DOI:10.11918/202510089

基于机器学习的 SANI 工艺除碳脱氮性能预测与优化

吴知静[1],江锦琦[1,2],毕鑫祺[3],王宗平[1],郭 刚[1]

(1.长江流域多介质污染协同控制湖北省重点实验室(华中科技大学),武汉 430074;
2.华中科技大学 人工智能研究院,武汉 430074;3.北京市城市规划设计研究院,北京 100045)

摘 要:硫酸盐还原-硫自养反硝化-硝化(SANI)工艺是一种具有污泥产量低、耐高温、耐盐等优势的新型污水生物脱氮处理技术。然而,该工艺内部的复杂微生物群落对运行参数极为敏感,导致系统性能易波动,亟需可靠的建模工具优化。传统机理模型构建繁琐且通用性有限,而机器学习方法又常受到小样本、数据不均衡及可解释性不足等限制。为此,提出一种融合生成对抗网络(GAN)与分阶段机器学习建模(包括极端梯度提升、人工神经网络和支持向量机算法)的数据驱动新方法,依据工艺生化机理构建厌氧硫酸盐还原和缺氧硫自养反硝化两阶段预测模型,并综合驯化与运行数据以拓展模型适用性。结果表明,该模型对 COD 与氮去除率的预测均方根误差(RMSE)分别低至6.8%与5.6%,展现出优异预测精度。可解释性分析进一步明确了工艺的关键环境因子及最优调控区间:进水 COD 340~500 mg/L、进水硫酸盐(以 S 计)300~450 mg/L、厌氧出水硫化物(以 S 计)50~100 mg/L、厌氧工作容积5~6.5 L 和水力停留时间6~8 h。本研究为 SANI 工艺的精准调控提供了可靠工具,更构建了一套可迁移、可扩展的"数据增强—分阶段建模—机理解释"融合范式,为复杂污水生物处理过程的智慧优化提供了全新的方法论基础。

关键词:硫酸盐还原;硫自养反硝化;机器学习;生成对抗网络;可解释性
中图分类号:X703 **文献标志码**:A **文章编号**:0367-6234(2026)06-0120-10

Prediction and optimization of carbon and nitrogen removal performance in SANI process based on machine learning

WU Zhijing[1], JIANG Jinqi[1,2], BI Xinqi[3], WANG Zongping[1], GUO Gang[1]

(1. Hubei Key Laboratory of Multi-Media Pollution Control in Yangtze Basin (Huazhong University of Science and Technology), Wuhan 430074, China; 2. Institute of Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, China; 3. Beijing Municipal Institute of City Planning & Design, Beijing 100045, China)

Abstract: The sulfate reduction-sulfur autotrophic denitrification-nitrification (SANI) process is a novel biological wastewater treatment technology for nitrogen removal, which has advantages such as low sludge yield, high temperature tolerance, and salt tolerance. However, the complex microbial communities within this process are extremely sensitive to operational parameters, leading to easy fluctuation of system performance, which urgently requires reliable modeling tools for optimization. Traditional mechanistic models are cumbersome to construct and have limited universality, while machine learning methods are often limited by small sample size, data imbalance, and insufficient interpretability. To this end, this paper proposed a new data-driven method integrating generative adversarial network (GAN) and staged machine learning modeling (including XGBoost, Artificial Neural Network, and Support Vector Machine algorithms), constructed two-stage prediction models for anaerobic sulfate reduction and anoxic sulfur autotrophic denitrification according to the biochemical mechanism of the process, and integrated acclimation and operation data to expand the applicability of the models. The results show that the root mean square error (RMSE) of the model for the prediction of COD and nitrogen removal rates are as low as 6.8% and 5.6%, respectively, exhibiting excellent prediction accuracy. The interpretability analysis further clarifies the key environmental factors and optimal control ranges of the process: influent COD of 340-500 mg/L, influent sulfate (as S) of 300-450 mg/L, anaerobic effluent sulfide (as S) of 50-100 mg/L, anaerobic working volume of 5-6.5 L, and hydraulic retention time of 6-8 h. This paper provides a reliable tool for the precise regulation of the SANI process and further constructs a transferable and scalable integration paradigm of "data augmentation, staged modeling, and mechanism interpretation", which provides a brand-new methodological foundation for the intelligent

收稿日期:2025-10-27;**录用日期**:2025-11-27;**网络首发日期**:2026-01-09
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20260109.1319.002
基金项目:国家重点研发计划(2023YFC3207203)
作者简介:吴知静(2002—),女,硕士;王宗平(1969—),男,教授,博士生导师;郭 刚(1987—),男,副研究员,博士生导师
通信作者:江锦琦,jiang0912@hust.edu.cn

optimization of complex biological wastewater treatment processes.

Keywords: sulfate reduction; sulfur autotrophic denitrification; machine learning; generative adversarial network; interpretability

硫酸盐还原-硫自养反硝化-硝化(SANI)组合工艺是一种硫源(S)介导的除碳(C)脱氮(N)新工艺,适用于沿海城市的市政污水及含硫、含盐工业废水处理,具有污泥产量低和耐高温、耐盐等优势[1]。然而,SANI 工艺在实际应用中仍面临严峻挑战,工艺涉及多种微生物群落(如硫酸盐还原菌和硫氧化菌等)的协同作用,这些菌群对高维水质因子和操作参数较敏感,导致工艺性能不稳定,需要借助数学模型的手段对工艺进行优化调控[2-3]。活性污泥模型(ASM)和稳态模型等曾被应用于 SANI 工艺的建模、预测和优化中,但是传统机理模型需要精确的化学计量学及动力学参数,构建复杂且普适性差,任何环境因子的改变都需要进行模型的重新校准[4]。机器学习技术为破解这一难题提供了崭新的数据驱动范式。Xiong 等[5]曾将基于树的模型应用于硫酸盐还原菌去除重金属的预测,并识别了影响系统的关键环境因子。通过从海量历史运行数据中学习复杂的非线性关系,机器学习模型有望准确预测 SANI 工艺在不同进水条件和操作参数下的出水水质,从而为优化调控提供决策支持。然而,污水生物处理工艺采集数据周期长,高质量、大规模的标注数据集获取成本高昂,存在小数据集及数据不平衡的问题,容易导致模型训练效果不佳[6]。同时,SANI 工艺是一个包含物理、化学及生物反应的复杂动态系统,单一的"黑箱"模型难以全面捕捉其内在的阶段性反应机理,预测结果的可靠性和可解释性有待提升[7]。

因此,提出了一种创新的机器学习建模方案:利用生成对抗网络(GAN)合成数据,以有效弥补小样本与数据不平衡的缺陷[8-9];并采用分阶段建模策略,依据工艺内在生化过程(硫酸盐还原与自养反硝化)构建专用子模型,从而同步提升预测精度与可解释性。特别地,模型同时引入了驯化期与运行期数据,使其能广泛应用于工艺启动与稳定运行等多种场景。本研究不仅为 SANI 工艺的精准调控提供了可靠工具,也为数据驱动下的其他污水生物处理工艺建模提供了可借鉴的新范式。

1　实验方法

1.1　数据收集及预处理

数据来源于实验室运行约 180 d 的 SANI 生物反应器,包括驯化阶段(89 d)和稳定运行阶段(97 d)。其中,收集到的环境因子包括水质参数厌氧段进出水 COD(Inf/Eff COD)、进水硫酸盐(Inf SO_4)、出水硫化物和缺氧段进出水硫酸根、进出水硝酸盐(Inf/Eff NO_3)、进水硫化物(Inf S)、出水亚硝酸盐等和工艺参数厌氧内回流比($Reflux_{ana}$)、工作容积(Volume)、水力停留时间(HRT)等。整体数据集存在约 1% 的缺失值(主要为部分水质参数),采用均值填补法处理缺失值,并且采用 StandardScaler 对数据进行归一化[6]。

