



农业环境科学学报

Journal of Agro-Environment Science

ISSN 1672-2043, CN 12-1347/S

《农业环境科学学报》网络首发论文

题目：基于红外光谱的猪粪反应器堆肥氨排放浓度速测模型构建
作者：贾远昊，杨增军，Cheryl Marie Cordeiro, Erik Sindhøj, Heinz W.Siesler, 秦文佼，赵润，张克强
收稿日期：2026-04-29
网络首发日期：2026-07-07
引用格式：贾远昊，杨增军，Cheryl Marie Cordeiro, Erik Sindhøj, Heinz W.Siesler, 秦文佼，赵润，张克强. 基于红外光谱的猪粪反应器堆肥氨排放浓度速测模型构建[J/OL]. 农业环境科学学报.
<https://link.cnki.net/urlid/12.1347.s.20260706.1926.006>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

基于红外光谱的猪粪反应器堆肥氨排放浓度速测模型构建

贾远昊^{1,2}, 杨增军², Cheryl Marie Cordeiro³, Erik Sindhöj³, Heinz W. Siesler⁴, 秦文俊², 赵润^{2*}, 张克强^{1,2*}

(1.东北农业大学资源与环境学院, 哈尔滨 150030; 2.农业农村部环境保护科研监测所, 天津 300191, 3.Division of Environment and Natural Resources, Norwegian Institute of Bioeconomy Research (NIBIO), Ås, NO-1431, Norway; 4.Department of Physical Chemistry, University Duisburg-Essen, Essen,45117, Germany)

摘要: 针对猪粪反应器堆肥氨排放监测过程样品采集时效性不足、有效样本易损失、浓度预测稳定性弱等问题, 本研究综合运用光声光谱和红外光谱技术搭建了氨气原位在线监测装备系统; 开发了基于对数分箱的多指标迭代异常筛除方法(A), 同传统整体阈值单一指标法(B)比对了筛除结果; 采用偏最小二乘、随机森林、极端随机树等方法分别构建了不同浓度分箱的氨浓度定量预测模型, 并基于多指标评价方法对模型性能进行了系统评价。结果表明: 相较于方法B, 方法A的样本留存率提高22.4%, 动态范围损失降低96.2%。在模型构建中, ET模型在各分箱中 R^2_{Pred} 达0.894~0.973, $RPD \geq 3.0$, 平均全量程误差(MAE%FS)为 $\pm 2.19\% \sim \pm 2.68\%$, 优于RF和PLS。在后续验证中, ET模型3个分箱的 R^2_{Val} 分别为0.908、0.978、0.989, RPD 分别为3.217、6.184、6.157, MAE%FS分别为 $\pm 2.98\%$ 、 $\pm 3.20\%$ 、 $\pm 3.59\%$, 满足复杂场景下高精度定量分析要求。研究表明, 基于对数分箱的多指标迭代异常筛除方法与ET模型相结合, 可实现氨浓度的快速定量反演, 为反应器堆肥过程氨排放水平的精准监测提供技术支撑。

关键词: 猪粪; 反应器堆肥; 氨; 定量预测分析; 红外光谱; 异常样本筛除; 预测模型

Development of a rapid ammonia concentration prediction model for pig manure reactor composting based on infrared spectroscopy

Jia Yuanhao^{1,2}, Yang Zengjun², Cheryl Marie Cordeiro³, Erik Sindhöj³, Heinz W. Siesler⁴, Qin Wenjiao², Zhao Run^{2*}, Zhang Keqiang^{1,2*}

(1. College of Resources and Environment, Northeast Agricultural University, Harbin 150030, China; 2. Agro-Environmental Protection Institute, Ministry of Agriculture and Rural Affairs, Tianjin 300191 China; 3. Division of Environment and Natural Resources, Norwegian Institute of Bioeconomy Research (NIBIO), Ås, NO-1431, Norway; 4. Department of Physical Chemistry, University Duisburg-Essen, Essen, 45117, Germany)

Abstract: To address the issues of poor timeliness in sample collection, susceptibility to loss of valid samples, and instability of concentration prediction during the monitoring of ammonia emissions from pig manure reactor composting, this study established an in-situ online monitoring system for composting gases using infrared spectroscopy. A logarithmic binning-based multi-metric iterative outlier removal method (Method A) was developed and compared with a conventional global-threshold single-indicator method (Method B). Partial least squares (PLS), random forest (RF), and extremely randomized trees (ET) were employed to construct quantitative prediction models for ammonia concentration across different concentration bins, and the model performance was systematically evaluated using a multi-metric assessment approach. The results showed that Method A achieved an overall removal rate of only 7.86% with a dynamic range loss of 2.69%, whereas Method B had a removal rate of 30.23% with a dynamic range loss of 98.86%. In model construction, the ET model achieved R^2_{Pred} values of 0.8938~0.9729 across

收稿日期: 2026-04-29 录用时间: 2026-06-19

作者简介: 贾远昊(2000—), 男, 河南濮阳人, 硕士研究生, 研究方向为环境保护与修复。E-mail: 13949707854@163.com

*通讯作者: 赵润 E-mail: zhaorun01@caas.cn; 张克强 E-mail: kqzhang68@126.com

基金项目: 国家重点研发计划项目(2023YFD1702000)

Project supported: National Key Research and Development Program(2023YFD1702000)

different bins, $RPD \geq 3.0$, and Mean Absolute Error as a percentage of Full Scale ($MAE\%FS$) of $\pm 2.19\% \sim \pm 2.68\%$, outperforming RF and PLS. In subsequent validation, the ET model yielded R^2_{val} values of 0.907 7, 0.978 0, and 0.9891 for the three bins, with RPD values of 3.217 0, 6.183 7, and 6.157 3, and $MAE\%FS$ of $\pm 2.98\%$, $\pm 3.20\%$, and $\pm 3.59\%$, respectively, meeting the requirements for high-precision quantitative analysis in complex scenarios. This study demonstrates that the combination of the logarithmic binning-based multi-metric iterative outlier removal method and the ET model enables rapid quantitative prediction of ammonia concentration, providing technical support for precise online monitoring of ammonia emission levels in reactor composting processes.

Keywords: pig manure; reactor composting; ammonia; quantitative prediction analysis; infrared spectroscopy; outlier removal; rapid prediction model

规模化猪场粪便堆肥处理过程氨等有害气体排放量大、强度高、引发恶臭投诉等问题突出,成为制约企业可持续运营和产业绿色发展的重要瓶颈^[1,2]。适宜的堆肥工艺可有效减少氨排放,其中反应器堆肥占地面积小、发酵效率高,近年来被越来越多地用于猪场粪污处理。但由于实际生产中通气方式不同引起的氨排放浓度跨度大、变异性强且受环境因子交叉干扰,难以实时捕捉到完整的气体动态排放特征,有必要建立场景适配、快速精准的原位监测方法,为堆肥全周期氨排放水平溯源、跟踪监测及减排效果评估等提供有效技术支撑^[3-6]。

目前,原位监测手段主要包括电化学传感法、电子鼻及光谱学技术^[7]。其中,电化学传感器在高粉尘、高水汽的复杂环境中易被腐蚀,稳定性差、使用寿命有限;电子鼻易受交叉气体及温湿度干扰,稳定性差、精度有限^[7-9]。相比之下,光谱学技术时效性强、精准度高、稳定性好、多组分气体同步反演,融合现代化学计量学方法建模对氨浓度快速预测,面向反应器密闭堆肥气体组分原位监测的应用前景广阔^[7,10-12]。

传统的氨排放监测多依赖于不同位点特定时段采样取平均值,操作繁琐、时效性差、样本代表性不足,难以反映气体浓度的连续动态变化特征;且由于堆肥现场干扰因素众多、排放情况复杂,易影响氨浓度的精准测量。如水汽具有宽谱吸收特性,会干扰多种目标气体的光谱信号^[13,14];粉尘颗粒物会引发红外光的散射和折射,易导致采样管道堵塞^[15,16];气体之间的交叉吸收也是光谱采集的重要干扰源^[17]。此外,代表性样本的获取是构建高精度校正模型的重要前提。堆肥各阶段时间分布不均会导致气体样本呈偏态分布,在样本筛选中易造成有效样本丢失,降低校正模型的鲁棒性^[7,18,19];机器学习凭借其复杂数据样本集中强大的特征提取与训练能力,可有效降低干扰因素对模型性能的影响,实现模型的精准定量预测。搭载原位连续监测系统,构建适用于偏态数据的高精度预测模型,是实现堆肥氨排放精准监测的关键路径。因此,本研究通过创建高频在线监测装备系统,快速、全面捕捉堆肥全周期氨浓度样本;融合光谱学技术与机器学习模型,提高建模分析效率与预测精度;开发适用于工况条件的异常样本筛选方法,减少有效样本丢失提升样本代表性,综合“三步曲”实现猪粪反应器堆肥氨排放的快速精准速测。

针对上述的技术瓶颈,本研究自主创建了气体高频在线监测装备系统,实现了堆肥全过程氨浓度与红外光谱的连续、高密度原位采集;开发了对数分箱与多指标迭代融合的异常样本筛选方法,比对了基于整体浓度单一指标(学生化残差)方法筛选后的样本保留情况;运用极端随机树(Extremely Randomized Trees)、偏最小二乘(Partial Least Squares)、随机森林(Random Forest)等多元算法建立了两种筛选方法下的定量预测模型^[20-23],并对模型性能进行了比对验证,以破解传统监测方法样本代表性差、有效样本易丢失、模型鲁棒性弱等问题,为反应器堆肥场景氨排放浓度的快速反演、排放通量的准确核算、排放环节的靶向减控提供关键技术支撑。

