移动扫码阅读
彭黄湖，姜勇，杨帆，等. 基于机器学习的含油污泥热解残渣含油率预测［J］. 能源环境保护，2025，39(6)：188−198.
PENG Huanghu，JIANG Yong，YANG Fan，et al. Prediction of Oil Content in Pyrolysis Residues of Oily Sludge Based on Machine Learning［J］. Energy Environmental Protection，2025，39(6)：188−198.

基于机器学习的含油污泥热解残渣含油率预测

彭黄湖[1]，姜　勇[1]，杨　帆[1,*]，陈泽洲[1]，吴圣姬[1]，车　磊[2]
（1. 湖州师范学院 工学院，浙江 湖州 313000；2. 浙江宜可欧环保科技有限公司，浙江 湖州 313000）

摘要：为快速预测含油污泥热解后残渣含油率的变化规律，指导含油污泥热解工艺参数优化，选取热解终温，热解时间，升温速率，氮气流量，含油污泥的含油率、含水率和含渣率作为输入变量，热解残渣含油率作为输出变量，采用梯度提升决策树(GBDT)、极端梯度提升(XGB)、支持向量机(SVM)及随机森林(RF)算法分别建立了含油污泥热解残渣含油率的预测模型。通过 228 组数据进行训练和测试，结果表明，GBDT、XGB、SVM 以及 RF 4 种含油率预测模型在测试集上的决定系数 R^2 分别为 0.871 6、0.866 7、0.835 6 和 0.917 1。经过贝叶斯优化算法(BOA)超参优化后，该 4 种含油率预测模型的测试集决定系数 R^2 分别提升至 0.901 2、0.900 1、0.896 5 和 0.920 4。其中，贝叶斯优化的随机森林(BO-RF)模型预测效果更佳，能更准确地预测含油污泥热解残渣含油率的动态变化规律。

关键词：含油污泥；热解；含油率预测；特征重要性分析；机器学习；贝叶斯优化算法
中图分类号：X502; X705　　**文献标识码：**A　　**文章编号：**2097−4183(2025)06−0188−11

Prediction of Oil Content in Pyrolysis Residues of Oily Sludge Based on Machine Learning

PENG Huanghu[1]，JIANG Yong[1]，YANG Fan[1,*]，CHEN Zezhou[1]，
WU Shengji[1]，CHE Lei[2]
（1. *School of Engineering*, *Huzhou University*, *Huzhou* 313000, *China*；
2. *Zhejiang Eco Environmental Technology Co.*, *Ltd.*, *Huzhou* 313000, *China*）

Abstract：To rapidly predict changes in residual oil content after the pyrolysis of oily sludge and to guide the optimization of pyrolysis process parameters, this study collected a dataset comprising 228 samples and employed machine learning methods to predict the oil content in the oily sludge pyrolysis residues. Several factors were used as input variables, including final pyrolysis temperature, pyrolysis time, heating rate, nitrogen flow rate, initial oil content, water content, and residue content of the oily sludge. The oil content in the pyrolysis residues was used as the output variable. Methodologically, this study applied four advanced machine learning algorithms in an innovative manner: Gradient Boosting Decision Trees (GBDT), eXtreme Gradient Boosting (XGB), Support Vector Machines (SVM), and Random Forests (RF), to construct high-precision prediction models for the oil content in pyrolysis residues. These models were rigorously trained and cross-validated on a dataset comprising 228 samples to ensure their generalization ability and prediction accuracy. The results showed that the coefficients of

收稿日期：2024−08−04　　**修回日期：**2024−09−10　　**接受日期：**2024−09−12　　**DOI**: 10.20078/j.eep.20240908
基金项目：国家自然科学基金资助项目(52100143)
第一作者：彭黄湖(1988—)，男，浙江湖州人，副教授，主要研究方向为化工过程预测与参数优化。E-mail: penghh@zjhu.edu.cn
***通讯作者：**杨　帆(1987—)，女，浙江湖州人，讲师，主要研究方向为有机固废热解工艺参数优化。E-mail: 02567@zjhu.edu.cn

determination (R^2) for the GBDT, XGB, SVM, and RF models on the test set reached 0.871 6, 0.866 7, 0.835 6, and 0.917 1, respectively, providing initial validation of the effectiveness of these machine learning models in predicting oil content in oily sludge pyrolysis residues. To further improve the predictive performance of the models, this study introduced the Bayesian Optimization Algorithm (BOA) to fine-tune the hyperparameters of the models. After BOA optimization, the R^2 values of the four models significantly increased to 0.901 2, 0.900 1, 0.896 5, and 0.920 4, respectively. Among them, the Bayesian-Optimized Random Forest (BO-RF) model exhibited the best predictive performance, demonstrating high consistency on the test set and extremely high accuracy in predicting the dynamic trends of oil content in oily sludge pyrolysis residues. Furthermore, through feature importance analysis, it was found that the final pyrolysis temperature, initial oil content in the sludge, and pyrolysis duration were the most critical factors influencing the oil content in the residues. In summary, by introducing advanced machine learning algorithms combined with a Bayesian optimization strategy, this study successfully constructed high-precision prediction models for the oil content in oily sludge pyrolysis residues. The BO-RF model, in particular, offers an effective and accurate approach for predicting oil content. This achievement contributes to enhancing the pyrolysis process of oily sludge, boosting resource utilization efficiency, and advancing sustainable waste treatment methods. It provides strong support for the pyrolysis treatment of oily sludge at both theoretical and practical levels, opening up new perspectives and approaches for environmental management and resource recovery research.

