陈亚松,刘家雯,赵云鹏,等.基于机器学习的人工湿地出水水质预测与影响因素 [J]. 中国环境科学, 2025,45(6):3161-3170.
Chen Y S, Liu J W, Zhao Y P, et al. Prediction of effluent water quality and analysis of influencing factors in constructed wetlands based on machine learning [J]. China Environmental Science, 2025,45(6):3161-3170.

基于机器学习的人工湿地出水水质预测与影响因素

陈亚松[1],刘家雯[2],赵云鹏[1],周英萍[2],沈秋实[1],肖 琳[2*],钱 新[2] (1.中国长江三峡集团有限公司,长江经济带生态环境国家工程研究中心,湖北 武汉 430010；2.南京大学环境学院,污染控制与资源化国家重点实验室,江苏 南京 210023)

摘要：基于水质指标、气候指标、湿地运行参数 3 个方向,收集以往研究文献数据,通过 3 种机器学习模型预测人工湿地出水氨氮(NH_4^+-N)、COD、磺胺甲噁唑(SMX)以及部分重金属的浓度.结果表明,随机森林(Random Forest)在整体性能上略优于 XGBoost 和 LightGBM,其决定系数(R^2)和均方根误差(RMSE)的表现更为稳定,尤其是在 NH_4^+-N 和 SMX 的预测上取得更高精度(NH_4^+-N 预测的 R^2 分别为 0.93、0.89 和 0.87).相比之下,在 COD 的预测中,3 种模型的表现相对较弱,R^2 分别为 0.71、0.61、0.64.通过引入 SMOTE 数据扩充技术,模型的预测性能和精度得到了显著的提升,尤其是对 COD 的预测性能提升幅度达 7.04%~26.23%.本研究将数据分析与机器学习算法相结合,可为实际工程应用提供可行方法.

关键词：机器学习；人工湿地；氨氮；COD；重金属
中图分类号：X703 **文献标识码**：A **文章编号**：1000-6923(2025)06-3161-10

Prediction of effluent water quality and analysis of influencing factors in constructed wetlands based on machine learning.

CHEN Ya-song[1], LIU Jia-wen[2], ZHAO Yun-peng[1], ZHOU Ying-ping[2], SHEN Qiu-shi[1], XIAO Lin[2*], QIAN Xin[2] (1.National Engineering Research Center of Eco-Environment in the Yangtze River Economic Belt, China Three Gorges, Wu Han 430010, China；2.State Key Laboratory of Pollution Control and Resource Reuse, School of Environment, Nanjing University, Nanjing 210023, China). *China Environmental Science*, 2025,45(6)：3161~3170

Abstract：Based on water quality indicators, climate indicators, and wetland operation parameters, data from previous studies were collected to predict the effluent concentrations of ammonia nitrogen (NH_4^+-N), COD, sulfamethoxazole (SMX), and some heavy metals in constructed wetlands using three machine learning models. The results showed that the Random Forest model slightly outperformed XGBoost and LightGBM in overall performance, demonstrating more stable R^2 and RMSE values. In particular, it achieved higher accuracy in predicting NH_4^+-N and SMX concentrations, with R^2 values of 0.93, 0.89, and 0.87, respectively, for NH_4^+-N. In contrast, the models performed relatively weaker in COD predictions, with R^2 values of 0.71, 0.61, and 0.64, respectively. By incorporating the SMOTE data augmentation technique, the prediction performance and accuracy of the models were significantly enhanced, especially for COD, where improvements ranged from 7.04% to 26.23%. This study combines scientific data analysis with machine learning algorithms, providing a feasible approach for practical engineering applications.

Key words：machine learning；constructed wetland；ammonium；COD；heavy metal

人工湿地作为一种基于自然的污水处理方式,已经被广泛应用于污水处理.然而,由于其复杂性、废水组成的多样性以及各种相互依存的物理、化学和生物因素,废水的性能变化很大,给人工湿地的设计带来挑战.为了应对这一挑战,许多数学模型如一级动力学模型[1]、Monod 方程[2]、CWM1[3]模型已被研究用于湿地系统的建模,但是这些线性动力学模型无法描述观察到的非线性机制,因此不能用于人工湿地这种比较复杂的场景[4-6].

机器学习(ML)可以处理输入和输出之间的非线性关系,并具有良好的预测性能.基于支持向量回归(SVR)的模型在防止过拟合和对数据噪声的鲁棒性方面优于人工神经网络,特别是用于有限的数据集.传统的训练测试划分方法可能使得模型在特定数据集上的性能表现过于乐观,从而增加了过拟合的风险.因此,K 折交叉验证(K-fold)得到了广泛的应用,能够有效消除因数据倾斜分布造成的偏倚估计,并最小化测试数据的不确定性和过拟合问题[7].有研究使用粒子群优化技术应对样本数量缺少的挑战,本文将使用 SMOTE (Synthetic Minority Over-sampling Technique, 合成少数类过采样技术)算法

收稿日期：2024-11-02
基金项目：中国长江三峡集团公司科研项目(NBWL202300014)
* 责任作者, 教授, xiaolin@nju.edu.cn

用于增强样本的多样性,以提高模型在不平衡数据集上的性能[8].机器学习内部的学习与训练过程是未知的,为了对这一“黑箱”模型有更深入的理解,重要性通常归因于每个输入特征,对这些特征的解释与模型训练和其他许多任务同样重要.有关特征重要性的计算方式已经有很多,但都不具备一致性.SHAP (Shapley)值作为一种特征选择机制具有较强的一致性,能够很好地为机器学习这一黑箱模型提供解释.SHAP 受一个来自博弈论和局部解释的概念 Shapley 值的启发,计算给定的特征数据集的每个子集对模型的边际贡献度,其思想是通过一系列局部相似性来近似全局相似性[9].

机器学习方法已被大量用于对水质水量的预测[10-12].目前研究主要集中在湿地中氮磷等常规指标的预测[13-14],而对重金属、抗生素等新型污染物的预测研究相对较少.为了全面评估人工湿地的处理效能,本文选取了 3 个不同方面相关的 15 个人工湿地属性,比较 3 种模型对 CWs 降解 NH_4^+-N、COD、磺胺甲噁唑(SMX)和几种常见重金属的预测能力,量化并探索影响人工湿地性能的因素,旨在为人工湿地的构建和运行参数提供参考.

