

文章编号 1004-924X(2025)20-3192-11

改进黑翅鸢算法优化的XGBoost可解释模型在转基因棉籽油太赫兹光谱鉴别中的应用

陈涛*, 赵利

(桂林电子科技大学 电子工程与自动化学院, 广西 桂林 541004)

摘要: 为实现对转基因和非转基因棉籽油的准确分类鉴别, 本研究提出一种基于改进黑翅鸢算法优化极端梯度提升(XGBoost)模型的可解释分类模型。首先, 应用太赫兹时域光谱(THz-TDS)系统采集转基因和非转基因棉籽油样品在0.3~1.8 THz频段的太赫兹吸收光谱。然后, 通过引入双目标适应度函数优化策略、反向学习初始化种群策略和瑞利分布函数控制Lévy飞行策略对传统黑翅鸢算法(BKA)进行改进, 并利用改进的黑翅鸢算法(DLBKA)对XGBoost模型的树深度、学习率和最大迭代次数进行双目标超参数优化, 构建出DLBKA-XGBoost分类模型。最后, 应用该模型对转基因棉籽油进行鉴别, 并结合SHAP方法对模型鉴别结果进行了可解释性分析。结果表明, 改进黑翅鸢算法优化的XGBoost可解释分类模型不仅提升了对转基因和非转基因棉籽油鉴别的准确率(其测试集准确率达97.78%, 较传统黑翅鸢算法优化模型提升了4.45%, 较传统鲸鱼算法(WOA)优化模型提升了14.45%), 还对模型给出了解释, 明确了关键特征频率对鉴别结果的正向影响机制, 提升了模型的透明度与可信度。因此, 本研究为转基因棉籽油的鉴别提供了一种快速准确的分析方法, 也为其他转基因物质的鉴别提供了有价值的参考。

关键词: 太赫兹光谱; 转基因棉籽油; 极端梯度提升; 改进黑翅鸢算法; 可解释性分析

中图分类号: O439 文献标识码: A

doi: 10.37188/OPE.20253320.3192

CSTR: 32169.14.OPE.20253320.3192

Application of XGBoost explainable model improved by black-winged kite algorithm optimization in the identification of genetically modified cotton seed oil by terahertz spectroscopy

CHEN Tao*, ZHAO Li

(School of Electronic Engineering and Automation, Guilin University of Electronic Technology,
Guilin 541004, China)

* Corresponding author, E-mail: tchen@guet.edu.cn

Abstract: To achieve accurate classification and identification of genetically modified and non-genetically modified cottonseed oil, this study proposes an explainable classification model based on an improved black-winged kite algorithm optimized extreme gradient boosting (XGBoost) model. First, a terahertz time-domain spectroscopy (THz-TDS) system was used to collect terahertz absorption spectra of genetically modified and non-genetically modified cottonseed oil samples in the 0.3-1.8 THz frequency range.

收稿日期: 2025-06-25; 修订日期: 2025-08-17.

基金项目: 国家自然科学基金资助项目(No. 62261012, No. 62261012)

Then, the traditional Black-winged Kite algorithm (BKA) was improved by introducing a dual-objective fitness function optimization strategy, a reverse learning initial population strategy, and a Rayleigh distribution function to control the Lévy flight strategy. The improved Black-winged Kite algorithm (DLBKA) was used to perform dual-objective hyperparameter optimization of the tree depth, learning rate, and maximum iteration count of the XGBoost model, thereby constructing the DLBKA-XGBoost classification model. Finally, the model was applied to identify genetically modified cottonseed oil, and the model's identification results were analyzed for interpretability using the SHAP method. The results showed that the improved Black-winged Kite Algorithm-optimized XGBoost interpretable classification model not only improved the accuracy of identifying genetically modified and non-genetically modified cottonseed oil (with a test set accuracy as high as 97.78%, an improvement of 4.45% over the traditional Black-winged Kite algorithm-optimized model, an improvement of 14.45% over the traditional Whale Optimization Algorithm (WOA)-optimized model), but also provided explanations for the model, clarifying the positive influence mechanism of key feature frequencies on the identification results, thereby enhancing the model's transparency and credibility. Therefore, this study provides a fast and accurate analytical method for the identification of genetically modified cottonseed oil and offers valuable references for the identification of other genetically modified substances.

Key words: Terahertz spectroscopy; genetically modified cottonseed oil; extreme gradient boosting; improved black-winged kite algorithm; explainability analysis

1 引言

棉花作为全球重要的纤维作物,在80多个国家广泛种植,年产量达2 500万吨左右。其副产品棉籽是重要的食用油原料,从中提取的棉籽油(Cottonseed Oil, CS-O)兼具优异的烹饪适应性与营养特性:其高达50%的亚油酸含量显著优于传统食用油(如橄榄油约30%),而亚油酸作为人体必需的不饱和脂肪酸,可抑制冠状动脉硬化;且棉籽油的光稳定性和高烟点使其适用于高温烹饪场景^[1]。其无毒且营养均衡的特性,使其成为全球饮食结构中的重要油脂来源。

随着基因组编辑技术的发展,转基因棉花通过定向修饰获得抗逆、抗病虫害等性状^[2],但其副产品转基因棉籽油的安全性引发关注。外源基因插入可能导致宿主产生毒性物质、过敏原或抗营养成分^[3],存在潜在的过敏、中毒风险及长期遗传效应^[4]。当前检测技术中,基于核酸^[5]和蛋白质^[6]的方法仅能靶向已知转基因成分^[7],高通量测序(High-Throughput Sequencing, HTS)虽具有全面性但受限于设备成本高、操作复杂等问题,难以满足实际检测需求^[8]。因此,亟须建立一种准确、高效且经济的转基因棉籽

