渔业科学进展
Progress in Fishery Sciences
ISSN 2095-9869,CN 37-1466/S

《渔业科学进展》网络首发论文

题目：基于脂肪酸谱和机器学习的野生鲢鉴别方法研究
作者：秦雷，钮冰，吕继洲，陈沁
DOI：10.19663/j.issn2095-9869.20250915001
收稿日期：2025-09-15
网络首发日期：2026-01-05
引用格式：秦雷，钮冰，吕继洲，陈沁．基于脂肪酸谱和机器学习的野生鲢鉴别方法研究[J/OL]．渔业科学进展. https://doi.org/10.19663/j.issn2095-9869.20250915001

DOI: 10.19663/j.issn2095-9869.20250915001　　http://www.yykxjz.cn/

秦雷, 钮冰, 吕继洲, 陈沁. 基于脂肪酸谱和机器学习的野生鲢鉴别方法研究. 渔业科学进展, 2026, 47
QIN L, NIU B, LÜ J Z, CHEN Q. Discrimination of wild *Hypophthalmichthys molitrix* based on fatty acid profiling and machine learning. Progress in Fishery Sciences, 2026, 47

基于脂肪酸谱和机器学习的野生鲢鉴别方法研究*

秦　雷[1]　钮　冰[1]　吕继洲[2①]　陈　沁[1①]

(1. 上海大学生命科学学院　上海　200444；2. 中国质量检验检测科学研究院　北京　100176)

摘要　我国自 2020 年起实施长江“十年禁渔”计划，全面禁止商业性捕捞。为应对长江流域商业性捕捞禁令背景下野生与养殖水产品难以区分的问题，本研究以鲢(*Hypophthalmichthys molitrix*)为研究对象，系统比较其肌肉组织脂肪酸谱差异，并基于机器学习算法构建鉴别模型。首先，通过检测分析野生与养殖个体脂肪酸组成特征，建立包含 6 种机器学习算法的鉴别体系；随后，应用特征选择对原始数据进行降维处理，筛选得到 7 种最具区分度的特征脂肪酸，并以此构建优化模型。结果显示，降维处理显著提升了不同算法的鉴别性能，其中自适应提升算法 M1 (AdaBoost.M1)表现最佳，训练集与测试集的判别准确率分别达到 90.5%与 81.0%。研究结果表明，脂肪酸谱结合特征选择与机器学习算法可实现野生与养殖鲢的高精度区分，为水产品来源鉴别提供了可行技术路径，对长江流域渔业资源保护与禁捕政策实施具有重要的支撑价值。

关键词　鲢；野生；养殖；特征选择；机器学习

中图分类号 TS207.3; O657　**文献标识码** A　**文章编号** 2095-9869(2026)00-0000-12

长江流域复杂的地貌和多样的气候使其成为了全球淡水水生生物多样性最丰富的地区之一。据统计，长江流域分布各种鱼类多达 378 种，其中 149 种为长江流域特有的鱼种(刘飞等, 2019)。近几十年来，由于人类对其生态系统的过度干预与破坏行为已超出了其自我修复的能力，导致长江流域渔业资源锐减，生物多样性持续下降，部分珍稀物种濒临灭绝(Chen *et al*, 2020)。为遏制这一趋势，我国自 2020 年起实施长江“十年禁渔”计划，并自 2021 年 1 月 1 日起全面禁止商业性捕捞(董芳等, 2023)。

作为“四大家鱼”之一，鲢(*Hypophthalmichthys molitrix*)是我国主要淡水经济鱼类之一。鲢以浮游植物与有机碎屑为食(朱思嫚等, 2024)，在净化水质及维持水域生态系统的稳定中具有重要作用(Xia *et al*, 2022)。然而，近年来长江流域生态系统破坏严重，野生鲢种群数量骤减(Du *et al*, 2025)。尽管“禁渔期”内严禁商业性捕捞，但仍有部分商家为谋取利益非法捕捞，并将其冒充养殖水产品进行售卖。鲢非法捕捞案件频发，对长江流域的资源保护与生态环境恢复造成了严重威胁(Ma *et al*, 2018; Liu *et al*, 2024)。由于野生与养殖鲢在生物学性状上高度相似，高精度甄别方法的缺失为监管部门执法带来不便(Zhang *et al*, 2025)。这不仅影响执法效率，也阻碍了长江流域生态系统的恢复，因此，区分野生与养殖鲢对执法与监管至关重要。

目前用于区分野生与养殖水产品的方法主要包括光谱法(Duarte *et al*, 2022)、元素分析法(Camin *et al*, 2018)、稳定同位素法(Liu *et al*, 2020)、耳石微化学法(许颖等, 2024)以及组学分析法(Duarte *et al*, 2025; Yang *et al*, 2024)等。组学分析能够揭示代谢物、

* 国家“十四五”重点研发计划专项——长江流域水产品监测技术研究及示范应用(2022YFF0608200)资助。秦　雷, Email: 1716390986@shu.edu.cn
① 通信作者：吕继洲，Email: Ljzffff@163.com；陈　沁，Email: chenqincc@shu.edu.cn
收稿日期: 2025-09-15, 收修改稿日期: 2025-11-18

蛋白质或肠道菌群等方面的系统性差异，但其检测成本高、周期长且数据处理复杂，限制了其在日常监管中的应用。耳石微化学可反映个体生长与栖息环境特征(翟东东等, 2025)，然而耳石样本的获取及前处理操作繁琐，使其难以推广至实际场景。元素分析与稳定同位素技术则易受环境因素干扰，在相邻水域或相似生境下的判别效果不理想。光谱方法虽然简便，但检测灵敏度较低，难以实现对野生与养殖水产品的高精度区分。相比之下，水产品肌肉组织的脂肪酸组成受水域流速、水生生物运动能力(Zhu *et al*, 2023)、食物结构及来源(Fonseca *et al*, 2022)等多因素影响，能够较好地反映野生与养殖水产品间生态与营养差异。养殖水产品在高营养饲喂条件下往往出现脂肪过量沉积，而野生水产品通常不存在此情况(曾霖等, 2023)，二者在肌肉组织脂肪酸谱上呈现显著差异。这一特征为构建野生与养殖水产品的甄别模型提供了潜在的分析价值与应用前景。鉴于此，本研究基于鲢肌肉组织脂肪酸谱，结合不同机器学习算法构建长江流域野生与养殖鲢鉴别模型，以期为长江流域“十年禁渔”政策提供技术支持，助力长江流域的生态环境恢复。

1 材料与方法

1.1 样品与试剂

本研究收集野生鲢 29 条(采集自长江不同江段)和养殖鲢 20 条，均由中国水产科学研究院淡水渔业研究中心提供；并于上海本地超市获取养殖鲢 14 条。所有实验用鱼体重范围为 2.0~3.5 kg 之间。样本在低温条件下运送至实验室后，采用极速降温法对鲢进行安乐死处理。随后使用无菌手术刀取其背部肌肉组织 50 g，液氮速冻 30 min，随后置于–80 ℃超低温保存，待后续分析。

实验所用试剂包括甲醇(色谱级，纯度为 99.9%，CNW)、正己烷(色谱级，纯度为 99.9%，CNW)、氯仿(分析纯，纯度为 99.7%)和氢氧化钠(分析纯，纯度为 99.7%)，均购自国药集团化学试剂有限公司。

1.2 实验仪器

实验使用的主要仪器设备：Millipore Synergy®超纯水系统，DHG-9005 电热恒温干燥箱(上海精宏实验设备有限公司)，Tissuelyser-24 多样品组织研磨仪(上海净信科技有限公司)，N-EVAP 112 型氮气吹干仪(Organomation Associates Inc.，美国)，气相色谱–质谱联用仪(7890A-5975C) (Agilent Technologies，美国)。