1 材料与方法

1.1 数据收集

表 1 数值型变量的汇总统计信息

Table 1 Summary statistics of numerical variables

变量类型	变量集	范围(均值)
湿地构造参数	基质层厚度(cm)	5~92 (34.886)
湿地运行参数	电压(V)	0~15 (1.036)
	水力停留时间(h)	2~840 (90.556)
气候	气温(℃)	4.2~30 (23.372)
水质	C/N	0~128.69 (7.92)
	pH 值	2~10.2 (7.312)
	DO(mg/L)	0.3~9.8 (3.651)
	进水 COD(COD-i)(mg/L)	2.24~2907 (319.582)
	进水 NH_4^+-N(NH_4^+-N-i)(mg/L)	0~360.4 (33.187)
	进水 NO_3^--N(NO_3^--N-i)(mg/L)	0.12~49.84 (12.116)
	进水 SMX(SMX-i)(mg/L)	0.005~100 (4.166)
	进水重金属(mg/L)	0~20.27
	出水 COD(COD-e)(mg/L)	1.434~980 (92.549)
	出水 NH_4^+-N(NH_4^+-N-e)(mg/L)	0.01~192.8 (12.967)
	出水 NO_3^--N(NO_3^--N-e)(mg/L)	0.01~54.44 (8.063)
	出水 SMX(SMX-e)(mg/L)	0.00001~23.9 (1.505)
	出水重金属(mg/L)	0~14.11

在 Web of Science、Google Scholar、Elsevier ScienceDirect、知网文献平台调研有关湿地的中英文文献,关键词为“constructed wetlands”、“人工湿地”、“水质预测”、“nitrogen”、“heavy metal”、“antibiotics”,文献搜索范围为 2000~2024 年.共收集了 96 篇文献,选取了 410 条数据.本文一共确定了 15 个特征(自变量),以及出水 COD、NH_4^+-N、SMX、重金属浓度作为因变量.这些特征包括进水 pH 值、溶解氧、氨氮、硝态氮、总氮、COD、重金属等水质指标,以及水利停留时间、湿地的类型、进水方式、植物类型等.为了有更深的认识,本文将这些特征进行分类(表 1).

1.2 数据预处理

为了避免引入无意义的数值顺序,对于离散型的分类特征,本文选择了频率编码.然后对每列数据进行特征缩放,防止样本中不同特征的数值差别较大而影响模型的学习效率.由于文献研究内容的差异,导致不可避免的数据缺失,因此本文采用岭回归算法结合多重插补来填充缺失值,从而提高后续模型的稳定性和预测性能.此外,灵敏度分析用于评估输入参数对输出变量的影响,有助于提高模型的预测精度[15].

SMOTE 是一种处理数据不平衡问题的技术,通过在少数样本之间进行插值,生成新的少数类样本,从而达到扩充数据集、平衡类分布的目的.对于每一个少数类样本, SMOTE 算法使用 k 近邻算法找到其 k 个最邻,这些邻居也是少数类样本.对于每个少数类样本和其邻居, SMOTE 算法会通过插值生成新样本.本文使用此技术进行数据的扩充,来达到使模型更加健壮的目的.具体新的合成样本 x_{new} 的计算公式如下:

$$x_{\text{new}} = x_i + \lambda \cdot (x_j - x_i) \tag{1}$$

式中:x_i 为当前的少数类样本; x_j 为当前少数类样本的一个邻居;λ为一个随机值,取值范围为 0~1,用于决定插值的程度.

1.3 模型的选择

在模型选择中,选用随机森林(RF)、XGBoost (XGB)和 LightGBM(LGB).这 3 种模型都具备出色的鲁棒性和处理复杂数据的能力.随机森林具备强大的准确性和稳定性,XGBoost 在处理缺失值和复杂数据时表现优异,LightGBM 则在高维度数据上具有更高的计算效率.通过对这些模型的超参数进行

调优,优化整体的预测性能.

1.4　模型训练及超参数调优

为了较好地拟合本研究的机器学习模型,建模过程主要包括两个阶段:模型训练和验证(图 1).经过数据的预处理,将 80%的数据划分为训练集用来训练模型,20%作为测试集用于检验模型的准确性.为了防止模型的过拟合问题,对训练集再使用 K 折交叉验证(K-Fold Cross-Validation)方法划分交叉验证集[16].本研究使用 10 折交叉验证,将训练集随机分成 10 个大致相等的部分(折),然后依次将每个折作为验证集,其余 9 个折作为训练集,进行 10 次迭代.记录每次迭代的模型性能,包括准确性或其他评估指标,然后计算模型性能的平均值,得出模型的综合性能.最后,使用验证集对模型的预测性能进行最终的评价.使用 python 3.6 的第三方库 sklearn 实现模型的构建,并用网格搜索的方式配合反复交叉验证对模型尝试使用超参数的不同组合.3 种模型需要优化的超参数配置如表 2 所示.

采用均方根误差(RMSE)和相关系数(R^2)衡量人工湿地预测值的偏差和预测器的性能. RMSE 越小,R^2 越接近 1,表明模型的预测效果越好.

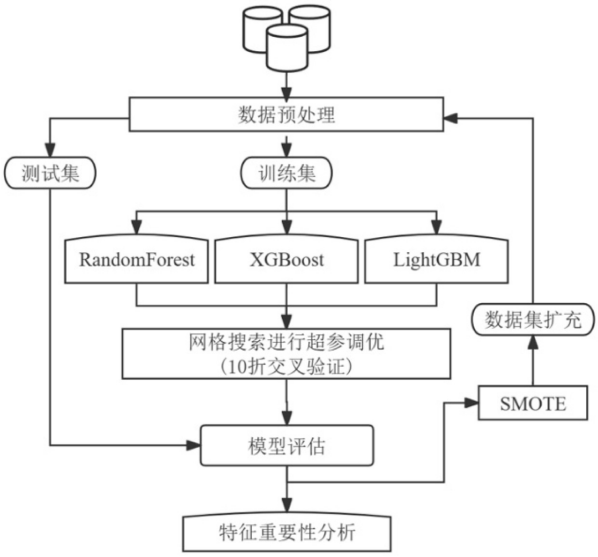

图 1　模型训练与预测流程

Fig.1　Model training and prediction process

表 2　通过网格搜索的 RandomForest、XGBoost、LightGBM 超参数优化

Table 2　Hyperparameter optimization of RandomForest, XGBoost, and LightGBM using Grid Search

模型	超参数	范围	描述
RandomForest	max_depth	[2,30]	决策树的最大深度
	min_samples_leaf	[1,12]	叶子节点的最小容量
	n_estimators	[50,1500]	决策树的数量
	min_samples_split	[2,25]	分裂节点的最小容量
XGBoost	eta	[0.8,0.9]	弱分类器,决定 XGBoost 复杂性的最大深度
	max_depth	[1,4]	树的最大深度
	Reg_lambda	[0.75,1]	正则化参数
	min_child_weight	[0,0.1]	最小叶子节点样本的总重量
	n_estimators	[60,90]	决策树的数量
	subsample	[0.9,1]	每棵树所使用的训练子样本占整个样本的比例,为了防止过拟合
LightGBM	num_leaves	[20,70]	每个弱学习,器的最大叶子数
	learning_rate	[0.01,0.1]	学习率,也称步长
	num_iterations	[100,500]	增强迭代的次数
	min_child_samples	[20,50]	决策树的数量
	subsample	[0.7,1]	每次树构建迭代使用的行的百分比
	colsample_bytree	[0.7,1]	在构建每颗树时,从全部特征中随机采样的比例

1.5　特征重要性评估

模型自带的特征重要性分析通常是全局性的,无法捕获局部信息. SHAP 是一种用于解释机器学习模型预测的工具. SHAP 提供了一种统一的方法来量化每个特征对预测结果的影响,使得解释复杂模型变得更加透明和直观.根据不同特征组合的模型输出,计算每个特征的 Shapley 值,即特征对预测结果的平均边际贡献.

