doi：10.12452/j.fxcsxb.26020201

基于能量特征融合机器学习和高丰度代谢组分定量的蓖麻子溯源区分研究

袁孟韬[1,2]，张媛圆[2]，侯　畅[1,2]，黄　旋[2]，尤　巍[2]，陈　佳[2]，谭美莲[3]，郭　磊[2*]，李开开[1*]，谢剑炜[2]

（1. 中国人民公安大学　侦查学院，北京　100038；2. 军事科学院军事医学研究院，北京　100850；3. 中国农业科学院油料作物研究所，湖北　武汉　430062）

摘　要：在蓖麻毒素相关公共安全事件的应急响应工作中，迫切需要开展蓖麻子地域来源的溯源分类工作。该研究建立了基于气相色谱-质谱的蓖麻子中脂肪酸分析方法和酶联免疫吸附分析的蓖麻毒素定量方法，对中国23个不同省份来源的100批样品进行了6种脂肪酸和蓖麻毒素的定量测定。以植物生理能量代谢特征为指导，引入3个能量特征参数(不饱和指数、油酸/亚油酸含量比值、蓖麻毒素/脂肪酸含量比值)，有效地结合了3种机器学习算法进行来源判别。模型评估表明，L1-正则化支持向量机模型在引入能量特征参数后及明确最小特征簇条件下的受试者工作特征曲线下面积分别为73.63%和71.95%，两者的测试准确率分别为70.37%和68.52%，外部验证集准确率均为75.00%，显示出良好的泛化性能。该研究证实了高丰度代谢组分在蓖麻子溯源方面的应用潜力，引入能量特征参数的机器学习为基于生物学特征的数据挖掘提供了可行路径。

关键词：蓖麻子；蓖麻毒素；脂肪酸；地理溯源；机器学习；能量特征参数

中图分类号：O657.7；R991　　**文献标识码：**A　　**文章编号：**1004-4957(2026)06-0001-08

Provenance Discrimination of Castor Beans Based on Energy Feature Fused Machine Learning and High-abundance Metabolites Quantification

YUAN Meng-tao[1,2]，ZHANG Yuan-yuan[2]，HOU Chang[1,2]，HUANG Xuan[2]，YOU Wei[2]，CHEN Jia[2]，TAN Mei-lian[3]，GUO Lei[2*]，LI Kai-kai[1*]，XIE Jian-wei[2]

（1. College of Investigation，People's Public Security University of China，Beijing　100038，China；2. Academy of Military Medical Sciences，Beijing　100850，China；3. Oil Crops Research Institute，Chinese Academy of Agricultural Sciences，Wuhan　430062，China）

Abstract：In the emergency response work for public safety incidents involving ricin，it is imperative to perform tracing and source classifications of castor beans. This research established a gas chromatography-mass spectrometric method for fatty acid determination and a sandwich enzyme-linked immunosorbent assay for ricin measurement. Afterwards it quantified six major fatty acids and ricin in 100 batches of castor bean samples collected from 23 domestic origins in China. Guided by plant psychological energy and metabolic mechanisms，it introduced three energy feature parameters，i.e.，unsaturation index，oleic/linoleic acid ratio，ricin/fatty acid ratio，and integrated with three machine learning algorithms to enhance the classification of origins. Model evaluation indicated that L1-regularized support vector machine model achieved area under the receiver operating characteristic curve(ROC-AUC) values of 73.63% with introduced energy feature parameters and 71.95% using the minimal feature cluster. The corresponding test accuracy were 70.37% and 68.52%，respectively. Both external validation accuracies reached 75.00%，demonstrating robust generalization performance. This research confirms the feasibility and application potential of high-abundance metabolites

收稿日期：2026-02-02；**修回日期：**2026-03-11
基金项目：国家重点研发计划课题(2023YFC3303905)
*通讯作者：李开开，博士，副教授，研究方向：理化物证检验，E-mail：zlkk77@163.com
郭　磊，博士，研究员，研究方向：毒物药物分析，E-mail：guolei@bmi.ac.cn

in the traceability of castor beans, and the energy-based feature integrated machine learning provides a reliable route for biological feature-based data mining.

Key words: castor beans; ricin; fatty acids; provenance discrimination; machine learning; energy feature parameters

20世纪90年代以来，蓖麻毒素(Ricin)白色粉末信件事件屡有发生，对公共安全构成严峻挑战。在蓖麻毒素相关事件处置、涉毒物证溯源及法庭证据链构建中，确定毒素来源具有重要意义。蓖麻毒素的源植物为非食用工业油料作物蓖麻，提取来源通常为蓖麻的种子—蓖麻子[1-2]。蓖麻毒素是蓖麻子中的高丰度、高毒性蛋白，占种子总重量的1%~5%，对小鼠的半数致死剂量(LD_{50}，静脉注射)为5~10 μg/kg，成年人口服8~12颗即可导致死亡[3]。在相关公共安全事件应急检测中，现有的分析检测策略多为针对蓖麻毒素自身的理化检测及活性测定，在建立物证样本与植物产地/种质来源之间关联的技术手段方面尚未形成共识，这也是事件判别和案件侦破的难点之一。