1 材料与方法

1.1 堆肥过程氨排放原位监测系统创建

试验创建了由堆肥反应器发酵单元、气体浓度实时测定、气体光谱原位采集三件套装置串接而成的氨排放连续自动在线监测系统。其中，堆肥反应器有效容积 100 L，物料总装填料量 27 kg。系统通气管道采用内径为 6 mm 的特氟龙管，每 3 d 翻堆一次，堆肥过程持续发酵至物料腐熟。在线监测系统如图 1 所示。

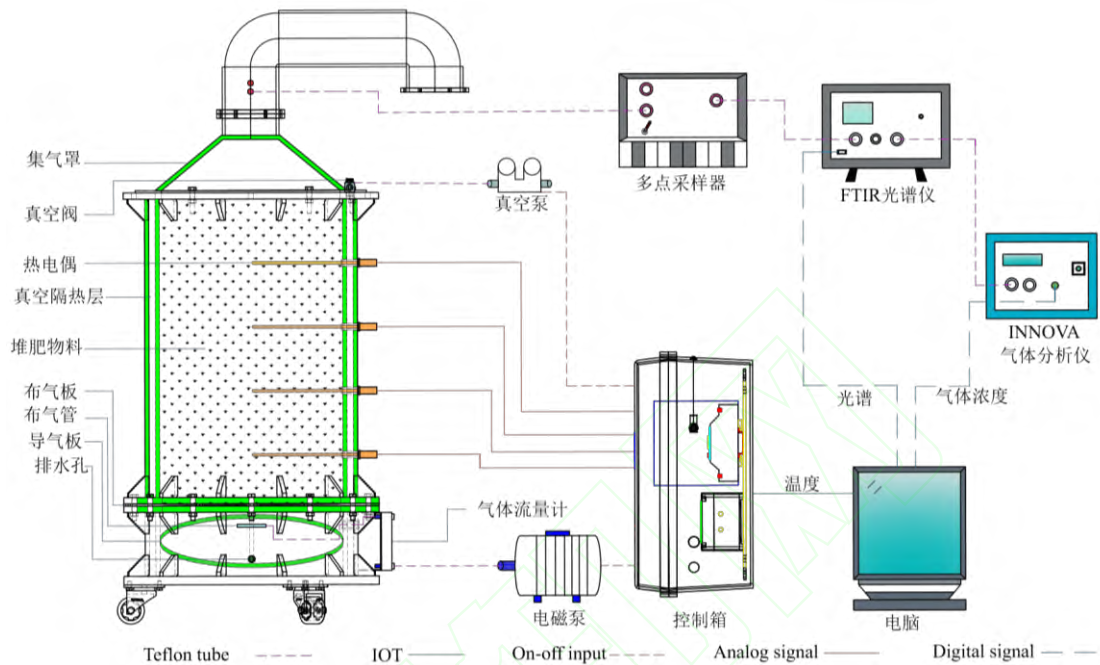


图 1 猪粪反应器堆肥过程氨气在线监测系统示意图

Figure 1 Schematic Diagram of the Compost Gas Monitoring System

监测系统全天候运行，每日 08:00 至 24:00 连续自动采集气体浓度与红外光谱数据。系统可实现反应器堆肥气体排放的高频监测，检测时长平均 $\leq 90\text{ s}\cdot\text{样}^{-1}$ 。

堆肥原料为新鲜猪粪和经粉碎处理后的干燥玉米秸秆（长度 3~5 cm），猪粪与秸秆的混合比例为 2.12:1。混合物料装填至反应器总容积的 90%~100%。新鲜猪粪初始含水率为 78.7%，通过添加玉米秸秆将混合物料含水率调整至 58%^[24]。通气量设定为 $30\text{ L}\cdot\text{min}^{-1}$ ，采用实际生产中习惯常用的间歇通气方式：通气 30 min，停止通气 30 min，循环往复进行。

1.2 样本采集与指标测定

1.2.1 氨浓度测定

氨浓度测定采用基于光声光谱（Photoacoustic Spectroscopy, PAS）技术原理的 INNOVA-1512 多功能气体分析仪及多点空气采样器。待测气体进入分析仪前，由聚四氟乙烯（PTFE）材质的过滤器去除粉尘颗粒物及部分水汽。采样频率 $2\text{ min}\cdot\text{样}^{-1}$ 。仪器相关参数如表 1 所示。

表 1 INNOVA-1512 型多功能气体分析仪主要技术参数

Table 1 Main technical specifications of INNOVA-1512 multi-gas analyzer

参数 Parameter	氨气 Ammonia
检出限	$0.08\text{ }\mu\text{mol}\cdot\text{mol}^{-1}$
准确度	$\pm 2\%\text{ FS}$
重复性	$\leq 0.03\text{ }\mu\text{mol}\cdot\text{mol}^{-1}$
量程	$0.08\sim 900\text{ }\mu\text{mol}\cdot\text{mol}^{-1}$

气体分析仪测定的氨浓度作为参考标准值，用于后续光谱定量模型的校正与验证。该仪器精度高、量程覆盖广，但设备昂贵、维护复杂，对工况条件要求严苛，难以长期连续在线运行。

1.2.2 气体光谱采集

相比 PAS，傅里叶变换红外光谱（Fourier Transform Infrared Spectroscopy，FTIR）具有响应速度快、原位光谱连续采集等优势，为后续氨气浓度校正模型的构建与应用奠定了基础。本试验中使用 BRUKER OMEGA 5 型便携式傅里叶变换红外光谱仪采集氨气红外光谱。检测前待测气体通过预处理装置过滤粉尘颗粒物及部分水汽；进入光谱仪后，在 FTIR 自带气体池内进一步加热并保持 50 °C 恒温，防止水汽在设备内部结露。光谱采集频率为 0.5 min·样⁻¹，可同步获取每个样本的平行光谱，满足氨气排放过程的连续监测需求。红外光谱仪相关参数如表 2 所示。

表 2 BRUKER OMEGA 5 型便携式傅里叶变换红外光谱仪主要技术参数
Table 2 Main technical specifications of the BRUKER OMEGA 5 portable Fourier transform infrared Spectrometer

参数 Parameter	规格 Specification
干涉仪类型	迈克尔逊干涉仪
分辨率	0.5~16 cm ⁻¹ （可调）
光谱范围	820~4 500 cm ⁻¹
检测器类型	MCT 检测器（碲镉汞）
光源	高稳定性红外光源
响应时间	≤1 min

在 20 d 的连续监测中共采集到 5 633 个气体样本，24 890 条光谱。

1.3 光谱预处理

原始光谱数据在承载目标气体组分特征信息的同时，亦混杂基线漂移、散射效应及随机噪声等多种干扰因素，因此需选用适配的预处理方法消除上述干扰，以提升模型的预测性能。光谱预处理方法主要包括：基线校正（Baseline）、标准正态变量变换（Standard Normal Variate，SNV）、Savitzky-Golay 平滑及导数变换等。其中，基线校正可消除由样品状态和测量条件差异引起的基线漂移；SNV 能够校正因颗粒度分布不均导致的散射效应；Savitzky-Golay 一阶导数平滑可有效抑制随机噪声并锐化重叠峰^[25-27]。在去除光谱中含有空白值吸光度的波段后，本研究设计了四种预处理组合方案：A 方案（仅基线校正）、B 方案（基线校正+SNV）、C 方案（基线校正+Savitzky-Golay 一阶导数平滑）、D 方案（基线校正+SNV+Savitzky-Golay 一阶导数平滑）。为全面评估不同预处理组合对各类模型的适用性，预处理筛选阶段同时考察了线模型“岭回归（Ridge）与偏最小二乘（PLS）”、树模型“随机森林（RF）与极端随机树（ET）”，以及基于神经网络的模型“卷积神经网络（CNN）”的预测性能，通过校正决定系数（ R^2_{cal} ）、预测决定系数（ R^2_{pred} ）、校正均方根误差（RMSEC）、预测均方根误差（RMSEP）、平均绝对误差（MAE）和相对分析误差（RPD）等指标以识别出对模型类型不敏感、稳健性较好的预处理策略。

表 3 不同预处理与建模方法的预测性能比对
Table 3 Comparison of prediction performance using different preprocessing methods and modeling algorithms

模型 Model	方案 Scheme	R^2_{cal}	RMSEC	RMSECV	R^2_{Pred}	RMSEP	RPD
Ridge	A	0.945 8	9.105 4	13.077 8	0.700 3	11.682 8	4.418 6
Ridge	B	0.945 6	10.035 6	13.129 3	0.774 3	11.287 7	4.573 3
Ridge	C	0.934 7	11.222 5	15.473 3	0.708 4	12.809 1	4.030 1
Ridge	D	0.937 2	12.197	16.719 8	0.689 4	14.385 2	3.588 6
PLS	A	0.905 2	15.788 1	17.206 2	0.661 9	14.245 5	3.623 8
PLS	B	0.893 4	18.587 7	19.667 5	0.700 9	14.039 2	3.677 0
PLS	C	0.902 3	17.195 0	19.639 3	0.684 5	15.224 8	3.390 7
PLS	D	0.893 3	21.400 2	23.523 7	0.677 4	16.360 3	3.155 3
RF	A	0.969 3	3.087 7	7.380 3	0.909 3	4.588 1	11.251 4