2　结果与讨论

2.1　数据前处理

通过岭回归算法对缺失值进行填补,并结合箱线图和 KS 检验进行验证.填补前后的数据集箱线图如图 2 所示,中位数或均值均无显著偏移.其次,假设两

组数据来自相同的分布,通过 KS 检验, C/N、pH 值、DO、COD−i、COD−e、NO_3^-−N−i、NO_3^-−N−e 变量的 *P* 分别为 0.10445、0.99965、0.55992、0.99829、0.55843、1.0、0.99999,均大于 0.05,表明无法拒绝原假设,即两组数据的分布一致.表明通过岭回归填补的数据并未扭曲原始数据集的真实分布.

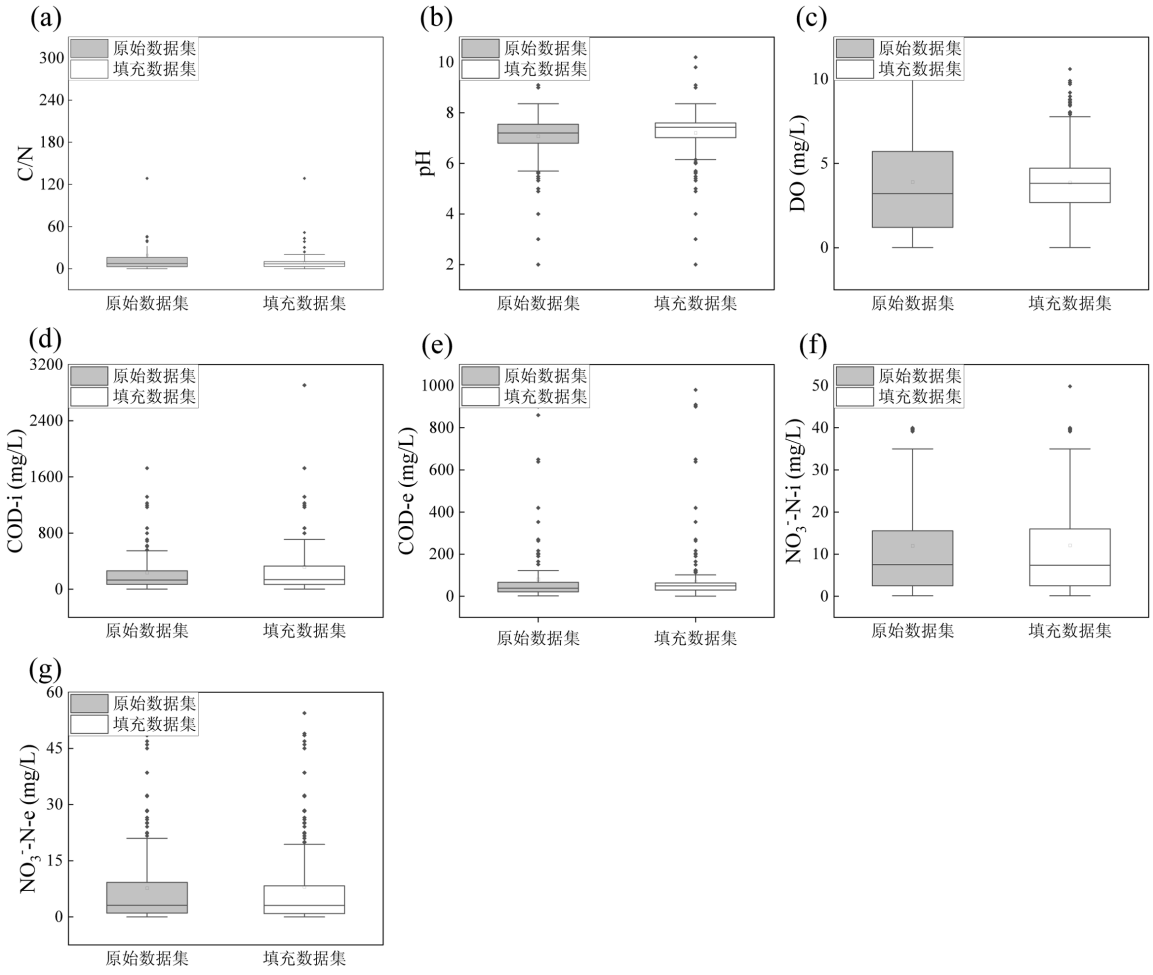

图 2 缺失值填补前后数据集的分布对比

Fig.2 Comparison of the distributions between the original and imputed datasets

通过绘制 NH_4^+−N、COD 和 SMX 的相关性热图检查无多重共线性(图 3),发现相关性在−0.58~0.5,只有进出水 SMX 和 NH_4^+−N 浓度的相关性上升到 0.91 和 0.72,呈高度正相关性,本文在后续模型训练与预测时将此变量剔除.而其他变量之间的相关性均保持在−0.5~0.5,意味着数据之间的多重共线性相对较低.将数据集的数据用箱线图绘制出来进行异常点检查,将部分异常点去除更有利于机器学习的效率.根据图 4 得知,本数据集异常值数量较少.

为了进行敏感性分析,在保持另一个输入参数不变的情况下,将输入变量的数据改变 10%,并使用机器学习方法计算出水参数(NH_4^+−N、COD、SMX)的百分比变化,计算结果如表 3 所示[17].由于温度、DO、pH、填料层厚度和进水 COD 的变化, NH_4^+−N−e、COD−e 和 SMX−e 的变化百分比最高.灵敏度分析旨在评估和理解模型输出对输入变量的响应程度,确定哪些输入变量对模型预测结果有显著的影响,以及评估模型在输入变量发生变化时的稳定性,即模型是否对微小的输入变化产生过大的输出波动.温度变化导致 NH_4^+−N 和 SMX 参数发生显著变化,这是因为温度升高会提高生物反应从而提高去除效率[18].填料层厚度、DO 和 pH 值改变 10%也会影响出水参数,这对氨氮的去除有必然影响,填料层厚度也间接地影响了人工湿地 DO 的含量.