1.3 实验方法

脂肪提取：采用经适当改良的 Bligh-Dyer 法(Bligh & Dyer, 1959)。称取 50 mg 肌肉组织样本，加入 1 mL CM 溶液(甲醇：氯仿=2：1, *V*/ *V*)，使用组织研磨仪研磨 3 min 后，震荡提取 1 h，并离心分层。向所得上清液中加入 250 μL 0.9% NaCl 溶液，再次震荡混匀并离心，取下层氯仿相，于氮吹仪中吹干并称重，获得总脂肪质量。

脂肪甲酯化：将提取得到的脂肪样品溶解于正己烷，配制浓度为 10 mg/mL 的溶液；随后加入等体积的 0.4 mol/L KOH-甲醇溶液，震荡混匀后于 37 ℃烘箱中静止 30 min 进行甲酯化。反应结束后加入等体积的去离子水，震荡混匀并静置分层，取上层有机相，经过 0.22 μm 有机滤膜过滤后转入进样瓶，待气相色谱–质谱分析。

1.4 鲢肌肉组织脂肪酸检测

鲢肌肉组织中脂肪酸的定性与定量分析采用气相色谱–质谱联用技术(GC-MS)进行。采用面积归一化法对脂肪酸相对含量进行定量分析。

色谱条件：使用 DB-WAX 毛细管色谱柱(30 m×250 μm×0.25 μm, Agilent Technologies)。程序升温条件如下：初始柱温为 70 ℃，保留 2 min；以 20 ℃/min 升温至 200 ℃，保留 0 min；再以 1 ℃/min 升至 220 ℃，保留 0 min；随后以 2 ℃/min 升至 232 ℃，保留 0 min；最后以 20 ℃/min 升至 240 ℃，保留 5 min。进样口温度设为 240 ℃，进样体积 1 μL，采用分流进样方式，分流比为 10：1。载气为高纯氦气(纯度>99.99%)，流速为 1.0 mL/min。

质谱条件：采用电子轰击(Electron Impact, EI)电离方式，电子能量为 70 eV。离子源温度设为 230 ℃，质量扫描范围为 30~550 *m*/*z*，全扫描采集模式，溶剂延迟时间设为 3 min。

1.5 数据分析与数据集的构建

所有实验均重复 3 次。数据整理与初步统计分析在 Excel 和 IBM SPSS Statistics 27 软件中完成。组间差异的显著性检验采用配对样本 *t* 检验(Paired samples *t*-test)进行。进一步，通过 Weka 3.6.15 软件进行特征脂肪酸的筛选，并构建了基于 6 种不同机器学习算法的识别模型，以区分野生与养殖鲢个体。本研究中鲢样本共 63 条，其中，野生 29 条、养殖 34 条。

样本按 2∶1 比例随机划分为训练集(n=42)与测试集(n=21)，并采用 10 折交叉验证与独立测试集相结合的方式对模型性能进行评估。

1.6　机器学习算法的选择

为系统评估机器学习算法在野生与养殖鲢鉴别任务中的适用性，本研究中基于贝叶斯分类器、函数分类器、惰性分类器、元分类器、集成学习分类器以及规则分类器共选取了 6 种具有不同建模机制的代表性算法：贝叶斯网络(Bayes Net)、逻辑回归(Logistic)、K 近邻算法(K-Nearest Neighbors, K-NN)、自适应提升算法 M1 (AdaBoost.M1)、随机森林(Random Forest)和决策表(Decision Table)用于基于肌肉组织脂肪酸谱的野生与养殖鲢的鉴别。本研究所有模型的迭代次数均统一设置为 100。

Bayes Net：是一种基于概率图模型的工具，由有向无环图(Directed Acyclic Graphs, DAGs)和条件概率表(Conditional Probability Tables, CPTs)组成。相较于传统回归方法，其在处理变量间复杂的依赖关系和不确定性信息等方面具备显著优势(Tian *et al*, 2023)，有助于从概率角度区分野生与养殖群体。本研究采用 K2 算法进行结构学习。

Logistic：是一种分类问题的统计模型，旨在找到最佳拟合线性模型的系数，以描述二元因变量与一个或多个自变量之间的关系(Dinh *et al*, 2019)。本研究中正则化参数设为 1.0×10^{-8}。

K-NN：作为一种惰性的机器学习分类算法，其能够计算新数据点与数据集中各点的距离(通常为欧氏距离)，并通过选取最邻近的 k 个样本的类别相似程度进行分类的一种监督学习方法(Cai *et al*, 2023)。本研究设定近邻数 k=8，并采用线性搜索方法。

AdaBoost.M1：是一种迭代的集成学习方法，通过迭代训练多个弱分类器并赋予不同样本不同权重并将其组合构建强分类器，在每轮迭代中，增加被误分类样本的权重，从而使模型逐步修正错误，最终提升鉴别精度(Yang *et al*, 2023)。本研究以 LMT 为基本分类器。

Random Forest：属于一种集成学习算法，在分类和回归任务中应用广泛。它以决策树为基本分类器，分类结果由多数投票决定，能够很好地处理高维数据，具有抗噪能力强、精度高等优势(Lösel *et al*, 2024)。本研究设定树的数量为 200，且不限制每棵树的深度。

Decision Table：是一种用于描述条件逻辑的结构化工具，通常由条件部分和动作部分组成，核心思想是将复杂的决策过程转化为易于理解和操作的表格形式(Arnold *et al*, 2018)。本研究采用 Best First 搜索算法进行特征选择。

1.7　模型性能评估

准确率 Accuracy：

$$ACC=(TP+TN)/(TP+TN+FP+FN) \quad (1)$$

灵敏度 Sensitivity：

$$SN=TP/(TP+FN) \quad (2)$$

特异性 Specificity：

$$SP=TN/(TN+FP) \quad (3)$$

其中，TP 为真阳性数，TN 为真阴性数，FP 为假阳性数，FN 为假阴性数

2　结果与分析

2.1　野生与养殖鲢肌肉组织脂肪酸含量检测

本研究采用气相色谱–串联质谱(GC-MS)技术检测野生与养殖鲢肌肉组织脂肪酸组成，野生与养殖鲢共检测出 18 种脂肪酸。脂肪酸组成结果显示(表 1)，养殖鲢在单不饱和脂肪酸(MUFA)总含量显著高于野生鲢，具体体现在 C16:1、C18:1n9c 等关键 MUFA 组分上。野生鲢肌肉组织中的多不饱和脂肪酸(PUFA)含量显著高于养殖鲢，尤其是 C22:6n3、C20:5n3 及 C20:4n6 等组分。而饱和脂肪酸(SFA)在野生与养殖鲢群体间无显著性差异。对各脂肪酸组分的显著性分析结果表明，野生与养殖鲢肌肉组织脂肪酸除 C14:0、C18:0、C18:3n6、C20:0 及 C22:2 不存在显著性差异(P>0.05)，其余脂肪酸均存在统计学差异。

野生与养殖鲢肌肉组织脂肪酸含量的聚类热图结果如图 1 所示：由于野生鲢样本具有 PUFA 含量较高、SFA 含量较低的特征，而养殖鲢则表现出较高的 SFA 与 MUFA 含量，因此，在聚类分析中，两类样本整体上能够被有效区分为两个独立的聚类簇。然而，有 3 个养殖鲢样本(Farmed 1、2、3)被划分至野生鲢所在的聚类簇中。进一步分析发现，驱动野生鲢聚类的关键因素在于其较高的 C20:4n6、C22:6n3 和 C17:0 等脂肪酸含量以及较低的 C14:0、C16:0、C18:0 和 C20:0 等 SFA 含量，而养殖鲢的聚类特征则与此相反。对 Farmed 1、2、3 三个养殖鲢样本的脂肪酸相对含量分析发现，虽然它们的 C20:4n6 和 C22:6n3 含量较低，但其 C17:0、C20:3n3、C20:5n3 脂肪酸含