Keywords：Oily sludge；Pyrolysis；Oil content prediction；Feature importance analysis；Machine learning；Bayesian optimization algorithm

0　引　　言

原油在开采、储存、运输和精炼过程中会不可避免地产生含油污泥，导致世界范围内含油污泥大量积累[1−2]。目前我国含油污泥的年产量和存量较大，并且年产量呈现增长的趋势，存量更是超过 1 000 万 t[3]。

含油污泥含有高浓度的石油烃（PHCs）、重金属以及其他有毒成分，因此被许多国家列为危险废物[4]。长期以来，我国对含油污泥的处置方式以焚烧及填埋为主，造成了极大的资源浪费和严重的环境污染。现阶段，热解技术越来越受到关注。含油污泥热解技术指在无氧气氛中对含油污泥加热，生成不同热解产物的技术，这一过程中还可以回收其中的热解油、热解气及热解残渣，且不易产生有害物质，实现了含油污泥处理的资源化和无害化[5]。针对含油污泥热解残渣的资源化，可以通过添加不同的复合固化剂将热解残渣制成路基材料[6]或者制成烟气脱硫剂[7]，但前提是残渣的含油率降至一定范围。然而残渣含油率检测较为繁琐耗时，需使用索氏提取器进行萃取，因此快速确定热解残渣含油率是实现含油污泥热解残渣资源化的关键。

包括含油污泥在内的大部分有机固废，其组分多样且热解条件较为复杂。相较于传统的实验和理论计算，使用机器学习预测有机固废处理效果正受到广泛关注。在有机固废处置中，常见的机器学习回归模型包括经典模型、树模型、神经网络模型以及种群模型[8]。其中，经典模型在处理线性问题时效果较好，树模型和神经网络模型在处理复杂的非线性问题时表现较好，比如有机固废热解产物产率及特性预测，而种群模型主要用于优化算法。SU 等[9]使用极端梯度提升（Extreme Gradient Boosting, XGB）模型，根据不同的热解条件和生物质特性建立了生物油含氧组分预测模型，为提高生物油质量提供了新思路。NGUYEN 等[10]以 10 个城市固废产生特性作为输入，通过支持向量机（Support Vector Machine, SVM）、随机森林（Random Forest, RF）等算法对城市固废产量进行预测，发现 RF 模型对城市固废产量预测效果最好，预测 R^2 大于 0.96，这有助于城市固废的规划管理。PATHY 等[11]使用 XGB 模型对藻类生物炭产量及其组成进行预测，生物炭产量预测模型的 R^2 达到了 0.84，这为理解输入参数对预测藻类

生物炭产量的影响提供了新的见解。LENG 等[12]使用梯度提升决策树(Gradient Boosting Decision Tree, GBDT)、RF、SVM 3 种模型，建立了三相产物分布和生物油热值预测模型，借助模型分析得出生物质在流化床中热解制油的最佳温度为 480 ℃。RF、GBDT、XGB 和 SVM 模型也可以用于预测固废热解的热重分析(TG)曲线[13−14]，还可以用于固废气化过程的建模[15−16]，并利用优化算法对模型进行优化，预测合成气的产率，研究发现优化模型具有较高的精度。近年来预测算法应用于有机固废的报道较多，但鲜见关于采用预测模型预测含油污泥热解过程的研究报道。为此，采用树模型预测含油污泥热解残渣含油率，将极大地节省含油污泥热解试验时间并指导含油污泥热解工艺参数优化。

为探究不同热解条件下含油污泥热解残渣含油率的变化规律，分别采用 GBDT、XGB、SVM 以及 RF 4 种机器学习算法建立含油污泥热解残渣含油率的预测模型，评估各预测模型的优劣势和可靠性并进行超参数优化。模型结果可实现含油污泥热解残渣含油率的快速预测，为指导含油污泥热解过程的智能控制提供依据。

1 材料和方法

1.1 数据收集和处理

本研究收集了 228 组数据样本[3, 5, 7, 17−31]构建含油污泥热解残渣含油率的预测模型。数据集包含不同原料类型，如落地含油污泥、罐底含油污泥、炼厂含油污泥等。原始数据收集过程中主要考虑含油污泥的自身理化特性和含油污泥热解过程的工艺参数，包括热解终温、热解时间、升温速率、氮气流量以及含油污泥的含油率、含水率和含渣率 7 种特征值。

表 1 为所有变量的平均值、标准差(SD)、最小值(Min)、四分位数(Q)以及最大值(Max)。由表 1 可以看出，热解终温、油泥含油率、油泥含水率、油泥含渣率特征变量标准差值较大，表明数据覆盖范围广泛，因此模型具有良好的泛化性能。各个分位数展示了数据的集中程度，进一步验证了数据的有效性。同时，图 1 对比了各变量的直方图和核密度图，直观地展示了数据的分布情况。直方图通过将数据划分为多个区间，展示了每个区间内数据点的频数，通过观察直方图的形状，可以识别数据的集中趋势、离散程度，这些特征有助于理解数据的基础分布情况。例如，直方图数据呈现多个波峰，这可能意味着数据来源于多个不同的子群体。核密度图是直方图的平滑版本，通过核函数对数据进行处理，生成一条连续的曲线来估计数据的概率密度，核密度图能够更清晰地展示数据的分布形状，直观地显示数据在不同区域的密度变化。图 1 可以看出，各个输入变量的特征大致呈现正态分布，同时各变量的核密度分布显示出明显的波峰，这些波峰代表了数据的集中区域。多个波峰意味着数据可能存在多个高点，这种分布特性在模型训练中非常重要，因为它提示数据集可能包含多个不同类别或来源的子群体。核密度图具有波峰状特征，通常意味着数据中存在某种非线性关系，简单的线性模型无法有效地拟合数据，预测效果较差。在这种情况下，使用非线性模型会有更好的预测效果，如多项式

表 1 数据集中数值特征的统计分析

Table 1 Statistical analysis of numerical features in the dataset

变量	特征	平均值	SD	Min	1/4Q	1/2Q	3/4Q	Max
X_1	热解终温/℃	524.09	129.20	100.00	470.00	500.00	600.00	900.00
X_2	热解时间/h	1.72	1.13	0.17	0.83	1.50	2.50	6.00
X_3	升温速率/(℃·min^{-1})	10.16	9.60	0	8.00	10.00	10.16	100.00
X_4	氮气流量/(mL·min^{-1})	125.50	57.23	20.00	100.00	100.00	125.79	400.00
X_5	油泥含油率/%	17.71	12.12	2.96	10.14	14.71	21.47	60.64
X_6	油泥含水率/%	32.45	15.94	3.29	20.17	29.51	36.75	78.70
X_7	油泥含渣率/%	49.83	20.02	6.83	37.29	55.50	66.00	86.57
Y	残渣含油率/%	0.84	0.87	0.07	0.24	0.37	1.20	3.40