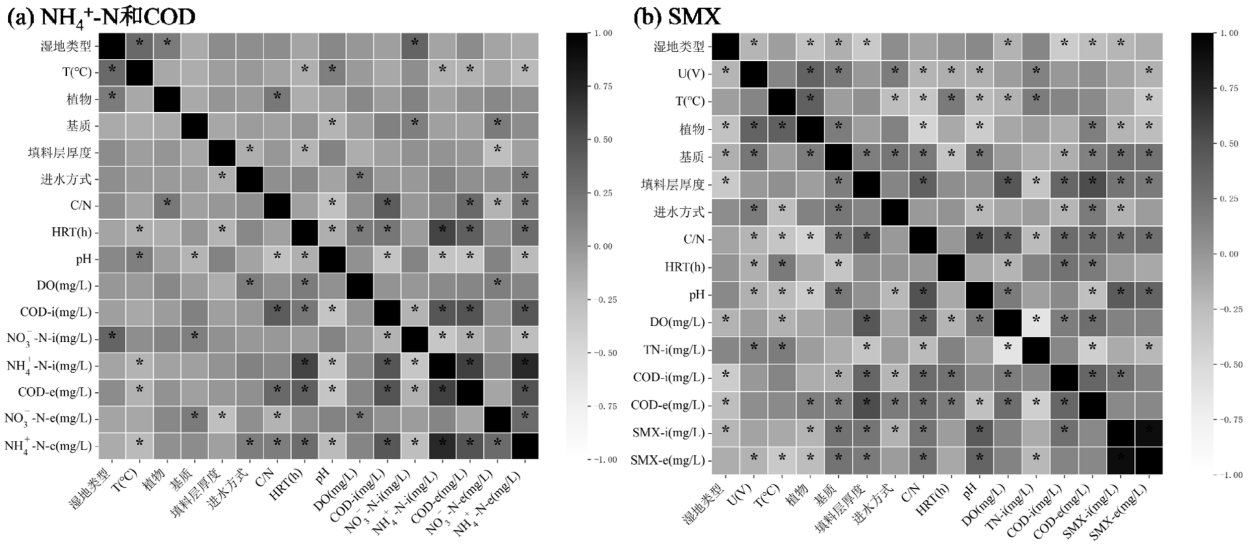

图 3　输入和输出数据集之间的特征相关热图

Fig.3　Feature correlation Heatmap between input and output datasets

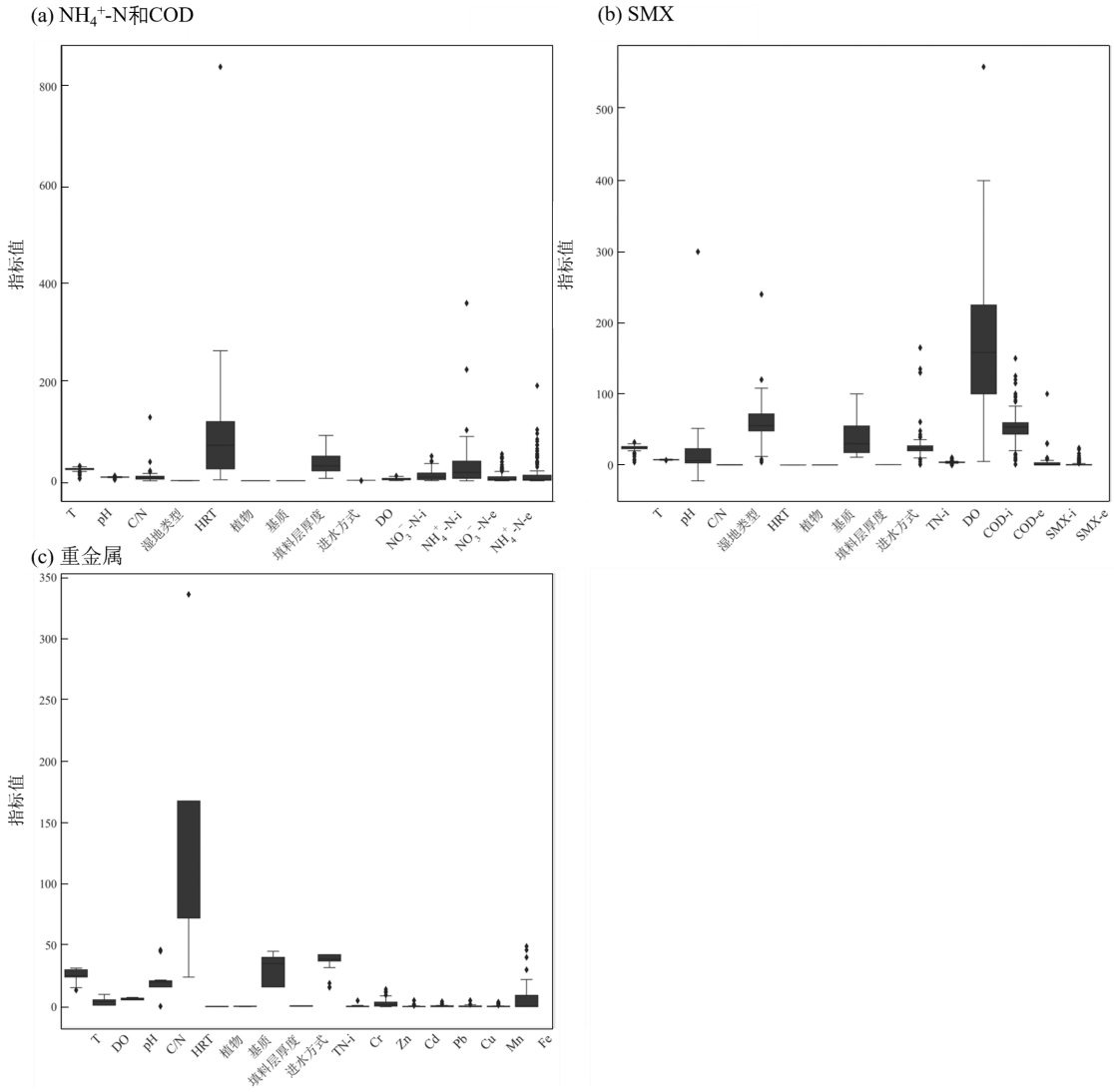

图 4　变量的箱线图

Fig.4　Box plot for variables

表 3 NH_4^+-N-e、COD-e 和 SMX-e 对其他输入变量的敏感性分析

Table 3 Sensitivity analysis of NH_4^+-N-e, COD-e and SMX-e with other variables

参数	输入变量变化 10%		
	NH_4^+-N-e 变化百分比(%)	COD-e 变化百分比(%)	SMX-e 变化百分比(%)
T(℃)	4.155	0.709	3.168
填料层厚度(cm)	1.543	3.58	0.887
C/N	2.734	1.689	0.283
HRT(h)	0.085	0.163	0.388
DO(mg/L)	0.009	1.739	8.255
COD-i(mg/L)	9.739	2.477	4.665
pH 值	1.81	0.732	6.619

2.2 模型参数调优

对于模型超参数的调优本文都使用网格搜索的方式,统一借助于 python 的 sklearn 第三方库实现,对各模型的参数进行组合.所有模型都使用随机抽样从所有超参数的组合到一定次数的迭代进行优化.训练集被分成几个相等的集,其中一个部分用于模型验证,剩下的集合用于模型训练[10].对于 Random Forest, n_estimators 和 min_samples_leaf 是主要需要优化的参数,分别表示森林中树的数量和每个叶子节点上最少数据数目.对于 n_estimators,从 50~1500 以 50 为步长进行递增.当 n_estimators 和 min_samples_leaf 分别取 150 和 1 时,模型的训练效果最好.但是叶节点较小的样本数可能导致过拟合,因此 RMSE 稍微增高. max_depth 代表树的最大深度,随着它的增加, RMSE 值呈现出较小的波动,整体趋势相对平稳,这表明深度的增加对模型性能影响有限. min_samples_split 影响较大,随着最小分裂样本数增加,RMSE 值有一定波动性,反映了分裂标准对模型复杂性的影响.