量分别达到 0.8%~1.4%、0.3%~0.6%、3.1%~4.8%，而 C14:0、C18:0、C16:0 脂肪酸含量分别为 5.1%~5.2%、15.8%~17.8%、9.3%~11.6%。这些独特的脂肪酸组成特征使其更接近于野生鲢，从而导致其聚类结果偏向于野生类群。总体而言，基于脂肪酸谱的聚类分析能够较好地区分野生与养殖鲢，初步验证了该方法在两类个体鉴别中的可行性与有效性。

表 1 野生与养殖鲢肌肉组织脂肪酸组成

(%，平均值±标准差)

Tab.1 Fatty acid composition of muscle tissue in wild and farmed *H. molitrix* (%, Mean±SD)

脂肪酸 Fatty acid	野生鲢 Wild *H. molitrix*	养殖鲢 Farmed *H. molitrix*
C14:0	4.75±1.89	5.71±1.93
C16:0	19.35±1.34^{a}	21.78±4.80^{b}
C16:1	8.00±1.66^{a}	12.36±2.56^{b}
C17:0	1.04±0.26^{a}	0.76±0.40^{b}
C17:1	0.97±0.24^{a}	1.43±0.54^{b}
C18:0	11.16±2.41	11.63±5.16
C18:1n9c	16.59±4.89^{a}	22.84±6.45^{b}
C18:1n9t	3.71±0.39^{a}	4.15±1.05^{b}
C18:2n6c	2.85±0.44^{a}	2.04±1.20^{b}
C18:3n6	1.01±1.58	0.82±0.77
C18:3n3	5.55±6.41^{a}	1.91±2.15^{b}
C20:0	1.03±1.07	0.96±1.05
C20:1	1.09±0.29^{a}	1.76±0.97^{b}
C20:2	0.48±0.16	0.50±0.47
C20:3n3	0.61±0.17^{a}	0.33±0.23^{b}
C20:4n6	6.39±1.38^{a}	2.32±1.38^{b}
C20:5n3	6.47±1.11^{a}	3.77±1.89^{b}
C22:6n3	11.78±1.99^{a}	4.93±2.04^{b}
饱和脂肪酸 SFA	37.33±5.50	40.83±11.39
单不饱和脂肪酸 MUFA	30.35±6.42^{a}	42.54±10.02^{b}
多不饱和脂肪酸 PUFA	35.12±6.94^{a}	16.63±7.25^{b}

注：同行数据肩标不同小写字母表示差异显著(P<0.05)。

Note: in the same row, values with different lowercase letters superscripts are significantly different (P<0.05).

2.2 基于脂肪酸谱–机器学习的野生鲢鉴别模型的构建

2.2.1 *基于不同机器学习算法的野生鲢鉴别模型的构建* 为系统评估机器学习算法的鉴别能力，本研究基于野生与养殖鲢肌肉组织脂肪酸 18 种脂肪酸的相对含量，利用 Bayes Net、Logistic、K-NN、AdaBoost.M1、Random Forest 和 Decision Table 6 种机器学习算法，分别构建了野生与养殖鲢鉴别模型。结果显示，在训练集中，K-NN、AdaBoost.M1 和 Random Forest 的判别准确率最高，均可达到 85.7%，展现出对训练数据较强的拟合能力。在测试集中，Bayes Net、K-NN、Random Forest 和 AdaBoost.M1 的识别准确率均为 76.2%。其余 2 种机器学习算法预测准确率仅为 71.4%，表明其泛化能力较弱，可能存在过拟合现象。总体而言，K-NN 与 AdaBoost.M1 在训练集与测试集表现出较好的判别性能，为野生鲢鉴别的候选模型。

受试者工作特征曲线(Receiver Operating Characteristic, ROC)通过绘制真阳性率(y 轴)与假阳性率(x 轴)之间的关系，用于评估鉴别模型在不同判别阈值下的性能。曲线下面积(AUC)作为一个重要量化指标，反映了模型的整体鉴别能力，其取值范围为 0~1，值越高表明模型的稳定性和鲁棒性越强。为进一步验证模型的性能，本研究对 AdaBoost.M1、Random Forest 和 K-NN 等 6 种机器学习算法所构建的野生鲢鉴别模型进行 ROC 曲线的绘制(图 2)并计算 AUC 值(表 3)。结果显示，Bayes Net、AdaBoost.M1

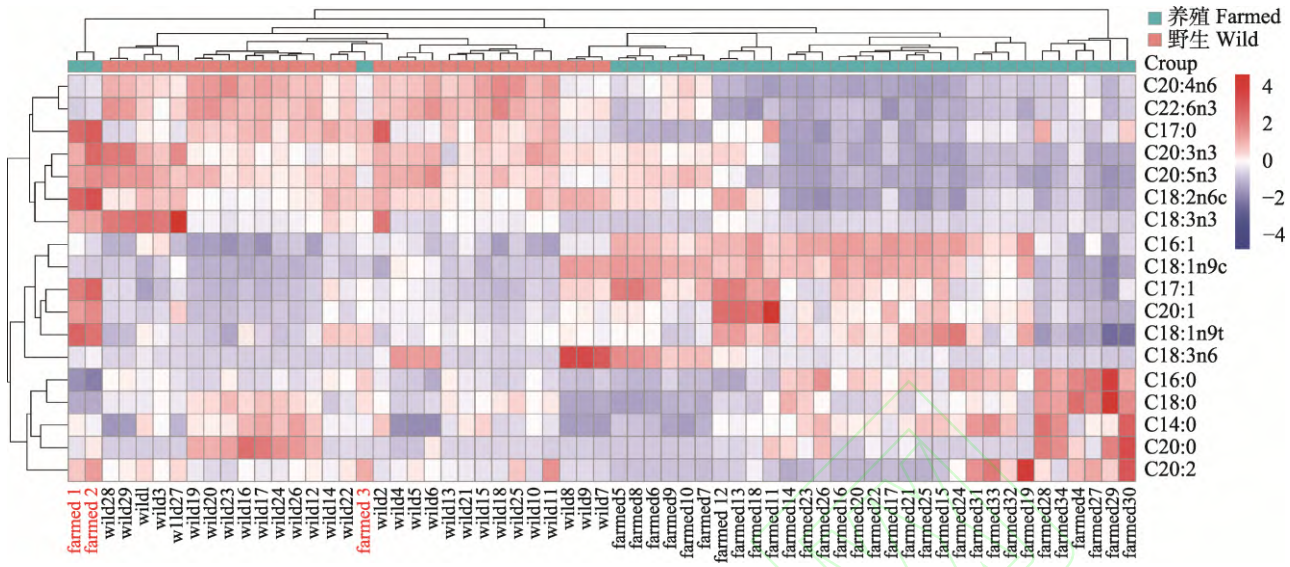

图 1　野生与养殖鲢脂肪酸组成聚类热图

Fig.1　Cluster heat map of fatty acid composition in wild and farmed *H. molitrix*

表 2　采用不同机器学习算法构建野生与养殖鲢鉴别模型的准确率

Tab.2　Accuracy of different machine learning algorithms in constructing models for discriminating wild from farmed *H. molitrix*

模型 Model	训练集 Training set			测试集 Test set		
	灵敏度 SN/%	特异性 SP/%	准确率 ACC/%	灵敏度 SN/%	特异性 SP/%	准确率 ACC/%
Bayes Net	78.9	82.6	81.0	90.0	63.6	76.2
Logistic	84.2	60.9	71.4	80.0	63.6	71.4
K-NN	89.5	82.6	85.7	80.0	72.7	76.2
AdaBoost.M1	84.2	87.0	85.7	90.0	63.6	76.2
Decision Table	72.7	82.6	78.6	80.0	63.6	71.4
Random Forest	84.2	87.0	85.7	80.0	72.7	76.2