回归、树模型或神经网络。同时为了捕捉波峰的特征，回归模型需要引入更多的特征，这虽然提高了模型的拟合能力和普适性，但也增加了过拟合的风险。

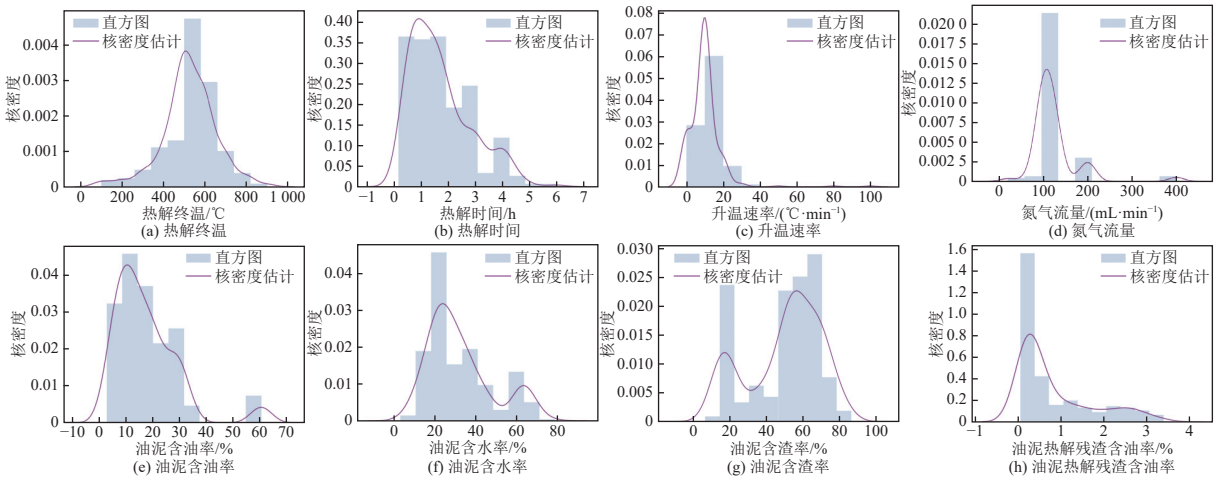

图 1　变量直方图与核密度图

Fig. 1　Histogram and kernel density plot of variables

由于各参数的数值范围不同，采用数据标准化处理，使其符合标准正态分布，从而解决各指标之间的可比性问题。正态分布函数(Z_i)见式(1)。

$$Z_i = \frac{X_i - \rho}{\sigma} \tag{1}$$

式中：X_i 为原始数据，ρ 为样本的均值，σ 为样本的标准差。

1.2　算法基本原理

(1) 梯度提升决策树。GBDT 是一种梯度提升算法，通过逐步建立一系列决策树，每棵树纠正前一棵树的错误，从而提高整体模型的准确性。具体来说，GBDT 采用加法模型的形式，通过连续的迭代优化每一棵树，使其在预测目标变量时的误差最小。

(2) 极端梯度提升。XGB 是一种高效、灵活且可扩展的梯度提升算法，结合 GBDT 的思想，同时在模型训练速度和性能上进行了大量优化。相比于 GDBT 算法，XGB 算法引入了正则化项，有效防止了过拟合。尽管 XGB 算法中的树与树之间存在串行关系，但同一层级的节点可以并行处理，从而充分利用多核处理器的计算能力，加速模型训练。此外，XGB 算法采用最大增益剪枝方法，而不是传统的预剪枝或后剪枝，这进一步提升了模型的性能。

(3) 支持向量机。SVM 是一种既可以完成分类任务，又可以实现回归任务的机器学习模型。SVM 对于非线性回归问题具有强大的处理能力，并且对异常值具有较强的鲁棒性。

(4)随机森林。RF 是一种基于决策树的集成学习方法，每棵树都是一个独立的回归模型[32]。此外，由于每棵树的训练数据和特征是随机的，RF 模型具有较强的抗过拟合能力，在回归问题上表现出色。最终模型通过集成所有树的计算结果，选出最佳结果作为 RF 后输出。

(5) 贝叶斯优化。BOA 是一种用于全局优化黑箱函数的策略。BOA 在机器学习模型的超参数调优、实验设计和其他需要高效寻优的场景中得到了广泛应用。BOA 的核心思想是通过构建目标函数的概率模型(即代理模型)，并使用该模型在可能的搜索空间中选择最有希望的采样点，然后根据分布选择下一个采样的超参数组合，从而以最少的目标函数评估次数找到全局最优解[33]。BOA 通常只需较少的采样次数就可以达到传统优化算法效果。BOA 算法的流程如图 2 所示。

(6) 贝叶斯优化算法设置。本研究将利用 BOA 算法优化 4 种模型的超参，以模型作为寻优的适应度函数，训练后模型的 R^2 作为寻优目标，找到模型的最佳超参。所有模型待优化参数见表 2，未被优化的参数保持之前的设置。

(7) 超参对模型性能的影响。因为 GBDT、XGB、RF 均属于树模型，所以它们的超参较为类似。这 3 种模型中，学习率(learning_rate)决定每棵树对最终模型的贡献大小，较低的学习率通常

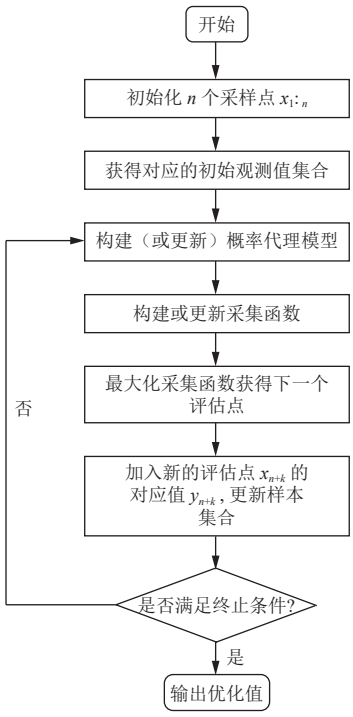

图 2 贝叶斯优化算法

Fig. 2 Bayesian optimization algorithm

表 2 各个模型待寻优超参范围

Table 2 Hyperparameter optimization ranges for each model

模型	待寻优超参	寻优范围
GBDT	learning_rate	[0.01, 0.30]
	max_depth	[3, 10]
	subsample	[0.6, 1.0]
	min_samples_split	[2, 10]
XGB	colsample_bytree	[0.1, 1.0]
	learning_rate	[0.01, 0.30]
	max_depth	[3, 15]
	n_estimators	[50, 300]
	subsample	[0.1, 1.0]
SVM	gamma	[0.01, 1.00]
	C	[1, 1 000]
	epsilon	[0.01, 1.00]
RF	max_depth	[0, 50]
	min_samples_split	[2, 30]
	min_samples_leaf	[1, 30]