RF	B	0.971 6	3.279 1	8.511 5	0.922 2	4.256 4	12.128 1
RF	C	0.968 0	3.357 8	8.378 9	0.907 9	5.072 4	10.177 1
RF	D	0.968 3	3.117 3	8.166 4	0.911 5	5.178 5	9.968 5
ET	A	0.984 8	1.860 6	6.885 3	0.922 3	3.486 6	14.806 0
ET	B	0.983 6	1.895 6	6.572 2	0.929 4	3.065 8	16.837 9
ET	C	0.982 9	2.054 7	7.813 9	0.915 4	3.417 6	15.105 0
ET	D	0.982 6	1.731 6	6.632 9	0.923 8	3.028 3	17.046 5
CNN	A	0.569 8	64.916 3	68.256 3	0.359 8	34.502 4	1.496 2
CNN	B	0.762 8	32.395 4	33.645 2	0.630 8	19.278 6	2.677 7
CNN	C	0.557 6	66.966 4	68.536 7	0.301 5	36.850 0	1.400 9
CNN	D	0.563 5	66.912 4	67.736 3	0.285 2	37.446 0	1.378 6

1.4 异常样本筛除与分集

由于氨浓度跨度较大 ($1.55\sim904.33\ \mu\text{mol}\cdot\text{mol}^{-1}$) 且呈右偏分布, 若直接使用全局异常阈值会误删高浓度样本。对此, 本研究开发了一套基于偏最小二乘回归 (PLS) 的分箱迭代异常样本筛除方法, 通过对数分箱策略将浓度数据划分为若干子区间, 在每个分箱内进行独立异常识别, 避免了全局统一阈值对稀疏区间正常样本的系统性误判。筛除流程包含浓度对数分箱、稳健主成分分析 (Robust-PCA) 初筛、PLS 多指标迭代精筛三个步骤。同时将该方法与传统的整体阈值单一指标 (学生化残差) 法对异常样本的筛除结果及对有效样本的保留情况进行比对分析。

1.4.1 浓度对数分箱

浓度对数变换: 设原始浓度值为 c , 为避免原始浓度值为 0 时对数计算无意义, 引入极小值 ε (取 10^{-6}), 对原始浓度进行常用对数变换, 变换后浓度值 c' 为:

$$c' = \log_{10}(c + \varepsilon) \quad (1)$$

等宽分箱边界计算: 设定分箱数为 n , 计算变换后浓度值的最小值 $c'_{\min}$ 与最大值 $c'_{\max}$, 按等宽原则计算各分箱的变换后边界值 c'_i :

$$c'_i = c'_{\min} + i \cdot \frac{c'_{\max} - c'_{\min}}{n}, (i = 0, 1, \dots, n) \quad (2)$$

基于上述计算结果, 将所有样本按原始浓度值分配至对应浓度分箱, 输出分箱标签、分箱边界及各分箱样本统计信息, 为后续分箱迭代精筛提供数据基础。

考虑到氨浓度度较大 (接近 3 个数量级) 且在对数空间近似均匀分布, 结合物理分段合理性 (低、中、高三段对应堆肥不同阶段)、单箱样本量均衡性 (每箱约 100 个以上) 及实际建模效果 ($n=3$ 时 RMSECV 最低), 将分箱数设为 3。

1.4.2 稳健主成分分析 (Robust-PCA) 异常初筛

在迭代精筛之前, 首先采用 Robust-PCA 对每个分箱内的样本进行异常初筛, 以快速识别并剔除光谱严重畸变、采样污染或仪器波动引起的极端异常样本。Robust-PCA 相比传统 PCA 对极端值具有更强的鲁棒性。

对于每个分箱内的光谱数据矩阵 $\mathbf{X}$ (样本 $\times$ 波数), 首先进行稳健标准化 (减去中位数并除以中位数绝对偏差 MAD), 然后采用奇异值分解提取主成分。计算每个样本的重构误差:

$$Q_i = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2 \quad (3)$$

其中 $\mathbf{x}_i$ 为第 i 个样本的原始光谱向量, $\hat{\mathbf{x}}_i$ 为利用保留主成分重构的光谱向量。重构误差 Q_i 越大, 表明样本光谱特征偏离主流模式越显著。

为确保光谱重构精度与抗噪能力的平衡, 分析显示, 前 10 个主成分的累计方差贡献率已超过 95%, 能够有效分离正常与异常样本的重构误差, 且该取值符合红外光谱领域有效化

学信息维度 5~15 的经验范围, 因此将 Robust-PCA 保留主成分数设为 10。

采用中位数绝对偏差法确定异常判定阈值。设分箱内所有样本重构误差的中位数为 $\text{median}(Q)$, 则中位数绝对偏差 (MAD) 为:

$$\text{MAD}(Q) = \text{median}(|Q - \text{median}(Q)|) \quad (4)$$

则阈值设定为:

$$\text{threshold} = \text{median}(Q) + k \cdot \text{MAD}(Q) \quad (5)$$

通常取 $k=3$ 。当 Q_i 大于该阈值时, 将样本标记为初筛异常。同时结合重复测量一致性校验计算的可疑度评分 (样本光谱与其重复组均值光谱的欧氏距离), 经双重判定后剔除初筛异常样本。初筛阶段采用相对保守的阈值策略, 以避免正常样本被误删, 主要目的是减轻后续精筛阶段的计算负担。

为降低重尾分布光谱数据中正常样本被误删的概率, 学生化残差、Q 残差及 KNN 距离的异常阈值均采用中位数加 5 倍 MAD 即 $k=5$, 该阈值较常规的 3 倍 MAD 更为保守。

1.4.3 PLS 多指标迭代精筛

初筛完成后, 对每个分箱内的样本执行迭代精筛。该步骤基于 PLS 回归模型的建模结果推导学生化残差 (Studentized Residual)、霍特林 T^2 (Hotelling T^2)、杠杆值 (Leverage)、Q 残差 (Q residual) 与 K 近邻距离 (K-Nearest Neighbors distance) 5 个核心异常指标, 各指标从残差拟合、特征空间影响力、数据重构、局部密度等不同维度综合评估每个样本的异常程度, 并通过多指标融合投票机制做出判定。迭代过程持续至无新增异常样本为止。

为平衡统计稳定性与局部敏感性, 避免个别噪声干扰或范围过大模糊局部结构, 在 PLS 得分空间中计算每个样本与其 5 个最近邻的平均欧氏距离, 该取值符合化学计量学 KNN 异常检测的通行标准, 因此将 K 值设为 5。

为避免 PLS 模型过拟合, 采用 K 折交叉验证 ($K=5$) 根据 RMSECV 自动优选潜变量数, 并将潜变量数的上限设为 10。

(1) 多指标融合投票判定

设定五项指标对样本 i 的异常标记结果为 $F_1^{(i)}, F_2^{(i)}, \dots, F_5^{(i)}$, 其中 $F_j^{(i)}=1$ 表示判定为异常, $F_j^{(i)}=0$ 表示判定为正常。本研究采用投票逻辑进行融合, 设定投票阈值为 4, 即至少 4 个指标判定为异常时, 综合标记该样本为异常:

$$F_{\text{final}}^{(i)} = 1 \left\{ \sum_{j=1}^5 F_j^{(i)} \geq 4 \right\} \quad (6)$$

本研究所采用的 5 个异常指标中, 学生化残差与 Q 残差属于残差类, 霍特林 T^2 与杠杆值属于杠杆/位置类, K 近邻距离属于局部密度类。为防止单一类型异常指标偏差导致样本过度剔除, 保证全面考察样本的异常情况, 要求 5 个异常指标中至少 4 个同时判定为异常时才筛除样本, 因此将多指标投票阈值设定为 4 (共 5 个指标)。

(2) 迭代终止条件

每轮迭代中剔除所有综合标记为异常的样本, 然后利用剩余样本重新构建 PLS 模型, 计算新一轮的 5 项指标, 再次进行投票判定。重复上述过程, 直至某一轮次中没有新样本被标记为异常时迭代终止, 输出精筛后的纯净样本集。