油鉴别方法。

太赫兹波(0.1~10 THz,对应波长3 mm~30 μ m)位于电磁频谱的红外(Infrared Spectroscopy, IR)与微波(Microwave, MW)频段的过渡区域,兼具穿透能力强、光子能量低、光谱带宽宽等交叉特性^[9],基于该特性的太赫兹时域光谱(Terahertz Time-Domain Spectroscopy, THz-TDS)技术作为当前最成熟的太赫兹技术^[10],已在转基因植物油检测中展现应用潜力。Zhou等^[11]将太赫兹光谱与极限学习机结合,对转基因大豆油进行了成功鉴别;Chen等^[12]通过优化支持向量机,实现了对转基因菜籽油的有效分类。然而,现有研究对转基因棉籽油的鉴别关注较少,且鉴别模型存在鉴别准确度偏低,尤其是可解释性差的问题。针对这一问题,本研究拟引入极端梯度提升模型(Extreme Gradient Boosting, XGBoost)对转基因和非转基因棉籽油进行鉴别。XGBoost作为集成学习算法,在表格分类任务中优势显著^[13],但模型性能非常依赖超参数的优化。为此,本研究采用黑翅鸢算法(Black-Winged Kite Algorithm, BKA)优化其超参数,并针对传统BKA易陷局部最优的缺陷,通过引入双目标适应度函数、反向学习策略初始化种群和

瑞利分布函数控制 Lévy 飞行策略对其进行改进,形成改进的黑翅鸢算法(Double-Objective Lévy Flight Black-Winged Kite Algorithm, DLBKA)^[14]。最后,将DLBKA应用于XGBoost超参数优化,构建出DLBKA-XGBoost模型,结合太赫兹光谱实现对转基因棉籽油的快速准确鉴别。同时,本研究还利用SHAP方法对模型鉴别结果进行了可解释性分析,可视化模型特征贡献,揭示了模型决策逻辑。

2 实验部分

2.1 样品制备

本研究所用非转基因棉籽油与转基因棉籽油均符合现行国家标准,且通过官方正规渠道采购以保障样本可靠性,其品种与来源信息具体如下:非转基因棉籽油购自上海阿拉丁生化科技股份有限公司,品种为中棉 113;转基因棉籽油购自吉安市锦煜生物科技有限公司,品种为携带 Bt 抗虫基因的鲁棉研 37 号。同种棉籽油样品均来自同一批次,且所有样品在实验前均储存于室温避光环境中,以抑制氧化反应及光降解过程对样品成分的影响,确保其化学性质稳定。鉴于聚四氟乙烯在太赫兹波段(0.1~10 THz)的低吸收特性,实验采用由聚四氟乙烯材料制成的可拆卸式液体样品池作为光谱采集载体。制样时,使用注射器沿池壁缓慢注入 1 mL 油样,避免气泡产生。注液完成后立即密封样品池,确保样品均匀填充通光孔区域。按照上述方法,对每种棉籽油独立制备 225 个样品,累计获得 450 个实验样品。

2.2 数据采集

本研究采用的太赫兹光谱实验测量设备为中国华讯方舟科技有限公司生产的 CCT-1800 型 THz-TDS 系统,系统原理图如图 1 所示,该系统主要由飞秒激光器、时间延时控制模块、太赫兹波产生装置以及太赫兹波探测装置 4 个部分构成。数据采集前,向系统内通入干燥氮气 15 分钟,去除空气中的水蒸气,降低水蒸气对太赫兹光谱的吸收干扰。

对于每种棉籽油中的每一个样品,使用 THz-TDS 系统获取太赫兹光谱数据,每种棉籽油获取 225 组数据,其中 180 组作为训练集用于

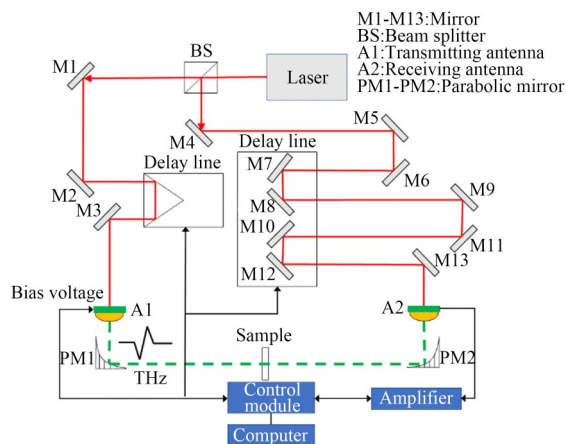


图 1 CCT-1800 型 THz-TDS 系统原理图

Fig. 1 Schematic diagram of CCT-1800 THz-TDS system

模型参数学习,45 组作为测试集评估模型泛化能力,两种棉籽油共计 450 组数据,并采用太赫兹吸光度进行光谱数据分析,其定义如下:

$$Absorbance(\omega) = -\log_{10} \left| \frac{E_{sam}(\omega)}{E_{ref}(\omega)} \right|^2, \quad (1)$$

其中: $E_{ref}(\omega)$ 为频域参考信号, $E_{sam}(\omega)$ 为频域样品信号, ω 为角频率。

2.3 模型评价指标

混淆矩阵是可视化二分类模型分类结果的核心工具,其具体包含 4 个基础指标:真正例(True Positive, TP):转基因油样被正确鉴别为转基因的样品数量;假正例(False Positive, FP):非转基因油样被误判为转基因的样品数量;真负例(True Negative, TN):非转基因油样被正确鉴别为非转基因的样品数量;假负例(False Negative, FN):转基因油样被误判为非转基因的样品数量。