可以提高模型的泛化能力，但需要更多的迭代次数；树的最大深度（max_depth）控制每棵树的最大深度，较大的深度可以提高模型的拟合能力，但可能导致过拟合；子采样率（subsample）决定每棵树训练时使用的数据子集大小，较低的值可以增加模型的多样性，减少过拟合；最小样本分裂数（min_samples_split）决定一个节点的最小样本数，以便进一步分裂，较大的值可以防止模型过拟合；列采样率（colsample_bytree）控制每棵树或每层使用的特征子集大小，增加模型的多样性；树的数量（n_estimators）决定了模型中树的数量，更多的树可以提高模型的拟合能力，但也会提高过拟合的风险；最小样本叶子数（min_samples_leaf）决定了叶子节点的最小样本数，较大的值可以防止过拟合。

SVM 模型中，正则化参数（C）控制模型的松弛变量，较小的 C 值会让模型更倾向于泛化，较大的 C 值更倾向于精确分类训练数据；核函数系数（gamma）主要用于径向基函数（RBF），控制单个训练样本的影响范围，较小的 gamma 值会使模型更泛化，较大的 gamma 值可能导致过拟合；松弛变量（epsilon）可以平衡模型的复杂度和对噪声的容忍度，从而影响最终的预测性能。

1.3 模型性能评估

为了科学、直观地表现模型的预测效果，因此引入决定系数（Coefficient of Determination, R^2）、均方根误差（Root Mean Squared Error, RMSE）、平均绝对误差（Mean Absolute Percentage Error, MAPE）作为拟合评价指标。

R^2 是用于评估回归模型拟合性能的指标，表示模型能够解释数据方差的比例，通常用于比较不同模型的表现。R^2 越接近 1，表明模型的拟合效果越好。R^2 计算公式见式（2）。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{2}$$

RMSE（E_{RMS}）是用于衡量预测模型在连续性数据上的预测精度的指标。它衡量了预测值与真实值之间的均方根差异，表示预测值与真实值之间的平均偏差程度，对偏离实际值较大的误差比较敏感。E_{RMS} 计算公式见式（3）。

MAPE（E_{MAP}）是评估预测模型准确性的指标。它通过计算预测值与实际值之间的绝对百分比误差来评估模型的预测精度。MAPE 的值越小，说明模型的预测精度越高。E_{MAP} 计算公式见式（4）。

$$E_{\mathrm{RMS}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{3}$$

$$E_{\mathrm{MAP}}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{4}$$

式中：y_i代表实际值，$\hat{y}_i$代表预测值，n 为样本数。

2　结果分析

2.1　数据特征重要性分析

首先将 7 个输入变量（热解终温、热解时间、升温速率、氮气流量以及含油污泥的含油率、含水率和含渣率）和输出（热解残渣含油率）的数据集进行归一化，随后引入了皮尔逊相关系数（Pearson′s Correlation Coefficient, PCC）度量各个变量之间的相关程度，其计算公式见式（5）。

$$C_{\mathrm{PC}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\sum_{i=1}^{n}(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{5}$$

式中：C_{PC}表示皮尔逊相关系数，x_i、y_i为数据集中的 2 个变量，$\bar{x}$、$\bar{y}$为所有 x 和 y 数据点的平均值。

图 3（a）展示了 7 种输入变量的皮尔逊相关系数热力图。PCC 的取值范围为 [−1, 1]，其中C_{PC}>0 或C_{PC}<0 表示 2 个变量之间分别存在正相关或负相关关系，而C_{PC}=0 则表示 2 个变量之间没有关联。PCC 绝对值大小反映了变量对输出结果的重要性和相关性强弱。

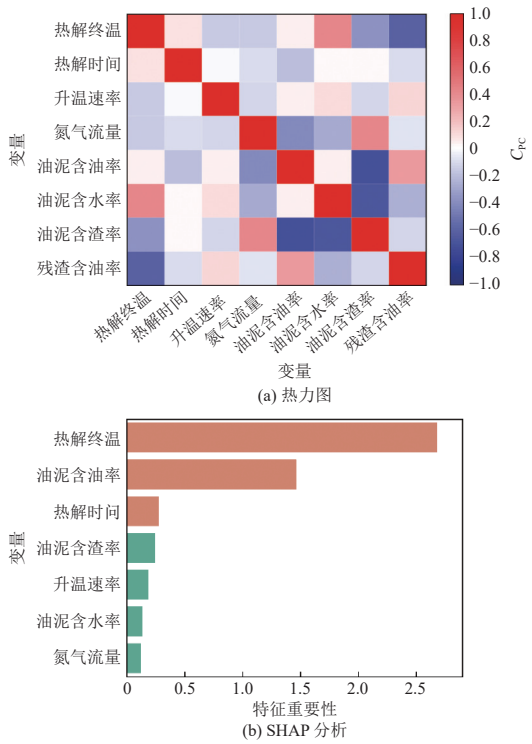

图 3　含油污泥热解残渣含油率的特征重要性分析

Fig. 3　Feature importance analysis of oil content in oily sludge pyrolysis residues

由图 3（a）可知，热解终温对含油污泥热解残渣含油率具有较为显著的负相关，这说明热解温度越高，热解残渣含油率越低，而含油污泥的含油率对含油污泥热解残渣含油率具有正相关影响。此外，使用 SHAP 分析进一步对输入变量与输出数据之间的关系进行可视化分析（图 3（b））。结果显示，影响含油污泥热解后残渣含油率的前 3 个主要因素依次是热解终温、油泥含油率和热解时间。这 2 种特征重要性分析的结果一致，与宋薇等[34]研究结论一致。其中，热解终温是首要影响因素，因为温度可促进含油污泥中的石油烃类发生挥发、裂解、缩合反应，生成热解油和热解气；并且提高热解终温，可以使含油污泥中较难裂解的重油成分转化为轻质油组分[23]。随着热解终温的进一步提高，热解油组分还会发生裂解和缩聚，转化成氢气和甲烷等小分子气体[35]，进一步降低含油率。