1.2　模型搭建

模型搭建流程如图 1 所示,选择极端梯度提升(XGBoost)、人工神经网络(ANN)和支持向量回归(SVR)3 种基于树的机器学习算法进行模型搭建。这些算法的选择基于 SANI 工艺数据的特征及其在环境领域表现出的较好性能[6,10],XGBoost 善于处理参数间的非线性交互作用并控制过拟合;ANN 擅长学习高度复杂的非线性模式,适合模拟微生物反应过程;SVR 则对小样本数据有较好的适应性。

将数据集的 85% 划分成训练集,余下的 15% 划分成测试集用于模型评估,采用贝叶斯优化算法和五折交叉验证进行超参数调优[11]。具体的超参数搜索空间设置如下:XGBoost 模型主要调整 n_estimators(10 ~ 500)、max _ depth(3 ~ 15)、subsample(0.5 ~ 1.0)、learning_rate(0.1 ~ 1.0)及 gamma(0 ~ 0.5);ANN 模型涉及隐藏层层数(1 ~ 4)、各层神经元数(8 ~ 256,步长 8)、激活函数、L2 正则化项 alpha(1E-5 ~ 1E-2,对数尺度)与初始学习率(1E-4 ~ 1E-2,对数尺度);SVR 模型则重点优化惩罚参数 C(0.1 ~ 20.0)、核函数(在'linear'/'rbf'/'poly'中选择)、epsilon(0.001 ~ 1.0)以及核系数 gamma(0.001 ~ 1.0)。

1.3　数据增强

针对污水处理生物连续流反应器常见的小数据集及数据不平衡问题,模型搭建采用了以下策略:引入驯化数据;分阶段搭建;数据增强。其中,分阶段搭建指的是分别针对厌氧硫酸盐还原反应器和缺氧硫自养反硝化反应器分别搭建模型,可以增强模型性能及可解释性[7]。在厌氧阶段模型中,输入变量包括进水硫酸盐质量浓度、进水 COD、碳硫质量比(C/S)、厌氧区有效容积、内回流比及水力停留时间(HRT);输出变量为厌氧 COD 去除率与硫化物质量浓度。选择厌氧 COD 去除率作为主要输出指标的

原因在于若 SANI 工艺在厌氧阶段未能有效去除 COD,将直接影响后续缺氧阶段的硫自养反硝化过程。缺氧阶段模型的输入变量包括进水硫化物、进水硝酸盐、进水硫氮质量比(S/N)及水力停留时间;输出变量则为出水硫酸盐、亚硝酸盐及脱氮率。

为缓解小数据集的问题,采用了数据增强策略,选用条件生成对抗网络(CTGAN)进行数据合成[12]。CTGAN 是一种针对表格数据特征设计的生成对抗网络,用于在保持原始数据统计特征的前提下生成高质量的合成样本,通过对生成器与判别器的对抗训练,在保持原始数据集中复杂统计特征与分布规律的前提下,生成高质量、高保真的合成样本。根据数据分布特征(见 2.1 节),厌氧段多数样本的 COD 去除率处于稳定运行阶段(去除率≥90%,161 组),而去除率低于 90% 的样本相对较少(25 组),导致数据分布不平衡。为了提升厌氧模型训练的泛化能力与稳定性,利用 CTGAN 对去除率 <90% 的数据进行补充生成,以平衡整体数据集。训练过程中,模型参数设置为训练轮数 300、批大小 64,PAC 参数为 1。模型训练完成后,基于学习到的潜在分布生成 150 个合成样本,合成样本仅加入至训练集。

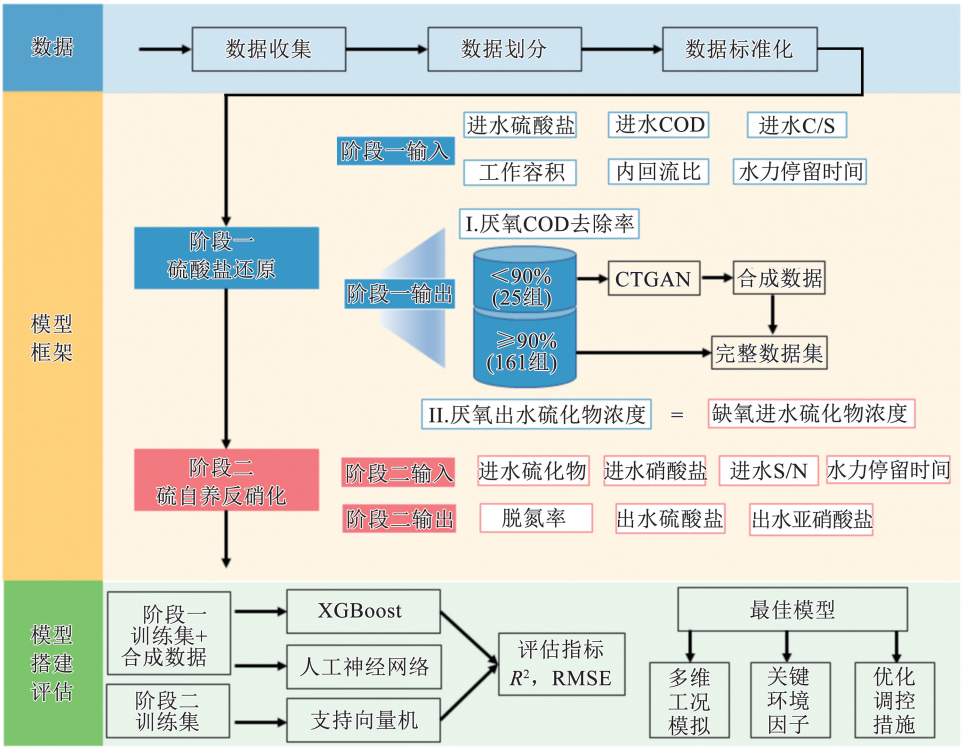


图 1 模型搭建流程

Fig. 1 Model construction process

1.4 模型性能评估

选择决定系数 R^2 和均方根误差(RMSE,E_{RMS})对两阶段机器学习回归模型进行评估[6,13],即

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{\mathrm{p}}^i - y_{\mathrm{t}}^i)^2}{\sum_{i=1}^{N}(y_{\mathrm{t}}^i - y_{\mathrm{m}})^2} \tag{1}$$

$$E_{\mathrm{RMS}} = \sqrt{\frac{\sum_{i=1}^{N}(y_{\mathrm{p}}^i - y_{\mathrm{t}}^i)^2}{N}} \tag{2}$$

式中:y_{p}^i 为预测值,y_{t}^i 为真实值,y_{m} 为样本平均值,N 为样本数。

1.5 模型可解释性

采用 SHAP 可解释性方法对模型进行特征重要性分析,以量化各环境因子在不同反应阶段中的贡献度[11,14]。SHAP 是一种基于合作博弈论的可解释性方法,其核心思想是通过计算每个特征在所有特征组合中对模型输出的贡献度平均值(即 Shapley 值),来量化该特征的全局重要性。对于一个样本预测值,SHAP 通过以下加性模型进行解释:

$$\phi_i(f,x) = \sum_{S\subset M\setminus i}\frac{|S|!(M-|S|-1)!}{M!}\cdot [f_x(S\cup\{i\}) - f_x(S)] \tag{3}$$

式中:ϕ_i 为特征 i 的 Shapley 值,即该特征的贡献度;M 为全部特征的集合;S 为不包含特征 i 的特征子集;$f_x(S)$ 为使用子集 S 中的特征得到的预测值;$[f_x(S\cup\{i\}) - f_x(S)]$ 为表示特征 i 加入子集 S 前后所产生的预测值变化。

