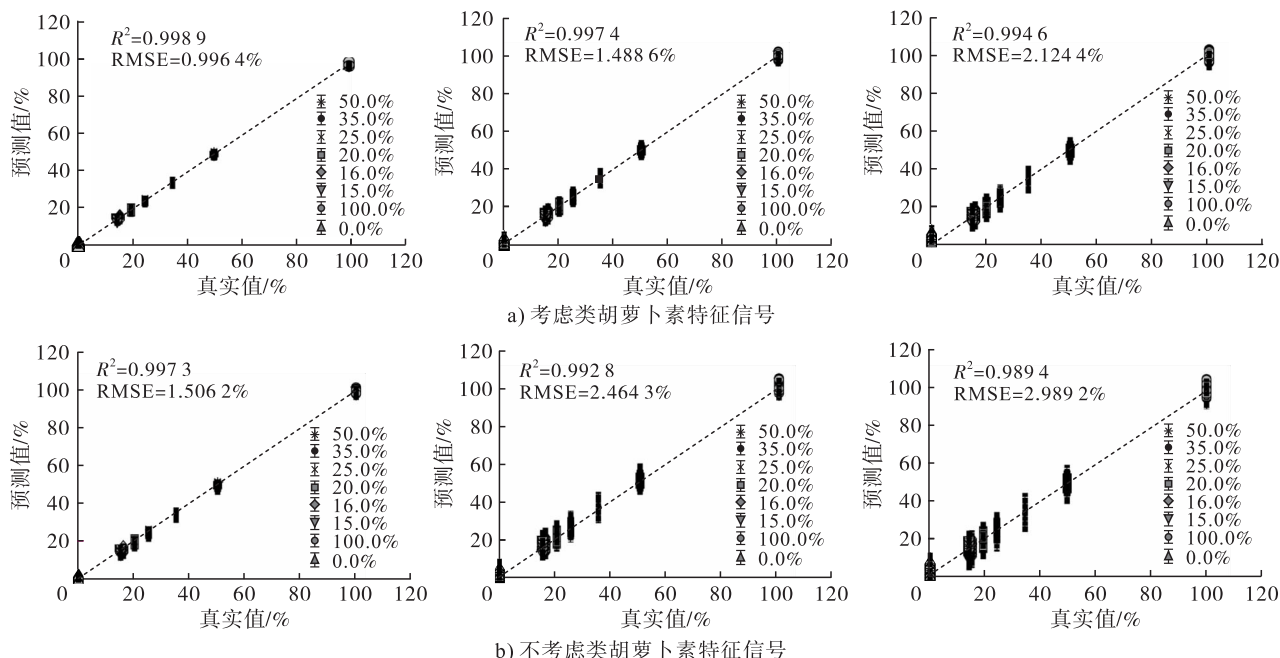


图5 PCA得分图

考虑类胡萝卜素特征信号建立的PCA - SVM模型实际值和预测值之间的关系如图6a所示。由图6a可以看出,PCA - SVM模型数据稳定分布在回归线上,即该模型较好地预测了结果。该模型构建的前提是明确使用低温压榨菜籽油(含有天然类胡萝卜素)掺入特级初榨橄榄油,但在绝大多数情况下很难确认掺入的低价值植物油是否含有类胡萝卜素,而引入的 $\beta$ -胡萝卜素的特征拉曼光谱可能会影响构建模型的预测效果。为降低 $\beta$ -胡萝卜素对模型应用的约束,本文在预处理后的拉曼光谱中剔除 $\beta$ -胡萝卜素的特征拉曼光谱峰, $\beta$ -胡萝卜素的特征峰范围在 $1\ 490 \sim 1\ 560\ \text{cm}^{-1}$ 、 $1\ 130 \sim 1\ 180\ \text{cm}^{-1}$ 和 $990 \sim 1\ 040\ \text{cm}^{-1}$ 范围内(即该谱峰范围内的权重设为0),重新构建PCA - SVM定量分析模型,结果如图6b所示。



注:图a、图b中从左到右依次为训练集、验证集和测试集

图6 PCA - SVM模型真实值和预测值之间的关系

由图6b可知,在不考虑类胡萝卜素特征信号的情况下,训练集、验证集和测试集的 $R^2$ 有所降低,而RMSE均有所升高,但模型评价指标仍在可接受的

范围内。结果说明重新构建的PCA - SVM模型的应用范围可不受低价植物油光敏色素的干扰,从而提高了定量分析模型的适用范围。

## 2.4 基于 PCA - SVM 掺伪定量分析模型的检出限

PCA - SVM 掺伪定量分析模型的检出限问题是该技术应用推广的关键因素之一。按 1.2.1 方法, 配制 12 种低掺伪浓度的混合油样, 采集其拉曼光谱, 并与纯特级初榨橄榄油进行对比, 结果如图 7 所示。