基于上述指标,衍生出 6 个模型评价指标,即准确率(Accuracy)、精确率(Precision)、召回率(Recall), F1 分数(F1-Score)、特异度(Specificity)和 AUC(Area Under the ROC Curve)值。其数学定义分别为^[15]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (2)$$

$$Precision = \frac{TP}{TP + FP}, \quad (3)$$

$$Recall = \frac{TP}{TP + FN}, \quad (4)$$

$$F1-Score = \frac{Precision \times Recall}{Precision + Recall}, \quad (5)$$

$$Specificity = \frac{TN}{TN + FP}, \quad (6)$$

其中, AUC 值通过计算受试者工作特征曲线 (Receiver Operating Characteristic, ROC) 下的面积得到, 它可以综合评估模型在不同分类阈值下的泛化能力, 计算步骤如下:

(1) 对模型输出的预测概率按降序排序。

(2) 计算各阈值下的真正例率 (True Positive Rate, TPR) 和假正例率 (False Positive Rate, FPR):

$$\begin{cases} TPR = \frac{TP}{TP + FN} \\ FPR = \frac{FP}{FP + TN} \end{cases}, \quad (7)$$

其中, TP, FP, TN, FN 均随阈值变化而动态调整。

(3) 采用梯形积分法计算面积:

$$AUC = \frac{\sum_{i=1}^n (FPR_i - FPR_{i-1})(TPR_i + TPR_{i-1})}{2}, \quad (8)$$

其中, $FPR_0 = TPR_0 = 0, FPR_n = TPR_n = 1$ 。

3 分类模型

3.1 极端梯度提升模型

极端梯度提升模型 (XGBoost) 作为机器学习领域的代表性模型, 本质上是通过连续残差拟合生成多个决策树的提升集成算法。该算法由华盛顿大学机器学习专家陈天奇提出, 其核心思想源于对传统梯度提升树 (Gradient Boosting Decision Tree, GBDT) 模型的系统性改进^[16]。该算法的基本原理如下:

对于包含 m 个样品的棉籽油吸收光谱数据 $D = \{(x_i, y_i)\} (|D| = m, x_i \in R^n, y_i \in R)$, 树集成模型使用 K (树的数量) 个基础模型组合来预测输出最终结果。该模型对第 i 个样品的预测值可表示为:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i). \quad (9)$$

XGBoost 的预测精度是由偏差和方差共同决定的, 其中, 损失函数反映模型的偏差。为了控制方差并简化模型, 添加了正则项以抑制模型复杂性。XGBoost 算法的目标函数由损失函数

和正则项构成, 如式 (10) 所示:

$$\begin{cases} Obj^{(t)} = \sum_i l(\hat{y}_i, y_i) + \sum_i \Omega f_k \\ \Omega(f) = \gamma T + \frac{1}{2} \lambda(\omega)^2 \end{cases}, \quad (10)$$

其中: $Obj^{(t)}$ 是第 t 轮迭代模型的目标函数, l 是平衡预测值 $\hat{y}_i$ 和真实值的 y_i 可微凸损失函数。 Ω 是避免过拟合的正则项, T 是叶节点数量, γ 是叶节点数的正则化参数, ω 是叶节点权值, λ 是叶节点权值的正则化参数。

在实际应用中, 虽然 XGBoost 通过上述优化机制展现出强大性能, 但其性能受树深度 (max_depth)、学习率 (learning_rate) 和最大迭代次数 (n_estimators) 3 个超参数的影响显著^[17]。手动调参不仅需要深厚的专业知识储备, 过程耗时耗力, 且极易陷入次优解。因此, 本研究采用黑翅鸢算法对 XGBoost 的树深度、学习率和最大迭代次数进行优化, 以提升模型在转基因棉籽油鉴别任务中的准确性与泛化能力。

3.2 黑翅鸢算法

黑翅鸢算法是 2024 年由 Wang 等人受黑翅鸢生物行为启发提出的群体智能优化算法^[14]。该算法通过模拟黑翅鸢的狩猎策略与迁徙机制, 结合柯西分布与动态领导权更替策略, 实现全局搜索能力与收敛效率的协同优化。算法主要包含攻击行为和迁徙行为两部分:

攻击行为: 模拟黑翅鸢狩猎过程, 通过概率选择机制执行不同攻击模型, 平衡全局探索与局部开发:

$$X_i(t+1) = \begin{cases} X_i(t) + n(1 + \sin r)X_i(t), & p < r \\ X_i(t) + n(2r - 1)X_i(t), & \text{else} \end{cases}, \quad (11)$$

$$n = 0.05e^{-2\left(\frac{t}{Max_{it}}\right)^2}, \quad (12)$$

其中: t 和 $t+1$ 表示迭代次数, $X_i(t)$ 是第 t 次迭代的第 i 个解, $X_i(t+1)$ 是第 $t+1$ 次迭代的第 i 个解, r 是 $[0, 1]$ 随机数, 常数 $p=0.9$, Max_{it} 表示最大迭代次数。 n 是随迭代次数衰减的系数, 其值由公式 (12) 定义, 其随着迭代次数 t 的增加呈指数衰减, 使得算法在前期侧重全局探索, 后期逐步聚焦局部开发。

迁徙行为: 模仿黑翅鸢迁徙中的领导权更替

机制,根据适应度动态调整解的位置:

$$X_i(t+1)=\begin{cases} X_i(t)+C(\eta)(X_i(t)-L(t)), & F_t < F_{ri} \\ X_i(t)+C(\eta)(L(t)-mX_i(t)), & \text{else} \end{cases}, \quad (13)$$

$$m=2\sin\left(r+\frac{\pi}{2}\right), \quad (14)$$

$$C(\eta)=\frac{1}{\pi(\eta^2+1)}, -\infty < \eta < +\infty, \quad (15)$$

其中: $X_i(t)$ 是第 t 次迭代的第 i 个解, $X_i(t+1)$ 是第 $t+1$ 次迭代的第 i 个解, m 由公式(14)定义, $L(t)$ 是种群在第 t 次迭代得到的全局最优解, r 是 $[0,1]$ 随机数。 F_t 为当前解的适应度, F_{ri} 为当前种群内随机解的适应度。 $C(\eta)$ 由公式(15)定义符合柯西分布,能产生较大的随机扰动,有助于算法在迁徙过程中跳出局部最优解。