基于两阶段最佳机器学习模型系统模拟 SANI 工艺的复杂运行工况，揭示工艺的应用场景。通过平均效应分析进一步确定关键环境因子对除碳脱氮性能的适宜阈值范围，为工艺参数的优化与协同调控机理建立提供了数据支撑与理论依据。该方法通过将某个特征的值在其数据范围内进行均匀分箱，然后，对于每个箱位值将数据集中所有样本的该特征值替换为此箱位值，并用训练好的模型进行预测，最后，计算所有样本在该箱位值下预测结果的平均值，即

$$\mathrm{ME}(x_j) = \frac{1}{N}\sum_{i=1}^{N} f(x_j, x_{i,-j}) \tag{4}$$

式中：$\mathrm{ME}(x_j)$为特征 j 在取某个特定箱位值时的平均效应，f 为训练好的机器学习模型，x_j为特征 j 的特定箱位值，$x_{i,-j}$为代表数据集中第 i 个样本除特征 j 外的所有其他特征的值，N 为总样本数。

2　结果与讨论

2.1　厌氧硫酸盐还原模型搭建

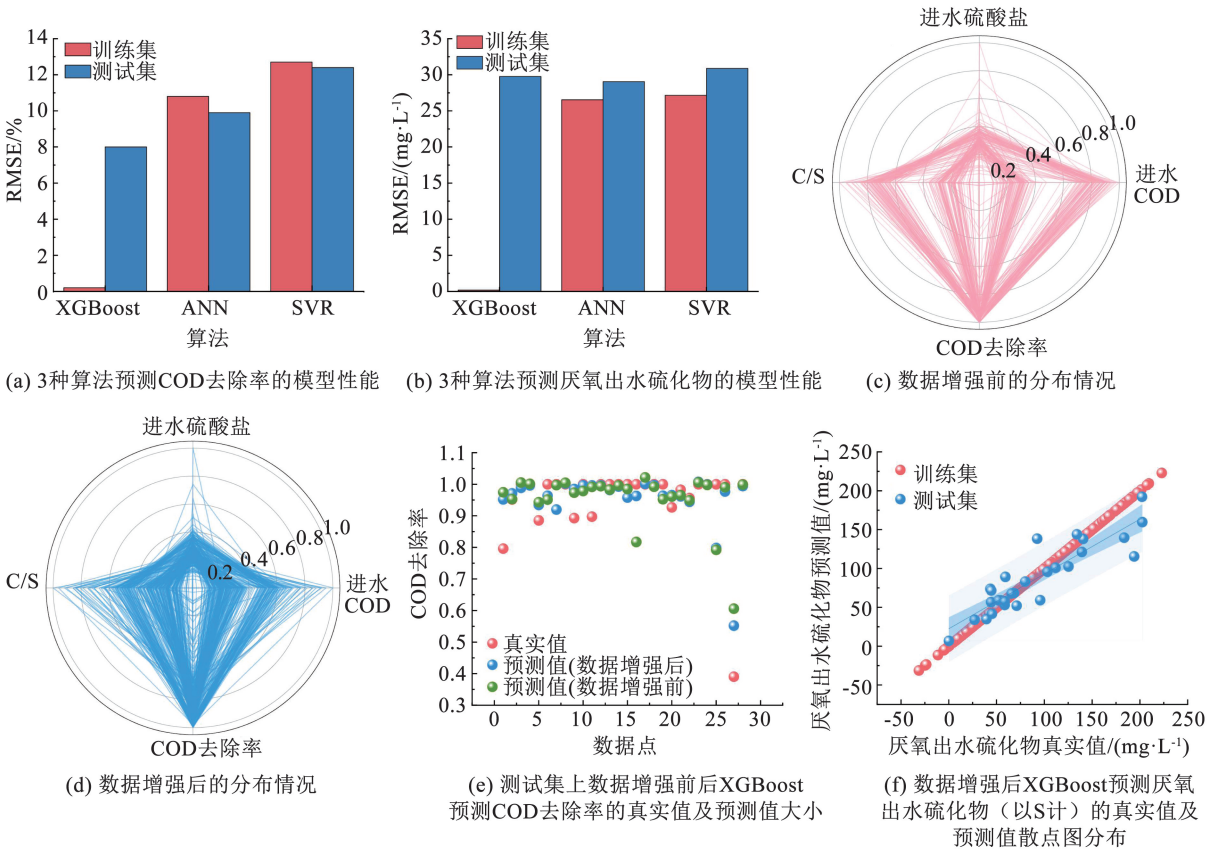
采用 XGBoost、ANN、SVR 3 种算法对厌氧硫酸盐还原反应器进行建模与训练，超参数均设为默认值，相关结果的 RMSE 如图 2(a)、(b)所示。针对出水硫化物(以 S 计)，XGBoost、ANN、SVR 在测试集上的 RMSE 分别为 29.76、29.05 和 30.89 mg/L，XGBoost 和 ANN 在出水硫化物预测中性能相近。针对厌氧 COD 去除率，3 种算法在测试集上的整体 RMSE 分别为 8%、9.9% 和 12.4%。按数据区间进一步评估发现，当 COD 去除率 <90% 时，RMSE 分别增大至 13.9%、20.3% 和 29.0%，说明在工艺不稳定阶段，XGBoost 具有更强的预测稳健性；而当 COD 去除率 ≥90% 时，RMSE 分别为 6.1%、5.4% 和 1.6%，三者差异相对较小。尽管 XGBoost 在高去除率区间的 RMSE 略高，但其在整体预测，尤其是在 COD 去除率 <90% 的不稳定运行阶段中表现最优，体现了其在多工况条件下的稳定性。综合来看，认为 XGBoost 在多任务预测中的总体性能最优($R^2_{\mathrm{COD}}=0.54$，$R^2_{\text{硫化物}}=0.70$)。

数据分布(图 2(c))显示机器学习预测任务中，COD 去除率存在明显的数据不平衡现象，由于实验室数据主要来源于稳定运行期间，约 85% 数据的 COD 去除率高于 90%，即使纳入了驯化阶段数据仍然会导致模型泛化能力受限[15]。因此，采用 CTGAN 对低 COD 去除率样本进行数据增强，关键变量增强后数据分布总体形态未发生明显变化(表 1、图 2(d))，但补充了低 COD 去除率区间的样本特征，从而提升了模型训练的有效性[16]。增强后的 $\mathrm{RMSE_{COD}}$(7.8%)和 $\mathrm{RMSE_{硫化物}}$(27.15 mg/L)均有所下降。进一步地，采用贝叶斯算法和交叉验证对增强后的 XGBoost 模型进行超参数调优(图 2(e)、(f))：最终模型在 COD 去除率预测任务，训练集的 R^2 与 RMSE 分别为 0.99 和 0.1%，测试集为 0.68 和 6.8%($\mathrm{RMSE_{去除率<90\%}}=5\%$，$\mathrm{RMSE_{去除率\geq 90\%}}=11.9\%$)；在出水硫化物预测任务的训练集上表现为 0.99 的 R^2 和 0.001 mg/L 的 RMSE，在测试集为 0.79 和 25.28 mg/L。