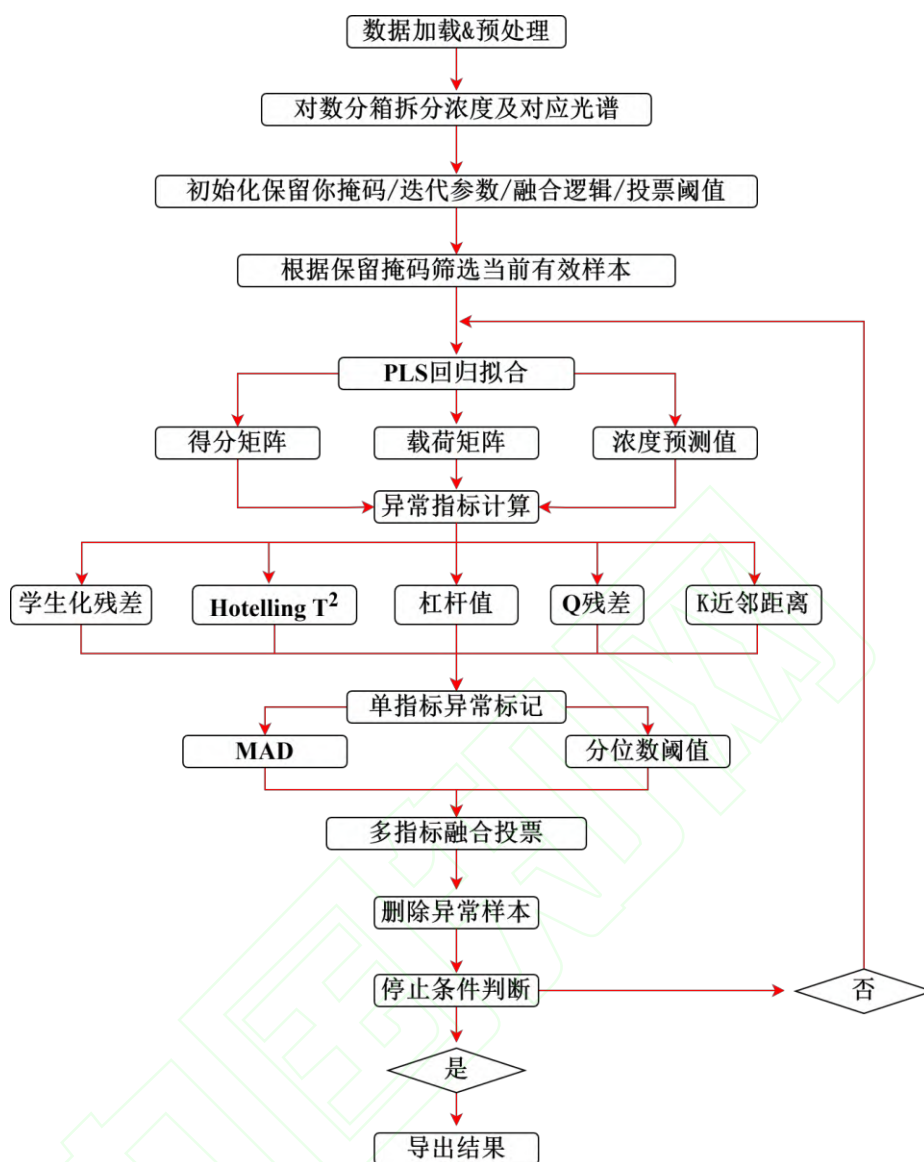


图 2 分箱 PLS 多指标迭代精筛核心算法流程示意图

Figure 2 Flowchart of the core algorithm for binning-based PLS multi-metric iterative refinement
1.4.4 样本集划分 (SPXY 法)

异常样本筛除后, 采用光谱-理化值共生矩阵法 (SPXY) 将纯净样本集划分为校正集与预测集。SPXY 算法在 Kennard-Stone 算法基础上同时考虑光谱变量 ($\mathbf{X}$) 和浓度变量 ($\mathbf{y}$) 的距离^[28]。本研究设定预测集样本数占总样本数的 30%, 其余 70% 作为校正集, 并确保预测集样本的浓度范围不超出校正集的浓度范围, 以避免模型外推预测。

表 4 校正集与预测集样本信息

Table 4 Sample information of prediction set and validation set

分箱 Bin	分类 Classification	样品数 Sample amount	浓度范围 Concentration range/($\mu\text{mol}\cdot\text{mol}^{-1}$)	均值 Mean value/($\mu\text{mol}\cdot\text{mol}^{-1}$)
bin0	校正集	3 151	1.552~12.886	2.999
	预测集	1 350	1.672~12.097	2.834
bin1	校正集	184	13.011~106.760	40.493
	预测集	79	13.181~106.760	41.358
bin2	校正集	298	114.726~904.338	586.614
	预测集	128	115.324~903.996	579.54

1.5 建模方法及评判标准

考虑到不同浓度区间内气体浓度与光谱特征之间可能呈现不同的线性或非线性关系, 本

研究采用分箱建模策略，即在各浓度子区间内独立构建定量预测模型，以避免全局模型被低浓度主体样本主导而忽略高浓度稀疏区间的特异性。模型算法方面，鉴于 1.3 节光谱预处理中 Ridge 作为线模型预测能力相对偏低，CNN 模型复杂度高、可解释性弱且拟合效果差，而 PLS（线模型）和 RF（树模型）预测效果较好；因此选择更具代表性和工程实用性的 PLS（线模型），以及极端随机树（ET）和 RF 等树模型参与后续的模型构建。为避免单一指标评价造成偏差，研究建立了由基础指标与派生指标组成的联合评价体系，其中基础指标直接反映模型拟合与预测精度，派生指标用于评价模型泛化能力与稳定性。

为客观比较偏最小二乘（PLS）、随机森林（RF）和极端随机树（ET）的预测性能，本研究对树模型（RF、ET）的关键超参数进行了系统优化。超参数优选采用网格搜索（Grid Search）结合 3 折交叉验证（3-fold cross-validation）的策略，以交叉验证决定系数（ $CV-R^2$ ）作为评分指标，从预设的候选参数组合中选取最优配置。为避免测试集信息泄露，整个建模流程嵌套在 SPXY 样本划分框架内：内层在校正集（占总样本 70%）上执行 3 折 CV 网格搜索以确定最优超参数，外层以独立预测集（30%）的 RMSEP、 R^2 和 RPD 作为模型泛化性能的无偏估计。

RF 的超参数搜索空间如下：树数量（ $n_estimators$ ）设为 220（经预试验确认模型误差已收敛）；最大深度（ max_depth ）取 8~10；节点分裂所需的最小样本数（ $min_samples_split$ ）设为 20~40；叶节点所需的最小样本数（ $min_samples_leaf$ ）设为 8~16；随机特征比例（ $max_features$ ）取 0.25~0.35；代价复杂度剪枝系数（ ccp_alpha ）取 1×10^{-4} ~ 5×10^{-4} ；bootstrap 抽样比例（ $max_samples$ ）取 0.6~0.8。

ET 模型的超参数搜索范围与 RF 基本一致，但树数量设为 300（因为 ET 采用随机分裂阈值，单棵树的偏差较高，需要更多的树来降低集成方差），同时额外比对了是否启用 bootstrap 两种情形。

PLS 模型的潜变量数同样纳入网格搜索，候选值为 2、4、6、8、10、12，采用相同的 3 折交叉验证决定系数（ $CV-R^2$ ）准则确定最优潜变量数，以保证 3 种算法在超参数优化方式上具有可比性。

候选值范围设定遵循高维光谱数据建模原则是：红外光谱特征维度高、相邻波长间强相关，因此通过限制树深度、节点最小样本数及采用 0.25~0.35 的随机特征比例，可有效抑制过拟合并打破特征共线性；同时引入代价复杂度剪枝进一步控制模型复杂度，增强泛化能力。

1.5.1 基础评价指标

基础指标包括校正决定系数（ R^2_{cal} ）、预测决定系数（ R^2_{pred} ）、校正均方根误差（RMSEC）、预测均方根误差（RMSEP）、平均绝对误差（MAE）和相对分析误差（RPD）。其中，决定系数（ R^2 ）衡量模型解释能力，均方根误差（RMSE）衡量预测精度，MAE 对异常值不敏感，RPD 则综合误差与样本差异，按阈值（>2.0 优秀，1.6~2.0 良好，<1.6 不可用）判断模型实用价值^[29,30]。

1.5.2 派生评价指标

ΔR^2 （校正-预测拟合差）：由 R^2_{cal} 与预测集 R^2_{pred} 计算得到，公式为：

$$\Delta R^2 = R^2_{cal} - R^2_{pred} \quad (7)$$

该指标衡量训练性能与验证性能的一致性，用于识别过拟合风险。 ΔR^2 越小，表明训练与验证越一致，泛化越稳定； ΔR^2 明显为正且较大时，提示存在过拟合。

RMSEP/RMSEC（误差放大比值）：由验证误差与校正误差相除得出。该指标衡量模型从训练到验证阶段误差的放大程度，越接近 1 表示误差迁移越稳定，显著大于 1 说明验证阶段误差放大，泛化能力偏弱。

ΔR^2 侧重反映拟合能力的下降程度，RMSEP/RMSEC 侧重反映误差的放大程度，二者互为补充。

MAE%FS（平均全量程误差）：为评价模型在各量程段的预测精度，引入平均全量程误差（MAE%FS）作为评价指标。该指标将模型的平均绝对误差归一化至对应量程范围，消除了不同浓度区间因量程差异导致的误差不可比性，便于在分段模型间进行横向对比。计算公式为：

$$\text{MAE\%FS} = \frac{\text{MAE}}{\text{预测上限} - \text{预测下限}} \times 100\% \quad (8)$$

MAE%FS 越小，表明模型在该量程段内的预测偏差相对于满量程越小，精度越高。相较于传统以最大绝对误差定义的 $\pm\text{FS}$ 指标，MAE%FS 采用平均误差代替极值，能够更稳健地反映模型的整体预测水平，避免因个别异常点导致评价结果失真。

1.5.3 模型评价综合排序方法

为在同一分箱内对不同模型进行客观比较，本研究建立了多指标综合排序规则。在每个分箱内，分别对 R^2_{pred} 、 R^2_{cal} （越大越优）以及 RMSEP、 ΔR^2 、RMSEP/RMSEC、MAE%FS（越小越优）进行单项排名（排名数值越小表示在该指标上表现越好）。将各单项排名求和，得到每个模型的综合得分，再按综合得分升序确定最终排名。综合得分最低者排名第一，其模型质量等级定为“高”；排名第二者定为“中”；排名第三者定为“低”。该排序策略将拟合能力、预测精度、过拟合风险和稳定性纳入统一框架，确保模型选择的客观性和可复现性。