针对以上问题，2010年以来逐渐出现了针对不同品种、种质和产地的蓖麻子溯源探索。多采用一种或多种分析技术联用，例如气相色谱-质谱(GC-MS)、液相色谱-质谱、核磁共振、激光剥蚀电感耦合等离子体质谱、质谱成像等，挖掘代谢物(蓖麻碱[4]、糖及氨基酸[5]、脂肪酸[6])、微量元素[7]、蓖麻防御肽[8]等特征对区分品种或产地方面的可行性[9]。本团队前期建立了基于蓖麻防御肽的质谱成像空间分布特征分析新方法，初步揭示了中国南北和海拔分类的线索[10-11]。但以上研究的局限性在于样品来源种类较少，测定指标有限，多采用简单的化学计量学分类，无法得到解释性强的结果等，且均未进行外部验证，溯源模型的外部泛化能力不明确[4-9]。特别的，Webster等[6]曾试图利用脂肪酸指纹对蓖麻子进行产地溯源。但认为无法实现产地判别，这一结论与选择特征指标的合理性和缺乏对数据的深入挖掘直接相关。

蓖麻子中的脂肪酸以甘油三酯的形式储存，其中蓖麻油酸为主要成分，约占脂肪酸总量的80%以上，油酸、亚油酸、亚麻酸等同为不饱和脂肪酸，棕榈酸、硬脂酸则为饱和脂肪酸。另外还含有含量低于0.5%的庚酸、肉豆蔻烯酸、芥酸等微量成分。本研究以收集自23个不同省市地理来源的100份蓖麻种质资源为基础，繁种后以其种子籽粒作为研究对象，选取6种高丰度的脂肪酸(棕榈酸、硬脂酸、油酸、亚油酸、亚麻酸、蓖麻油酸等)进行GC-MS定量；同时对蓖麻毒素进行酶联免疫吸附分析(ELISA)定量测定。进一步从植物内在机制的角度引入3个能量特征参数，并关联毒素和脂肪酸含量特征，比较了最小绝对收缩和选择算子算法(LASSO)、支持向量机(SVM)及L1-正则化支持向量机(L1-SVM)3个高解释性的机器学习算法，实现了准确率较高的外部验证。置换特征重要性分析则进一步给出最小特征决定簇，可为蓖麻子的南北地理来源溯源及相关的法庭科学物证鉴定提供新的数据及理论依据。

1 实验部分

1.1 实验材料

本研究从国家油料作物种质资源中期库(https://www.cgris.net/home)引种了100份不同地理来源的、具有代表性的蓖麻种质资源样品，由中国农业科学院油料作物研究所提供，样品覆盖23个省市来源，并以中国传统南北地理分界线(秦岭-淮河一线)为基准进行划分，包括北方13个来源(黑龙江、吉林、辽宁、甘肃、新疆、陕西、内蒙古、河北、北京、天津、山西、河南、山东)、南方10个来源(江西、湖南、安徽、湖北、贵州、四川、云南、广东、广西、浙江)。所用蓖麻子样本为引种蓖麻材料繁种收获的蓖麻籽粒，样本均为干燥种子形态，置于常温下避光保存。

1.2 仪器、试剂与耗材

6890N-5975型气相色谱-质谱联用仪(配备电子轰击电离源EI，美国Agilent公司)；HH3400型多功能酶标仪(美国PerkinElmer公司)；TP-24型组织细胞破碎仪(天津杰灵仪器制造有限公司)；PC-420D加热磁力搅拌器(美国Corning公司)；N-1210旋转蒸发仪(日本EYELA公司)；3-18KS台式高速冷冻离心机(德国Sigma公司)。

HP-INNOWAX和HP-5MS色谱柱(30 m×0.25 mm，0.25 μm)购自美国Agilent公司；定性滤纸(Φ9 cm)购自杭州特种纸业有限公司；HRP标记试剂盒(MD010A)购自星宝生物科技有限公司。

7种参考品（壬酸甲酯、棕榈酸、硬脂酸、油酸、亚油酸、亚麻酸、蓖麻油酸）、三氟化硼乙醚购自上海麦克林生化科技股份有限公司；mg级蓖麻毒素参考品由本课题组自行制备[12]，单克隆抗体MIL50获赠于军事医学研究院冯建男研究员课题组，去唾液酸胎球蛋白（ASF）及吐温20（Tween 20）购自Sigma-Aldrich（美国）；磷酸盐（PBS）缓冲液（干粉）购自北京酷来搏科技有限公司；无水硫酸钠、石油醚、甲醇、正己烷、氢氧化钾、氯化钠、浓硫酸、碳酸钠及碳酸氢钠购自国药集团试剂有限公司（上海）；3，3′，5，5′-四甲基联苯胺（TMB）显色液购自北京索莱宝科技有限公司；超纯水（18.2 MΩ·cm，25 ℃）由Milli-Q Advantage A10超纯水仪（美国Millipore公司）制备。