2.2　预测模型比较分析

将 228 组数据按 8∶2 的比例划分，获得训练集 182 组，测试集 46 组。为了消除输入变量之间的单位和标度对模型的影响，对样本进行归一化处理。本研究选择 Z-score 来标准化特征$X_{i'}$，即通过移除平均值并缩放到单位方差来进行归一化，其公式见式（6）。

$$X_{i'}=\frac{X_i-\mu}{\sigma_\omega} \tag{6}$$

式中：X_i、μ和σ_ω分别代表原始数据、数据的均值以及数据的标准差。完成数据的标准化处理后，建立 GBDT 算法、XGB 算法、SVM 算法及 RF 算法 4 个非线性模型。为了得到较好的预测效果，对模型的超参进行了设置（表 3）。

图 4 为 4 种模型对含油污泥热解后残渣含油率的预测结果，图中红线区域代表 95% 置信区间，偏离红线的预测点可视为波动性大的样本。

由图 4 可知，GBDT、XGB、SVM 以及 RF 4 种算法模型的训练集和预测集都存在向直线 y=x 靠拢的趋势，其预测结果的可靠性较高。其中，RF 模型的预测精度最高，R^2=0.917 1。这是因为 RF 模型通过集成多个决策树，有效减少单个模

表 3　各个模型的超参设置

Table 3　Hyperparameter settings for each model

模型	超参名称	参数值
GBDT	n_estimators	100
	learning_rate	0.1
	max_depth	3
	random_state	4 082
XGB	colsample_bytree	0.9
	learning_rate	0.1
	max_depth	3
	n_estimators	300
	subsample	0.5
SVM	kernel	rbf
	gamma	0.1
	C	100
	epsilon	0.1
RF	n_estimators	100
	max_depth	0
	min_samples_split	2
	min_samples_leaf	1

注：random_state为模型的随机种子；kernel为模型的核函数；rbf为高斯径向基函数。

型可能存在的过拟合问题，从而提高整体模型的泛化能力。同时，RF 模型通过在每棵树的训练过程中随机选择样本和特征，保证了树与树之间的多样性，这种多样性是提高模型性能的关键。同时可以看出，RF 模型的 RMSE 波动最小，而 XGB、SVM 模型中 RMSE 波动较大，且出现了偏离置信区间的异常样本。

GBDT、XGB、SVM 以及 RF 的模型预测结果见表 4。GBDT 模型在训练集上 R^2 的表现要优于 SVM 和 RF 模型，达到了 0.984 2，而测试集效果一般。此外，GBDT 模型的测试集 RMSE 和 MAPE 分别为 1.951 和 4.319%，优于 XGB 和 SVM 模型。GBDT 模型在训练集上的良好表现可能归功于其强大的预测性能，它可以通过逐步建立多个弱学习器，并结合弱学习器的预测结果来提升整体模型的性能。相比于单一的决策树，GDBT 的预测性能显著提高，特别是在处理复杂非线性关系时表现尤为出色。然而针对其测试集效果一般的问题，仍未找到其最适合的超参。GDBT 的性能对部分超参数非常敏感，不合适的参数选择就会导致模型过拟合或欠拟合。如果能够找到更合适的超参，GDBT 模型可能会得到更好的预测效果。

XGB 模型在测试集上 R^2 的表现是 4 个模型

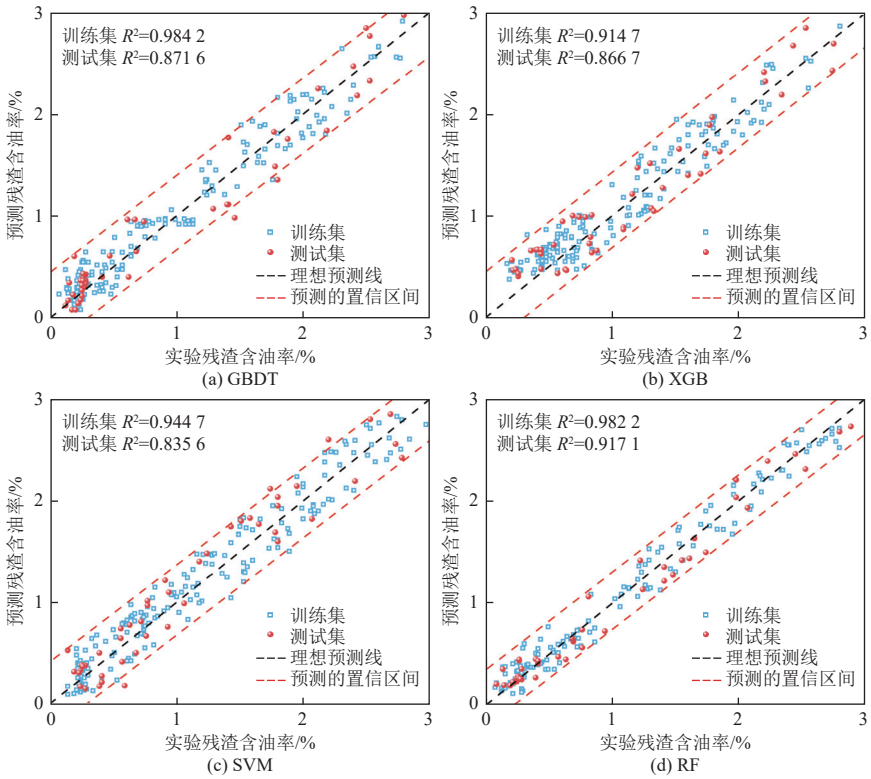

图 4　20% 测试集划分下 GBDT、XGB、SVM、RF 对含油污泥热解残渣含油率预测

Fig. 4　Prediction of oil content in oily sludge pyrolysis residues by GBDT, XGB, SVM, and RF on a 20% test set

表 4　GBDT、XGB、SVM、RF 的模型评价指标

Table 4　Evaluation metrics for GBDT, XGB, SVM, and RF models

模型	训练集	测试集		
	R^2	R^2	RMSE	MAPE/%
GBDT	0.984 2	0.871 6	1.951	4.319
XGB	0.914 7	0.866 7	2.849	5.046
SVM	0.944 7	0.835 6	2.385	4.922
RF	0.982 2	0.917 1	1.693	4.487

中最差的，达到了 0.914 7，测试集表现也一般，为 0.866 7。XGB 模型在测试集里的其他指标也较差，这说明模型的波动较大。XGB 模型的超参较多，调参过程较为复杂，找到最佳参数组合需要大量的时间并且 XGB 模型可能在小数据集上出现过拟合，需要谨慎调参和验证，所以如果能够找到合适的超参，XGB 会具有更加优秀的效果。