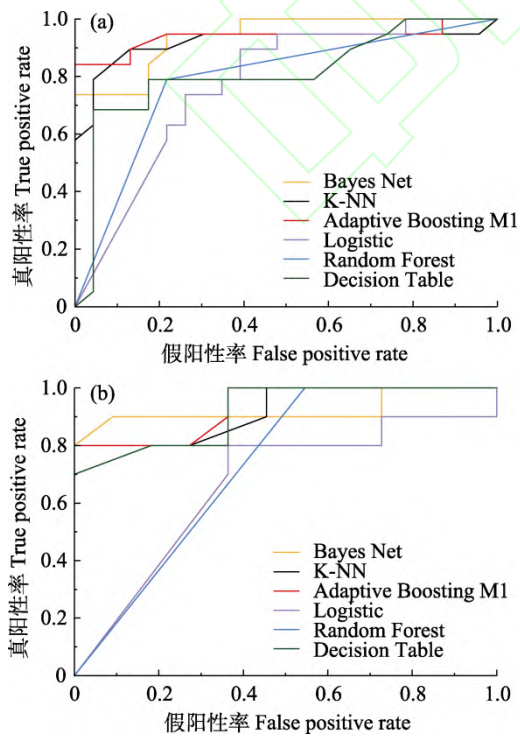

图 2　不同机器学习算法构建野生与养殖鲢鉴别模型的 ROC 曲线

Fig.2　ROC curves of different machine learning algorithms for constructing models to discriminate wild from farmed *H. molitrix*

a：训练集；b：测试集。

a: Training set; b: Test set.

和 K-NN 在训练集与测试集均表现出较高的 AUC 值(超过 0.9)，显示出良好的稳定性。其中，AdaBoost.M1 在训练集保持最高预测准确率的同时，AUC 值在所有模型中仍为最高，训练集与测试集 AUC 分别达到 0.94 和 0.93。结合表 2 所示的灵敏度、特异性及准确率结果可见，AdaBoost.M1 在训练集与测试集的 ACC 值和 AUC 值在 6 种机器学习算法中均为最高，表明其具有较高的综合识别性能，也表明在数据降维前，该算法已能够较为有效地区分野生与养殖鲢。但当前 6 种机器学习算法预测准确率仍有提升空间，尤其

表 3　不同机器学习算法构建野生与养殖鲢鉴别模型 ROC 曲线的 AUC 值

Tab.3　AUC values of ROC curves for different machine

learning algorithms in constructing models to discriminate wild from farmed *H. molitrix*

模型 Model	曲线下面积 AUC	
	训练集 Training set	测试集 Test set
Bayes Net	0.94	0.92
Logistic	0.77	0.66
K-NN	0.92	0.92
AdaBoost.M1	0.94	0.93
Decision Table	0.81	0.91
Random Forest	0.79	0.77

Logistic 与 Random Forest 在训练集和测试集上的 AUC 值均低于 0.8,表明其模型稳定性与鲁棒性较差,需通过数据降维、特征清洗等进一步方法优化模型性能。

2.2.2 降维后基于不同机器学习算法的野生鲢鉴别模型的构建 为提高模型的鉴别效率与泛化能力,本研究采用相关性特征选择子集评估(Correlation-based Feature Subset Selection, CFS Subset Eval)算法进行特征筛选,最终筛选出 7 个特征脂肪酸,分别为 C16:1、C17:0、C20:2、C20:3n3、C20:4n6、C20:5n3 和 C22:6n3。基于以上 7 个特征脂肪酸,使用 **2.2.1** 所述的 6 种机器学习算法对野生与养殖鲢进行模型构建。结果显示,特征筛选后,所有模型的鉴别性能均显著提升(表 4)。对于训练集,各算法的准确率均超过 85%;测试集中,除 Decision Table 以及 Logistic 外,其余模型预测准确率均超过 80%。其中,K-NN 与 AdaBoost.M1 表现优于其他算法,训练集的准确率达到 90.5%;测试集的准确率达到 81.0%。该结果表明,通过特征筛选可以去除冗余变量、降低噪声干扰、优化特征空间结构,从而提升模型的泛化能力与稳定性,增强了鉴别结果的可靠性。

表 4 特征筛选后采用不同机器学习算法构建野生与养殖鲢鉴别模型的准确率

Tab.4 Accuracy of different machine learning algorithms in constructing models for discriminating wild from farmed *H. molitrix* after feature selection

模型 Model	训练集 Training set			测试集 Test set		
	灵敏度 SN/%	特异性 SP/%	准确率 ACC/%	灵敏度 SN/%	特异性 SP/%	准确率 ACC/%
Bayes Net	84.2	87.0	85.7	90.0	72.7	81.0
Logistic	94.4	78.3	85.7	90.0	63.6	76.2
K-NN	89.5	91.3	90.5	90.0	72.7	81.0
AdaBoost.M1	89.5	91.3	90.5	80.0	81.8	81.0
Decision Table	89.5	87.0	88.1	80.0	63.6	71.4
Random Forest	94.7	82.6	88.1	100.0	63.6	81.0

为进一步评估模型性能,本研究基于特征筛选后的野生鲢鉴别模型绘制了 Bayes Net、Logistic、K-NN 等算法的 ROC 曲线(图 3),并计算了相应的 AUC 值(表 5)。结果显示,经特征筛选后所有模型 AUC 值均有所提升,除 Logistic 与 Random Forest 外,其余模型训练集与测试集的 AUC 值均超过 0.90,表明特征筛选显著增强了模型的整体判别能力,剔除无关变量有效提高了模型的稳定性与预测可靠性。结合表 5,降维后模型的预测准确率分析可知,尽管 K-NN 与 AdaBoost.M1 在训练集和测试集上的准确率均高于其他算法,但 AdaBoost.M1 在训练集和测试集上的 AUC 值分别达到 0.97 和 0.94,显示出优于 K-NN 的稳定性与鲁棒性。因此,在 6 种机器学习算法中,AdaBoost.M1 整体表现最佳,故采用 AdaBoost.M1 模型为野生与养殖鲢最终判别模型。

3 讨论

3.1 不同因素对野生与养殖鲢肌肉组织脂肪酸组成的影响

水产品肌肉组织的脂肪酸组成受多种因素影响,其中饲料成分是最主要的因素之一(Ren *et al*, 2020; 肖昌伦等, 2025)。在 SFA 中,野生与养殖鲢的 C16:0 与 C18:0 含量较高。作为一种滤食性鱼类,野生鲢主要以浮游植物和有机碎屑为食,而养殖鲢常与其他鱼类混养,其摄食来源包括商业饲料、浮游植物及有机碎屑,这些食源中通常富含 C16:0、C18:0 及其合成前体物质(Mukherjee *et al*, 2010; Acar *et al*, 2018; Wang YH *et al*, 2024),这可能是二者体内 C16:0 与 C18:0 含量较高的原因。在 MUFA 组成中,野生与养

殖鲢均表现出较高的 C18:1n9c 含量，且养殖鲢显著高于野生鲢。这一差异可能与养殖过程中投喂的人工饲料富含植物油脂(如菜籽油)密切相关。研究表明，水产品肌肉组织脂肪酸组成受脂质成分的摄入影响较大(Yu *et al*, 2018; Liu *et al*, 2022)。菜籽油由于其产量大、价格低、MUFA 含量高等优势，常被用作水产饲料的脂质来源(闫春为等, 2018)。此外，菜籽油富含 C18:1n9c (López-Marcos *et al*, 2024)，因此，野生鲢相较于养殖鲢 C18:1n9c 含量较低。在 PUFA 组成方面，野生鲢肌肉组织脂肪酸 PUFA 含量显著高