1.3　脂肪酸组分的测定

参考《中国药典》（2025年版）一部“蓖麻油”及通则0713[12]进行脂肪酸的测定。每个来源取5粒饱满种子，去壳称重后研磨，采用索氏提取法提取蓖麻油，并经0.5 mol/L氢氧化钾-甲醇溶液（5 mL，60 ℃回流30 min）和三氟化硼乙醚-甲醇溶液（4 mL，60 ℃回流5 min）进行甲酯化衍生后，进行GC-MS分析检测。方法学验证依据《中国药典》（2025年版）四部通则9101[12]。色谱条件为HP-INNOWAX柱，氦气流速0.8 mL/min，进样体积1 μL。不分流模式，溶剂延迟4 min；程序升温为：50 ℃保持2 min，以16 ℃/min升至250℃，保持5 min。质谱条件为EI正离子模式，源温230 ℃，能量70 eV；扫描方式为全扫描模式及选择离子监测（SIM）模式，其中，全扫描模式的质量检测范围为*m/z* 40~500；SIM模式的特征目标离子为*m/z* 74。每个样品平行测定2次。

1.4　蓖麻毒素的测定

蓖麻毒素的提取步骤主要参照Cook等[13]报道的方法，并结合本实验具体条件进行改进。每个来源取5粒饱满种子，去壳称重后研磨，加入5 mL丙酮于冰浴中，振荡脱脂2 h，离心（1 000×g，5 min）弃上清液，重复脱脂2次。向得到的沉淀中加入1 mL PBS缓冲液，于4 ℃浸提过夜，离心（14 000×g，20 min）后收集上清液，即为蓖麻粗毒提取液，测定时稀释30 000倍。

制备HRP标记的检测探针，检测蛋白选用ASF。参照Xiao等[14]报道的方法进行制备，先将1.59 g碳酸钠、2.94 g碳酸氢钠溶于1 000 mL超纯水中制备碳酸盐缓冲液（CBS），再将ASF用CBS缓冲液稀释至1 g/L，与HRP标记试剂反应，4 ℃孵育过夜，调节pH值终止反应，即得HRP-ASF探针，4 ℃保存备用。

蓖麻毒素定量采用夹心ELISA[15]。以10 μg/mL MIL50包被酶标板，4 ℃过夜；用含0.5%吐温20的PBS洗涤后37 ℃封闭2 h。加入系列浓度蓖麻毒素（0.5~1 000 ng/mL，电泳纯，自行微量制备[15]）或经稀释的样品，37 ℃孵育2 h。洗涤后加入1 000倍稀释的HRP-ASF探针，37 ℃孵育1 h。洗涤后加入TMB显色液显色，37 ℃避光反应5 min，以1 mol/L硫酸终止。于450 nm测定吸光度值，根据标准曲线计算样品中的毒素浓度，每个样品设置复孔测定。方法学验证依据《中国药典》（2025年版）四部通则9101[12]。

1.5　安全操作注意事项

蓖麻毒素具有极高毒性，根据生物安全操作指南，实验过程需严格遵守实验室生物安全操作规范。操作人员全程佩戴防护口罩、护目镜及乳胶手套，并在生物安全2级环境中进行。实验产生的废液及接触毒素的固体废弃物，均需经高压灭菌处理使其失活，统一分类回收处理[16]。

1.6　数据统计、处理与机器学习算法

所有测定数据采用Origin软件（2024版，美国Origin Lab公司）进行线性回归与统计计算。数据处理方面，首先引入油脂营养品质特征参数—不饱和指数（UI）。其中，油酸和蓖麻油酸为C18∶1类型，即结构中含有1个双键；亚油酸为C18∶2类型；亚麻酸为C18∶3类型；以及引入评价油脂氧化稳定性的重要指标—油酸/亚油酸含量比值（O/L）；另自行拟定蓖麻毒素/脂肪酸含量比值（R/FAs），计算公式如下所示。

$$\mathrm{UI} = w_{\text{Oleic acid}} \times 1 + w_{\text{Linoleic acid}} \times 2 + w_{\text{Linolenic acid}} \times 3 + w_{\text{Ricinoleic acid}} \times 1 \tag{1}$$

其中w为不饱和脂肪酸在蓖麻油中的质量分数。

$$\mathrm{R/FAs} = \frac{\text{Ricin}}{\sum \text{Fatty Acids}} \tag{2}$$

$$O/L = \frac{w_{Oleic\ acid}}{w_{Linoleic\ acid}} \quad (3)$$

主成分分析(PCA)使用Python 3.9完成；机器学习采用LASSO、SVM和L1-SVM三种分类算法，均在Python 3.9环境下实现，并基于网格搜索进行参数优化。所有数据预处理阶段使用Standard Scaler进行Z-score标准化，采用按南北样品比例独立数据集划分策略：首先随机抽取10%的数据作为外部验证集；剩余90%的数据按7∶3的比例划分为训练集与测试集，最终形成训练集∶测试集∶外部验证集=63∶27∶10的划分比例。主要数值处理依托Pandas与NumPy库完成。分类模型通过Scikit-learn库(v1.5.2)构建，其中LASSO采用L1正则化实现特征选择与线性回归；SVM采用径向基函数核；L1-SVM采用L1正则化损失函数，结果可视化由Matplotlib实现。

2 结果与讨论

2.1 脂肪酸检测方法的建立及验证

蓖麻子中6种脂肪酸的结构式与分子量如图1A所示。考虑到脂肪酸的高极性及末端羧基基团易产生分子间氢键作用的特性，本研究采用三氟化硼催化的甲酯化衍生策略，将甘油三酯水解后的脂肪酸的末端羧基转化为化学性质稳定的甲酯，从而有效屏蔽氢键作用，提高分析物的挥发性与热稳定性，较好地适用于GC-MS分析检测。