SVM 模型适用于高维数据和复杂非线性关系，但对数据的异常值和噪声较为敏感。从预测集结果来看（图 4(c)），大部分数据点都分布在红线区域范围内，还出现了偏离置信区间的异常样本，这可能是数据中的噪声对预测结果造成了影响，也有可能是 SVM 超参设置不合适。

RF 是通过集成多个决策树来提高预测性能，对于非线性关系和数据中的噪声有较好的鲁棒性。对比 RF 的训练集与测试集预测结果（图 4(d)），大部分蓝点（训练集）分布在拟合线上，少部分红点（测试集）偏离较远，这可能是因为 RF 对某些复杂非线性关系或特定数据分布不够敏感。从测试集的评价指标来看，RF 相较于其他 3 个模型效果最好，R^2 高达 0.917 1，说明模型的拟合度高。RF 模型的 RMSE 数值最小，说明 RF 模型抗干扰能力较强。这是因为 RF 模型由多个树的集成，即使某些数据中存在噪声，RF 也能通过多数表决的方式减少噪声对最终决策的影响。同时 RF 模型对部分模型参数的选择并不敏感，即使模型参数不完美，仍然能够提供较好的预测结果，故 RF 模型具有一定泛化能力。

4 种机器学习模型中，GBDT 模型和 RF 模型在训练集上的表现是最好的，R^2 超过了 0.98，基于训练集上的性能排序为 GBDT>RF>SVM>XGB。然而在测试集上 RF 模型的表现最好，R^2 超过了 0.91，RMSE 和 MAPE 也比其他 3 种模型低，基于测试集上的性能排序为 RF>GBDT>XGB>SVM。为了达到更好的预测效果，后续对模型进行超参优化，以达到更高的预测精度。

2.3　模型优化结果

4 种模型的超参数寻优结果见表 5，4 种优化后模型的预测结果如图 5 所示。RF 模型（图 5(d)）优化效果最好，大部分数据点都集中在拟合线上；其次是 GBDT 模型（图 5(a)），除了个别测试集数据偏离较大，大部分数据在拟合线附近浮动；再次是 XGB 模型（图 5(b)），其大部分数据也在拟合线附近浮动；优化效果最差的是 SVM 模型（图 5(c)），其训练集和测试集优化都有部分数据点偏离拟合线。为对比优化后具体情况，本研究计算了优化后模型的评价指标（表 6）。

表 5　4 种模型的超参数寻优结果

Table 5　Hyperparameter optimization results for the 4 models

模型	待优化超参	最优值
GBDT	learning_rate	0.1
	max_depth	7
	subsample	0.9
	min_samples_split	9
XGB	colsample_bytree	0.7
	learning_rate	0.2
	max_depth	9
	n_estimators	182
	subsample	0.8
SVM	gamma	0.97
	C	708
	epsilon	0.03
RF	max_depth	19
	min_samples_split	2
	min_samples_leaf	1

从表 6 可知，优化效果变化最大的是 XGB 模型，训练集与测试集 R^2 分别达到了 0.997 3 和 0.900 1，测试集 RMSE 下降 38.85%，预测精度得到一定的提高。GBDT 模型的训练集 R^2 变化不大，但是测试集 R^2 达到了 0.901 2。SVM 模型优化后 RMSE 下降了 20.62%，测试集 R^2 提高了 6.79%。相反，RF 经过优化后，R^2 前后差别并不明显，尤其是在训练集上几乎没有变化，优化后的测试集 R^2 有小幅度提升，由原来的 0.917 1 提升至 0.920 4。

4 种模型经过优化后含油污泥热解残渣含油率预测效果提升最明显的是 XGB 模型，RF 模型

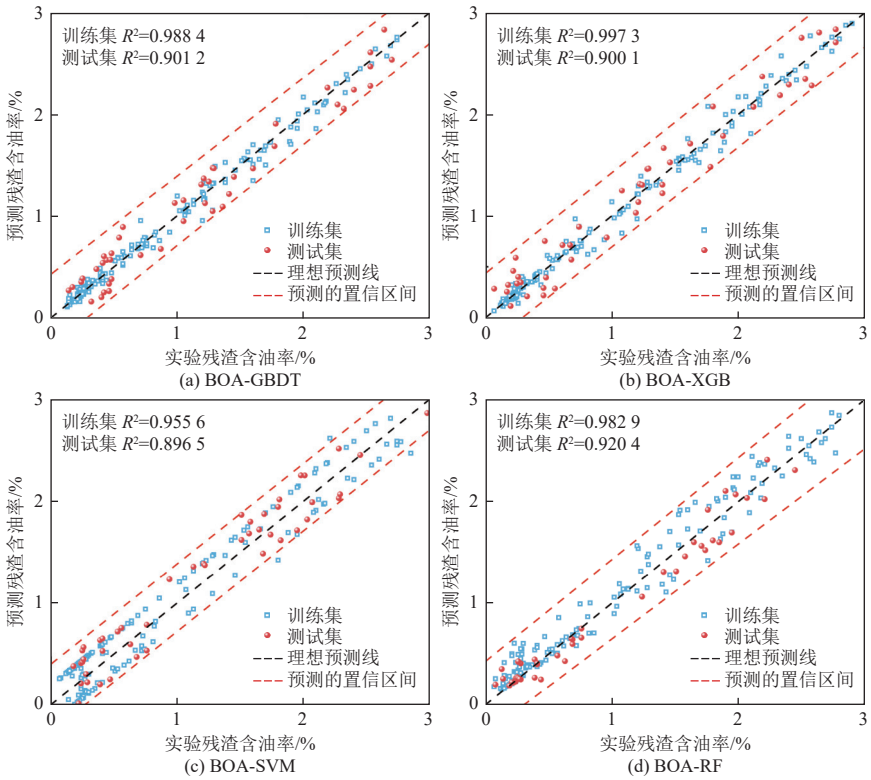

图 5　20% 测试集划分比例下 BOA-GBDT、BOA-XGB、BOA-SVM、BOA-RF 分别对含油污泥热解残渣含油率预测

Fig. 5　Prediction of oil content in oily sludge pyrolysis residues by BOA-GBDT, BOA-XGB, BOA-SVM, and BOA-RF on a 20% test set