表 1　数据增强前后的分布变化

Tab. 1　Data distribution changes before and after data augmentation

数据	参数	单位	均值	标准差	最小值	25%	50%	75%	最大值
增强前 158 组	Inf $\mathrm{COD_{ana}}$	mg/L	647.44	183.46	329.70	485.21	549.66	835.61	955.25
	Inf $\mathrm{SO_{4ana}}$	mg/L	303.16	23.03	234.28	293.84	302.01	313.03	456.53
	Reflux	—	11.32	3.51	8.00	8.00	8.00	15.00	15.00
	Volume	L	6.55	3.01	3.40	3.40	9.40	9.40	9.40
	$\mathrm{HRT_{ana}}$	h	8.37	2.50	6.00	6.00	6.00	11.00	11.00
	C/S	—	2.14	0.60	1.14	1.62	1.86	2.73	3.33
	$\mathrm{COD_{rem}}$	—	0.95	0.12	0.27	0.95	1.00	1.00	1.00
增强后 308 组	Inf $\mathrm{COD_{ana}}$	mg/L	657.72	185.49	329.70	516.45	600.62	814.56	1 266.16
	Inf $\mathrm{SO_{4ana}}$	mg/L	294.90	20.56	234.28	282.31	294.23	305.93	456.53
	Reflux	—	13.27	3.25	8.00	8.00	15.00	15.00	17.00
	Volume	L	4.87	2.93	1.43	2.82	3.40	9.40	10.58
	$\mathrm{HRT_{ana}}$	h	9.61	2.30	6.00	6.00	11.00	11.00	13.00
	C/S	—	2.24	0.64	1.14	1.75	2.11	2.68	4.48
	$\mathrm{COD_{rem}}$	—	0.87	0.16	0.20	0.81	0.91	1.00	1.00



(a) 3种算法预测COD去除率的模型性能 (b) 3种算法预测厌氧出水硫化物的模型性能 (c) 数据增强前的分布情况

(d) 数据增强后的分布情况 (e) 测试集上数据增强前后XGBoost预测COD去除率的真实值及预测值大小 (f) 数据增强后XGBoost预测厌氧出水硫化物（以S计）的真实值及预测值散点图分布

图 2 厌氧硫酸盐还原模型搭建及数据增强前后性能对比

Fig. 2 Construction of anaerobic sulfate reduction model and performance comparison before and after data augmentation

综上，优化后的 XGBoost 模型结合了驯化和稳定运行期数据，并通过数据增强有效补偿了运行效果较差阶段的样本不足。因此，该模型不仅适用于稳定运行阶段 SANI 反应器厌氧段的性能预测，也可用于硫酸盐还原单体反应器驯化与稳定运行过程的分析与预测。

2.2 缺氧硫自养反硝化模型搭建

图 3(a) ~ (c)为 XGBoost、ANN、SVR 经过超参数优化后在脱氮率预测中训练集和测试集上的散点图分布。XGBoost 在训练集/测试集上的 R^2 分别为 0.92/0.87，RMSE 分别为 5.6%/6.3%；ANN 在训练集/测试集上的 R^2 分别为 0.88/0.83，RMSE 分别为 6.6%/7.1%；SVR 在训练集/上测试集上的 R^2 分别为 0.70/0.75，RMSE 分别为 10.50%/8.7%。因此，针对脱氮率预测任务，XGBoost 在测试集上具有最大的 R^2 和最小的 RMSE。针对出水亚硝酸盐(图 3(d)、(e))/硫酸盐(图 3(f)、(g))预测，3 种模型经过超参数优化后训练集和测试集上的 R^2 和 RMSE 显示，SVR 在出水亚硝酸盐预测中表现最佳，XGBoost 在出水硫酸盐预测中表现最佳。但由于该模型为多输出多任务预测模型，通过同一模型同时预测 3 种目标的任务，XGBoost 在缺氧硫自养反应器中表现最佳（$R^2_{平均}$ =0.78 >0.75(SVR) >0.70(ANN)）[17]。

2.3 基于最佳 XGBoost 模型的 SANI 工况模拟

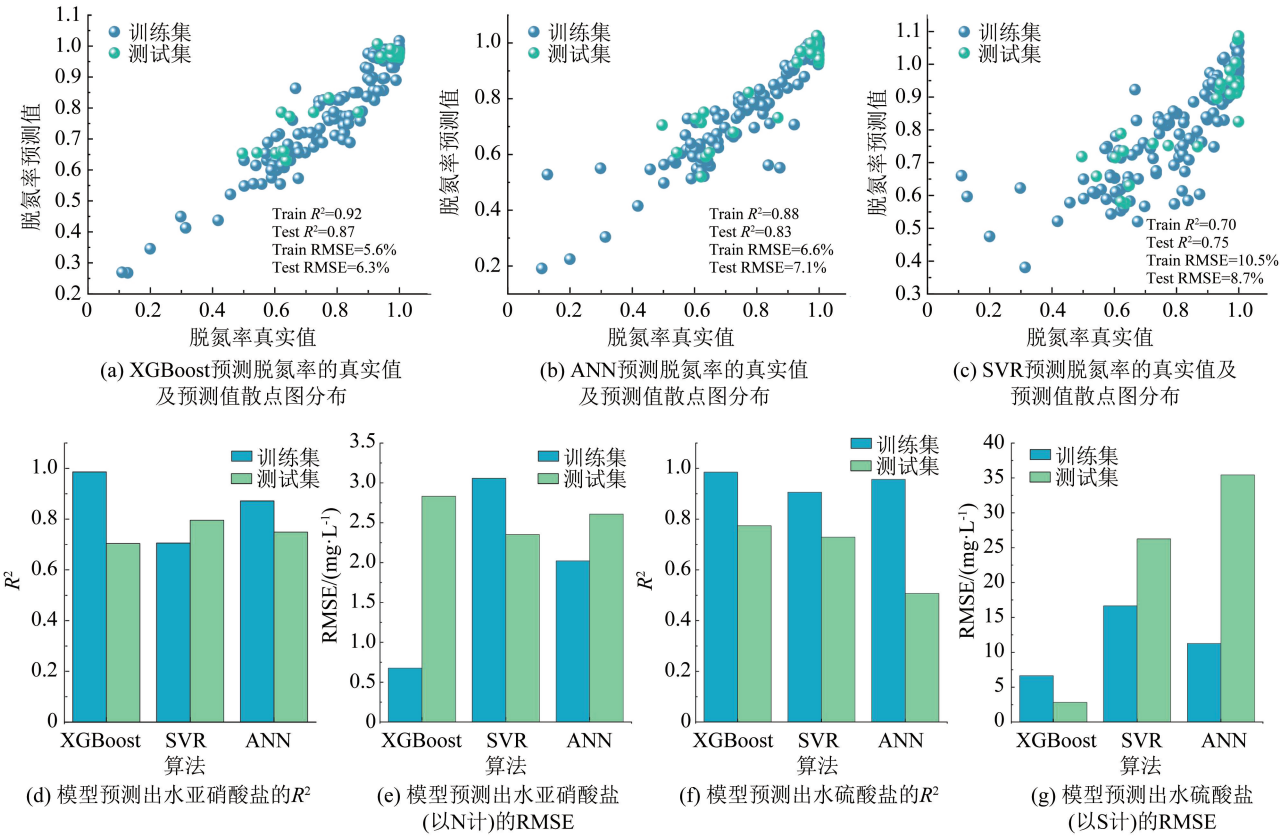
前述结果显示，XGBoost 在硫酸盐还原和硫自养反硝化中均表现最佳，为进一步验证分阶段模型相比整体模型(直接使用所有输入参数预测除碳脱氮率)的优势，也采用 XGBoost 对相同连续流训练数据建模。然而，其预测结果不理想(COD 与氮去除率的 R^2 分别仅为 0.27 和 0.09)，性能显著低于分阶段模型，原因可能是连续流训练数据规模较小。相反，分阶段模型不仅预测性能更优，其适用性也更广，可扩展至微生物驯化阶段。

SANI 工艺两阶段 XGBoost 模型联合使用，可对应 SANI 工艺的厌氧–缺氧运行阶段。基于阶段一模型预测的硫化物生成可以用于阶段二模型的输入，从而模拟 SANI 工艺高维环境因子下的复杂工况。借助原始数据集各环境因子的最大值与最小值，随机生成 1 000 组不同进水水质指标和操作参数构成的数据组合，通过两阶段模型预测 COD 去除率、N 去除率、厌氧硫化物生成量、出水硫酸盐/亚硝酸盐质量浓度，进行了 SANI 工艺的多工况模拟分析。