以硬脂酸作为代表性化合物，采用内标法评价了该甲酯化的衍生化效率[12]，内标选用壬酸甲酯。得到硬脂酸甲酯的加标回收率为98.7%±1.2%，证明了该衍生化步骤具有较好的转化效率，满足美国分析化学家协会Appendix F的要求[17]。

基于衍生化得到的6种脂肪酸甲酯，建立了相应的GC-MS方法。首先对HP-5MS与HP-INNOWAX两种色谱柱的分离性能进行比较(图1B)。HP-5MS色谱柱上C_{18}型的4种脂肪酸甲酯均未实现基线分离，原因在于分析物和中等极性固定相间的作用较为相似；而HP-INNOWAX色谱柱的固定相为聚乙二醇，实现了6种脂肪酸甲酯的基线分离，原因可能为分析物的甲酯、不同数量的不饱和双键和聚乙二醇固定相间产生了不同程度的弱氢键、偶极-偶极作用。其中硬脂酸(C18∶0)、油酸(C18∶1)、亚油酸(C18∶2)、亚麻酸(C18∶3)的保留时间与不饱和度呈正相关。6种脂肪酸甲酯在0.1~25 μg/mL范围内线性关系良好(r^2为0.993~0.998)，定量下限为0.1~1.0 μg/mL(油酸甲酯为0.1 μg/mL，亚油酸甲酯为0.2 μg/mL，其余4种脂肪酸的定量下限均为1.0 μg/mL)；各分析物在低、中、高3个浓度水平下的精密度(以相对标准偏差(RSD)表示，为0.36%~13%)与准确度(以加标回收率表示，为94.2%~112%)均符合定量分析的方法学要求。

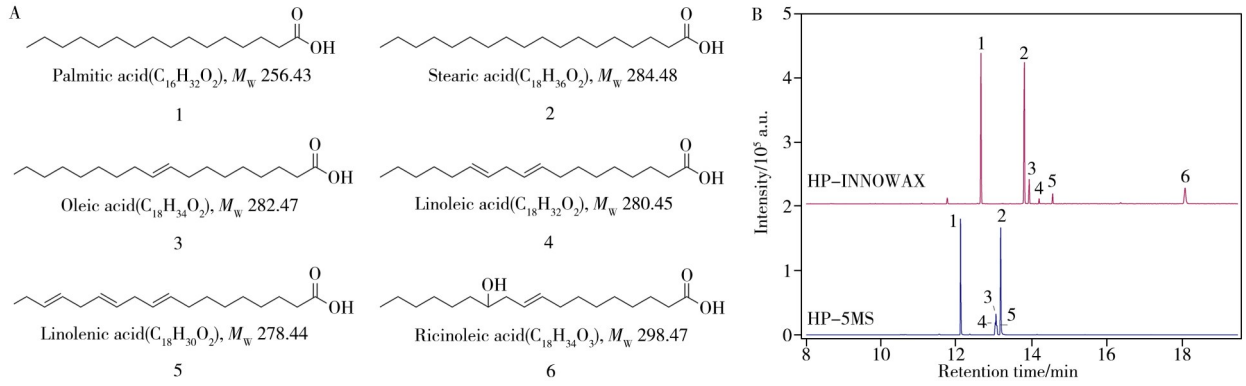

图1 6种脂肪酸的结构式(A)；6种脂肪酸甲酯的GC-MS色谱图(SIM模式，*m/z* 74)(B)

Fig. 1 Chemical structures of six fatty acids(A)；GC-MS chromatograms of six fatty acid methyl esters(SIM mode，*m/z* 74)(B)

2.2 蓖麻毒素ELISA检测方法的验证

首先采用所建立的ELISA法测定蓖麻毒素参考品，绘制标准曲线。该方法在10~400 ng/mL范围内线性良好(r^2=0.997)，定量下限为10 ng/mL。精密度RSD为4.5%(<10%)，准确度(以加标回收率表示)为113%±4.97%，表明方法准确、重复性好，符合定量分析要求。

2.3　样品中脂肪酸及蓖麻毒素含量测定

基于前述建立的脂肪酸及蓖麻毒素含量测定方法，分别对100份蓖麻子样品进行了定量测定。统计分析结果显示(图2A)，脂肪酸中以蓖麻油酸为主要成分(含量为86.69%±5.33%)，同时含有棕榈酸、硬脂酸、油酸、亚油酸、亚麻酸等其他组分。蓖麻毒素的含量为3.18%±1.11%，上述7种分析物在100份样品中基本呈正态分布，但各组分在不同样品间的含量存在差异，提示样品自身的种质或来源可能对差异造成一定影响。PCA(图2B)表明，23个蓖麻子样品未见明显区分，和现有文献的结论[6]基本一致。