3.3 改进的黑翅鸢算法

传统黑翅鸢算法易陷入局部最优的问题,本研究通过双目标适应度函数、反向学习初始化及瑞利分布函数控制 Lévy 飞行策略 3 项策略对其进行改进,以提升算法全局寻优能力。

3.3.1 双目标适应度函数优化策略

本研究构建双目标适应度函数实现关键指标的协同优化,具体如下:

(1)分类准确率主导优化:将分类准确率设为首要优化目标,通过模型超参数空间搜索,筛选出准确率最高的超参数组合作为候选解集;

(2)AUC 值辅助决策:考虑到分类准确率相同的超参数组合可能存在模型泛化能力的差异,针对候选解集中准确率相同的超参数组合,引入 AUC 值作为次优指标,选取 AUC 值最大的超参数配置作为最终优化结果。

优化流程遵循分层决策机制:首先基于分类准确率对模型超参数进行初步筛选,形成候选解集;随后在候选空间内,以 AUC 值为评判标准进行二次择优,确定全局最优超参数配置。

3.3.2 反向学习初始化种群策略

为有效提升初始种群的质量与多样性,本研究引入反向学习机制,通过构建双重种群的方式进行种群初始化,具体实施步骤如下:

随机种群生成:基于问题的解空间,采用均匀分布随机采样策略,生成规模为 S 的初始种群 $X=\{X_1, X_2, \dots, X_S\}$,其中 X_i 表示第 $i(i=$

$1, \dots, S)$ 个候选解。

反向解构造:依据反向学习原理,对初始种群中的每个个体 X_i ,通过如下映射关系构造其反向解 $\hat{X}_i$:

$$\hat{X}_i=Lb+Ub-X_i, \quad (16)$$

其中: $\hat{X}_i$ 表示 X_i 的反向解, Lb, Ub 表示寻优超参数的上界和下界;

精英种群筛选:将初始种群 X 与反向学习生成的种群 $\hat{X}=\{\hat{X}_1, \hat{X}_2, \dots, \hat{X}_S\}$ 进行合并,形成规模为 $2S$ 的混合种群。随后,基于预先定义的双目标适应度函数,从混合种群中筛选出前 S 个综合性能最优的个体,构成最终的初始种群,完成种群初始化过程。

3.3.3 瑞利分布函数控制 Lévy 飞行策略

为增强算法在解空间的全局探索与局部开发能力,本研究在迁徙行为中引入 Lévy 飞行机制,并结合瑞利分布函数进行动态调控,构建如下改进迁徙行为模型:

$$X_i(t+1)=\begin{cases} X_i(t)+C(\eta)(X_i(t)-L(t))+ \\ r(t)levy(\varphi), & F_t < F_{ri} \\ X_i(t)+C(\eta)(L(t)-mX_i(t))+ \\ r(t)levy(\varphi), & \text{else} \end{cases}, \quad (17)$$

其中: $X_i(t)$ 是第 t 次迭代的第 i 个解, $X_i(t+1)$ 是第 $t+1$ 次迭代的第 i 个解, F_t 为当前解的适应度, F_{ri} 为当前种群内随机解的适应度, $L(t)$ 是种群完成在第 t 次迭代后的全局最优解, $C(\eta)$ 由公式(15)定义符合柯西分布, r 是 $[0,1]$ 随机数。

上述模型中核心改进参数定义如下:

Lévy 飞行机制: $levy(\varphi)$ 遵循 Lévy 分布,通过幂律步长特性赋予算法跨越局部最优的长程搜索能力,有效拓展解空间探索范围^[18]。

瑞利分布函数控制:

$$r(t)=\frac{t}{\sigma^2}e^{-\frac{t^2}{\sigma^2}}, \sigma=1. \quad (18)$$

该函数呈现前期(t 值较小)高强度扰动特性,促使算法快速探测解空间未知区域;随着迭代进程推进, $r(t)$ 呈指数衰减趋势,逐步降低搜索随机性,驱动算法向最优解方向精细收敛。

通过双目标函数平衡分类精度与模型泛化

能力,反向学习初始化拓展初始解空间以增强种群多样性,以及 Lévy 飞行结合瑞利调控策略在迭代前期利用长距离跳跃跳出局部最优,后期通过扰动衰减确保收敛精度。三项改进策略协同作用,显著提升算法在复杂优化问题中的求解效能。由此构建得到的 DLBKA-XGBoost 模型,解决了 BKA 在 XGBoost 超参数寻优过程中容易陷入局部最优解的问题,增强了对 XGBoost 最优超参数的搜索精度,提升了模型的整体性能。

3.3.4 算法时间复杂度分析

BKA 的时间复杂度主要与种群规模(S)、最大迭代次数(Max_{it})、问题维度(d ,即 XGBoost 关键超参数数量)及单次适应度评估时间($Time$)有关,则 BKA 的时间复杂度为 $O(S \times Max_{it} \times (d + Time))$ 。

DLBKA 的双目标适应度函数、反向学习初始化种群及瑞利分布函数控制 Lévy 飞行 3 项策略均不改变原算法的时间复杂度量级:反向学习初始化种群策略为一次性初始化操作,筛选排序复杂度远小于后续迭代开销。单轮迭代中攻击行为与改进的迁徙行为(瑞利分布函数控制 Lévy 飞行)复杂度均为 $O(S \times d)$,无额外维度或计算步骤,单轮搜索总复杂度与 BKA 保持一致。双目标适应度仅增加常数项时间 $Time_{AUC}$,仍属 $O(S \times Time)$ 量级。