图 4(a)展示了工况模拟数据集中各环境因子的皮尔逊相关性分析，运行特性、进水特性与系统去除效果之间存在明显的耦合关系。其中，进水硫酸盐质量浓度与 COD 去除率呈正相关，表明适量的硫

酸盐能为 C-S 协同过程提供电子受体，从而促进有机物降解[18]。S/N 与脱氮率之间同样表现出正相关，说明较高的硫供体比例有助于强化硫自养反硝化过程。HRT 与 COD_{nem}、N_{nem} 则呈现负相关，说明过度延长反应时间不利于碳、氮的协同去除[7]。此外，C/S 与部分进水参数之间的负相关反映出碳源与硫源在系统中存在一定的竞争关系。

为进一步分析 SANI 工艺在复杂工况中更适合处理哪种类型的废水，引入了结合 C/N、C/S、S/N 与除碳脱氮之间的聚类分析，结果如图 4(b)所示。聚类结果显示，SANI 系统可分为 3 类典型运行工况：Cluster 0 为低碳源、高硫源（$C/S_{平均}$ = 1.39，$C/N_{平均}$ = 13.44，$S/N_{平均}$ = 2.42）的硫酸盐还原-自养反硝化协同型工况，表现出最高的 COD 与氮去除率（94%、90%）；Cluster 1 为碳主导型工况（$C/S_{平均}$ = 2.45，$C/N_{平均}$ = 20.99，$S/N_{平均}$ = 2.05），硫循环作用减弱，脱氮效率相对较低；Cluster 2 具有最高的 S/N（4.69）和 C/N（30.36），体现出强化的 C-N-S 耦合特征，系统脱碳脱氮性能均较优。总体而言，低 C/S-高 S/N 条件有利于促进硫酸盐还原与自养反硝化过程的协同，实现 SANI 工艺的高效稳定运行，也说明 SANI 工艺更适合运用于低碳源含硫废水处理中[19-20]。



(a) XGBoost预测脱氮率的真实值及预测值散点图分布　(b) ANN预测脱氮率的真实值及预测值散点图分布　(c) SVR预测脱氮率的真实值及预测值散点图分布

(d) 模型预测出水亚硝酸盐的R^2　(e) 模型预测出水亚硝酸盐(以N计)的RMSE　(f) 模型预测出水硫酸盐的R^2　(g) 模型预测出水硫酸盐(以S计)的RMSE

图 3　缺氧硫自养反硝化模型搭建性能

Fig. 3　Construction performance of anoxic sulfur autotrophic denitrification model

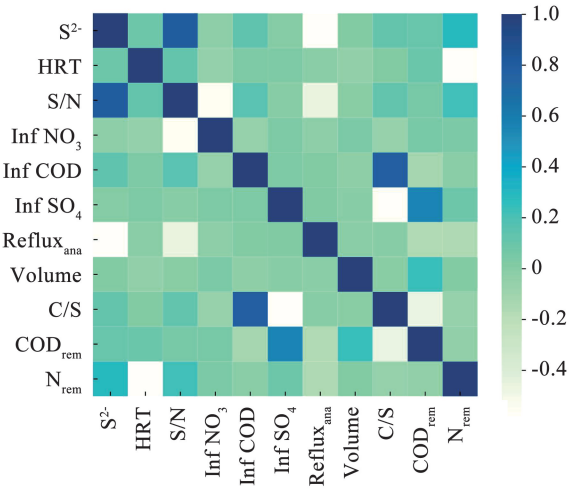
2.4　识别影响工艺 C-N-S 循环的关键环境因子

训练好的机器学习模型虽然在 SANI 工艺复杂工况下的除碳脱氮性能预测任务中表现优异，但是缺乏直接优化和调控 SANI 工艺的能力，需要借助可解释性分析技术识别影响工艺 C-N-S 循环的关键环境因子[21]。

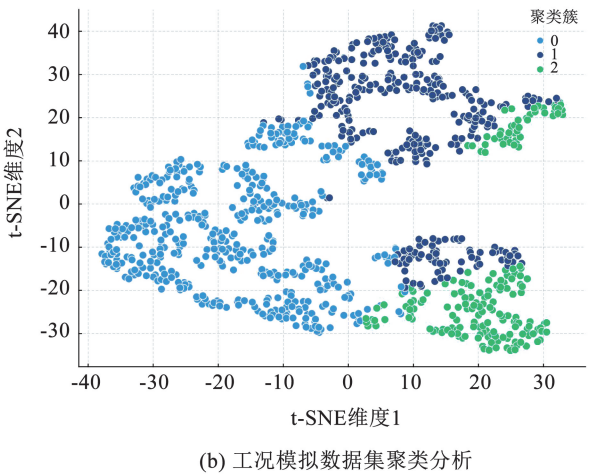
图 5(a)采用 SHAP 分析解释影响厌氧 COD 去除率（红色）和缺氧脱氮率（蓝色）的关键环境因子贡献度，图 5(b)、(c)则进一步显示了各因素对除碳脱氮性能的方向性及影响机理。在厌氧硫酸盐还原阶段，对有机物去除贡献较大的因子依次为进水硫酸盐（34.01%）、工作容积（19.53%）和进水 COD（16.73%）。结果表明，硫酸盐作为电子受体对还原过程的驱动作用最强，且呈现正向促进作用，这和复杂工况分析中的热图分析结果一致[3]。反应器体积与进水 COD 较高的贡献度说明负荷强度与反应空间直接影响有机物降解效率[22]，而其具体的影响机理（正向/逆向）并没有直接体现，需要进一步的分析；C/S 与回流比等作用也反映了碳源分配及系统循环对厌氧阶段稳定运行的促进作用[15]。

图 5(d)展示了阶段一模型中厌氧段各环境因子对出水硫化物的影响。其中，内回流比占比最高，回流会把反应器上端的物质（包括未转化的硫或硫化物）带回反应器底部，增加厌氧段中可被还原为

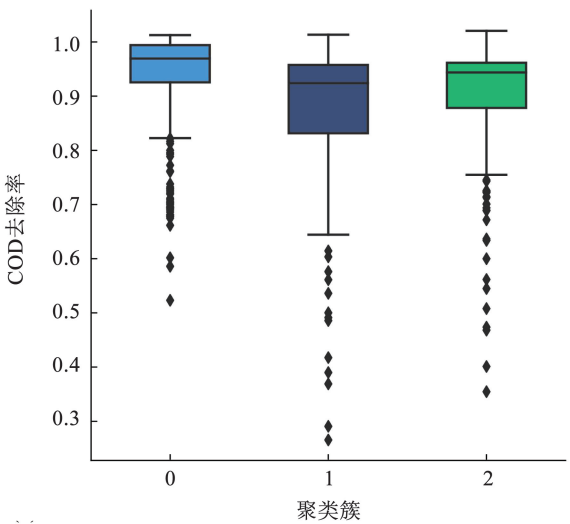
硫化物的底物,SHAP 方向性显示回流增大会倾向于降低硫化物生成。进水有机/硫源负荷次之,高 C/S 比倾向于提升硫化物生成,这体现了有机物作为电子供体的重要性[23]。反应器体积(或停留空间)和厌氧 HRT 对硫化物累积也有影响,但其相对贡献小于回流与进水负荷。