表 6　优化前后的模型评价指标

Table 6　Model evaluation metrics before and after optimization

模型	训练集	测试集		
	R^2	R^2	RMSE	MAPE/%
GBDT	0.984 2	0.871 6	1.951	4.319
XGB	0.914 7	0.866 7	2.849	5.046
SVM	0.944 7	0.835 6	2.385	4.922
RF	0.982 2	0.917 1	1.693	4.487
BOA-GBDT	0.988 4	0.901 2	1.729	3.862
BOA-XGB	0.997 3	0.900 1	1.742	4.845
BOA-SVM	0.945 6	0.896 5	1.893	4.693
BOA-RF	0.982 9	0.920 4	1.659	4.141

提升最不明显。GBDT 模型尽管在训练集上的含油污泥热解残渣含油率预测提升不大，但在测试集上有改进，其 RMSE 下降了 38.85%，从而降低了模型的波动，SVM 模型也是如此。优化后的模型在训练集上的性能排序为 BOA-XGB>BOA-GBDT>BOA-RF>BOA-SVM；在测试集上，BOA-RF>BOA-GBDT>BOA-XGB>BOA-SVM。综上所述，优化后 RF 模型更适合含油污泥热解残渣含油率的预测。

3　结　　论

针对含油污泥热解残渣含油率预测，分别建立了 4 种模型。通过对比发现，在训练集上模型预测性能排序为 GBDT>RF>SVM>XGB，在测试集上为 RF>GBDT>XGB>SVM。模型训练集 R^2 均超过了 0.9，测试集 R^2 均超过了 0.8。

使用 BOA 算法对模型进行超参数优化，预测性能进步最大的是 XGB 模型，测试集上的 R^2 达到了 0.900 1，RMSE 下降 38.85%；GBDT 模型的进步也比较显著；SVM 模型的测试集 RMSE 也下降了 20.62%；性能变化较小的是 RF 模型，其测试集的 R^2 仅从 0.917 1 提升至 0.920 4，但其预测精度仍然最高。BOA-RF 模型更适合含油污泥热解残渣含油率预测。

参考文献 (References) :

[1]　DA SILVA L J，ALVES F C，DE FRANÇA F P. A review of the technological solutions for the treatment of oily sludges from petroleum refineries[J]. Waste Management & Research，2012，30(10)：1016–1030.

[2]　HU Guangji，LI Jianbing，ZENG Guangming. Recent

development in the treatment of oily sludge from petroleum industry: A review[J]. Journal of Hazardous Materials, 2013, 261: 470–490.

[3] 齐加胜，杜长星，赵建平. 油泥热解工艺参数优化分析及应用 [J]. 安徽师范大学学报(自然科学版)，2019，42(6)：544–549.
QI Jiasheng, DU Changxing, ZHAO Jianping. Optimization analysis and application on process parameter for oil sludge pyrolysis[J]. Journal of Anhui Normal University (Natural Science), 2019, 42(6): 544–549.

[4] WANG Xiang, WANG Qunhui, WANG Shijie, et al. Effect of biostimulation on community level physiological profiles of microorganisms in field-scale biopiles composed of aged oil sludge[J]. Bioresource Technology, 2012, 111: 308–315.

[5] 彭涛，刘雪东，郭文元，等. 基于残渣测定的含油污泥热解过程及工艺选择 [J]. 化学工程，2021，49(6)：4–8+25.
PENG Tao, LIU Xuedong, GUO Wenyuan, et al. Pyrolysis process and process selection of oily sludge based on residue determination[J]. Chemical Engineering (China), 2021, 49(6): 4–8+25.

[6] 曹蕊，韩冬云，朱涛，等. 含油污泥热解残渣特性及其资源化利用 [J]. 化工环保，2023，43(3)：353–358.
CAO Rui, HAN Dongyun, ZHU Tao, et al. Characteristics and resource utilization of pyrolysis residue from oily sludge[J]. Environmental Protection of Chemical Industry, 2023, 43(3): 353–358.

[7] 侯影飞，张建，祝威，等. 油田含油污泥热解制备烟气脱硫剂 [J]. 环境污染与防治，2010，32(1)：51–55.
HOU Yingfei, ZHANG Jian, ZHU Wei, et al. Preparation of flue gas desulphurizer from oil sludge by pyrolysis[J]. Environmental Pollution & Control, 2010, 32(1): 51–55.

[8] 张子杭，许丹，胡艳军，等. 机器学习在有机固废全链条处置中的应用进展 [J]. 能源环境保护，2023，37(1)：184–193.
ZHANG Zihang, XU Dan, HU Yanjun, et al. Application progress of machine learning in the whole chain disposal of organic solid waste[J]. Energy Environmental Protection, 2023, 37(1): 184–193.

[9] SU Sheng, WANG Juan. Machine learning prediction of contents of oxygenated components in bio-oil using extreme gradient boosting method under different pyrolysis conditions[J]. Bioresource Technology, 2023, 379: 129040.

[10] NGUYEN X C, NGUYEN T T H, LA D D, et al. Development of machine learning - based models to forecast solid waste generation in residential areas: A case study from Vietnam[J]. Resources, Conservation and Recycling, 2021, 167: 105381.

[11] PATHY A, MEHER S, BALASUBRAMANIAN P. Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods[J]. Algal Research, 2020, 50: 102006.

[12] LENG Erwei, HE Ben, CHEN Jingwei, et al. Prediction of three-phase product distribution and bio-oil heating value of biomass fast pyrolysis based on machine learning[J]. Energy, 2021, 236: 121401.

[13] ZHANG Junhui, LIU Jingyong, EVRENDILEK F, et al. TG-FTIR and Py-GC/MS analyses of pyrolysis behaviors and products of cattle manure in CO_2 and N_2 atmospheres: Kinetic, thermodynamic, and machine-learning models[J]. Energy Conversion and Management, 2019, 195: 346–359.

[14] LI Yize, GUPTA R, YOU Siming. Machine learning assisted prediction of biochar yield and composition via pyrolysis of biomass[J]. Bioresource Technology, 2022, 359: 127511.

[15] KARDANI N, ZHOU Annan, NAZEM M, et al. Modelling of municipal solid waste gasification using an optimised ensemble soft computing model[J]. Fuel, 2021, 289: 119903.

