DOI：10.19789/j.1004-9398.2026.02.003　　**Open Access**
文献引用：张玉虎，李洁．机器学习在环境科学与工程中的应用进展[J]．首都师范大学学报（自然科学版），2026，47(2)：18-27. ZHANG Y H，LI J. Advances in the applications of machine learning in environmental science and engineering[J]. Journal of Capital Normal University (Natural Science Edition)，2026，47(2)：18-27.

机器学习在环境科学与工程中的应用进展*

张玉虎**，李　洁
（首都师范大学资源环境与旅游学院，北京　100048）

摘要：随着环境科学与工程研究的深入，该领域已逐步迈入大数据时代。机器学习作为人工智能的重要分支，凭借高效的数据处理和信息挖掘能力被广泛应用于环境科学与工程领域。本文系统梳理了机器学习在环境科学与工程中的相关研究和应用进展，涉及空气污染、水污染、土壤污染、固体废弃物以及噪声领域，并归纳总结出环境科学与工程领域常见机器学习算法的优缺点。进一步探讨了当前研究中存在的问题和挑战，并对其未来发展进行了展望。整体来看，神经网络和随机森林凭借其自身特性，是环境科学与工程研究中应用普遍的机器学习算法。而创建数据与模型共享平台、探索集成学习模型及跨学科技术融合等是未来基于机器学习的环境科学与工程研究有效方案。

关键词：环境科学与工程；机器学习；集成学习
中图分类号：X5　　**文献标志码**：A

Advances in the applications of machine learning in environmental science and engineering*

ZHANG Yuhu**，LI Jie
（College of Resources Environment and Tourism，Capital Normal University，Beijing　100048）

Abstract：With the continuous advancement of environmental science and engineering，this field has increasingly entered the era of big data. As a pivotal branch of artificial intelligence，machine learning has been widely employed in environmental science and engineering owing to its superior capabilities in data processing and knowledge discovery. This review systematically examines the recent research progress and practical applications of machine learning in environmental science and engineering，encompassing diverse domains such as air pollution，water pollution，soil pollution，solid waste，noise pollution. The advantages and limitations of commonly used machine learning algorithms are comprehensively summarized. In addition，this paper discusses the existing challenges in current research and outlines prospects for future development. Among various algorithms，neural networks and random forests have gained significant traction due to their unique strengths and adaptability. Looking forward，the establishment of data and model sharing platforms，the development of ensemble learning approaches，and the integration of interdisciplinary technologies are identified as promising strategies to advance machine learning applications in environmental science and engineering.

收稿日期：2025-04-08
*国家自然科学基金重点国际（地区）合作研究项目（42220104004）；国家重点研发计划项目（2023YFC3306400）；国家自然科学基金面上项目（42477440）
**通信作者：zhang_yuhu@163.com

Keywords:environmental science and engineering;machine learning;ensemble learning
CLC:X5 **DC**:A

0 引 言

联合国环境署在2021年发布《与自然和平相处》,强调气候变化、环境污染和生物多样性丧失是全球面临的共同挑战。当前,环境问题日益严峻,影响着全球生态平衡、社会可持续发展和人类健康。随着对环境问题研究的逐步深入,在技术方法和时空尺度上发生变化,使环境数据量大幅增加[1]。数据质量存在不确定性且异质性增加,仅靠传统的分析处理方法不仅效率低而且准确性也难以保证。如何高效处理、分析多样化数据,人工智能为研究环境科学与工程领域复杂问题提供了新的思路[2]。

以人工智能为主导的"第四次工业革命"成为推动科技发展和社会进步的重要力量,在医疗[3]、金融[4]和地质[5]等领域具有广阔应用前景。ChatGPT、DeepSeek及文心一言等大模型在近年来的迅速发展,显著提升了人工智能的影响力,进一步推动了其在各领域的普及与应用。联合国环境署在其发布的《数字生态系统构建的理由》报告中明确,人工智能、大数据等数字技术是实现2030年可持续发展目标的关键。奥地利于2021年推出"AI for Green"创新资助计划,目标在于利用人工智能技术应对气候变化、提高能源效率、实现可持续发展。中国2023年发布的《数字中国建设整体布局规划》中强调,2025年数字中国建设的重要目标是数字生态文明建设取得积极进展。在此背景下,推动人工智能与环境领域的协同融合,使环境科学与工程步入数据驱动、多学科交叉的新时代显得尤其迫切和重要。