DLBKA 总时间复杂度最终简化为 $O(S \times Max_{it} \times (d + Time + Time_{AUC}))$,与 BKA 的时间复杂度量级完全一致。其仅通过少量常数项开销,实现超参数寻优能力提升,达成复杂度与寻优能力平衡,适配转基因棉籽油太赫兹光谱鉴别的高效优化需求。

3.4 模型可解释性分析

由于机器学习模型的“黑盒”特性,解释模型输出至关重要。为了进一步分析特征频率在预测模型中的重要性,本研究采用 SHAP(Shapley Additive exPlanations)方法来评估关键特征频率在 DLBKA-XGBoost 分类模型中的贡献。

SHAP 方法是由 Lundberg 和 Lee 于 2017 年提出,用于解释机器学习模型的输出^[19]。SHAP 方法基于博弈论和局部解释,提供了量化特征贡献的方法。在 SHAP 中,每个特征(ϕ_i 表示特征 i 的贡献)对模型预测结果的贡献是根据其边际贡

献分配的。SHAP 方法可以使用线性解释模型和 Shapley 值来估计 DLBKA-XGBoost 分类模型。SHAP 的解释模型如下:

$$h(Z) = \phi_0 + \sum_{i=1}^M \phi_i Z_i, \quad (19)$$

其中: $h(Z)$ 表示解释模型的输出, M 表示输入特征的数量, ϕ_i 表示特征 i 对预测结果的贡献, Z 表示简化输入(其中 Z_i 通常是一个二元变量,表示特征 i 是否存在于解释模型的输入中), ϕ_0 表示基准值(所有特征都不存在时的解释模型输出)。特征贡献 ϕ_i 通过下式计算:

$$\phi = \sum_{K \subseteq N \setminus \{i\}} \frac{|K|!(N - |K| - 1)!}{N!} \cdot [p_x(K \cup \{i\}) - p_x(K)], \quad (20)$$

$$p_x(K) = E[p(x) | x_K], \quad (21)$$

其中: N 表示所有输入特征的集合, K 表示不包含特征 i 的任意子集, $E[p(x) | x_K]$ 表示仅使用子集 K 的模型预测值。

该方法通过枚举所有可能的特征子集组合,综合考虑每个特征加入不同子集时对预测结果的影响,确保每个特征的贡献分配满足公平性和局部一致性原则。

4 结果及分析

本研究的数据分析、模型构建及结果可视化均在统一环境完成:硬件为配备英特尔®酷睿™i9-13900HX 处理器(基础频率 2.20 GHz)及 16 GB 内存的笔记本电脑,操作系统为 Windows 11,软件基于 Python 3.10 开发。优化算法层面,传统黑翅鸢算法(BKA)、改进黑翅鸢算法(DLBKA)及鲸鱼优化算法(WOA)均为自主编程实现。工具库应用上,通过 XGBoost(v2.0.3)库搭建分类模型,利用 SHAP(v0.47.2)库量化特征贡献度;借助 Scikit-learn(v1.6.1)库完成数据集划分、指标计算与验证;通过 Matplotlib(v3.10.0)库实现混淆矩阵、全局 SHAP 解释图等可视化,保证实验规范与结果可复现。

4.1 光谱分析

通过使用 THz-TDS 系统获取 2 种棉籽油共计 450 个样品的太赫兹光谱数据,2 种棉籽油及参考信号的太赫兹时域光谱如图 2 所示。图中

Reference 表示参考信号,为系统稳定后实验舱中样品架空载时的测量值,GMO(Genetically Modified Oil)表示转基因棉籽油样品信号,non-GMO(non-Genetically Modified Oil)表示非转基因棉籽油样品信号。

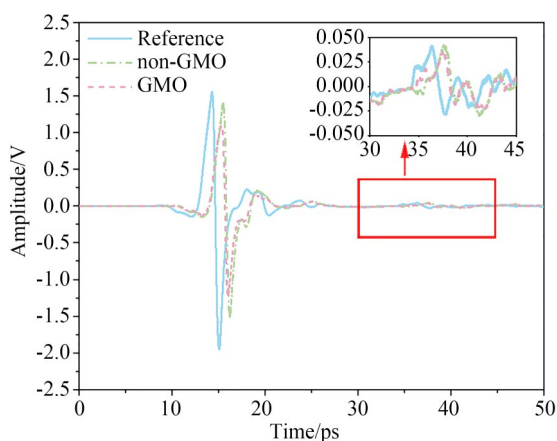


图2 两种棉籽油及参考信号的太赫兹时域光谱

Fig. 2 THz time-domain spectra of two rapeseed oils and reference signal

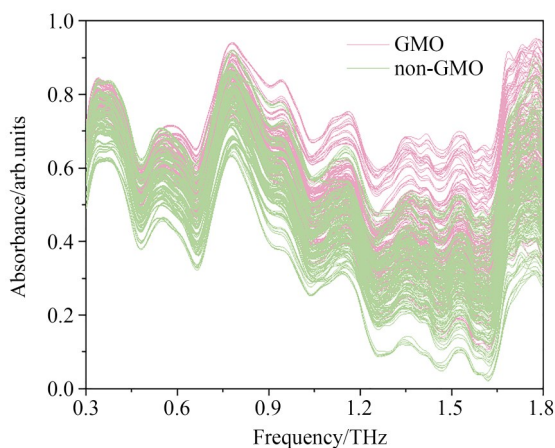


图3 450个棉籽油样品在0.3~1.8 THz波段的太赫兹吸收光谱

Fig. 3 THz absorption spectra of 450 rapeseed oil samples in 0.3~1.8 THz

为抑制 THz-TDS 系统中由太赫兹波在平行界面样品内部多次反射引起的法布里-珀罗(Fabry-Perot, FP)效应,本研究采用时域截断法进行数据预处理:通过截取时域信号主脉冲段(图2中第一个回波出现位置33 ps之前的波段),再对截断后的信号进行傅里叶变换,以消除FP效应导致的频域光谱周期性振荡,确保最终得到的太赫