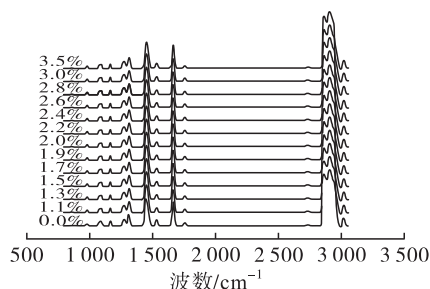


图 7 不同掺伪油样的拉曼光谱

由图 7 可知, 低掺伪浓度混合油样的拉曼光谱图差异较小, 很难通过肉眼辨别出光谱的细微差异。

基于 PCA - SVM 建立低浓度掺伪量化分析模型, 其预测效果如图 8 所示。

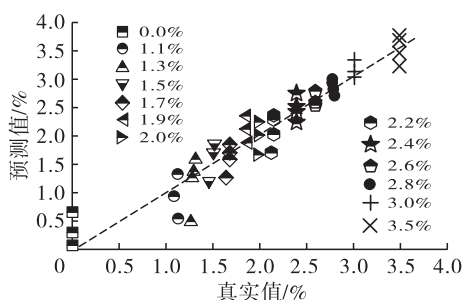


图 8 PCA - SVM 模型的检出限

由图 8 可知, 当掺伪低温压榨菜籽油的体积分数高于 2% 时, 模型具有较好的预测效果, 而低于 2% 时, 模型预测效果较差。在低浓度掺伪量化预测中, 掺入的低温压榨菜籽油浓度越低, 预测的效果越差, 这主要受两种因素的干扰: 一是低浓度掺伪所引起的光谱间差异的减小, 使得模型提取的特征相近, 在定量分析时出现边界模糊的现象; 二是低浓度掺伪的过程中, 人为制备样品时会存在一定的误差, 这种误差受样品制备者、移液枪等主客观因素的影响和干扰。

## 3 结论

本文实现了基于 PCA - SVM 的机器学习算法结合共聚焦拉曼光谱对特级初榨橄榄油掺伪低温压榨菜籽油的定量建模分析, 并讨论了类胡萝卜素特征谱峰对模型的影响及模型的适用范围。考虑类胡萝卜素时, PCA - SVM 模型测试集评价指标决定系数和均方根误差分别为 0.994 6 和 2.124 4%; 不考虑类胡萝卜素时, PCA - SVM 模型测试集评价指标决定系数和均方根误差分别为 0.989 4 和 2.989 2%。相比之下, 考虑类胡萝卜素的特征拉曼光谱峰在掺

伪定量分析中有助于提高预测精度, 但实际应用中往往受限于掺伪的低价植物油是否含有类胡萝卜素。本文研究表明, 在剔除类胡萝卜素特征拉曼光谱的情况下, 采用 PCA - SVM 模型依然可以实现较好的定量预测结果。

## 参考文献:

- [1] STARK A H, MADAR Z. Olive oil as a functional food: Epidemiology and nutritional approaches [J]. *Nutr Rev*, 2002, 60(6): 170 - 176.
- [2] 曹晨鹏, 郝仕国, 罗宁宁, 等. 基于受激布里渊散射及紫外 - 可见吸收光谱的橄榄油等级鉴别与分析 [J]. *中国激光*, 2018, 45(9): 165 - 170.
- [3] SQUEO G, GRASSI S, PARADISO V M, et al. FT - IR extra virgin olive oil classification based on ethyl ester content [J]. *Food Control*, 2019, 102: 149 - 156.
- [4] 王泓鹏, 万雄. 橄榄油的聚集荧光猝灭及吸收光谱特性的研究 [J]. *光谱学与光谱分析*, 2019, 39(8): 2421 - 2424.
- [5] LI Y, CHEN S, CHEN H, et al. Effect of thermal oxidation on detection of adulteration at low concentrations in extra virgin olive oil: Study based on laser - induced fluorescence spectroscopy combined with KPCA - LDA [J/OL]. *Food Chem*, 2020, 309: 125669 [2023 - 04 - 12]. <https://doi.org/10.1016/j.foodchem.2019.125669>.
- [6] 马金鸽, 杨巧玲, 邓晓军, 等. 基于便携式拉曼光谱的进口散装橄榄油品质现场快速无损鉴别方法 [J]. *光谱学与光谱分析*, 2021, 41(9): 2789 - 2794.
- [7] PSOOMIADOU E, TSIMIDOU M. Pigments in Greek virgin olive oils: Occurrence and levels [J]. *J Sci Food Agric*, 2001, 81: 640 - 647.
- [8] FACI M, DOUZANE M, HEDJAL M, et al. Changes in secoiridoids content and chemical characteristics of cultivated and wild Algerian olive oil, in term of fruit maturation [J/OL]. *PLoS One*, 2021, 16(11): e0260182 [2023 - 04 - 12]. <https://doi.org/10.1371/journal.pone.0260182>.
- [9] GANDUL - ROJAS B, ROCA M, GALLARDO - GUERRERO L. Chlorophylls and carotenoids in food products from olive tree [M]//BOSKOU D, CLODOVEO M. *Products from olive tree*. Rijeka, Croatia: InTech, 2016: 67 - 87.
- [10] FANG P, WANG H, WAN X. Olive oil authentication based on quantitative  $\beta$  - carotene Raman spectra detection [J/OL]. *Food Chem*, 2022, 397: 133763 [2023 - 04 - 12]. <https://doi.org/10.1016/j.foodchem.2022.133763>.
- [11] JOLLIFFE I T. *Principal component analysis* [M]. New York: Springer, 2022: 513.
- [12] HEARST M A, DUMAIS S T, OSMAN E, et al. Support vector machines [J]. *IEEE Intel Syst Appl*, 1998, 13(4): 18 - 28.