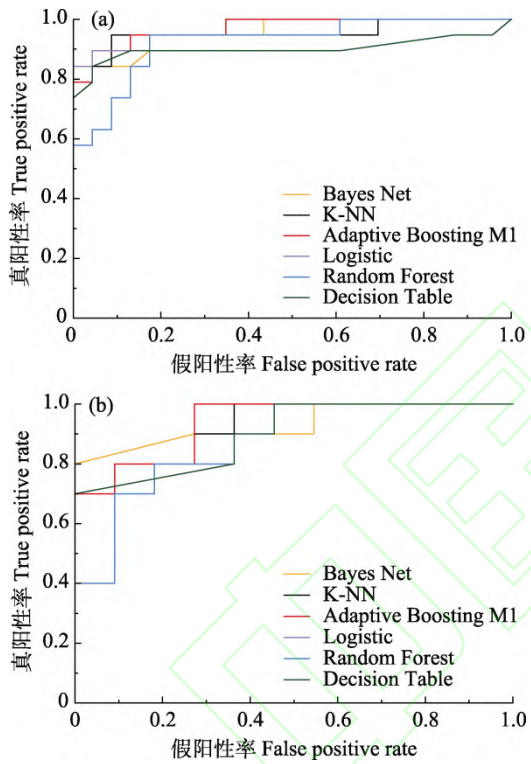

图 3　特征筛选后不同机器学习算法构建野生与养殖鲢鉴别模型的 ROC 曲线

Fig.3　ROC curves of different machine learning algorithms for constructing models to discriminate wild from farmed *H. molitrix* after feature selection

a：训练集；b：测试集。

a: Training set; b: Test set.

表 5　特征筛选后不同机器学习算法构建野生与养殖鲢鉴别模型 ROC 曲线的 AUC 值

Tab.5　AUC values of ROC curves for different machine learning algorithms in constructing models to discriminate wild from farmed *H. molitrix* after feature selection

模型 Model	曲线下面积 AUC	
	训练集 Training set	测试集 Test set
Bayes Net	0.96	0.93
Logistic	0.95	0.87
K-NN	0.95	0.93
AdaBoost.M1	0.97	0.94
Decision Table	0.90	0.90
Random Forest	0.92	0.88

于养殖鲢，尤其是 C22:6n3、C20:5n3 以及 C20:4n6 等高价值 PUFA 差距显著，因此，野生鲢具有更高的营养价值，这也部分解释了鲢非法捕捞案件频发的原因。造成这一差异的主要原因为饮食结构的不同，自然水体中的野生鲢主要以天然藻类和浮游生物为食，这些天然饵料中富含 PUFA 前体(Muhammad *et al*, 2024)，而养殖鲢所摄食的饲料中脂质来源较为单一，因此，养殖鲢 PUFA 含量较低。

除了饲料组成，水产品肌肉组织脂肪酸组成还受生存环境、游泳能力(Zhu *et al*, 2023)和季节变化(Scheuer *et al*, 2024)等因素影响。环境温度变化会促使水生生物调节体内脂肪酸饱和度以维持膜流动性(Yoon *et al*, 2022)，从而间接影响其食物组成及自身脂肪酸谱。盐度变化也会影响细胞膜渗透压平衡，通常导致 PUFA 含量升高，以维持膜稳定性、提供能量并抵抗盐度胁迫(Chen *et al*, 2022; Dildar *et al*, 2025)。光照强度通过影响水生植物的光合作用效率，进而影响其脂肪酸合成(Kumar *et al*, 2019; Maltsev *et al*, 2021)，最终沿食物链传递至鱼类。金属与微量元素也会干扰鱼类的生理生化过程，不同水域环境中元素含量的差异可能进一步影响水产品的脂肪酸组成(Jovičić *et al*, 2024)。另外，养殖鲢活动范围有限，运动强度较低，无需像野生个体为觅食或避敌而长距离游动，这进一步造成野生与养殖鲢脂肪酸组成的差异(Huang *et al*, 2022; 龙珍满等, 2023)。而本研究中的鲢样本并未完全涵盖季节性的变化，也是本研究的局限性之一。

脂肪酸组成是区分野生和养殖水产品的重要指标(David, 2019)，虽然鱼类肌肉组织脂肪含量较少，但其 PUFA 含量较高。已有研究显示，野生鲤(Wang LM *et al*, 2024)以及野生鲈鱼和鲷鱼(Amoussou *et al*, 2022)的 PUFA 和 SFA 含量较高，而养殖个体的 MUFA 含量较高，表明野生与养殖水产品在脂肪酸组成上存在显著差异。本研究通过气相色谱–串联质谱(GC-MS)分析野生与养殖鲢肌肉中的脂肪酸组成，发现野生鲢肌肉组织中 C20:4n6、C22:6n3、C20:3n3、C20:5n3

等 PUFA 含量较高，养殖鲢则是 C14:0、C16:0、C18:0 等 SFA 含量较高，与上述研究结果一致。

聚类分析结果表明，野生鲢与养殖鲢总体上可被清晰地划分为两个不同的聚类簇，说明脂肪酸组成在区分二者中具有重要潜力，初步验证了利用脂肪酸特征鉴别野生与养殖鲢的可行性。尽管有 3 个养殖鲢样本被归入野生鲢所在聚类簇中，但这可能是由于个体间脂肪酸组成的差异所致，由于这 3 个养殖鲢样本的脂肪酸组成特征更接近于野生鲢，导致这种现象的发生。后续研究应该增加来自 Farmed 11、12、13 地区的养殖鲢样本数量，以增强其在聚类分析中对养殖群体特征的代表性。

3.2 不同机器学习算法对野生鲢鉴别模型的影响

本研究采用气相色谱–串联质谱对野生与养殖鲢肌肉组织的脂肪酸组成进行分析，并通过面积归一化法测定各脂肪酸的相对含量。虽然绝对含量能够提供更丰富的信息维度，但相对含量能够直接反映各脂肪酸在总脂肪中的构成比例，更能有效揭示水产品在饮食结构及生活环境方面的差异。目前，基于脂肪酸谱的分析方法已在多种食品地理溯源与真实性鉴别研究中得到广泛应用(Grazina *et al*, 2020; Dou *et al*, 2024; Eugelio *et al*, 2025)。鉴于此，本研究基于鲢肌肉组织的脂肪酸谱，对野生与养殖鲢进行鉴别。为实现高效、准确的判别目标，本研究系统比较了 6 种机器学习算法在鉴别野生与养殖鲢方面的性能。为进一步提高模型判别效率与分类精度，采用 CFS Subset Eval 算法进行特征筛选。该方法通过评估变量与类别之间的相关性，基于预定评价函数筛选出最具代表性的特征变量(Chang *et al*, 2023)。最终确定了 7 个关键特征脂肪酸(C16:1、C17:0、C20:2、C20:3n3、C20:4n6、C20:5n3 和 C22:6n3)，这些脂肪酸多为 PUFA。已有研究指出，PUFA 在野生与养殖水产品中含量差异显著，更能反映二者在生活环境与饮食结构中的不同，从而为判别分析提供了可靠的生物学依据。基于特征筛选所构建的野生与养殖鲢鉴别模型性能显著提升，实现了对野生鲢的高精度判别。

在未进行特征筛选的情况下，不同机器学习算法的准确率差异较大，训练集的准确率为 71.4%~85.7%；测试集的准确率为 71.4%~76.2%。其中，K-NN、AdaBoost.M1 和 Random Forest 在训练集上的准确率均超过 85%，显示出较强的拟合能力。然而 Logistic 和 Decision Table 在测试集上的准确率均低于 75%，可能存在过拟合现象。值得注意的是，集成学习算法 AdaBoost.M1 和 Random Forest 均表现出了较好的预测能力。尽管 Random Forest 在训练集和测试集上的准确率分别达到 85.7%和 81.0%，但其训练集与测试集的 AUC 值仅分别为 0.79 和 0.77，显著低于 AdaBoost.M1，说明其鉴别稳定性较差，更容易受到噪声干扰。K-NN 作为一种惰性学习算法，在野生鲢鉴别任务中表现同样较好，这很可能得益于不同来源的鲢样本的脂肪酸谱类间分离度较高，有利于近邻机制发挥作用(Srisuradetchai *et al*, 2024)。未经特征降维的模型虽具备一定的判别能力，但预测准确率与稳定性仍有较大提升空间。