2 结果与讨论

2.1 氨浓度分布特征

研究计以四类统计诊断图的方式呈现分布形态与统计特征（图 3）。其中，箱线图（a）通过展示数据的中位数、四分位距及异常值分布，反映数据的集中趋势与离散程度；经验累积分布函数图（b）描绘样本浓度随分位数的累积概率变化，直观呈现数据在各浓度区间的分布密度；频率分布直方图及核密度估计曲线图（c）用于观察数据的整体分布形态，其中核密度曲线可平滑估计其概率密度函数；分位数-分位数图（d）将样本分位数与理论正态分布的分位数进行比对，用于检验数据是否服从正态分布。

氨浓度统计分析结果如表 5 和图 3 所示，浓度范围为 1.552~904.338 $\mu\text{mol}\cdot\text{mol}^{-1}$ 。综合判断，氨浓度呈现典型的右偏重尾分布，即“低值高频、少量极端高值拉升均值”的分布形态。统计表中均值（51.22 $\mu\text{mol}\cdot\text{mol}^{-1}$ ）显著高于中位数（2.58 $\mu\text{mol}\cdot\text{mol}^{-1}$ ）和众数（2.11 $\mu\text{mol}\cdot\text{mol}^{-1}$ ），表明绝大多数样本浓度处于较低水平，但存在突发性高浓度情景拉高了整体均值。分位数分析进一步印证了这一特征： P_{90} （538.26）表示 90% 的样本浓度不超过该值， P_{95} （538.2626）和 P_{99} （782.58）分别对应 95% 和 99% 分位数，最大值达 904.34 $\mu\text{mol}\cdot\text{mol}^{-1}$ ，说明尾部延伸极长。箱线图中上须远长于下须且上侧离群点密集，基于四分位距（IQR，即 Q_3 与 Q_1 之差，反映中间 50% 样本的离散程度）判定的异常值比例达 16.88%，同样证实高值异常频繁出现。QQ 图右端明显上翘。经验累积分布函数（ECDF）曲线前段陡升、后段缓慢拖尾，直观显示主体样本集中在低浓度区，少量高值贡献了主要的分布。

在好氧堆肥过程中，氨的排放浓度呈现典型的右偏重尾分布，这一特征与堆肥工艺的动态特性密切相关。氨在升温期有机物快速分解阶段发生爆发式释放，形成高浓度峰值；而在大部分平稳时段（如腐熟期、降温期后期及高温期平稳阶段），氨的排放浓度较低。统计显示，氨的均值（51.22 $\mu\text{mol}\cdot\text{mol}^{-1}$ ）远高于中位数（2.58 $\mu\text{mol}\cdot\text{mol}^{-1}$ ），高分位数（ $P_{95}=538.26$ $\mu\text{mol}\cdot\text{mol}^{-1}$ ）与最大值（904.34 $\mu\text{mol}\cdot\text{mol}^{-1}$ ）跨度大，异常值比例达 16.88%，表明低浓度样本占主体，但少量高浓度剧烈排放情景对整体负荷贡献显著。这种偏态分布是通过自动连续监测对堆肥全过程高频采样反映出的客观结果，并非仪器误差或操作问题导致的伪异常，但在样本筛除中易造成有效样本丢失，降低模型预测精准度^[7,18,19]。因此，在后续建模与统计分析中，应最大化保留各区间尤其高浓度氨样本，而应区分真实排放高峰与伪异常，采用能

够兼顾主体分布与尾部事件的建模策略，如浓度分层、稳健统计或分位数回归等方法，以准确反映堆肥氨排放的动态全貌^[31,32]。

表 5 氨浓度的分位数与四分位距统计
Table 5 Quantile and Interquartile Range Statistics for Ammonia Concentrations

气体 Gas	P ₉₀	P ₉₅	P ₉₉	IQR	IQR 下界 Lower Bound	IQR 上界 Upper Bound
氨	41.442 6	538.262 6	782.576 8	2.765	-1.967 5	9.092 5

注：P₉₀、P₉₅、P₉₉ 分别表示第 90、95、99 百分位数（Percentile），即分别有 90%、95%、99% 的样本浓度低于对应数值。IQR 为四分位距（Interquartile Range），定义为第三四分位数（Q₃）与第一四分位数（Q₁）之差，反映中间 50% 样本的离散程度。IQR 下界=Q₁-1.5IQR，IQR 上界=Q₃+1.5IQR，超出该上下界的样本按四分位距法判定为异常值（离群点）。

Note: P₉₀, P₉₅, and P₉₉ denote the 90th, 95th, and 99th percentiles, respectively, meaning that 90%, 95%, and 99% of the sample concentrations fall below the corresponding values. IQR refers to the interquartile range, defined as the difference between the third quartile (Q₃) and the first quartile (Q₁), reflecting the dispersion of the central 50% of the samples. The lower bound is defined as Q₁-1.5×IQR, and the upper bound as Q₃+1.5×IQR. Samples falling outside these bounds are classified as outliers (anomalies) according to the interquartile range method.

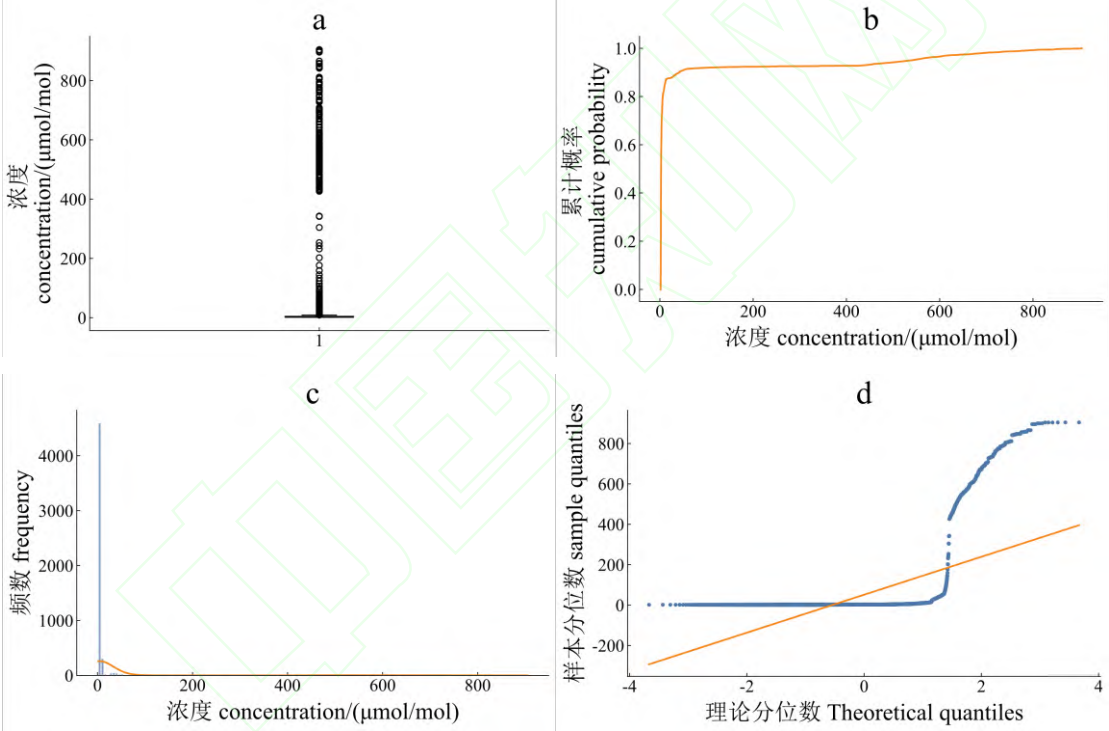


图 3 氨气浓度分布特征及概率分布分析
Figure3 Analysis of Ammonia Concentration Distribution Characteristics and Probability Distribution

2.2 异常样本剔除结果比对

将两种异常样本剔除方法处理后的结果分列于表 6 和表 7。基于对数分箱与多指标迭代的异常样本剔除方法处理后，3 个浓度分箱（bin0~bin2）的样本分别由 4 895、288、450 降至 4 501、263、426，总体保留 5190 个样本，累计剔除率 7.86%；动态范围损失极小（bin0 损失 0.54%、bin1 损失 1.42%、bin2 损失 0.73%，累计损失率 2.69%），原始高浓度样本（>100 μmol·mol⁻¹）完整保留。而传统整体阈值单一指标法将所有高于 11.31 μmol·mol⁻¹ 的样本判定为异常并剔除，动态范围从 1.552~904.338 μmol·mol⁻¹ 急剧压缩至 1.628~11.31 μmol·mol⁻¹，损失率高达 98.86%；累计剔除 1 703 个样本，总体剔除率达 30.23%，高浓度区间样本完全丢失。

上述结果表明，基于对数分箱与多指标迭代的异常样本剔除方法在保留原始浓度梯度、

避免有效样本误删方面具有显著优势,尤其适用于偏态严重、动态范围宽的光谱数据预处理。整体阈值单一指标法出现系统性误判的根本原因在于,其隐含了数据分布相对均匀的前提假设,而在宽动态范围、严重偏态的数据中,这一前提并不成立;稀疏区间的高浓度样本因远离由主体样本主导的全局中心,其距离或残差被系统性放大,进而被误判为异常^[33,34]。对于这类动态范围大、分布严重右偏的数据,整体阈值单一指标法将中高浓度样本全部剔除,导致关键高值信息大量流失,样本数据难以代表堆肥场景真实的排放情况,从而影响后续精准定量预测模型的构建。因此,基于对数分箱与多指标迭代的异常样本筛除方法保留原始浓度梯度、避免有效样本误删方面具有显著优势,尤其适用于偏态严重、动态范围宽的样本数据,能够为后续构建高精度定量预测模型提供高质量的数据基础。