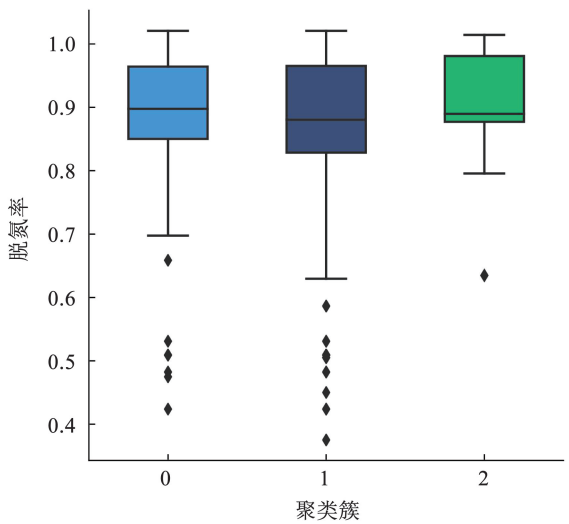
(a) 工况模拟数据集中各环境因子的皮尔逊相关性分析



(b) 工况模拟数据集聚类分析



(c) 各聚类簇的COD去除率箱型图分析



(d) 各聚类簇的脱氮率箱型图分析

图 4 基于最佳 XGBoost 模型的 SANI 工艺工况相关性及聚类分析

Fig. 4 Correlation and clustering analysis of SANI process conditions based on optimal XGBoost model

在缺氧自养反硝化阶段,主要影响脱氮性能的因子依次为 HRT(53.30%)、硫化物(21.10%)、S/N(13.36%)和进水硝酸盐(12.2%)。其中,HRT 贡献度最高,说明反应时间对硫自养反硝化效率具有决定性作用[1];硫化物与 S/N 的较高贡献反映了其在电子供体供给与反硝化动力学中的关键影响;而进水硝酸盐质量浓度贡献虽不比前述 3 项,但仍体现了氮负荷对脱氮潜力的贡献[19]。图 5(e)、(f)展示了阶段二模型中缺氧段环境因子对出水硫酸盐/亚硝酸盐的影响,环境因子的排序和脱氮率预测任务保持一致。

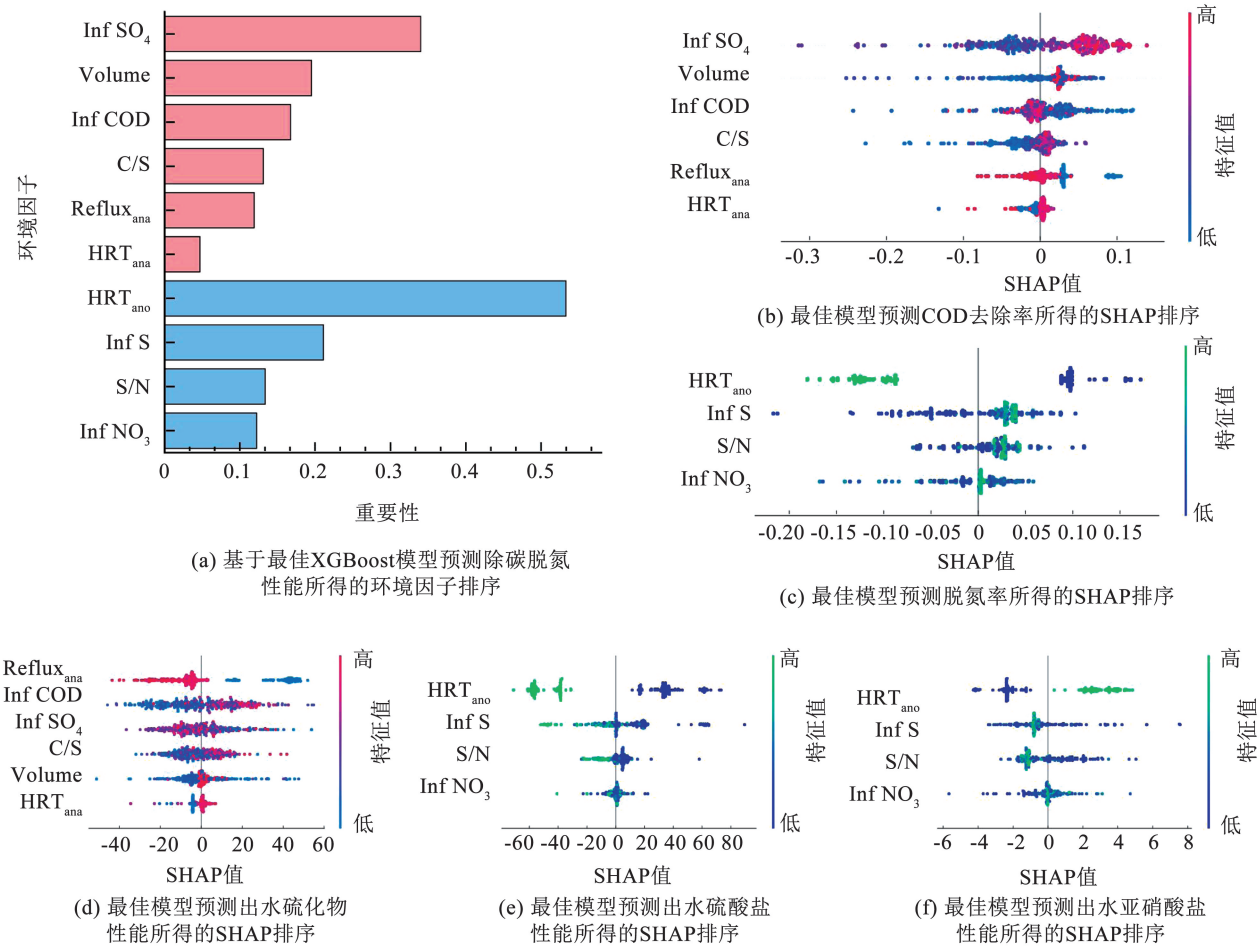
2.5 关键环境因子与除碳脱氮性能关联性分析

影响 SANI 工艺 C-N-S 循环的关键环境因子虽然被识别,但是缺乏适宜阈值的确定以直接优化调控 COD 和氮去除率。因此,通过对 SANI 工艺运行的 1 000 条工况模拟数据集的平均效应分析(如图 6 所示),可以得到适用于提升除碳脱氮性能的最佳范围,再进一步建立 SANI 工艺的优化调控机理[5]。

图 6(a)展现了进水 COD 和进水硫酸盐对 COD 去除率的协同影响,可以看出,发现进水 COD 为 340 ~ 500 mg/L 和进水硫酸盐(以 S 计)为 300 ~ 450 mg/L,COD 去除率维持在较高水平,这和聚类分析中显示的 SANI 工艺适配于低碳源含硫废水处理相匹配,更低的硫酸盐质量浓度和更高的 COD 水平会降低 COD 去除率。图 6(b)展现了厌氧工作容积对于 COD 去除率的影响,随着工作容积的增加

(3 L提升至 6.5 L)，厌氧 COD 去除率也有所增加，但当工作容积进一步增加时，COD 去除率稳定维持在最佳水平。因此，在实验室规模下，将工作容积维持在 5 ~ 6.5 L 是最经济高效的选择[15]。

图 6(c) ~ (e)分别展现了硫化物生成量、水力停留时间和进水硝酸盐质量浓度对脱氮率的影响。随着硫化物生成量的增加，脱氮率也随之增加；当硫化物生成量(以 S 计)小于 50 mg/L 时，脱氮率水平较低；当其高于 100 mg/L 时，脱氮率基本趋于稳定；但进一步提升至 175 mg/L 时，脱氮率有轻微的下降，可能是由于硫化物对微生物的毒性作用。因此，将硫化物生成量控制在 50 ~ 100 mg/L，脱氮率维持在较优水平。针对水力停留时间，当其范围控制在 6 ~ 8 h 时，脱氮率维持在较高水平，但大于 8 h 后，脱氮率明显降低。针对进水硝酸盐质量浓度，该项环境因子在脱氮率预测任务中排序靠后，平均效应分析也显示其对脱氮率的影响虽有一定约束作用，但在 20 ~ 50 mg/L，脱氮率始终维持在较高水平[1]。