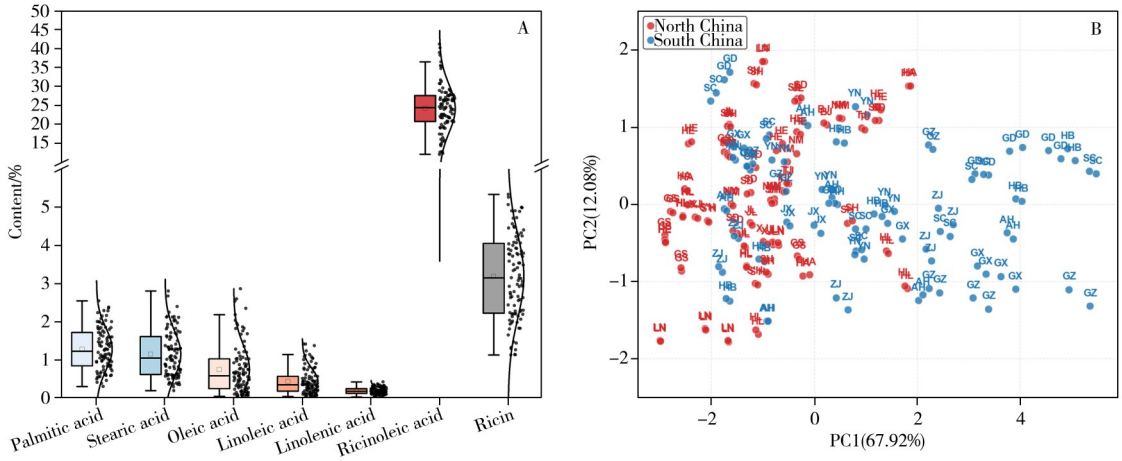

图2　100批次样品中脂肪酸与蓖麻毒素的含量分布(A)；主成分分析图(B)

Fig. 2　Distribution of fatty acid and ricin contents in 100 batches of samples(A)；principal component analysis plot(B)

2.4　基于机器学习的分类模型构建

2.4.1　基于测定数据集的机器学习模型　探讨了机器学习在来源溯源分类中的可行性。Elferjani等[18]在针对油菜生殖发育期的研究中揭示，油脂作物的含油量受温度生境变化的影响与调控，高温导致含油量显著下降并改变了脂肪酸的组成比例，油酸和饱和脂肪酸的相对含量升高，而多不饱和脂肪酸(主要是亚油酸和亚麻酸)的含量降低。植物生态学研究表明，植物在长期适应特定地理生境(如北方低温或南方高温)的过程中，其脂质代谢途径会发生适应性进化，通过调节不饱和脂肪酸比例以维持膜流动性与能量平衡，进而形成具有稳定生理特征的地理生态型。考虑到23个地理来源蓖麻子在中国境内广泛分布，在此仅确定与温度因素相关性强的南北方来源作为主要分类目标。

选择合适的机器学习模型非常重要。针对特征维度为7、200例样本定量数据的小数据集，过于简单的模型易欠拟合，过于复杂的模型则解释性差且易过拟合。参考Zhao等[19]提出的模型选择策略，本研究选择擅长特征筛选的LASSO模型(高解释性、低计算性能)、能有效挖掘非线性关系的SVM模型(低解释性、高计算性能)和兼顾解释性与泛化能力的L1-SVM嵌入式特征选择模型，以在模型可解释性与计算性能之间取得平衡。另外前期还使用了基于决策树的随机森林算法，但参数优化后仍存在大量的过拟合结果，因此，未将随机森林算法纳入进一步研究。3种机器学习算法的超参数如表1所示，其分类性能如表2所示。

表1　机器学习模型的超参数值

Table 1　Hyperparameter values of machine learning models

Category	Model	Hyperparameter
Linear model	LASSO	C=0.5；penalty=L1；solver=liblinear；class_weight=None；random_state=42
Kernel-based model	SVM	C=0.1；γ=0.1；kernel=RBF；probability=True；class_weight=balanced；random_state=42
Hybrid model	L1-SVM	Cselector=0.05；Csvm=0.2；γsvm=0.2；class_weight=balanced；random_state=42

针对6种脂肪酸的LASSO、SVM、L1-SVM模型呈现一定的分类效果，测试集准确率为中等水平(64%~70%)，但综合性能较差。仅依靠蓖麻毒素单一指标进行分类时，LASSO、SVM和L1-SVM的判别能力均有限。综合蓖麻毒素与6种脂肪酸组合时，LASSO仅体现了脂肪酸为主的分类特征；SVM兼顾了蓖麻毒素与脂肪酸特征的筛选，测试集准确率提高至64.81%，无明显过拟合；L1-SVM的测试集准确

率则略有提高，为68.52%；另外调优时发现，与线性核相比，使用非线性核的L1-SVM模型性能更佳。

表2 七维数据组的3种机器学习模型性能(单位：%)

Table 2 The performance of three machine learning models via seven-dimension original data(Unit：%)

Dataset	Model	Accuracy	Precision	F1_Score	Recall	ROC-AUC	Accuracy_Gap	ROC-AUC_Gap
Ricin	LASSO	59.26	53.85	38.89	30.43	51.82	-0.53	8.58
	SVM	57.41	0.00	0.00	0.00	47.89	0.53	-8.29
	L1-SVM	51.85	45.16	51.85	60.87	53.51	6.08	7.72
FAs	LASSO	70.37	65.22	65.22	65.22	70.41	3.43	1.63
	SVM	64.81	56.25	65.45	78.26	69.64	1.06	4.87
	L1-SVM	66.67	58.62	65.38	73.91	70.41	7.93	13.03
Ricin_FAs_	LASSO	66.67	63.16	57.14	52.17	69.85	3.97	6.01
combined	SVM	64.81	57.69	61.22	65.22	73.77	1.85	4.57
	L1-SVM	68.52	60.00	67.92	78.26	71.25	12.43	18.54