XGBoost 和 LightGBM 模型的调优结果表示,学习率(eta)对 RMSE 有显著影响,随着 eta 的变化,RMSE 值的变化范围明显.并且较低的学习率更适合 LGB 模型.与 RF 类似,树的数量增加对 RMSE 有一定影响,但并不是非常显著,而过多的树叶会带来计算资源的浪费,这个参数的调优需要考虑到模型复杂性和训练时间的平衡.树的最大深度对 XGBoost 的影响相对较小,RMSE 波动不大. reg_lambda 是正则化参数,可以帮助防止过拟合.子采样率(subsample)的变化对 XGB 也有明显影响,反映了模型在训练过程中样本选择的影响,而此参数对 LGB 似乎并没有显著的影响.较小的 min_child_samples 可能导致 LGB 更低的 RMSE. colsample_bytree 代表了 LGB 随机选择用于训练的特征比例,随着该参数值的增加, RMSE 的波动不大,这表明在这个范围内, colsample_bytree 对模型的性能影响相对较小.

2.3 模型对比分析

如图 5 所示,XGBoost 在测试集上的正则化系数(R^2)最高,为 0.93,超过了 Random Forest (R^2=0.89)和 LightGBM (R^2=0.87)模型,并且均方根误差也是最小,只有 2.93. 3 个模型的正则化系数均超过了 0.85,展现出对氨氮去除效果较强的预测能力.如图 5(a)所示,模型残差(实际值和预测值之间的差值)中的一些异常值意味着它们可能导致回归线倾斜[4].与 LGB 的范围为-12.5~7.21 和 XGB 的-8.94~11.76 相比, RF 的残差范围为-8.73~7.54. RF 能更好地预测数据集实际值中的极值点.以往的研究表明,机器学习在 CWs 氮预测中的应用研究均各不相同.例如 Akratos 等[19]使用具有 9 个输入特征的人工神经网络来预测 5 个试点出水的总氮去除率,其 R^2 为 0.69. Salem 等[20]也采用随机森林模型来预测污水处理厂的出水水质, BOD 和 COD 的去除率 R^2 分别为 0.66 和 0.68.

然而,在对 COD 的预测上,预测值和实际值的拟合均不如 NH_4^+-N, Random Forest、XGBoost、LightGBM 的 R^2 分别只有 0.71、0.61、0.64,均方根误差分别为 25.6、29.64、28.22.这可能是由于数据集中含有某些特殊类型的废水,如化工废水和农业废水,其 COD 含量显著高于其他生活污水或尾水,从而对模型的学习造成了干扰.此外,实际污水中 COD 的生化含量具有时效性变化,且 COD 和 BOD_5 的测量过程中存在固有误差,这些因素可能共同影响了预测结果的准确性.Dai 等[21]也使用基于 ASM、随机森林系统框架对人工湿地运行和设计进行优化,而本研究的模拟结果显示了更为有效的预测性能.总体来看, 3 个模型的表现相对接近,这可能与数据集的特性以及模型的调优过程有关. 3 种模型对 SMX 的预测结果分别为 0.97、0.98、0.97,均方根误差分别为 0.12、0.12、0.13. 3 个模型取得了比较一致的结果,说明该训练集具有比较强的规

律性.根据表 4 中的重金属测试集预测结果,模型表现出色,几乎所有指标均达到了较为理想的预测效果.综合考虑,随机森林在人工湿地水质预测建模中更具优势.

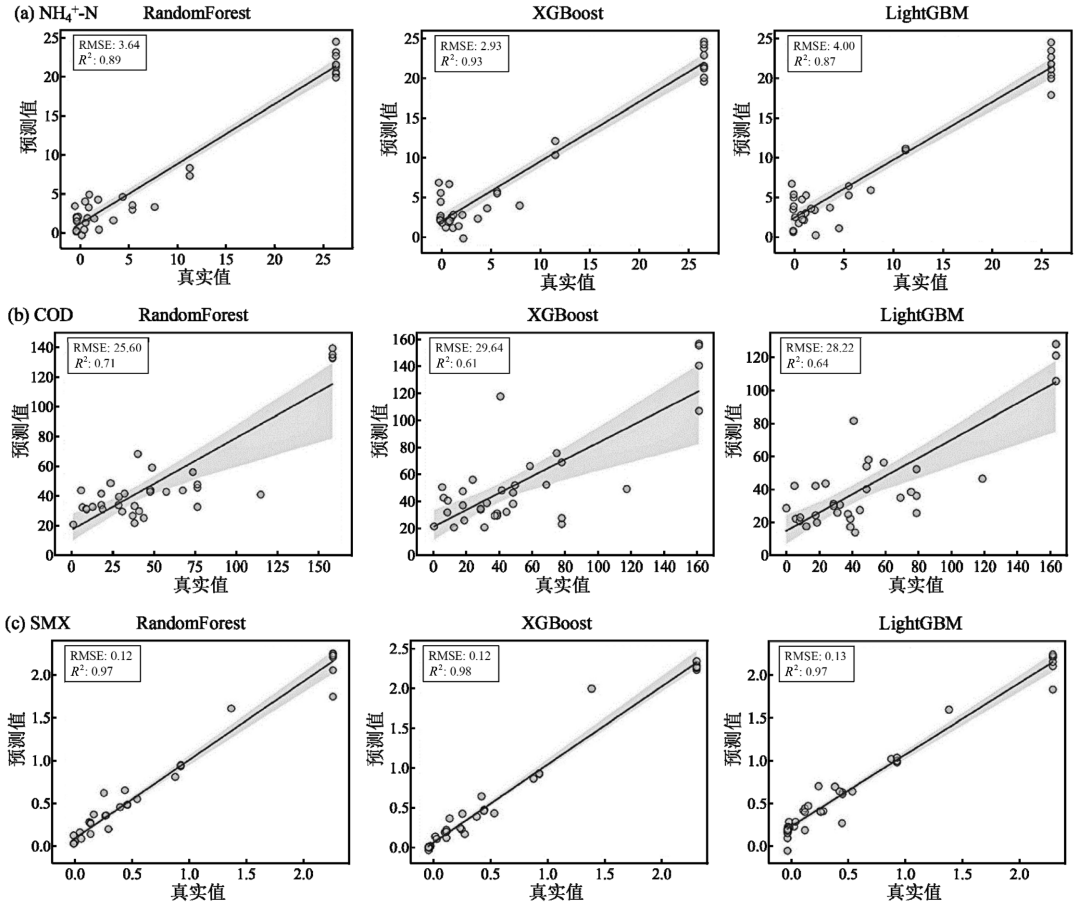

图 5　Random Forest、XGBoost、LightGBM 模型对出水 NH_4^+–N, COD 以及 SMX 的预测

Fig.5　Prediction of effluent NH_4^+–N, COD and SMX using Random Forest, XGBoost, and LightGBM

表 4　三种机器学习模型对重金属的预测结果

Table 4　Prediction results of heavy metals using three machine learning models

模型	Cd	Cr	Cu	Mn	Pb	Zn
RF	0.94	0.98	0.95	0.97	0.99	0.97
XGB	0.92	0.97	0.94	0.96	0.98	0.96
LGB	0.91	0.95	0.93	0.94	0.98	0.96

2.3　数据集扩充对 COD 预测优化

结果表明,基于 COD 的预测模型拟合效果不理想,因此选择 SMOTE 技术,进一步提升 3 种人工湿地水质预测模型精度.随机选择一个少数类样本,在该样本的 k 邻近中随机选择一个样本,然后沿着该样本和邻近样本的连线,随机选择一个点作为新生成的样本[22].然后使用虚拟样本和原来样本共同对模型进行再次训练、评估,结果如表 5 所示.为了验证 SMOTE 生成虚拟样本方法的可用性,本文同样以 R^2 和 RMSE 作为评估标准.在对 COD 的预测中, RF、XGB 和 LGB 3 个模型的预测 R^2 各自提升了 7.04%、26.23%、14.06%,都达到了 0.7 以上.均方根误差也同样得到降低,分别下降了 13.59%、26.01%、16.16%.其中, XGB 的改善效果最为显著,模型更佳稳定,如图 6 所示.