通过特征选择可以有效提高模型的预测准确率，提高模型的判别精度(Ekinci *et al*, 2023; Niu *et al*, 2013)。本研究通过 CFS Subset Eval 进行特征筛选后，模型性能得到全面改善，这进一步证明所筛选得到的 7 种关键脂肪酸在野生与养殖鲢鉴别模型中的重要作用。在野生与养殖鲢的鉴别模型中，特征筛选本质是对脂肪酸数据进行降维处理，旨在剔除与食源、环境等关键影响因素关联较弱的冗余信息，同时保留能够最大程度反映组间差异的核心脂肪酸特征。例如，C20:5n3 和 C22:6n3 是野生鲢从天然藻类与浮游生物中获取的标志性 PUFA，在养殖个体中含量显著偏低，因此具备较强的区分能力。通过关键特征提取，不仅提升了模型的判别精度，也增强了其泛化能力。实验结果显示，所有模型训练集的准确率均超过 85%。其中 Random Forest 和 AdaBoost.M1 的准确率均大于 90%，表明通过提取关键脂肪酸特征能有效去除冗余信息，降低模型复杂度，从而减少过拟合的风险(Sim *et al*, 2023)。

通过 CFS Subset Eval 特征筛选为不同算法提供了优化路径：其为 K-NN 构建了低维、高区分度的特征空间，使其执行转换欧氏距离的方式更加高效；同时为 AdaBoost.M1 提供了高质量的特征，增强了弱分类器的判别能力，使其更好地挖掘泛化规律。因此，K-NN 与 AdaBoost.M1 在经过特征选择后，在野生鲢的鉴别任务中同样展现出较好效果。综合准确率值与 AUC 值分析发现，AdaBoost.M1 表现优于其他模型，其训练集的准确率为 90.5% (SN=89.5%, SP=91.3%)，测试集的准确率为 81.0% (SN=100%, SP=63.6%)，训练集与测试集的 AUC 分别达到 0.97 和 0.94，均高于 K-NN，显示出较高的拟合精度和泛化能力。此外，Bayes Net、Logistic 和 K-NN 在训练集上的 AUC 均

大于 0.95，说明经特征优化后，多种模型均实现了高效判别。综上所述，在基于肌肉脂肪酸谱的野生与养殖鲢鉴别中，经 CFS Subset Eval 特征筛选后，AdaBoost.M1 在 6 种机器学习算法中表现出最佳的综合识别性能。

本研究验证了基于肌肉组织脂肪酸谱结合机器学习算法鉴别野生与养殖水产品的可行性。在评估的 6 种算法中，AdaBoost.M1 在处理小样本、高维度化学计量数据方面展现出独特优势，其训练集与测试集准确率分别为 90.5%和 81.0%，取得了较好的预测效果。目前，基于脂肪酸谱结合机器学习技术的野生与养殖水产品鉴别方法已在三文鱼研究中得到应用，训练集与测试集准确率均超过 90% (Grazina *et al*, 2020)，虽然本研究结果与其存在一定差距，但本研究中样本来源相对复杂，且部分产地的样本量相对有限，导致难以有效提取样本稀少区域的脂肪酸特征。这些因素在客观上增加了野生与养殖鲢鉴别模型的构建难度，但同时也表明该技术路径在复杂真实场景下仍具备良好的应用潜力。为实现产业化应用，后续研究应在扩大样本量的基础上，特别注重涵盖不同季节的样本，以系统验证模型的普适性与稳定性。未来，该技术可进一步与人工智能和自动化检测平台相结合，比如：构建野生与养殖水产品溯源系统；开发轻量化模型，集成于便携式检测设备中，从而拓展至不同地理来源水产品的判别，为市场监管与水产品认证提供快速、可靠的技术支持。

4　结论

本研究通过气相色谱–串联质谱(GC-MS)，分析了野生与养殖鲢肌肉组织脂肪酸组成的差异，并结合多种机器学习算法构建了野生与养殖鲢鉴别模型。研究表明，养殖鲢的饱和脂肪酸(SFA)含量较高，而野生鲢则是富含多不饱和脂肪酸(MUFA)。通过 CFS Subset Eval 算法筛选出的 7 种特征脂肪酸进行数据降维后，所有模型的识别性能显著提升，其中，AdaBoost.M1 表现优于其他算法，训练集与测试集的准确率分别提高至 90.5%与 81.0%。该方法不仅为野生与养殖鲢的鉴别提供了科学依据，未来也有希望为长江流域禁渔期内的生态环境修复以及执法部门监管提供技术服务，展现出广阔的应用前景。

参　考　文　献

董芳, 方冬冬, 张辉, 等. 长江十年禁渔后保护与发展. 水产学报, 2023, 47(2): 245–259 [DONG F, FANG D D, ZHANG H, *et al*. Protection and development after the ten-year fishing ban in the Yangtze River. Journal of Fisheries of China, 2023, 47(2): 245–259]

刘飞, 林鹏程, 黎明政, 等. 长江流域鱼类资源现状与保护对策. 水生生物学报, 2019, 43(S1): 144–156 [LIU F, LIN P C, LI M Z, *et al*. Situations and conservation strategies of fish resources in the Yangtze River basin. Acta Hydrobiologica Sinica, 2019, 43(S1): 144–156]

龙珍满, 朱峰跃, 郭杰, 等. 捕食胁迫对“四大家鱼”幼鱼生理反应的影响. 渔业科学进展, 2023, 44(3): 111–123 [LONG Z M, ZHU F Y, GUO J, *et al*. Effects of predation stress on the physiological responses of juvenile four major Chinese carps. Progress in Fishery Sciences, 2023, 44(3): 111–123]

肖昌伦, 孙云飞, 鹿珍珍, 等. 两种育肥方式与湖泊养殖中华绒螯蟹营养品质的比较. 水产学报, 2025, 49(1): 193–208 [XIAO C L, SUN Y F, LU Z Z, *et al*. Comparison of nutritional quality between two fattening methods and lake-cultured *Eriocheir sinensis*. Journal of Fisheries of China, 2025, 49(1): 193–208]

许颖, 姜涛, 杨健, 等. 长江安庆江段存在溯河洄游型和淡水定居型刀鲚实证研究. 渔业科学进展, 2024, 45(4): 1–14 [XU Y, JIANG T, YANG J, *et al*. Coexistence of freshwater resident and anadromous *Coilia nasus* in the Anqing section of the Yangtze River in Anhui Province, China. Progress in Fishery Sciences, 2024, 45(4): 1–14]

闫春为, 陈乃松, 李自强, 等. 大黄鱼幼鱼饲料中大豆磷脂油与菜籽油的适宜配比. 渔业科学进展, 2018, 39(4): 56–65 [YAN C W, CHEN N S, LI Z Q, *et al*. Suitable ratio of soybean phospholipid oil to rapeseed oil in the diet of juvenile large yellow croaker (*Larimichthys crocea*). Progress in Fishery Sciences, 2018, 39(4): 56–65]

曾霖, 宋炜, 谢正丽, 等. 基于代谢组解析大黄鱼对低温和饥饿胁迫的适应机制. 水产学报, 2023, 47(7): 88–99 [ZENG L, SONG W, XIE Z L, *et al*. Metabolomics-based analysis of adaptive mechanism of *Larimichthys crocea* to low temperature and starvation stresses. Journal of Fisheries of China, 2023, 47(7): 88–99]