表 6 两种方法筛除异常样本前后结果比对

Table 6 Comparison of sample retention before and after removal under different outlier removal methods				
分箱 Bin	筛除前样本数 Before removal	筛除后样本数 After removal	累计剔除样本数 Total removed	累计剔除率 Total removal rate/%
0	4 895	4 501	394	8.05
1	288	263	25	8.68
2	450	426	24	5.33
分箱汇总	5 633	5 190	443	7.86
单一阈值&指标	5 633	3 930	1 703	30.23

表 7 两种方法筛除异常样本前后动态范围比较

Table 7 Comparison of dynamic ranges before and after removal under different outlier removal methods				
分箱 Bin	筛除前/($\mu\text{mol}\cdot\text{mol}^{-1}$) Before removal	筛除后/($\mu\text{mol}\cdot\text{mol}^{-1}$) After removal	动态范围损失 Dynamic range loss(%)	累计损失 Cumulative loss(%)
0	1.55~12.96	1.55~12.89	0.54	
1	12.96~108.33	13.01~106.81	1.42	2.69
2	108.33~904.33	114.72~904.33	0.73	
单一阈值&指标	1.55~904.33	1.63~11.31	98.86	98.86

2.3 模型预测性能比对

光谱预处理结果(表 3)表明,经基线校正与 SNV 联合处理后,各模型均表现出良好的预测精度与稳定性,因此将该方案确定为最优预处理方法用于后续建模。根据浓度分布特征,将整体样本集划分为 bin0、bin1、bin2 三个分箱。评价时综合考察预测能力(R^2_{Pred} 、RMSEP)与拟合-泛化一致性(ΔR^2 、RMSEP/RMSEC)。图 4 中 ETbin0 为 bin0 分箱的 ET 模型,ETbin1 为 bin1 分箱的 ET 模型,ETbin2 为 bin2 分箱的 ET 模型,其他图像同理。3 个分箱的样本分集结果如表 4 所示。

在氨的 3 个浓度分箱中,3 种模型的预测性能差异明显(表 8)。在 bin0(低浓度区),ET 模型的 R^2_{Pred} 为 0.894、RMSEP 为 $0.434 \mu\text{mol}\cdot\text{mol}^{-1}$ 、MAE%FS 为 $\pm 2.19\%$,明显优于 RF ($R^2_{\text{Pred}}=0.870$ 、RMSEP= $0.480 \mu\text{mol}\cdot\text{mol}^{-1}$ 、MAE%FS= $\pm 2.29\%$)和 PLS ($R^2_{\text{Pred}}=0.764$ 、RMSEP= $0.646 \mu\text{mol}\cdot\text{mol}^{-1}$ 、MAE%FS= $\pm 4.25\%$)。在 bin1(中浓度区),ET 表现最佳, R^2_{Pred} 达 0.973、RMSEP 为 $3.769 \mu\text{mol}\cdot\text{mol}^{-1}$ 、 ΔR^2 仅 0.001、MAE%FS 为 $\pm 2.39\%$,训练与验证几乎一致;RF 和 PLS 的 RMSEP 分别为 $6.907 \mu\text{mol}/\text{mol}$ 和 $7.323 \mu\text{mol}\cdot\text{mol}^{-1}$ 。在 bin2(高浓度区),ET 仍最优, R^2_{Pred} 为 0.972、RMSEP 为 $30.118 \mu\text{mol}\cdot\text{mol}^{-1}$ 、MAE%FS 为 $\pm 2.68\%$,RF 和 PLS 的 RMSEP 分别为 $35.125 \mu\text{mol}\cdot\text{mol}^{-1}$ 和 $34.683 \mu\text{mol}\cdot\text{mol}^{-1}$,MAE%FS 分别为 $\pm 3.22\%$ 和 $\pm 3.29\%$ 。

综合各分箱的预测集 R^2_{Pred} 、RMSEP、 ΔR^2 及 RMSEP/RMSEC 等排名,ET 在所有分箱中均排名第一,质量等级均为“高”,模型效果优先序为 ET>RF>PLS。ET 在氨的 3 个浓

度区间均表现出最佳预测能力和泛化性,适用于偏态分布、动态范围宽的气体浓度定量预测。

ET 在各分箱中均获得最佳预测效果,这与其算法特性有关。首先,光谱信号与氨浓度之间存在复杂的非线性关系,在宽浓度范围内尤为显著^[35]。ET 通过构建多棵决策树并随机选择特征分裂点,能够有效捕捉此类非线性模式;而 PLS 作为线性模型处理非线性关系的能力有限^[36]。其次,ET 对极端值具有更强的鲁棒性,其双重随机性(每次训练随机抽取部分样本、每次分裂随机选取特征)降低了个别高浓度异常样本对整体模型的影响,避免 RF 因完全随机采样可能导致的局部过拟合^[37]。因此,在处理严重偏态且重尾数据时,ET 更能抵御稀疏尾部高值样本带来的过拟合风险,从而提升模型在尾部区间的稳定性和泛化能力。

值得注意的是,在预处理筛选中,CNN 在所有预处理方案下的预测性能均显著低于 RF 和 ET (R^2_{Pred} 仅为 0.285~0.631, $RPD \leq 2.68$)。这可能是因为 CNN 模型参数量大,而本研究的有效光谱样本仅 5 633 条,相对有限的数量难以支撑网络充分训练,易陷入过拟合^[38];气体红外光谱在去除水汽和基线漂移后,有效特征维度相对较低,光谱与氨浓度之间的映射关系更趋于线性或浅层非线性,CNN 的深层特征提取能力未能充分发挥^[39];CNN 对输入光谱的平移、缩放等变化较为敏感,尽管经过预处理,但现场采集的光谱仍存在微小波动,影响了 CNN 的泛化稳定性^[40]。相比之下,RF 和 ET 作为集成树模型,参数较少、对数据规模要求低,且天然具有抗过拟合能力,更适合本研究的有限样本场景。

表 8 氨模型预测结果比对
Table. 8 Comparison of Ammonia Model Performance

分箱 Bin	bin0	bin0	bin0	bin1	bin1	bin1	bin2	bin2	bin2
模型	ET	RF	PLS	ET	RF	PLS	ET	RF	PLS
校正集 R^2_{Cal}	0.960	0.949	0.886	0.974	0.925	0.934	0.980	0.972	0.976
校正集 RMSEC	0.379	0.425	0.622	3.128	5.266	4.843	23.871	30.592	25.312
校正集 MAE	0.203	0.217	0.417	2.163	3.482	3.752	16.003	21.963	20.430
预测集 R^2_{Pred}	0.894	0.870	0.764	0.973	0.890	0.879	0.972	0.965	0.96
RMSEP	0.434	0.480	0.646	3.769	6.907	7.323	30.118	35.125	34.683
预测集 MAE	0.231	0.242	0.449	2.243	4.408	5.569	21.142	25.395	26.004
RPD	3.026	2.738	2.034	5.422	2.959	2.791	5.689	4.878	4.940
ΔR^2	0.066	0.079	0.122	0.001	0.035	0.055	0.008	0.007	0.016
RMSEP/RMSEC	1.147	1.129	1.039	1.205	1.312	1.512	1.262	1.148	1.370
MAE%FS	2.19%	2.29%	4.25%	2.39%	4.70%	5.94%	2.68%	3.22%	3.29%
预测集 R^2_{Pred} 排名	1	2	3	1	2	3	1	3	2
RMSEP 排名	1	2	3	1	2	3	1	3	2
校正集 R^2_{Cal} 排名	1	2	3	1	3	2	1	3	2
ΔR^2 排名	1	2	3	1	2	3	2	1	3
RMSEP/RMSEC 排名	3	2	1	1	2	3	2	1	3
MAE%FS 排名	1	2	3	1	2	3	1	2	3
总体排名	1	2	3	1	2	3	1	2	3
质量等级	高	中	低	高	中	低	高	中	低