(a) 基于最佳XGBoost模型预测除碳脱氮性能所得的环境因子排序

(b) 最佳模型预测COD去除率所得的SHAP排序

(c) 最佳模型预测脱氮率所得的SHAP排序

(d) 最佳模型预测出水硫化物性能所得的SHAP排序

(e) 最佳模型预测出水硫酸盐性能所得的SHAP排序

(f) 最佳模型预测出水亚硝酸盐性能所得的SHAP排序

图 5　基于最佳 XGBoost 模型分析所得的环境因子排序

Fig. 5　Ranking of environmental factors analyzed based on optimal XGBoost model

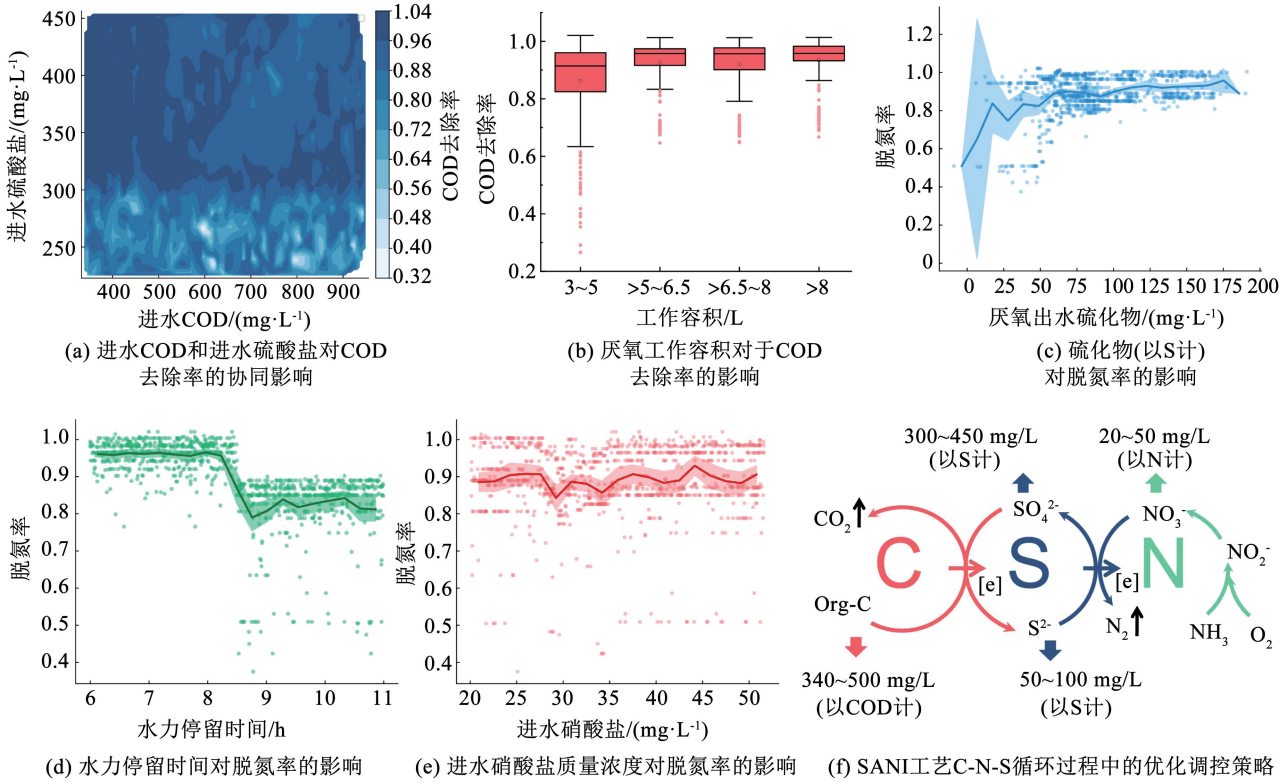
图 6(f)总结了 SANI 工艺 C-N-S 循环过程中的优化调控策略：进水 COD 为 340 ~ 500 mg/L、进水硫酸盐为 300 ~ 450 mg/L、厌氧反应器工作容积为 5 ~ 6.5 L 时，有利于提升 COD 去除率；厌氧出水硫化物为 50 ~ 100 mg/L、水力停留时间为 6 ~ 8 h 时，有利于提升脱氮率。

2.6　模型指导除碳脱氮性能情景分析

在实际 SANI 工艺运行过程中，最佳两阶段 XGBoost 模型可以应用于以下几方面：1) SANI 工艺、单体硫酸盐还原反应器、缺氧硫酸盐反应器在驯化和连续流运行过程中的除碳脱氮性能预测；2) SANI工艺出水中的硫循环优化和出水亚硝酸盐水平控制；3) 基于可解释性分析、工况模拟和平均效应分析所得的优化调控机理，可以辅助 SANI 工艺除碳脱氮性能的优化[4]。

图 7 展示了模型指导除碳脱氮性能的情景分析。初始选择工况模拟中的数据组合，最初的 COD 去除率为 54%、脱氮率为 66%，基于模型进行了以下 3 次优化：1) 将 HRT 由 10.5 h 降低至适宜阈值之内的 8 h，这导致了脱氮率升高至 97%，但 COD 去除率降低至 39%；2) 将厌氧工作容积由 3.75 L 提升至 6.5 L，这使得脱氮率维持在 84%、COD 去除

率维持在68%,二者相较于初始均有提高;3)除去针对操作条件的改变,也可以通过投加外源物质改变水质参数,促进除碳脱氮性能,如通过投加廉价硫素,进水硫酸盐水平从 229 mg/L 提升至 350 mg/L,这使得除碳脱氮性能进一步分别提升至 95% 和 92%,成功达到优化目的。该方法可以进一步应用到 SANI 工艺复杂工况优化中[24]。



(a) 进水COD和进水硫酸盐对COD去除率的协同影响　(b) 厌氧工作容积对于COD去除率的影响　(c) 硫化物(以S计)对脱氮率的影响

(d) 水力停留时间对脱氮率的影响　(e) 进水硝酸盐质量浓度对脱氮率的影响　(f) SANI工艺C-N-S循环过程中的优化调控策略

图 6　关键环境因子与除碳脱氮性能关联性分析及优化调控策略

Fig. 6　Correlation analysis and optimal regulation strategies between key environmental factors and carbon and nitrogen removal performance

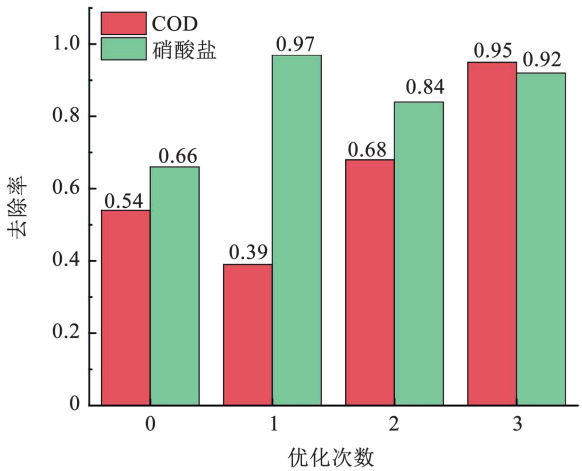


图 7　模型指导除碳脱氮性能情景分析

Fig. 7　Scenario analysis of carbon and nitrogen removal performance guided by model

3　结　论

1)成功构建了两阶段 XGBoost 机器学习模型用于预测 SANI 工艺的除碳脱氮性能,在小数据集和数据不均衡的背景下,针对 COD 去除率和脱氮率的 RMSE 分别达 6.8% 和 5.6%,表现出优异的准确性和鲁棒性。

2)基于两阶段 XGBoost 模型进行 SANI 工艺 C-N-S循环复杂工况模拟,聚类和箱型图分析验证了 SANI 工艺在低碳源含硫废水处理中的高效性(COD 与氮去除率均值分别为 94%、90%)。

3)通过 SHAP 分析和平均效应分析,总结影响 SANI 工艺的关键环境因子及适宜范围:进水 COD 为 340 ~ 500 mg/L、进水硫酸盐为 300 ~ 450 mg/L、厌氧反应器工作容积为 5 ~ 6.5 L 时,有利于提升 COD 去除率;厌氧出水硫化物为 50 ~ 100 mg/L、水力停留时间为 6 ~ 8 h 时,有利于提升脱氮率。

参考文献

[1] WANG Jin, LU Hui, CHEN Guanghao, et al. A novel sulfate reduction, autotrophic denitrification, nitrification integrated (SANI) process for saline wastewater treatment[J]. Water Research, 2009, 43: 2363. DOI: 10.1016/j.watres.2009.02.037

[2] ZOU Xu, GUO Hongxiao, JIANG Chukuan, et al. Physics-informed neural network-based serial hybrid model capturing the hidden kinetics for sulfur-driven autotrophic denitrification process [J].