翟东东, 李雯, 王东, 等. 三峡库区不同鲢群体耳石微化学差异分析. 水生生物学报, 2025, 49(8): 17–25 [ZHAI D D, LI W, WANG D, *et al*. Microchemical difference analysis of otolith in different silver carp (*Hypophthalmichthys molitrix*) populations in the Three Gorges Reservoir area. Acta Hydrobiologica Sinica, 2025, 49(8): 17–25]

朱思嫚, 胡红娟, 贾佳, 等. 鲢滤食不同藻类的磷吸收及排泄过程研究. 水生生物学报, 2024(5): 744–752 [ZHU S H, HU H J, JIA J, *et al*. Phosphorus cycle of silver carp feeding on different algae and its effect on the water nutrient cycle. Acta Hydrobiologica Sinica, 2024(5): 744–752]

ACAR Ü, TÜRKER A. Response of Rainbow trout (*Oncorhynchus mykiss*) to unrefined peanut oil diets: Effect on growth performance, fish health and fillet fatty acid composition. Aquaculture Nutrition, 2018, 24(1): 292–299

AMOUSSOU N, MARENGO M, IKO AFÉ O H, *et al*. Comparison of fatty acid profiles of two cultivated and wild marine fish from Mediterranean Sea. Aquaculture International, 2022, 30(3): 1435–1452

ARNOLD J G, BIEGER K, WHITE M J, *et al*. Use of decision tables to simulate management in SWAT+. Water, 2018, 10(6): 713

BLIGH E G, DYER W J. A rapid method of total lipid extraction and purification. Canadian Journal of Biochemistry and Physiology, 1959, 37(8): 911–917

CAI Z Y, HUANG Z H, HE M Y, *et al*. Identification of geographical origins of *Radix Paeoniae Alba* using hyperspectral imaging with deep learning-based fusion approaches. Food Chemistry, 2023, 422: 136169

CAMIN F, PERINI M, BONTEMPO L, *et al*. Stable isotope ratios of H, C, O, N and S for the geographical traceability of Italian rainbow trout (*Oncorhynchus mykiss*). Food Chemistry, 2018, 267: 288–295

CHANG L S, FUKUOKA Y, AOUIZERAT B E, *et al*. Prediction performance of feature selectors and classifiers on highly dimensional transcriptomic data for prediction of weight loss in Filipino Americans at risk for type 2 diabetes. Biological Research for Nursing, 2023, 25(3): 393–403

CHEN T G, WANG Y, GARDNER C, *et al*. Threats and protection policies of the aquatic biodiversity in the Yangtze River. Journal for Nature Conservation, 2020, 58: 125931

CHEN W W, LI X, QIN K X, *et al*. Effects of low salinity on fatty acid and free amino acid composition of muscle tissues in *Portunus trituberculatus*. Aquaculture Research, 2022, 53(5): 1627–1635

DAVID F. A worldwide reliable indicator to differentiate wild *vs*. farmed *Penaeid* shrimps based on 207 fatty acid profiles. Food Chemistry, 2019, 292: 247–252

DILDAR T, CUI W X, IKHWANUDDIN M, *et al*. Aquatic organisms in response to salinity stress: Ecological impacts, adaptive mechanisms, and resilience strategies. Biology, 2025, 14(6): 667

DINH A, MIERTSCHIN S, YOUNG A, *et al*. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 2019, 19(1): 211

DOU X J, WANG X F, MA F, *et al*. Geographical origin identification of *Camellia* oil based on fatty acid profiles combined with one-class classification. Food Chemistry, 2024, 433: 137306

DU J, TIAN H W, XIANG Z Y, *et al*. Impact of the fishing ban on fish diversity and population structure in the middle reaches of the Yangtze River, China. Frontiers in Environmental Science, 2025, 12: 1530716

DUARTE B, FEIJÃO E, CRUZ-SILVA A, *et al*. Reveal your microbes, and I'll reveal your origins: Geographical traceability *via Scomber colias* intestinal tract metagenomics. Animal Microbiome, 2025, 7(1): 43

DUARTE B, MAMEDE R, CARREIRAS J, *et al*. Harnessing the full power of chemometric-based analysis of total reflection X-ray fluorescence spectral data to boost the identification of seafood provenance and fishing areas. Foods, 2022, 11(17): 2699

EKINCI E, ÖZBAY B, OMURCA S İ, *et al*. Application of machine learning algorithms and feature selection methods for better prediction of sludge production in a real advanced biological wastewater treatment plant. Journal of Environmental Management, 2023, 348: 119448

EUGELIO F, MASCINI M, MARONE E, *et al*. Phenolic profile as a powerful machine learning tool for identification, traceability, and quality control of olive cultivars. Journal of Agricultural and Food Chemistry, 2025, 73(37): 23671–23683

FONSECA V F, DUARTE I A, MATOS A R, *et al*. Fatty acid profiles as natural tracers of provenance and lipid quality indicators in illegally sourced fish and bivalves. Food Control, 2022, 134: 108735

GRAZINA L, RODRIGUES P J, IGREJAS G, *et al*. Machine learning approaches applied to GC-FID fatty acid profiles to discriminate wild from farmed salmon. Foods, 2020, 9(11): E1622

HUANG Y Y, JIANG G Z, ABASUBONG K P, *et al*. High lipid and high carbohydrate diets affect muscle growth of blunt snout bream (*Megalobrama amblycephala*) through different signaling pathways. Aquaculture, 2022, 548: 737495

JOVIČIĆ K, DJIKANOVIĆ V, SANTRAČ I, *et al*. Effects of trace elements on the fatty acid composition in Danubian fish species. Animals, 2024, 14(6): 954

KUMAR G, NGUYEN D D, HUY M, *et al*. Effects of light intensity on biomass, carbohydrate and fatty acid compositions of three different mixed consortia from natural ecological water bodies. Journal of Environmental Management, 2019, 230: 293–300

LIU C S, ZHANG Z H, LI B Q, *et al*. Lipid metabolic disorders induced by organophosphate esters in silver carp from the middle reaches of the Yangtze River. Environmental Science & Technology, 2024, 58(11): 4904–4913

LIU Y L, YAN Y, HAN Z, *et al*. Comparative effects of dietary soybean oil and fish oil on the growth performance, fatty acid composition and lipid metabolic signaling of grass carp, *Ctenopharyngodon idella*. Aquaculture Reports, 2022, 22: 101002

LIU Z, YUAN Y W, ZHAO Y, *et al*. Differentiating wild, lake-farmed and pond-farmed carp using stable isotope and multi-element analysis of fish scales with chemometrics.

Food Chemistry, 2020, 328: 127115

LÓPEZ-MARCOS S, ESCOBEDO-FREGOSO C, PALACIOS E, *et al*. Muscle transcriptional response and fatty acid profile of Pacific white shrimp *Litopenaeus vannamei* fed dietary fish and canola oil: Insights into growth performance discrepancies. Aquaculture International, 2024, 32(6): 8479–8500

LÖSEL H, ARNDT M, WENCK S, *et al*. Exploring the potential of high-resolution LC-MS in combination with ion mobility separation and surrogate minimal depth for enhanced almond origin authentication. Talanta, 2024, 271: 125598

MA T T, ZHOU W, CHEN L K, *et al*. Concerns about the future of Chinese fisheries based on illegal, unreported and unregulated fishing on the Hanjiang River. Fisheries Research, 2018, 199: 212–217

MALTSEV Y, MALTSEVA K, KULIKOVSKIY M, *et al*. Influence of light conditions on microalgae growth and content of lipids, carotenoids, and fatty acid composition. Biology, 2021, 10(10): 1060

MUHAMMAD A M, YANG C, LIU B, *et al*. Comparative analysis of meat quality and hindgut microbiota of cultured and wild bighead carp (*Hypophthalmichthys nobilis*, Richardson 1845) from the Yangtze River area. Microorganisms, 2024, 13(1): 20