兹吸收光谱能真实反映样品本身的光学特性^[20]。通过计算获得450个棉籽油样品在0.3~1.8 THz频段的太赫兹吸收光谱如图3所示。由图可见,所有棉籽油样品在0.3~1.8 THz频段呈现出相似的波形和相近的幅值,无显著差异。这说明转基因棉籽油和非转基因棉籽油的太赫兹光谱差异非常小,采用直接观察法很难对其进行鉴别,因此需要结合机器学习方法来实现对转基因和非转基因棉籽油的快速有效鉴别。

4.2 超参数寻优及模型鉴别

为确保对比实验的公平性与结果可靠性,选取鲸鱼优化算法^[21](Whale Optimization Algorithm, WOA)、黑翅鸢算法(BKA)及改进的黑翅鸢算法(DLBKA)3种优化算法开展实验,统一设置种群规模为30、最大迭代次数为100,每种算法重复运行10次后取分类最优结果。实验中,在训练集中分别采用WOA、BKA和DLBKA对XGBoost模型进行超参数寻优,获取最佳的树深度(max_depth)、学习率(learning_rate)和最大迭代次数(n_estimators),优化过程中WOA、BKA以准确率作为寻优评价指标,DLBKA以准确率和AUC值作为其寻优评价指标,XGBoost采用五折交叉验证,XGBoost模型的超参数寻优结果如表1所示。由表可知,WOA总运行时间最短,BKA次之,DLBKA总运行时间最长。

表1 XGBoost的超参数寻优结果

Tab. 1 XGBoost hyperparameter optimization results

优化算法	树深度	学习率	最大迭代次数	总运行时间/s
WOA	2	0.147	14	768
BKA	2	0.480	192	887
DLBKA	4	0.424	87	1 012

将WOA、BKA和DLBKA的最优超参数寻优结果分别代入XGBoost中,并对测试集进行识别,最终得到WOA-XGBoost模型、BKA-XGBoost模型和DLBKA-XGBoost模型的分类结果混淆矩阵如图4所示。由图可见,采用WOA-XGBoost模型的测试集准确率为83.33%,其预测结果中存在15个误判,分别将7个GMO样品误判为non-GMO样品,将8个non-GMO样品误判为GMO样品,所得模型的精确率为82.61%,

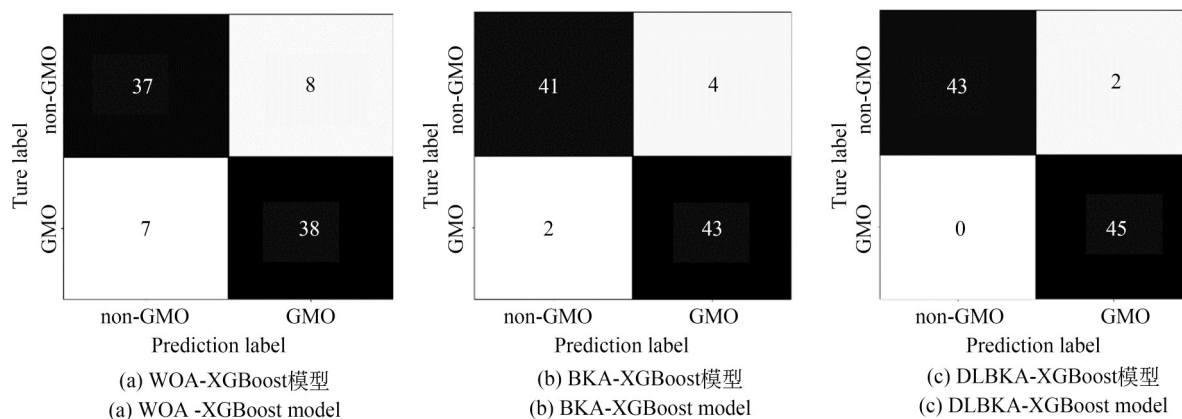


图 4 三种模型的分类结果混淆矩阵

Fig. 4 Confusion matrix of classification results for the three models

召回率为 84.44%, F1 分数为 83.52%, 特异度为 82.22%, AUC 值为 0.868 6。BKA-XGBoost 模型的测试集准确率为 93.33%, 其预测结果中存在 6 个误判, 分别将 2 个 GMO 样品误判为 non-GMO 样品, 将 4 个 non-GMO 样品误判为 GMO 样品, 所得模型的精确率为 91.49%, 召回率为 95.56%, F1 分数为 93.48%, 特异度为 91.11%, AUC 值为 0.987 7。DLBKA-XGBoost 模型的测试集识别准确率为 97.78%, 其预测结果存在两个误判, 将两个 non-GMO 样品误判为 GMO 样品。所得模型的精确率为 95.74%, 召回率为 100%, F1 分数为 97.83%, 特异度为 95.56%, AUC 值为 0.997 0。三种模型的多指标性能对比如图 5 所示。

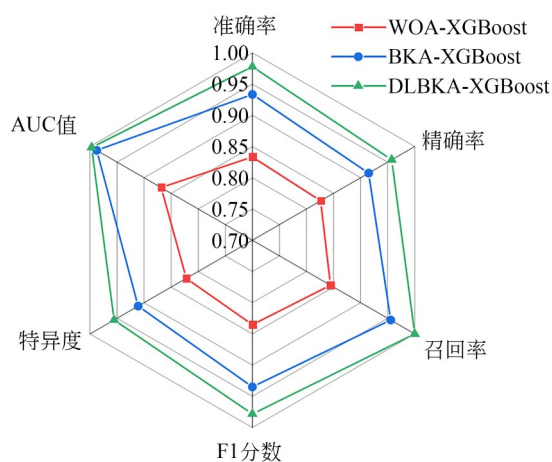


图 5 三种模型多指标性能对比

Fig. 5 Comparison of multi-indicator performance of three models

结果表明, 尽管 DLBKA 的总运行时间略长于 WOA 与 BKA, 但 DLBKA-XGBoost 模型在各项评价指标上均显著优于另外两种模型。这说明 DLBKA 能有效避免超参数寻优过程中陷入局部最优解, 增强全局搜索能力, 进而显著提升模型的性能。