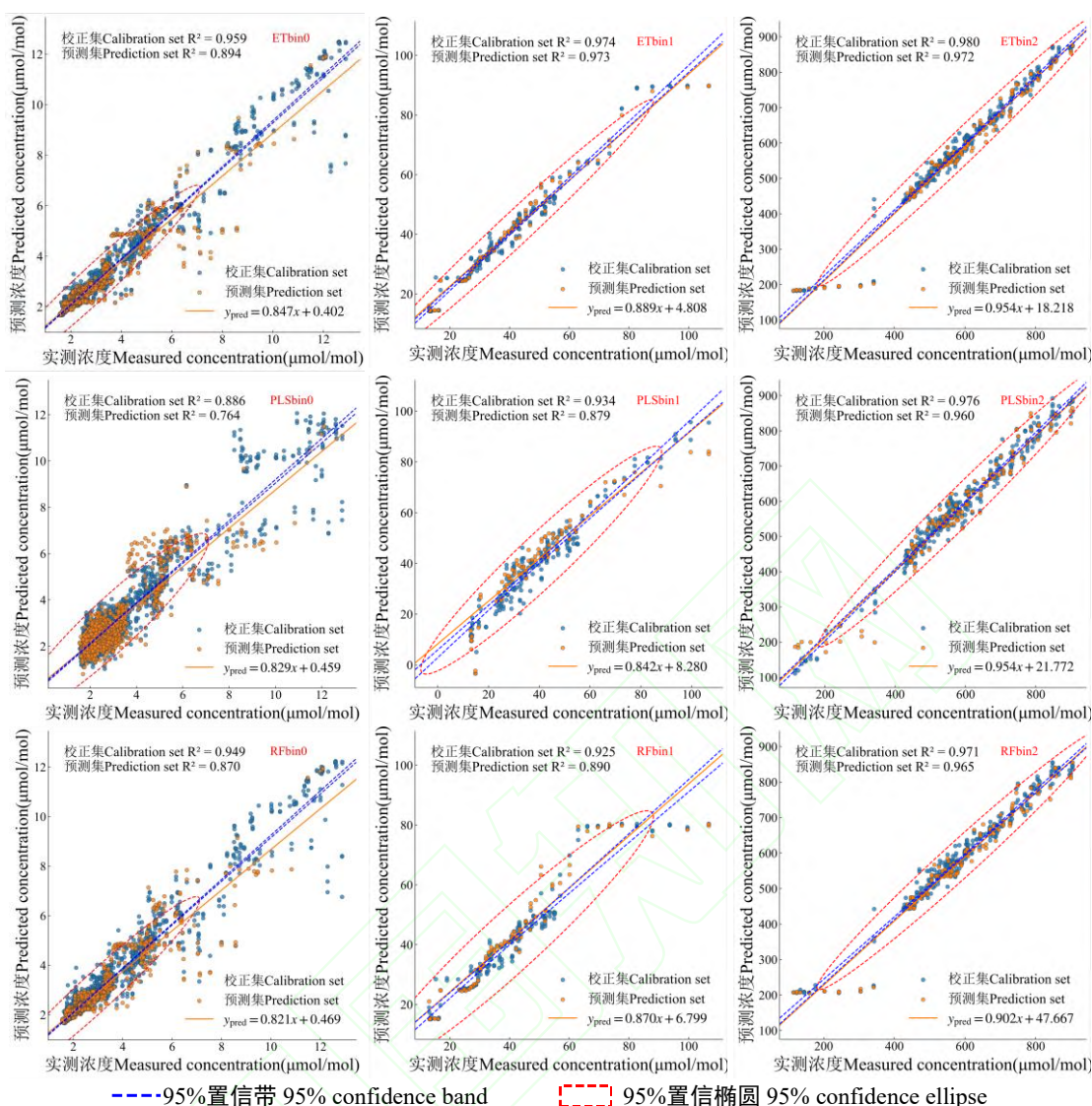


图4 氨浓度预测值与实测值拟合结果（95%置信区间）

Figure 4 Predicted versus Measured Ammonia Concentrations with 95% Confidence Bands

2.4 模型验证

为进一步验证所构建氨浓度速测模型的实际预测能力与泛化性能，本研究将独立验证集样本的红外光谱数据代入已构建的最优极端随机树（ET）定量模型。低浓度区间 bin0 模型验证决定系数 (R^2_{val}) 为 0.908，预测均方根误差 (RMSEP) 为 $0.672 \mu\text{mol}\cdot\text{mol}^{-1}$ ，相对分析误差 (RPD) 为 3.217；中浓度区间 bin1 模型 R^2_{val} 达 0.978，RMSEP 为 $4.444 \mu\text{mol}\cdot\text{mol}^{-1}$ ，RPD 为 6.184；高浓度区间 bin2 模型 R^2_{val} 为 0.989，RMSEP 为 $34.408 \mu\text{mol}\cdot\text{mol}^{-1}$ ，RPD 为 6.157。3 个分箱模型 RPD 均大于 3.0；bin0~bin2 的 MAE%FS 分别为 $\pm 2.98\%$ 、 $\pm 3.20\%$ 、 $\pm 3.59\%$ ，处于较低水平；结果表明，ET 模型能够准确捕捉猪粪好氧堆肥全周期氨浓度的动态变化特征，该模型在低、中、高全浓度区间均表现出可靠的预测能力与良好的适用性，成功实现了氨浓度的快速、准确、稳定速测，可为猪粪反应器堆肥过程氨排放通量的核算与减控提供技术支撑。

该模型目前仅基于单一反应器堆肥场景的气体排放数据开发，对其他堆肥工艺条件下氨浓度的预测泛化能力与精度尚需进一步验证与优化。然而，针对连续在线监测所产生的大体量样本集导致的数据偏态分布及样本代表性下降问题，本文提出的基于分箱策略的多指标迭代异常样本剔除方法及分箱建模思路，具有一定的普适性与借鉴价值。相关策略可为其他堆肥工艺场景中预测模型的构建、验证与优化提供方法学参考。

表 9 氨模型验证结果

Table 9 Validation results of the ammonia model

分箱 Bin	bin0	bin1	bin2
校正集 R^2_{cal}	0.959	0.974	0.980
RMSEC	0.379	3.128	23.871
校正集 MAE	0.203	2.163	16.003
验证集 R^2_{val}	0.908	0.978	0.989
RMSEP	0.672	4.444	34.408
验证集 MAE	0.307	2.993	25.027
RPD	3.217	6.184	6.157
MAE%FS	$\pm 2.98\%$	$\pm 3.20\%$	$\pm 3.59\%$

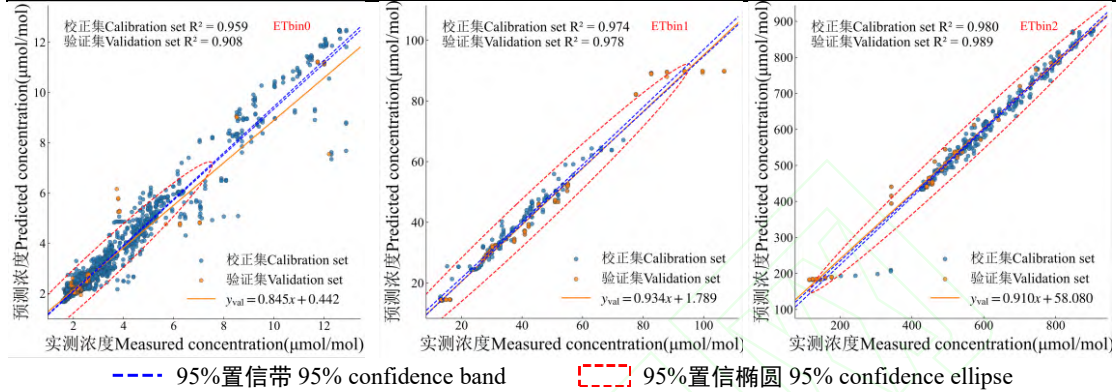


图 5 氨气浓度验证拟合结果

Figure. 5 Ammonia concentration validation fitting results

3 结论

(1) 猪粪好氧堆肥氨排放浓度呈右偏重尾分布，均值 ($51.22 \mu\text{mol}\cdot\text{mol}^{-1}$) 远高于中位数 ($2.58 \mu\text{mol}\cdot\text{mol}^{-1}$)， P_{95} 达 $538.26 \mu\text{mol}\cdot\text{mol}^{-1}$ ，异常值比例 16.88%。低浓度样本占主体，但少量高浓度事件对整体负荷贡献显著，该偏态是高频连续监测的客观结果，非仪器伪异常。

(2) 开发的基于对数分箱的多指标迭代异常筛除方法可有效保留原始浓度梯度，总体剔除率 7.86%，动态范围损失仅 2.69%；而传统整体阈值单一指标法剔除率达 30.23%，动态范围损失 98.86%，高浓度样本完全丢失。该方法能避免对稀疏尾部样本的系统性误删，为后续建模提供高质量数据基础。

(3) 光谱最优预处理方案为基线校正+SNV。分箱建模策略下，极端随机树模型 (ET) 在各分箱中均表现最优，预测集 R^2_{Pred} 为 0.894~0.973、RPD 为 3.026~5.689，平均全量程误差 MAE%FS 为 $\pm 2.98\%\sim\pm 3.59\%$ ，显著优于随机森林和偏最小二乘，且在后续验证中表现优秀，验证集 R^2_{val} 为 0.908~0.989，RPD 为 3.217~6.184。ET 模型可实现猪粪反应器堆肥全过程氨浓度的快速、准确预测，具有良好的泛化性能。

参考文献:

- [1] 孙耀威, 刘金松, 刘玉兰, 等. 规模化猪场臭气减排技术研究进展[J]. 中国畜牧兽医, 2024, 51(2): 591-600. Sun Y W, Liu J S, Liu Y L, et al. Research progress in odor reduction technology for large-scale pig farms[J]. China Animal Husbandry & Veterinary Medicine, 2024, 51(2): 591-600.
- [2] Wang Y, Xu H S, Wang Y X, et al. Achieving overall low greenhouse gas and ammonia emissions from digested pig slurry during storage and after field application by maintaining a micro-oxygen level at the storage stage[J]. Journal of Environmental Chemical Engineering, 2025, 13(3): 116712.
- [3] Nordahl S L, Preble C V, Kirchstetter T W, et al. Greenhouse gas and air pollutant emissions from composting[J]. Environmental Science & Technology, 2023, 57(6): 2235-2247.
- [4] Cao Y B, Wang X, Misselbrook T, et al. Quantification of N and C cycling during aerobic composting, including automated direct measurement of N_2 , N_2O , NO, NH_3 , CO_2 and CH_4 emissions[J]. Science of the Total Environment, 2024, 891: 166412.