表 5　数据扩充前后对 COD 预测的结果对比

Table 5　Comparison of COD prediction results before and after data augmentation

模型	COD			
	扩充前		扩充后	
	R^2	RMSE	R^2	RMSE
Random Forest	0.71	25.6	0.76	22.12
XGBoost	0.61	29.64	0.77	21.93
LightGBM	0.64	28.22	0.73	23.66

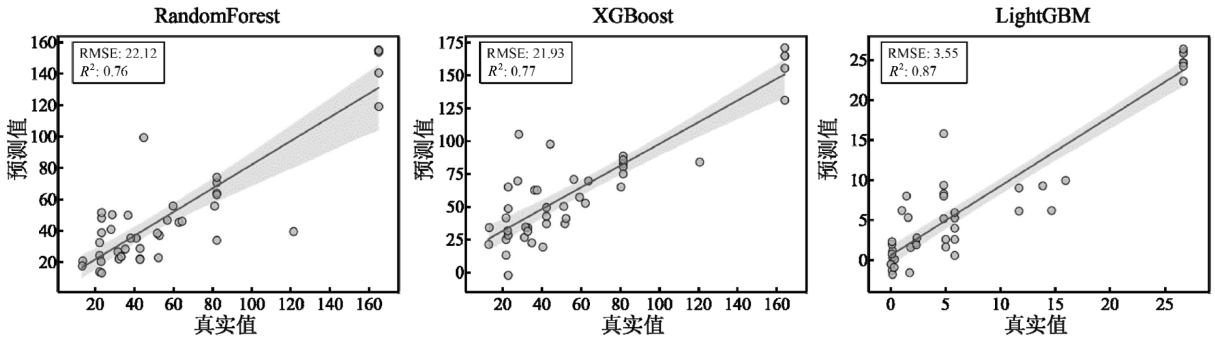

图 6 数据扩充后对 COD 的预测结果

Fig.6 Prediction results of COD after data augmentation

2.4 湿地条件变量重要性分析

对特征进行重要性评估是机器学习模型解释中的一个重要部分.通过评估特征的重要性,可以掌握哪些特征对模型的预测结果贡献最大.

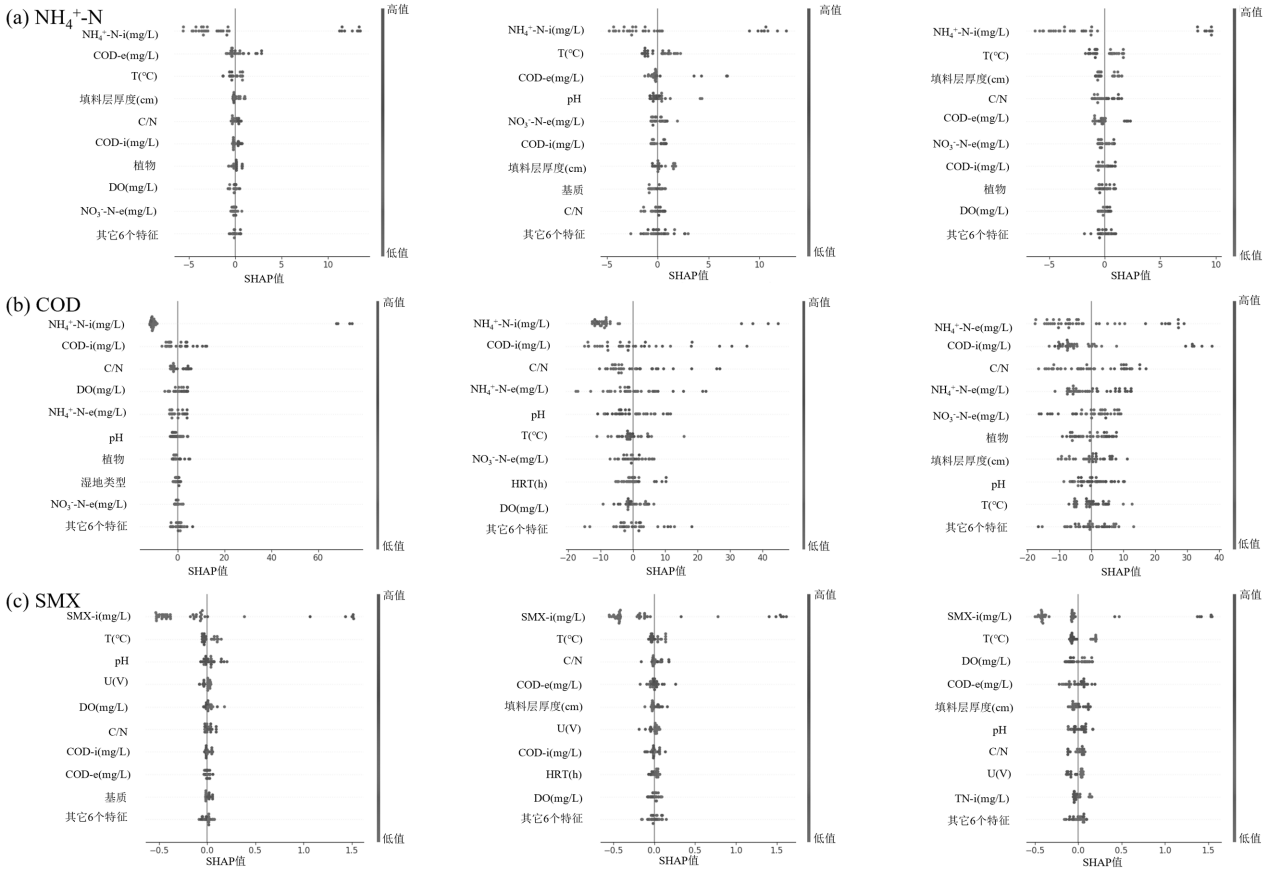

图 7 各模型预测 NH_4^+−N、COD 和 SMX 的 SHAP 特征重要性分析

Fig.7 Feature importance of design and operation parameters on the output of the three models for the predictions of NH_4^+−N, COD and SMX