本研究进一步分析了LASSO、SVM、L1-SVM模型在6种脂肪酸、单一蓖麻毒素、蓖麻毒素与6种脂肪酸组合的外部验证集上的分类性能(表3)，发现LASSO、SVM、L1-SVM模型的外部验证表现波动较大，缺乏稳定的泛化能力。以上结果提示，毒素和脂肪酸组合的数据集适用于南北方来源判别，但仍需进行深入的数据挖掘。

表3 七维数据组的3种机器学习模型外部验证(单位：%)

Table 3 External validation of three machine learning models via seven-dimension original data(Unit：%)

Dataset	Model	Accuracy	Precision	F1-Score	Recall	ROC-AUC
Ricin	LASSO	65.00	66.67	36.36	25.00	66.67
	SVM	60.00	0.00	0.00	0.00	33.33
	L1-SVM	60.00	50.00	60.00	75.00	68.75
FAs	LASSO	65.00	55.56	58.82	62.50	70.83
	SVM	60.00	50.00	55.56	62.50	72.92
	L1-SVM	60.00	50.00	55.56	62.50	78.13
Ricin_FAs_combined	LASSO	75.00	71.43	66.67	62.50	75.00
	SVM	70.00	62.50	62.50	62.50	73.96
	L1-SVM	60.00	50.00	60.00	75.00	80.21

2.4.2 基于能量特征融合的机器学习模型 蓖麻毒素是蓖麻为适应生存压力进化出的主动防御型次生代谢产物，其合成过程中，蓖麻子储存的甘油三酯经分解代谢提供能量和碳源，驱动蓖麻毒素的转录、翻译、加工及储存全流程，其中高含量的蓖麻油酸作为持续供能主力，而油酸/亚油酸具有高氧化效率，则支撑合成启动阶段的快速供能[20]。因此，本研究在数据集中引入3个能量特征参数，分别为：不饱和指数，代表油脂中的不饱和脂肪酸占比，植物通过调节不饱和脂肪酸的含量与双键数量，保证能量供给，维持细胞膜的流动性和稳定性，进而适应高低温胁迫[21]；蓖麻毒素与脂肪酸的比值，量化植物在次生防御代谢(高能耗毒素合成)与主要能量储备(脂肪酸积累)之间的关系[22]；油酸/亚油酸含量比值，是油脂氧化稳定性的核心指标，反映蓖麻子在抗逆环境下的能量供给持续性[23-24]。

引入能量特征参数后形成了十维特征的数据集。机器学习模型性能汇总如表4所示，与未含能量特征的数据集的模型参数对比，LASSO模型的测试集性能略有提升；SVM模型的准确率提升至66.67%，且其召回率(Recall)、F1值和ROC曲线下面积(ROC-AUC)均有所提升；L1-SVM模型的准确率提升至70.37%，且ROC-AUC值提升至73.63%。

表4 基于能量特征融合的机器学习模型性能(单位：%)

Table 4 The performance of energy feature fused machine learning models(Unit：%)

Model	Accuracy	Precision	F1_Score	Recall	ROC-AUC	Accuracy_Gap	ROC-AUC_Gap
LASSO	70.37	66.67	63.64	60.87	70.27	1.06	5.90
SVM	66.67	60.00	62.50	65.22	73.49	4.76	6.63
L1-SVM	70.37	66.67	66.67	60.87	73.63	7.41	14.61

进一步考察了该数据集的混淆矩阵和学习曲线(图3)，L1-SVM和LASSO展现出更好性能。学习曲线上，L1-SVM和LASSO的训练集和验证集的差值分别为0.024和0.015；SVM模型的表现相对较弱，其训练集和验证集的差值为0.059。结合表3及混淆矩阵结果可以发现，引入能量特征后，LASSO与L1-SVM模型具有较好的判别能力。