Water Research, 2023, 243: 120331. DOI: 10.1016/j.watres.2023.120331

[3] LU Hui, WANG Jin, LI Shen, et al. Steady-state model-based evaluation of sulfate reduction, autotrophic denitrification and nitrification integrated (SANI) process[J]. Water Research, 2009, 43: 3613. DOI: 10.1016/j.watres.2009.05.013

[4] YANG Yan, CALLEJA P P, LIU Yiwen, et al. Assessing intermediate formation and electron competition during thiosulfate-driven denitrification: an experimental and modeling study[J]. Environmental Science & Technology, 2022, 56(16): 11760. DOI: 10.1021/acs.est.2c03937

[5] XIONG Beiyi, CHEN Kai, KE Changdong, et al. Prediction of heavy metal removal performance of sulfate-reducing bacteria using machine learning[J]. Bioresource Technology, 2024, 397: 130501. DOI: 10.1016/j.biortech.2024.130501

[6] ZHU Junjie, YANG Meiqi, REN Z J. Machine learning in environmental research: common pitfalls and best practices[J]. Environmental Science & Technology, 2023, 57(46): 17671. DOI: 10.1021/acs.est.3c00026

[7] JIANG Jinqi, XIANG Xiang, ZHOU Qinhao, et al. Optimization of a novel engineered ecosystem integrating carbon, nitrogen, phosphorus, and sulfur biotransformation for saline wastewater treatment using an interpretable machine learning approach[J]. Environmental Science & Technology, 2024, 58(29): 12989. DOI: 10.1021/acs.est.4c03160

[8] LI Zilin, LIU Haixing, ZHANG Chi, et al. Generative adversarial networks for detecting contamination events in water distribution systems using multi-parameter, multi-site water quality monitoring[J]. Environmental Science and Ecotechnology, 2023, 14: 100231. DOI: 10.1016/j.ese.2022.100231

[9] 谭周亮，李响，王伟. 数据缺陷条件下支持污水处理厂智能管理的数据增强方法[J]. 中国环境科学，2024，44(5)：2101
TAN Zhouliang, LI Xiang, WANG Wei. Data augmentation method supporting intelligent management of wastewater treatment plants under data deficiency[J]. China Environmental Science, 2024, 44(5): 2101. DOI: 10.19674/j.cnki.issn1000-6923.2024.05.021

[10] 赵娜，陈阳，刘涛. 基于机器学习识别面向重金属分离的纳滤膜物化特性[J]. 膜科学与技术，2025，45(1)：89
ZHAO Na, CHEN Yang, LIU Tao. Identification of physicochemical properties of nanofiltration membranes for heavy metal separation based on machine learning[J]. Membrane Science and Technology, 2025, 45(1): 89. DOI: 10.16159/j.cnki.issn1007.2025.01.012

[11] HUANG Ruixing, MA Chengxue, MA Jun, et al. Machine learning in natural and engineered water systems[J]. Water Research, 2021, 205: 117666. DOI: 10.1016/j.watres.2021.117666

[12] YANG Yang, LI Weishuai, HUANG Jingang, et al. Tabular generative modeling framework for multi-property data synthesis of pyrolyzed biochar[J]. Bioresource Technology, 2025, 438: 133207. DOI: 10.1016/j.biortech.2025.133207

[13] WANG Rupeng, ZHANG Shiyu, CHEN Honglin, et al. Enhancing biochar-based nonradical persulfate activation using data-driven techniques[J]. Environmental Science & Technology, 2023, 57: 4050. DOI: 10.1021/acs.est.2c07073

[14] PALANSOORIYA K N, LI Jie, DISSANAYAKE P D, et al. Prediction of soil heavy metal immobilization by biochar using machine learning[J]. Environmental Science & Technology, 2022, 56(7): 4187. DOI: 10.1021/acs.est.1c08302

[15] ZHONG Shifa, ZHANG Kai, BAGHERI M, et al. Machine learning: new ideas and tools in environmental science and engineering[J]. Environmental Science & Technology, 2021, 55(24): 12741. DOI: 10.1021/acs.est.1c01339

[16] WEI Sizhou, CHEN Zhiyuan, ARUMUGASAMY S K, et al. Data augmentation and machine learning techniques for control strategy development in bio-polymerization process[J]. Environmental Science and Technology, 2022, 11: 100172. DOI: 10.1016/j.ese.2022.100172

[17] HOU Pengfei, HU Duofei, LIU Shiqi, et al. Generative deep learning model assisted multi-objective optimization for wastewater nitrogen to protein conversion by photosynthetic bacteria[J]. Bioresource Technology, 2025, 432: 132703. DOI: 10.1016/j.biortech.2025.132703

[18] 戴嘉敏，闫晓彤，李莫特，等. 基于硫循环的污水处理工艺——SANI 工艺研究进展[J]. 工业水处理，2017，37(2)：18
DAI Jiamin, YANG Xiaotong, LI Mete, et al. Research progress on sulfate reduction-autotrophic denitrification-nitrification integrated (SANI) process for wastewater treatment based on sulfur cycle[J]. Industrial Water Treatment, 2017, 37(2): 18

[19] HAO Tianwei, WEI Li, LU Hui, et al. Characterization of sulfate-reducing granular sludge in the SANI process[J]. Water Research, 2013, 47: 7042. DOI: 10.1016/j.watres.2013.07.052

[20] 武高明. 微生物电化学系统强化 SANI 工艺处理含硫酸盐高氨氮废水的研究[D]. 杭州：浙江大学，2020
WU Gaoming. Study on SANI process enhanced by bio-electrochemical system in treating high-strength ammonia wastewater containing sulfate[D]. Hangzhou: Zhejiang University, 2020. DOI:10.27461/d.cnki.gzjdx.2020.003622

[21] ZHI Wei, APPLING A P, GOLDEN H E, et al. Deep learning for water quality[J]. Nature Water, 2024, 2: 228. DOI:10.1038/s44221-024-00202-z

[22] HAO Tianwei, LUO Jinghai, WEI Li, et al. Physicochemical and biological characterization of long-term operated sulfate reducing granular sludge in the SANI process[J]. Water Research, 2015, 71: 74. DOI: 10.1016/j.watres.2014.12.051

[23] 远野，高君，张璐璐，等. 市政污水管道硫化氢生成影响因素及预测模型研究进展[J]. 环境工程，2023，41(11)：69
YUAN Ye, GAO Jun, ZHANG Lulu, et al. Research progress on influencing factors and prediction models of hydrogen sulfide generation in municipal sewage pipelines[J]. Environmental Engineering, 2023, 41(11): 69. DOI: 10.13205/j.hjgc.202311013

[24] WU Di, EKAMA G A, CHUI H K, et al. Large-scale demonstration of the sulfate reduction autotrophic denitrification nitrification integrated (SANI) process in saline sewage treatment[J]. Water Research, 2016, 100: 496. DOI: 10.1016/j.watres.2016.05.052

（编辑　刘　彤）