Environment, 2023, 857(Pt 2): 159177.

- [5] 湛卓越, 贺德春, 姜珊, 等. 鸡粪好氧堆肥过程氨气与温室气体排放特征及协同减排机制[J]. 农业环境科学学报, 2023, 42(11): 2582-2594. Zhan Z Y, He D C, Jiang S, et al. Emission characteristics and the mechanism of synergistic emission reduction for NH_3 and greenhouse gases during the aerobic composting of chicken manure[J]. Journal of Agro-Environment Science, 2023, 42(11): 2582-2594.
- [6] Cao Y B, Wang X, Zhang L, et al. Ammonium enrichment in livestock manure driven by ammonia-abatement practices can reduce nitrous oxide emissions[J]. One Earth, 2024, 7(4): 687-696.
- [7] Li T L, Wang C X, Ji W T, et al. Cutting-edge ammonia emissions monitoring technology for sustainable livestock and poultry breeding: a comprehensive review of the state of the art[J]. Journal of Cleaner Production, 2023, 428: 139387.
- [8] 汪开英, 吴捷刚, 赵晓洋. 畜禽场空气污染物检测技术综述[J]. 中国农业科学, 2019, 52(8): 1458-1474. Wang K Y, Wu J G, Zhao X Y. Review of measurement technologies for air pollutants at livestock and poultry farms[J]. Scientia Agricultura Sinica, 2019, 52(8): 1458-1474.
- [9] Hawko C, Verrielle M, Hucher N, et al. A review of environmental odor quantification and qualification methods: the question of objectivity in sensory analysis[J]. Science of the Total Environment, 2021, 795: 148862.
- [10] Ma W H, Ji X T, Ding L Y, et al. Automatic monitoring methods for greenhouse and hazardous gases emitted from ruminant production systems: a review[J]. Sensors, 2024, 24(13): 4423.
- [11] Bai M, Suter H, Lam S K, et al. Gaseous emissions from an intensive vegetable farm measured with slant-path FTIR technique[J]. Agricultural and Forest Meteorology, 2018, 258: 50-55.
- [12] Witkowska D. Volatile gas concentrations in Turkey houses estimated by Fourier Transform Infrared Spectroscopy (FTIR)[J]. British Poultry Science, 2013, 54(3): 289-297.
- [13] Tsiulyanu D, Stratan I, Tsiulyanu A, et al. Investigation of the oxygen, nitrogen and water vapour cross-sensitivity to NO_2 of tellurium-based thin films[J]. Sensors and Actuators B: Chemical, 2007, 121(2): 406-413.
- [14] Neftel A, Fischer C, Flechard C. Measurements of greenhouse gas fluxes from agriculture[J]. International Congress Series, 2006, 1293: 3-12.
- [15] Klug J C, Beuting M, Sanders S T. Toward inverse backscatter absorption gas imaging: imaging a dry gas plume with ambient H_2O absorption[J]. Optics Express, 2025, 33(11): 22416-22430.
- [16] Kang R Y, Liatsis P, Kyritsis D C. Emission quantification via passive infrared optical gas imaging: a review[J]. Energies, 2022, 15(9): 3304.
- [17] Kohl L, Koskinen M, Rissanen K, et al. Technical note: interferences of volatile organic compounds (VOCs) on methane concentration measurements[J]. Biogeosciences, 2019, 16: 3319-3336.
- [18] Cattaneo M, Tayà C, Burgos L, et al. Assessing ammonia and greenhouse gas emissions from livestock manure storage: comparison of measurements with dynamic and static chambers[J]. Sustainability, 2023, 15(22): 15987.
- [19] Xu Z, Zhao B, Wang Y Y, et al. Composting process and odor emission varied in windrow and trough composting system under different air humidity conditions[J]. Bioresource Technology, 2020, 297: 122482.
- [20] Sharma A, Goyal M K. A comparison of three soft computing techniques, Bayesian regression, support vector regression, and wavelet regression, for monthly rainfall forecast[J]. Journal of Intelligent Systems, 2017, 26(4): 641-655.
- [21] Feigenwinter I, Hörtnagl L, Buchmann N. N_2O and CH_4 fluxes from intensively managed grassland: the importance of biological and environmental drivers vs. management[J]. Science of the Total Environment, 2023, 903: 166389.
- [22] Wang H J, Liang Q X, Hancock J T, et al. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods[J]. Journal of Big Data, 2024, 11(1): 44.
- [23] 郭昱辰, 杨亮, 刘春红, 等. 基于 RF-LSTM 的鸡舍恶臭气体预测研究[J]. 中国环境科学, 2020, 40(7):

- 2850-2857. Guo Y C, Yang L, Liu C H, et al. Prediction of stench gas in chicken house based on RF-LSTM[J]. China Environmental Science, 2020, 40(7): 2850-2857.
- [24] 周谈龙, 尚斌, 董红敏, 等. 低碳氮比条件下猪粪堆肥氨气和温室气体排放[J]. 中国农业气象, 2017, 38(11): 689-698. Zhou T L, Shang B, Dong H M, et al. Emission characteristics of ammonia and greenhouse gas during the low C/N ratio swine manure composting[J]. Chinese Journal of Agrometeorology, 2017, 38(11): 689-698.
- [25] Munawar A A, Yunus Y, Devianti, et al. Calibration models database of near infrared spectroscopy to predict agricultural soil fertility properties[J]. Data in Brief, 2020, 30: 105469.
- [26] Tomuta I, Porfire A, Casian T, et al. Multivariate calibration for the development of vibrational spectroscopic methods[M]//Stauffer M T, ed. Calibration and Validation of Analytical Methods: A Sampling of Current Approaches. InTech, 2018.
- [27] Wahid R, Ward A J, Møller H B, et al. Biogas potential from forbs and grass-clover mixture with the application of near infrared spectroscopy[J]. Bioresource Technology, 2015, 198: 124-132.
- [28] Galvão R K, Araujo M C, José G E, et al. A method for calibration and validation subset partitioning[J]. Talanta, 2005, 67(4): 736-740.
- [29] 刘生博, 孙迪, 李梦婷, 等. 基于样品复配的奶牛场粪水近红外光谱氮磷速测模型构建[J]. 农业环境科学学报, 2023, 42(11): 2595-2603. Liu S B, Sun D, Li M T, et al. Rapid determination modeling of slurry nitrogen and phosphorus from dairy farm based on samples compounding[J]. Journal of Agro-Environment Science, 2023, 42(11): 2595-2603.
- [30] 孙迪, 李梦婷, 牟美睿, 等. 近-中红外光谱融合技术速测奶牛场粪水氮磷含量[J]. 光谱学与光谱分析, 2021, 41(10): 3092-3098. Sun D, Li M T, Mu M R, et al. Rapid determination of nitrogen and phosphorus in dairy farm slurry via near-mid infrared fusion spectroscopy technology[J]. Spectroscopy and Spectral Analysis, 2021, 41(10): 3092-3098.
- [31] Mukherjee R, Debattista K, Bashford-Rogers T, et al. Uniform color space based high dynamic range video compression[C]//Proceedings of the Color and Imaging Conference / IEEE International Conference on Image Processing (ICIP). Coventry, UK: University of Warwick, 2018: 1-12.
- [32] del Mar Rueda M, Martínez-Puertas S, Castro-Martín L. Methods to counter self-selection bias in estimations of the distribution function and quantiles[J]. Mathematics, 2022, 10(24): 4726.
- [33] Anonymous. Toward Secure and Comfortable Life in Emerging AI and Data-Driven Era!!: The 26th International Conference on Advanced Communications Technology: Phoenix Pyeongchang, Korea (South), Feb. 4-7, 2024[M]. Piscataway: IEEE, 2024.
- [34] Pimentel M A F, Clifton D A, Clifton L, et al. A review of novelty detection[J]. Signal Processing, 2014, 99: 215-249.
- [35] Shayanmehr M, Aarabi S, Ghaemi A, et al. A data driven machine learning approach for predicting and optimizing sulfur compound adsorption on metal organic frameworks[J]. Scientific Reports, 2025, 15: 3138.
- [36] Orhobor O I, Grinberg N F, Soldatova L N, et al. Imbalanced regression using regressor-classifier ensembles[J]. Machine Learning, 2023, 112(4): 1365-1387.
- [37] Ma X, Senko O V. Some approaches to improving prediction accuracy using ensemble methods[J]. Automatic Documentation and Mathematical Linguistics, 2025, 59(6): S542-S548.
- [38] Joshaghani M, Davari A, Hatamian F N, et al. Bayesian convolutional neural networks for limited data hyperspectral remote sensing image classification[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 5506305.
- [39] Silva R, Gramaxo Freitas O, Melo-Pinto P. Boosting the performance of SOTA convolution-based networks with dimensionality reduction: an application on hyperspectral images of wine grape berries[J]. Intelligent

Systems with Applications, 2023, 19: 200252.

- [40] Soysal D, García-Andrade X, Rodríguez L E, et al. Reevaluating convolutional neural networks for spectral analysis: a focus on Raman spectroscopy[J]. ACS Earth and Space Chemistry, 2025, 9(10): 2395-2412.