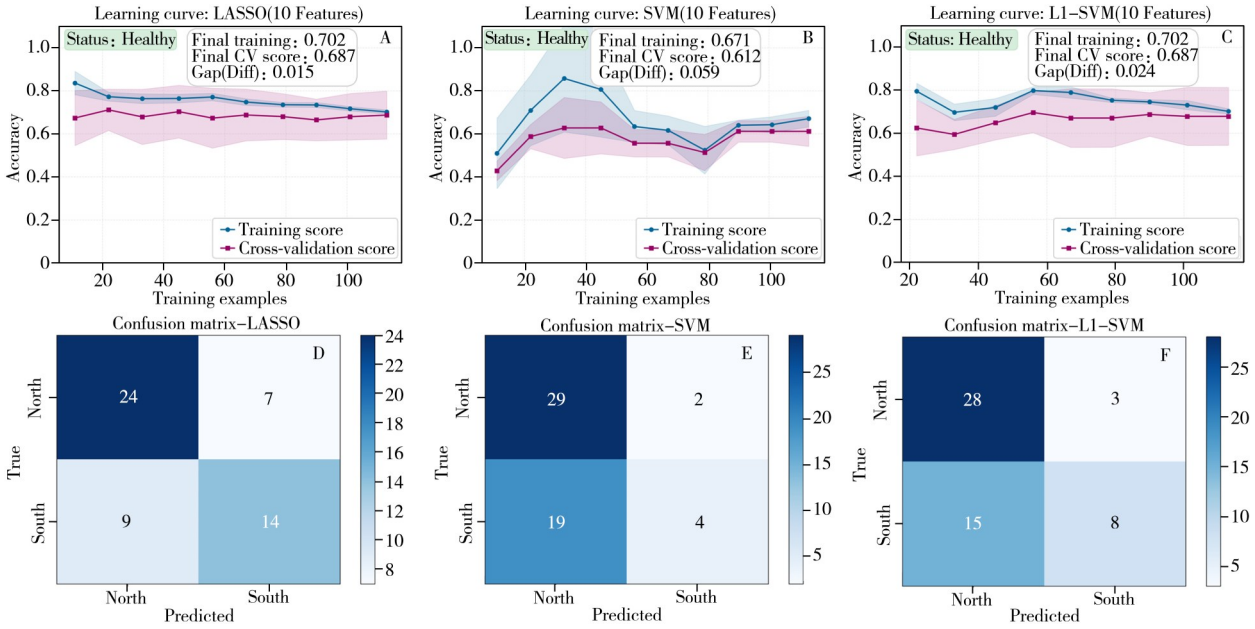

图 3　基于能量特征融合的 3 种机器学习模型的学习曲线(A~C)与混淆矩阵(D~F)

Fig. 3　Learning curves(A-C) and corresponding confusion matrices(D-F) of three machine learning models based on energy feature fusion

A. LASSO，B. SVM，C. L1-SVM，D. LASSO，E. SVM，F. L1-SVM

验证了上述模型在未知样本上的可迁移性，LASSO、SVM、L1-SVM 模型的外部验证结果如表 5 所示。L1-SVM 模型的准确率达 75.00%，ROC-AUC 为 78.13%，F1 值与精确度均有所提升；相较于未含能量特征参数时的波动状态，引入能量特征参数后的模型能够解决泛化不稳定问题。LASSO、SVM 模型的外部验证结果亦均有提升，说明所引入的能量特征参数对模型分类能力产生了正向影响。

表 5　基于能量特征融合的机器学习模型外部验证(单位：%)

Table 5　External validation of energy feature fused machine learning models(Unit：%)

Model	Accuracy	F1-Score	Precision	Recall	ROC-AUC
LASSO	75.00	66.67	71.43	62.50	72.92
SVM	70.00	62.50	62.50	62.50	72.92
L1-SVM	75.00	66.67	71.43	62.50	78.13

2.4.3　最小特征簇的机器学习模型性能评估　本研究在性能和外部验证最佳的 L1-SVM 模型中使用了非线性核，因此通过置换特征重要性(PFI)参数进行了特征筛选，通过计算随机打乱单一特征后模型准确率的下降幅度来评估各特征对整体分类性能的贡献。图 4 显示，蓖麻毒素含量在模型中均占主导地位，表现出最高的特征重要性，是核心分类指标。此外，不饱和指数被识别为第二重要的特征；油酸含量和亚油酸含量在模型中排名靠前，显示了其作为分类标志物的潜力；其余特征影响范围小于 0.03，对分类预测的贡献较低。因此，构建了由蓖麻毒素含量、不饱和指数、油酸含量和蓖麻油酸含量组成的最小特征簇。

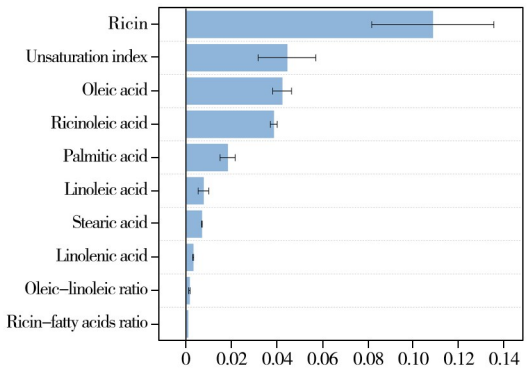

图 4　基于能量特征融合的 L1-SVM 机器学习模型置换特征重要性分析

Fig. 4　Permutation feature importance analysis of the L1-SVM machine learning model based on energy feature fusion

随后，应用 L1-SVM 模型对该最小特征簇进行训练。在测试集的性能评估中(表 6)，与能量融合后全部十维特征的数据集相比，最小特征簇的 L1-SVM 模型保持较高的准确度，F1 值保持不变。此外在外部验证中(表 6)，其准确率、F1 值与召回率均未改变，说明 L1-SVM 模型在四维特征簇与十维特征数据集中能保持较好的泛化能力。该四维特征簇具备进一步推广应用价值。

表6 基于最小特征簇的L1-SVM模型性能、外部验证结果汇总(单位：%)

Table 6 Summary of L1-SVM model performance based on minimal feature cluster combination and external validation results(Unit：%)

Subset category	Accuracy	Precision	F1-Score	Recall	ROC-AUC	Accuracy-Gap	ROC-AUC_Gap
Test set	68.52	63.64	62.22	60.87	71.95	3.70	8.98
External validation set	75.00	66.67	71.43	62.50	78.13	-	-