MUKHERJEE A K, KALITA P, UNNI B G, *et al*. Fatty acid composition of four potential aquatic weeds and their possible use as fish-feed neutraceuticals. Food Chemistry, 2010, 123(4): 1252–1254

NIU B, YUAN X C, ROEPER P, *et al*. HIV-1 protease cleavage site prediction based on two-stage feature selection method. Protein and Peptide Letters, 2013, 20(3): 290–298

REN H, ZHANG G Q, HUANG Y, GAO X C. Effects of different dietary lipid sources on fatty acid composition and gene expression in common carp. Czech Journal of Animal Science, 2020, 65(2): 51–57

SCHEUER F, STERZELECKI F C, WAGNER R, *et al*. Proximate and fatty acids composition in the muscle of wild and farmed sardine (*Sardinella brasiliensis*). Food Chemistry Advances, 2024, 4: 100637

SIM J, MCGOVERIN C, OEY I, *et al*. Stable isotope and trace element analyses with non-linear machine-learning data analysis improved coffee origin classification and marker selection. Journal of the Science of Food and Agriculture, 2023, 103(9): 4704–4718

SRISURADETCHAI P, SUKSRIKRAN K. Random kernel k-nearest neighbors regression. Frontiers in Big Data, 2024, 7: 1402384

TIAN T, KONG F, YANG R, *et al*. A Bayesian network model for prediction of low or failed fertilization in assisted reproductive technology based on a large clinical real-world data. Reproductive Biology and Endocrinology, 2023, 21(1): 8

WANG L M, XIONG J R, XU C C, *et al*. Comparison of muscle nutritional composition, texture quality, carotenoid metabolites and transcriptome to underling muscle quality difference between wild-caught and pond-cultured Yellow River carp (*Cyprinus carpio* Haematopterus). Aquaculture, 2024, 581: 740392

WANG Y H, SU N, LIAN E G, *et al*. Spatial heterogeneity of sedimentary organic matter sources in the Yangtze River estuary: Implications from fatty acid biomarkers. Marine Pollution Bulletin, 2024, 201: 116249

XIA Y G, LIU Q F, ZHU S L, *et al*. Do changes in prey community in the environment affect the feeding selectivity of silver carp (*Hypophthalmichthys molitrix*) in the Pearl River, China? Sustainability, 2022, 14(18): 11175

YANG B S, LI W J, WU X J, *et al*. Comparison of ruptured intracranial aneurysms identification using different machine learning algorithms and radiomics. Diagnostics, 2023, 13(16): 2627

YANG H, YUAN Q, RAHMAN M M, *et al*. Biochemical, histological, and transcriptomic analyses reveal underlying differences in flesh quality between wild and farmed ricefield eel (*Monopterus albus*). Foods, 2024, 13(11): 1751

YOON D S, BYEON E, KIM D H, *et al*. Effects of temperature and combinational exposures on lipid metabolism in aquatic invertebrates. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology, 2022, 262: 109449

YU H, ZHOU J, LIN Y, *et al*. Effects of dietary lipid source on fatty acid composition, expression of genes involved in lipid metabolism and antioxidant status of grass carp (*Ctenopharyngodon idellus*). Aquaculture Nutrition, 2018, 24(5): 1456–1465

ZHANG L, YE L T, SONG Z W, *et al*. Comparative metabolomics reveals biosignatures in wild *vs*. farmed *Hypophthalmichthys nobilis*: A UHPLC-MS/MS-based authentication strategy. Food Chemistry, 2025, 495: 146510

ZHU T Y, YANG R, XIAO R G, *et al*. Effect of swimming training on the flesh quality in Chinese Perch (*Siniperca chuatsi*) and its relationship with muscle metabolism. Aquaculture, 2023, 577: 739926

(编辑　冯小花)

Discrimination of Wild *Hypophthalmichthys molitrix* Based on

Fatty Acid Profiling and Machine Learning

QIN Lei[1], NIU Bing[1], LÜ Jizhou[2①], CHEN Qin[1①]

(1. *School of Life Science*, *Shanghai University*, *Shanghai* 200444, *China*;
2. *Chinese Academy of Quality and Inspection & Testing*, *Beijing* 100176, *China*)

Abstract The Yangtze River Basin, one of the world's most biodiverse freshwater ecosystems, has experienced a sharp decline in fishery resources and a continuous decrease in biodiversity in recent years owing to long-term, high-intensity human activities. To facilitate holistic conservation and restoration of the Yangtze River's ecological environment, China implemented a "Ten-Year Fishing Ban" policy in 2020, with a complete prohibition of commercial fishing from January 1, 2021. However, during the enforcement of this ban, some merchants engaged in illegal fishing of wild aquatic products and misrepresented them as farmed products for profit, severely undermining law enforcement and ecological recovery. *Hypophthalmichthys molitrix*, an economically important freshwater fish species in China, is frequently involved in illegal fishing. Owing to the high morphological similarity between wild and farmed individuals, traditional identification methods are inadequate to meet the demands of high-precision and high-efficiency regulatory oversight. Therefore, developing a scientific and reliable technique to distinguish between wild and farmed *H. molitrix* is critical and urgent.

To address this, this study focused on *H. molitrix*. The aim was to establish an accurate method for discriminating between wild and farmed individuals by systematically comparing differences in the fatty acid composition of their muscle tissues and integrating various machine-learning algorithms. Wild *H. molitrix* samples were collected from different sections of the Yangtze River and from commercially available farm samples. Fatty acids in the muscle tissue were analyzed using gas chromatography-mass spectrometry, and their relative contents were calculated using the area normalization method to obtain comprehensive fatty acid profile data. Subsequently, a discrimination model system incorporating six typical machine learning algorithms—Bayes net, logistic regression, K-nearest neighbors (K-NN), adaptive boosting M1 (AdaBoost.M1), random forest, and decision table—was constructed. The performance of each algorithm in discriminating between the wild and farmed *H. molitrix* strains was systematically evaluated.

A correlation-based feature selection subset evaluator (CFS subset interval) algorithm was introduced to screen fatty acid variables to further enhance the model performance and generalization capability. This process ultimately identified the seven most discriminative fatty acids: C16:1, C17:0, C20:2, C20:3n3, C20:4n6, C20:5n3, and C22:6n3. The six aforementioned machine learning models were reconstructed based on the selected key fatty acids. Their performance was systematically compared across multiple metrics, including accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC). The results indicated that feature dimensionality reduction significantly improved the overall discriminative ability of all models. The model built on the AdaBoost.M1 algorithm demonstrated optimal performance, achieving an accuracy of 90.5% (sensitivity, 89.5%; specificity, 91.3%) on the training set and 81.0% (sensitivity, 80.0%; specificity, 81.8%) on the test set. The AUC values for the development and test sets were as high as 0.97 and 0.94, respectively, indicating excellent fit and generalization stability. This model can be considered as the optimal choice for discriminating between wild and farmed *H. molitrix*. Other algorithms, such as K-NN and Bayes Net, also showed significant performance improvements after feature selection,

① Corresponding author: LÜ Jizhou, Email: Ljzffff@163.com; CHEN Qin, Email: chenqincc@shu.edu.cn

further validating the crucial role of the seven selected fatty acids in distinguishing wild and farmed *H. molitrix*.

In summary, we systematically analyzed the fatty acid profiles of muscle tissues from wild and farmed *H. molitrix*. By combining effective feature selection with multiple machine-learning algorithms, a reliable discrimination model for wild and farmed *H. molitrix* was successfully constructed. This method not only provides a robust scientific basis and practical tool for the traceability of aquatic products through technical pathways, but also offers technical support for the effective implementation of the fishing ban policy in the Yangtze River Basin and for the scientific conservation and ecological restoration of fishery resources, demonstrating the broad application prospects and practical applications.

Key words *Hypophthalmichthys molitrix*; Wild; Farmed; Feature selection; Machine learning