由 SHAP 对预测结果作出解释,可以进一步增加人们的理解.由蜂窝图(图 7(a))可以看出,各个模型对 NH_4^+−N 预测结果作出的解释基本相似,排在第一位的是进水氨氮,呈现出明显的正相关性.这可能由于收集的大部分数据都来自污水处理厂的尾水样本,其中 NH_4^+−N、COD 等都同时被去除[23].除此以外,温度对湿地的运行影响很大,呈负相关.具体而言,温度越高,出水的氨氮浓度就越低,这可能一部分是由于温度高有利于植物的生长,从而更进一步对氮的吸收[24].并且细菌对温度敏感,低温下往往更不利

于各种菌的生存[25].其次是填料层的高度和 C/N,填料层越厚,出水氨氮浓度越低,表明处理效果越好[26]. C/N 可能受一些如生物炭的基质影响而升高,这对反硝化菌进行反硝化脱氮是友好的,可以作为其外接碳源[27-28].RF 和 LGB 的特征性分析同时揭示了 DO 对 NH_4^+-N 的重要性,这或许归因于 DO 在促进 NH_4^+-N 去除反应中所扮演的核心角色,二者的浓度变化趋势呈现出明显的负相关关系[29].点越分散表明这个特征对预测结果的影响越大.温度的 SHAP 较总体均匀的分布在正方向上,说明温度升高对模型输出有积极作用.

进水水质如 NH_4^+-N、COD 浓度和 C/N 与出水 COD 浓度相关性最高,并呈现正相关性.温度对出水 COD 浓度呈强负相关性,温度越高, 出水 COD 越低.这主要由于低温严重降低了微生物的活性,阻碍了有机污染物的去除. DO 主要由天然供养以及植物根系释氧,对污染物的去除起到关键作用.并且,碳源的添加会导致植物氧气释放量变多,有利于氮和有机物的去除[30].此外, DO 势必也会受到温度的影响.有研究探讨了不同水生植物之间泌氧能力的差别,表明植物的泌氧率大小决定了 COD 削减量高低,并将其作为筛选湿地植物的一个重要指标.

在对出水 SMX 浓度的预测中,各模型作出的解释基本相同.进水 SMX 浓度对 SMX 的去除重要性最大,且呈负相关,较大初始浓度使得目标污染物更难去除[31].与 NH_4^+-N 和 COD 一样,温度的特征重要性排名第二,且呈负相关性.高温与植物通过吸收去除污染物直接相关,在相关文献综述中讨论了季节和温度,并指出在 15~25℃的温度下,后者活性达到最高[32].其次,电压在抗生素去除中的作用尤为重要,尤其是在低温条件(<15℃)下,传统方法的去除效果有限,通过电解一体化人工湿地,这一问题得到了有效改善[33-34].有研究表明,将电解池与潮汐流人工湿地结合后,温度对 SMX 的降解路径影响不大[35].

3　结论

3.1　对数据集进行预先数据分析,发现自变量之间的相关性并不高,相关性系数几乎都在-0.58~0.5,多重共线性相对较低.并且整个数据集数据较为规则,无明显异常值.而在对特征进行敏感性分析时发现,温度、pH 值、填料层厚度较为敏感,应着重考虑.

3.2　通过收集大量简易的水质指标及工艺参数,对比 3 种模型对两种出水指标的预测效果.为进一步减小模型预测误差,使用了数据扩充技术 SMOTE,大大提高了模型对 COD 的预测性能,提高了模型预测的精度,对此 3 种模型都具有较好的适应性.在预测效果对比中,3 种模型对 NH_4^+-N 的预测 R^2 都达到了 0.8 以上,对 COD 的预测达到 0.7.而对 SMX 和重金属的预测效果均达到了 0.97 及以上.综合来看,随机森林较其它模型拟合度 R^2 更高、RMSE 更小,更适合于此数据集预测.

3.3　对模型进行特征重要性分析,温度和水质初始浓度被识别为对模型预测性能影响最为显著的两个关键因子,这与敏感性分析结果对应.

参考文献：

[1] Nguyen X C, Chang S W, Nguyen T L, et al. A hybrid constructed wetland for organic-material and nutrient removal from sewage: Process performance and multi-kinetic models [J]. Journal of Environmental Management, 2018,222:378-384.

[2] Saeed T, Sun G. The removal of nitrogen and organics in vertical flow wetland reactors: predictive models [J]. Bioresource Technology, 2011, 102(2):1205-1213.

[3] Langergraber G, Rousseau D P, Garcia J, et al. CWM1: a general model to describe biokinetic processes in subsurface flow constructed wetlands [J]. Water Science and Technology, 2009,59(9):1687-1697.

[4] Nguyen X C, Nguyen T P, Lam V S, et al. Estimating ammonium changes in pilot and full-scale constructed wetlands using kinetic model, linear regression, and machine learning [J]. Science of the Total Environment, 2024,907, doi:10.1016/j.scitotenv.2023.168142.

[5] 董越洋,徐　波,王　鹏,等.一种基于Stella和R语言的湿地氮素动力学模型 [J]. 中国环境科学, 2020,40(1):198-205.
Dong Y Y, Xu B, Wang P, et al. An ecological kinetic model of nitrogen cycle in the wetland based on Stella and R language [J]. China Environmental Science, 2020,40(1):198-205.

[6] Yang B, Xiao Z, Meng Q, et al. Deep learning-based prediction of effluent quality of a constructed wetland [J]. Environmental Science and Ecotechnology, 2023,13, doi:10.1016/j.ese.2022.100207.

[7] Vu H L, Ng K T W, RichterA, et al. Analysis of input set characteristics and variances on k-fold cross validation for a Recurrent Neural Network model on waste disposal rate estimation [J]. Journal of Environmental Management, 2022,311,doi:10.1016/j.jenvman.2022.114869.

[8] Wang Z, Wu F, Hao N, et al. The combined machine learning model SMOTER-GA-RF for methane yield prediction during anaerobic digestion of straw lignocellulose based on random forest regression [J]. Journal of Cleaner Production, 2024,466,doi:10.1016/j.jclepro.2024.142909.

[9] Marcilio W E, Eler D M. From explanations to feature selection: assessing SHAP values as feature selection mechanism [Z]. 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). 2020,340(7),doi:10.1109/sibgrapi51738.2020.00053.

[10] 杨晓玲,王梦晓,李晓娟,等.机器学习-多元线性回归预测铜的水生态基准 [J]. 中国环境科学, 2024,44(7):3976-3985.

Yang X L, Wang M X, Li X J, et al. Predicting the water ecological criteria of copper using machine learning and multiple linear regression approach [J]. China Environmental Science, 2024,44(7): 3976−3985.

[11] 周 琪,于 洋,刘苗苗,等.基于机器学习和非参数估计的 $PM_{2.5}$ 风险评估 [J]. 中国环境科学, 2022,42(8):3554−3560.
Zhou Q, Yu Y, Liu M M, et al. Risk assessment of $PM_{2.5}$ pollution based on machine learning and nonparametric estimat [J]. China Environmental Science, 2022,42(8):3553−3560.