4.3 模型可解释性分析

在开发预测模型时, 高预测准确性的实现固然关键, 但明晰每个预测结果背后的成因, 同样是模型开发中不可忽视的环节。由于机器学习模型复杂的“黑盒”性质, 对其输出结果进行解释往往颇具挑战。为了进一步分析特征频率在预测模型中的作用机制, 本研究采用 SHAP 方法来评估关键特征频率在最佳预测模型中的贡献。图 6 呈现了全局 SHAP 解释结果, 图 7 展示了对应特征重要性分布。综合分析图 6 与图 7 可知, 0.307 THz 时的吸收系数对棉籽油预测模型的影响最显著, 其次是 0.389 THz 和 0.522 THz。这些频率下的吸收系数特征呈现显著正向预测效应, 即特征值升高时模型更倾向于判定为转基因棉籽油。从概率维度具体来看, 在上述频率下, 随着吸收系数的增大, 预测模型判定样品为转基因棉籽油的概率显著提升。

本研究的 SHAP 分析结果显示, 显著影响预测模型的关键特征频率主要集中在 0.3~0.8 THz 范围内, 其中 0.307 THz, 0.389 THz 和 0.522 THz 处的特征对模型决策的正向贡献尤为突出。这一发现不仅明确了特定太赫兹频段

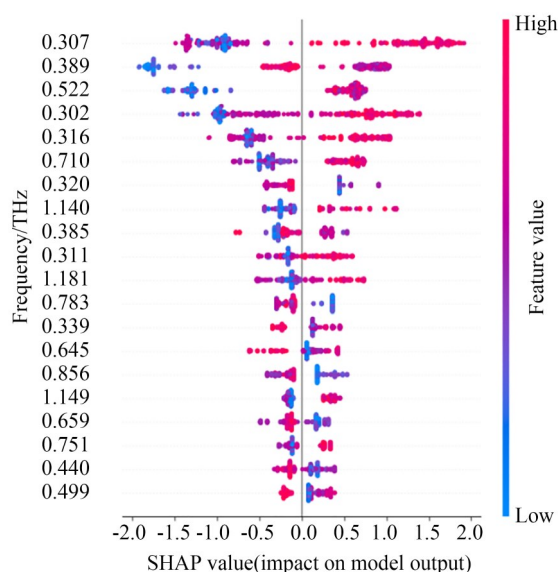


图6 全局SHAP解释

Fig. 6 Global SHAP explanation

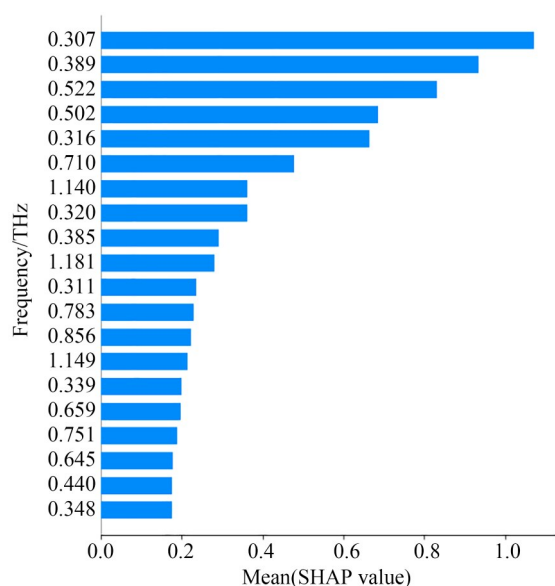


图7 全局SHAP解释的特征重要性

Fig. 7 Feature importance in global SHAP interpretation

在提升模型预测准确性中的关键作用,还揭示了0.3~0.8 THz太赫兹频段与转基因棉籽油分子结构差异的潜在关联性,即该频段的吸收系数变化极有可能是基因修饰引发化学键振动模式改变、分子构象差异的外在光谱表征(太赫兹频段

主要对应分子间低频振动跃迁)。因此,未来的研究可进一步结合红外光谱、拉曼光谱等技术,构建多频段光谱联合分析体系,从分子官能团角度深入解析0.3~0.8 THz特征的物质基础,进一步阐明太赫兹光谱特征与转基因棉籽油固有属性的作用机制。

5 结 论

本研究采用THz-TDS技术研究了转基因和非转基因棉籽油的太赫兹光谱,发现转基因与非转基因样品的光谱波形及幅值无显著差异,通过直接观察法无法有效鉴别,故引入WOA,BKA与DLBKA优化的XGBoost模型进行分类鉴别。结果表明,DLBKA-XGBoost模型测试集准确率达97.78%,较传统WOA-XGBoost模型准确率83.33%提升14.45%,误判数从WOA-XGBoost模型的15例降至2例,较常规BKA-XGBoost模型93.33%的准确率提升了4.45%,误判数从BKA-XGBoost模型的6例降至2例,且精确率、F1分数及AUC值均更优,其100%的转基因样品召回率证实DLBKA有效避免了局部最优解,增强了全局搜索能力;另外,应用SHAP方法分析表明0.307 THz,0.389 THz和0.522 THz处的吸收系数为棉籽油预测模型的关键影响因素,0.3~0.8 THz频段内的吸收系数对预测模型有显著正向影响,为太赫兹光谱在转基因检测中的应用提供了频率靶向依据,验证了特定频段对提升模型可解释性与预测精度的关键作用。因此DLBKA-XGBoost可解释分类模型为转基因棉籽油的分类鉴别提供了一种快速准确的分析方法,也为其他转基因油脂的快速检测提供了新的技术途径。