[16] LI Jie, LI Lanyu, TONG Y W, et al. Understanding and optimizing the gasification of biomass waste with machine learning[J]. Green Chemical Engineering, 2023, 4(1): 123–133.

[17] WANG Ziyi, GONG Zhiqiang, WANG Zhenbo, et al. Pyrolysis characteristics and products distribution of petroleum sludges[J]. Environmental Technology, 2022, 43(12): 1819–1832.

[18] LIN Bingcheng, HUANG Qunxing, CHI Yong. Co-pyrolysis of oily sludge and rice husk for improving pyrolysis oil quality[J]. Fuel Processing Technology, 2018, 177: 275–282.

[19] GONG Zhiqiang, WANG Zhentong, WANG Zhenbo, et al. Study on pyrolysis characteristics of tank oil sludge and pyrolysis char combustion[J]. Chemical Engineering Research and Design, 2018, 135: 30–36.

[20] WANG Zhiqi, GUO Qingjie, LIU Xinmin, et al. Low temperature pyrolysis characteristics of oil sludge under various heating conditions[J]. Energy & Fuels, 2007, 21(2): 957–962.

[21] 李彦，胡海杰，屈撑囤，等. 含油污泥催化热解影响因素研究及热解产物分析 [J]. 现代化工，2018，38(1)：67–71.
LI Yan, HU Haijie, QU Chengtun, et al. Influencing factors for catalytic pyrolysis of oily sludge and analysis of pyrolysis products[J]. Modern Chemical Industry, 2018, 38(1): 67–71.

[22] 丁安军，王雨辰，廖长君，等. 钻井含油污泥高温热解处理技术研究应用 [J]. 石油地质与工程，2018，32(5)：119–120.
DING Anjun, WANG Yuchen, LIAO Changjun, et al. Study and application of high temperature pyrolysis technology for drilling oily sludge[J]. Petroleum Geology and Engineering, 2018, 32(5): 119–120.

[23] 詹咏，张领军，谢加才，等. 热解终温对含油污泥三相产物特性的影响 [J]. 环境工程学报，2021，15(7)：2409–2416.
ZHAN Yong，ZHANG Lingjun，XIE Jiacai，et al. Effect of final pyrolysis temperature on characteristics of three-phase products of oily sludge[J]. Chinese Journal of Environmental Engineering，2021，15(7)：2409–2416.

[24] 孙丽，雍云乔，李来红. 含油污泥高温热解工艺参数优化及产物分析 [J]. 石油与天然气化工，2021，50(3)：122–126+133.
SUN Li，YONG Yunqiao，LI Laihong. Parameter optimization and product analysis of high temperature pyrolysis of oily sludge[J]. Chemical Engineering of Oil & Gas，2021，50(3)：122–126+133.

[25] 齐加胜，赵建平，黄欣欣. 长庆油田落地油泥处理技术研究 [J]. 当代化工，2018，47(11)：2326–2329+2333.
QI Jiasheng，ZHAO Jianping，HUANG Xinxin. Research on treatment technology of ground oil sludge in Changqing oilfield[J]. Contemporary Chemical Industry，2018，47(11)：2326–2329+2333.

[26] 祝威. 油田含油污泥热解产物分析及性能评价 [J]. 环境化学，2010，29(1)：127–131.
ZHU Wei. Analysis and performance mensuration of pyrolysis products for oil sludge[J]. Environmental Chemistry，2010，29(1)：127–131.

[27] 廉腾飞，唐龙飞，刘霞，等. 铁、钙氧化物对含油污泥热解特性的影响 [J]. 华东理工大学学报(自然科学版)，2024，50(3)：335–345.
LIAN Tengfei，TANG Longfei，LIU Xia，et al. Effect of iron and calcium oxides on pyrolysis characteristics of oilfield sludge[J]. Journal of East China University of Science and Technology，2024，50(3)：335–345.

[28] 胡志勇. 塔河油田含油污泥低温热解研究 [J]. 油气田环境保护，2015，25(3)：9–11+72.
HU Zhiyong. Experimental study on low temperature thermal pyrolysis of oily sludge in Tahe oilfield[J]. Environmental Protection of Oil & Gas Fields，2015，25(3)：9–11+72.

[29] 刘颖，杜卫东，程泽生，等. 含油污泥热解的影响因素初探 [J]. 油气田环境保护，2010，20(2)：7–9+28+60.
LIU Ying，DU Weidong，CHENG Zesheng，et al. A preliminary research on oily sludge pyrolysis influencing factors[J]. Environmental Protection of Oil & Gas Fields，2010，20(2)：7–9+28+60.

[30] 焦文超，郭晓丹. 异位热解技术处理塔河油田含油污泥 [J]. 化工环保，2021，41(3)：318–323.
JIAO Wenchao，GUO Xiaodan. Treatment of oily sludge in Tahe oilfield by heterotopic pyrolysis[J]. Environmental Protection of Chemical Industry，2021，41(3)：318–323.

[31] 李瑞娜，朱红霞，宋海燕，等. 含油污泥的热解及尾渣制陶粒工艺研究 [J]. 工业安全与环保，2023，49(4)：98–103.
LI Ruina，ZHU Hongxia，SONG Haiyan，et al. Study on technology of oily sludge pyrolysis and ceramsite preparation from pyrolysis residue[J]. Industrial Safety and Environmental Protection，2023，49(4)：98–103.

[32] ZHAO Jianhui，ZHANG Chenyang，MIN Lin，et al. Retrieval of farmland surface soil moisture based on feature optimization and machine learning[J]. Remote Sensing，2022，14(20)：5102.

[33] 崔佳旭，杨博. 贝叶斯优化方法和应用综述 [J]. 软件学报，2018，29(10)：3068–3090.
CUI Jiaxu，YANG Bo. Survey on bayesian optimization methodology and applications[J]. Journal of Software，2018，29(10)：3068–3090.

[34] 宋薇，刘建国，聂永丰. 含油污泥的热解特性研究 [J]. 燃料化学学报，2008，36(3)：286–290.
SONG Wei，LIU Jianguo，NIE Yongfeng. Pyrolysis properties of oil sludge[J]. Journal of Fuel Chemistry and Technology，2008，36(3)：286–290.

[35] YAZDANI E，HASHEMABADI S H，TAGHIZADEH A. Study of waste tire pyrolysis in a rotary kiln reactor in a wide range of pyrolysis temperature[J]. Waste Management，2019，85：195–201.