[12] 李涛宇,许秀春,杨 轩,等.利用机器学习预测华北地区冬小麦农田氮淋失 [J]. 中国环境科学, 2025,45(1):343−354.
Li T Y, Xu X C, Yang X, et al. Prediction of nitrogen leaching from winter−wheat production in North China based on random forest and XGBoost [J]. China Environmental Science, 2025,45(1):343−354.

[13] Li H, Xie L, Zhou B, et al. Machine learning−assisted prediction and identification of key factors affecting nitrogen metabolism for aerobic granular sludge [J]. Environmental Research, 2025,273,doi:10.1016/j.envres.2025.121158.

[14] Bao Y, Wang Y, Hu M, et al. Deciphering the impact of cascade reservoirs on nitrogen transport and nitrate transformation: Insights from multiple isotope analysis and machine learning [J]. Water Research, 2025,268, doi:10.1016/j.watres.2024.122638.

[15] Cosenza A, Mannina G, Vanrolleghem P A, et al. Global sensitivity analysis in wastewater applications: A comprehensive comparison of different methods [J]. Environmental Modelling and Software, 2013, 49:40−52.

[16] 苏子龙,严文亮,李慧敏,等.基于改进麻雀搜索算法优化BP神经网络的农业碳排放预测 [J]. 环境科学:1−15.
Su Z L, Yan W L, Li H M, et al. Prediction of Agricultural Carbon Emission Based on Improved BP Neural Networkwith Optimized Sparrow Search Algorith [J]. China Environmental Science: 1−15.

[17] Singh S, Kulshreshtha N M, Goyal S, et al. Performance prediction of horizontal flow constructed wetlands by employing machine learning [J]. Journal of Water Process Engineering, 2022,50,doi:10.1016/j.jwpe.2022.103264.

[18] Rampuria A, Gupta A B, Brighu U. Nitrogen transformation processes and mass balance in deep constructed wetlands treating sewage, exploring the anammox contribution [J]. Bioresource Technology, 2020,314,doi:10.1016/j.biortech.2020.123737.

[19] Akratos C S, Papaspyros J N E, Tsihrintzis V A. Total nitrogen and ammonia removal prediction in horizontal subsurface flow constructed wetlands: Use of artificial neural networks and development of a design equation [J]. Bioresource Technology, 2009,100(2):586−596.

[20] Salem M, El−sayed Gabr M, Mossad M, et al. Random Forest modelling and evaluation of the performance of a full−scale subsurface constructed wetland plant in Egypt [J]. Ain Shams Engineering Journal, 2022,13(6),doi:10.1016/j.asej.2022.101778.

[21] Dai W, Pang J−W, Zhao Y−j, et al. Machine learning assisted combined systems of wastewater treatment plants with constructed wetlands optimal decision−making [J]. Bioresource Technology, 2024,399, doi:10.1016/j.biortech.2024.130643.

[22] Fernandez A, Garcia S, Herrera F, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15−year anniversary [J]. Journal of Artificial Intelligence Research, 2018,61: 863−905.

[23] 何 强,胡书山,向泽毅,等.垂直流人工湿地系统净化污水厂尾水脱氮效果 [J]. 中国环境科学, 2023,43(8):3956−3965.
He Q, Hu S S, Xiang Z Y, et al. Study on the nitrogen removal ability of vertical flow constructed wetland treating tailwater of sewage plant [J]. China Environmental Science, 2023,43(8):3956− 3965.

[24] 王 旭,郭 雯,王明果,等.人类活动影响下异龙湖浮游植物碳氮稳定同位素变化特征 [J]. 中国环境科学, 2023,43(6):3087−3099.
Wang X, Guo W, Wang M G, et al. Pattern in changes of carbon and nitrogen stable isotopes of phytoplankton in Yilong Lake under the influence of human activities [J]. China Environmental Science, 2023, 43(6):3087−3099.

[25] Zhou L, Xiang X, Chen Y, et al. Enhanced nitrogen removal in modular moving bed constructed wetland at low temperature: Optimization of dissolved oxygen distribution and reconfiguration of core microbial symbiosis [J]. Environmental Research, 2025,276, doi:10.1016/j.envres.2025.121507.

[26] 刘 冰,郑煜铭,秦会安,等.填料对潮汐流人工湿地中 CANON 作用强化的影响 [J]. 环境科学, 2021,42(1):283−292.
Liu Bing, Zhen Y M, Qin H A, et al. Effect of filter medium on the enhancement of complete autotrophic nitrogen removal over nitrite process in a tidal flow constructed wetland [J]. Environmental Science, 2021,42(1):283−292.

[27] Wang X, Shen Z, Zhang Q, et al. Critical role of biochar in the production and emission of greenhouse gas N_2O in constructed wetlands: A comprehensive review [J]. Journal of Cleaner Production, 2025,505,doi:10.1016/j.jclepro.2025.145487.

[28] Ouyang B, Zhang Z, Chen F, et al. Energy production and denitrogenation performance by sludge biochar based constructed wetlands−microbial fuel cells system: Overcoming carbon constraints in water [J]. Water Research, 2025,273, doi:10.1016/j.watres.2024.123024.

[29] 苏建成.地表水中温度、溶解氧、氨氮的关系研究 [J]. 科技创新与应用, 2020,(14):44−55.
Su J C. Research on the relationship between temperature, dissolved oxygen, and ammonia nitrogen in surface water [J]. Technology Innovation and Application, 2020,(14):44−55.

[30] Feng L, He S, Yu H, et al. A novel plant−girdling study in constructed wetland microcosms: Insight into the role of plants in oxygen and greenhouse gas transport [J]. Chemical Engineering Journal, 2022,431, doi:10.1016/j.cej.2021.133911.

[31] Majumder A, Bhatnagar A, Kumar Gupta A. Simultaneous removal of sulfamethoxazole, 17β−estradiol, and carbamazepine from hospital wastewater using a combination of a continuous constructed wetland−based system followed by photocatalytic reactor [J]. Chemical Engineering Journal, 2023,466,doi:10.1016/j.cej.2023.143255.

[32] Gulkowska A, Leung H W, So M K, et al. Removal of antibiotics from wastewater by sewage treatment facilities in Hong Kong and Shenzhen, China [J]. Water Research, 2008,42(1/2):395−403.

[33] Chen H, Ailijiang N, Cui Y, et al. Enhanced removal of PPCPs and antibiotic resistance genes in saline wastewater using a bioelectrochemical−constructed wetland system [J]. Environmental Research, 2024,260, doi:10.1016/j.envres.2024.119794.

[34] Yu K, Hei S, Li P, et al. Removal of intracellular and extracellular antibiotic resistance genes and virulence factor genes using electricity−intensified constructed wetlands [J]. Journal of Hazardous Materials, 2024,475,doi:10.1016/j.jhazmat.2024.134749

[35] Liu Y, Liu X, Wang H, et al. Performance and mechanism of SMX removal in an electrolysis−integrated tidal flow constructed wetland at low temperature [J]. Chemical Engineering Journal, 2022,434, doi:10.1016/j.cej.2022.134494.

作者简介：陈亚松(1982−),男,湖北黄石人,正高级工程师,博士,研究方向为水环境治理技术研究和应用.发表论文 247 余篇.chen_yasong@ctg.com.cn.