3 结 论

基于所建立的脂肪酸GC-MS分析检测方法与蓖麻毒素的ELISA定量方法，本研究实现了对中国23个省市来源的100份蓖麻子样品的定量测定。对200组7种高丰度代谢物含量所构成的七维特征数据集以及引入3种能量特征后的十维特征数据集，分别进行了LASSO、SVM和L1-SVM的机器学习数据挖掘，融合能量特征后的机器学习算法在蓖麻子南北溯源中展现出较好的综合性能。进一步筛选构建了4个高贡献度特征(蓖麻毒素、不饱和指数、油酸和亚油酸)构成的最小特征簇，其在L1-SVM模型中仍保持较高的准确率和较好的泛化能力。但由于植物内在的生理能量代谢较为复杂，能量特征参数和地缘特征间的相关性不宜直接引申为因果关系。总体而言，引入能量特征参数的机器学习有效拓展了蓖麻子溯源归因技术的数据挖掘思路，利用最小特征簇进行溯源亦具有较好的可行性，在涉蓖麻类投毒案件的原料地理来源判定等相关公共安全事件处置和司法鉴定溯源方面具有一定的实际推广价值。

参考文献：

[1] Ogunniyi D S. *Bioresour. Technol.*, **2006**, 97(9): 1086-1091.
[2] Patel V R, Dumancas G G, Viswanath L C K, Maples R, Subong B J J. *Lipid Insights*, **2016**, 9: 1-12.
[3] Bolt H M, Hengstler J G. *Arch. Toxicol.*, **2023**, 97(4): 909-911.
[4] Ovenden S P B, Gordon B R, Bagas C K, Muir B, Rochfort S, Bourne D J. *Aust. J. Chem.*, **2010**, 63(1): 8-21.
[5] Pigott E J, Roberts W, Ovenden S P B, Rochfort S, Bourne D J. *Metabolomics*, **2012**, 8(4): 634-642.
[6] Webster L R, Ovenden P S, Yousef J. *Forensic Sci. Int. Rep.*, **2020**, 2: 100127.
[7] Bagas C K, Scadding R L, Scadding C J, Watling R J, Roberts W, Ovenden S P B. *Forensic Sci. Int.*, **2017**, 270: 46-54.
[8] Ovenden S P B, Pigott E J, Rochfort S, Bourne D J. *Phytochem. Anal.*, **2014**, 25(5): 476-484.
[9] Tres A, Ruiz-Samblas C, Veer D V G, Ruth V M S. *Food Chem.*, **2013**, 137(1/4): 142-150.
[10] Qin L Y, Han J S, Wang C, Xu B, Tan D Y, He S, Guo L, Bo X C, Xie J W. *Front. Plant Sci.*, **2022**, 13: 1083901.
[11] He W W, Wang C Y, Yang J W, Xu B, Guo L, Xie J W. *J. Instrum. Anal.*(何唯唯, 王晨钰, 杨捷威, 徐斌, 郭磊, 谢剑炜. 分析测试学报), **2021**, 40(4): 543-550.
[12] Chinese Pharmacopoeia Commission. *Pharmacopoeia of the People's Republic of China*. Beijing: China Medical Science and Technology Press(国家药典委员会. 中华人民共和国药典. 北京：中国医药科技出版社), **2025**.
[13] Cook D L, David J, Griffiths G D. *Toxicology*, **2006**, 223(1/2): 61-70.
[14] Xiao L, Luo L, Liu J, Liu L Y, Han H, Xiao R, Guo L, Xie J W, Tang L. *Toxins*, **2024**, 16(7): 312.
[15] Luo L, Yang J W, Li Z, Xu H, Guo L, Wang Y X, Luo L L, Wang J, Zhang P P, Yang R F, Kang W J, Xie J W. *Talanta*, **2022**, 238(P1): 122860.
[16] Centers for Disease Control and Prevention, National Institutes of Health. *Biosafety in Microbiological and Biomedical Laboratories*. 6th ed. Washington, D.C.: U.S. Department of Health and Human Services, **2020**: 343-357.
[17] AOAC International. *Appendix F: Guidelines for Standard Method Performance Requirements*. Official Methods of Analysis of AOAC International. 21st ed. Rockville, MD: AOAC International, **2019**.
[18] Elferjani R, Soolanayakanahally R. *Front. Plant Sci.*, **2018**, 9: 1224.
[19] Zhao C, Zhao T, Shen Q, Tang Z, Li X M, Huan T. *Anal. Chem.*, **2025**, 97(7): 2300-2313.
[20] Wu Z T, Xu F, Yu L L, Ouyang Y, Geng X X. *Biol. Plant.*, **2021**, 65: 273-282.
[21] Upchurch R G. *Biotechnol. Lett.*, **2008**, 30(6): 967-977.
[22] Huot B, Yao J, Montgomery L B, He S Y. *Mol. Plant*, **2014**, 7(8): 1267-1287.
[23] De Souza-Vieira Y, Felix-Mendes E, Valente-Almeida G, Felix-Cordeiro T, Corrêa R L, Jardim-Messeder D, Sachetto-Martins G. *Plants*, **2025**, 14(8): 1256.
[24] Browse J. *Vitam. Horm.*, **2005**, 72: 431-456.

(责任编辑：丁 岩)