作者贡献声明:

陈涛:方法的提出,论文构思,指导实验测量数据分析,论文编辑写作与审核;

赵利:光谱测量实验的设计及数据整理和分析,论文撰写。

参考文献:

[1] RIAZ T, IQBAL M W, MAHMOOD S, *et al.*

Cottonseed oil: a review of extraction techniques, physicochemical, functional, and nutritional proper-

- ties[J]. *Critical Reviews in Food Science and Nutrition*, 2023, 63(9): 1219-1237.
- [2] THANGARAJ A, KAUL R, SHARDA S, *et al.* Revolutionizing cotton cultivation: a comprehensive review of genome editing technologies and their impact on breeding and production[J]. *Biochemical and Biophysical Research Communications*, 2025, 742: 151084.
- [3] ZHANG C, WOHLHUETER R, ZHANG H. Genetically modified foods: a critical review of their promise and problems[J]. *Food Science and Human Wellness*, 2016, 5(3): 116-123.
- [4] KUMAR K, GAMBHIR G, DASS A, *et al.* Genetically modified crops: current status and future prospects[J]. *Planta*, 2020, 251(4): 91.
- [5] WANG J B, WANG Y, HU X W, *et al.* A dual-RPA based lateral flow strip for sensitive, on-site detection of CP4-EPSPS and Cry1Ab/Ac genes in genetically modified crops[J]. *Food Science and Human Wellness*, 2024, 13(1): 183-190.
- [6] WANG Y F, BEDNARCIK M, AMENT C, *et al.* Immunoassays and mass spectrometry for determination of protein concentrations in genetically modified crops[J]. *Journal of Agricultural and Food Chemistry*, 2024, 72(16): 8879-8889.
- [7] DEBODE F, HULIN J, CHARLOTEAUX B, *et al.* Detection and identification of transgenic events by next generation sequencing combined with enrichment technologies[J]. *Scientific Reports*, 2019, 9: 15595.
- [8] HE S T, FAN Y Y, TAO S M, *et al.* Application of next-generation sequencing in the detection of transgenic crop[J]. *Frontiers in Genetics*, 2024, 15: 1461115.
- [9] LI Q X, LEI T, CHENG Y L, *et al.* Predicting wheat gluten concentrations in potato starch using GPR and SVM models built by terahertz time-domain spectroscopy[J]. *Food Chemistry*, 2024, 432: 137235.
- [10] YU J X, PU H B, SUN D W. Stacked long and short-term memory (SLSTM) - assisted terahertz spectroscopy combined with permutation importance for rapid red wine varietal identification[J]. *Talanta*, 2025, 291: 127650.
- [11] ZHOU S L, ZHU S P, LIU G H, *et al.* Rapid detection of transgenic soybean oils by terahertz (THz) spectroscopy[J]. *Journal of Nanoelectronics and Optoelectronics*, 2017, 12(9): 956-960.
- [12] 陈涛, 李欣. 太赫兹光谱在转基因菜籽油鉴别中的应用: 基于改进蜉蝣算法的支持向量机模型[J]. *物理学报*, 2024, 73(5): 366-374.
- CHEN T, LI X. Application of terahertz spectroscopy in identification of transgenic rapeseed oils: a support vector machine model based on modified mayfly optimization algorithm[J]. *Acta Physica Sinica*, 2024, 73(5): 366-374. (in Chinese)
- [13] SHWARTZ-ZIV R, ARMON A. Tabular data: deep learning is not all you need[J]. *Information Fusion*, 2022, 81: 84-90.
- [14] WANG J, WANG W C, HU X X, *et al.* Black-winged kite algorithm: a nature-inspired meta-heuristic for solving benchmark functions and engineering problems[J]. *Artificial Intelligence Review*, 2024, 57(4): 98.
- [15] GEZIMATI M, SINGH G. Deep learning for multimodal breast cancer characterization with emergence of terahertz and infrared imaging[J]. *IEEE Transactions on Instrumentation and Measurement*, 2025, 74: 2511514.
- [16] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C]. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA. ACM*, 2016: 785-794.
- [17] SUN Z, LI Y L, YANG Y X, *et al.* Splitting tensile strength of basalt fiber reinforced coral aggregate concrete: Optimized XGBoost models and experimental validation[J]. *Construction and Building Materials*, 2024, 416: 135133.
- [18] ZHANG Z, WANG X K, CAO L. FOX optimization algorithm based on adaptive spiral flight and multi-strategy fusion[J]. *Biomimetics*, 2024, 9(9): 524.
- [19] LUNDBERG S M, LEE S A. Unified approach to interpreting model predictions[C]. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA*, 2017: 4765-4774.
- [20] 苏海霞, 张朝晖, 赵小燕, 等. THz-TDS测试中时域脉冲的本质段及其截取模型[J]. *光谱学与光谱分析*, 2013, 33(4): 921-925.
- SU H X, ZHANG Z H, ZHAO X Y, *et al.* Intrinsic section and its interception model for temporal

- pulse of THz-TDS [J]. *Spectroscopy and Spectral Analysis*, 2013, 33(4): 921-925. (in Chinese)
- [21] CHEN C W, YANG H Q, LI X Y, *et al.* Hyper-spectral estimation method for deterioration of rock carvings in the humid regions of Southern China [J]. *Heritage Science*, 2024, 12(1): 105.

作者简介:



陈 涛(1984—),男,广西桂林人,博士,教授,博士生导师,2013年于西安电子科技大学获得博士学位,主要从事太赫兹科学技术及应用方面的研究。E-mail:tchen@guet.edu.cn