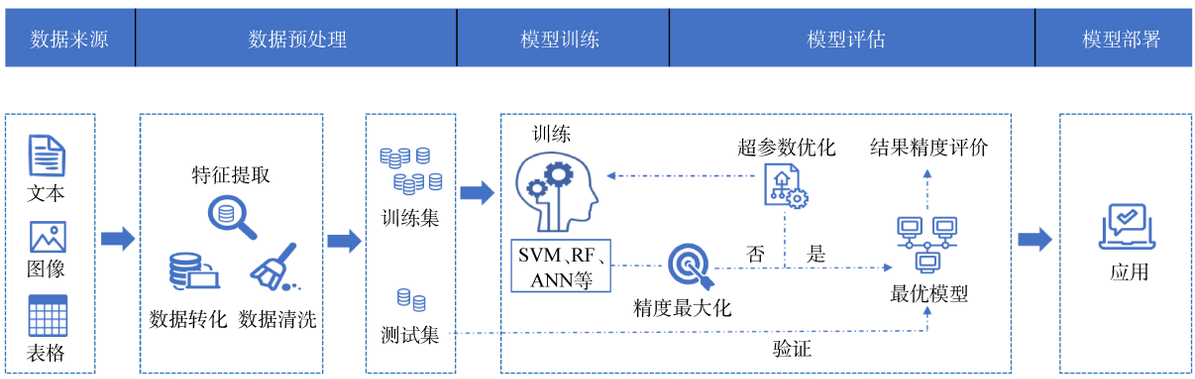
机器学习作为人工智能的重要分支,通过算法和统计学方法使计算机能从数据中自主学习,并做出预测或执行相关的任务,不需要明确的编程指令,能够有效地提高编程效率[6-7]。机器学习技术具有计算效率高、自动特征提取能力优、成本低、适用性广等特点。相较于传统统计方法,在处理高维、非线性数据方面具有独特优势[8]。此外,机器学习因其独特的学习机制和算法特性,所构建的模型稳定性较高、可重复性强,从而高效完成判断与预测[9-10]。近年来,机器学习已经在环境科学与工程的诸多领域中得到广泛的应用,凭借其强大的数据处理与建模能力,加快了污染监测、修复等方面的研究进程。

本文对机器学习在环境科学与工程领域的应用、挑战及未来展望进行综述。首先介绍了机器学习的基本原理方法,然后梳理了机器学习在空气污染、水污染、土壤污染、固体废弃物、噪声污染等方面的研究进展,最后分析了机器学习在环境科学与工程应用中存在的问题及未来展望,旨在推动机器学习在环境领域研究的深入发展。

1 机器学习概述

机器学习是一门融合统计学、概率论、数据挖掘和算法分析的交叉学科[11],其核心是通过算法对部分数据深入分析和学习,从中获取潜在规律。随后对另一部分数据进行预测或判断。机器学习的核心概念包含数据、模型、算法和评估4个部分。(1)机器学习模型的有效性取决于研究数据的品质和数量,数据来源包括表格的结构化数据和文本、图像和音频的非结构化数据[12],分为训练集和测试集2个部分;(2)模型是机器学习的核心,用于海量数据中挖掘出数据特征间存在的潜在规律;(3)算法则通过数据训练模型,调整模型参数来使预测误差最小化[13];(4)评估训练好的最优模型的性能,对输出结果进行精度评价[14]。机器学习的工作流程见图1。

根据数据类型和学习方式的不同,机器学习目前主要分为监督学习、无监督学习、强化学习和半监督学习等[15]。(1)监督学习通过有标签的数据来训练模型,通过学习输入与输出之间的映射关系,预测未出现的数据[16]。监督学习常用算法包括线性回归(linear regression,LR),人工神经网络(artificial neural networks,ANN),支持向量机(support vector machine,SVM),决策树(decision tree,DT)和随机森林(random forest,RF)[17-18]等。(2)无监督学习是通过无标签的数据来训练模型,目的是发现无标签数据中的潜在结构或模式[19]。K均值算法、主成分分析等是典型的无监督学习算法。(3)强化学习通过在试错过程中,基于环境反馈进行调整,实



注:SVM为支持向量机;RF为随机森林;ANN为人工神经网络。

图1 机器学习的工作流程

Figure 1 workflow of machine learning

现累积奖励的最大化[3]。(3)半监督学习是一种结合了有标签数据和无标签数据的学习方式。通过无标签数据来弥补有标签数据的不足,提升了模型的泛化能力[20]。在实际应用中,通过决定系数(R^2),平均绝对误差(mean absolute error,MAE),均方根误差(root mean squared error,RMSE)和平均绝对百分比误差(mean absolute percentage error,MAPE)等指标来评估模型性能,从而根据不同任务需求选取最优模型[1]。

2 机器学习在环境科学与工程应用的主要领域

2.1 机器学习在空气污染中的应用

空气污染是全球环境科学和公共健康领域的核心问题之一,不仅直接威胁人类健康,还对生态系统以及社会经济发展造成深远影响。近年来,随着大数据和传感器网络的发展,机器学习成为解决空气污染问题的重要方法。相比传统的统计方法和物理化学模型,机器学习在处理大规模、高维非线性数据方面具有显著优势,能够有效捕捉污染物浓度与排放源特征、气象条件及人类活动之间的复杂关系[13]。

(1)空气污染物监测。传统监测方法受限于成本和空间分布,难以满足高分辨率、广覆盖的空气质量数据需求。基于机器学习的空气污染物监测手段不断丰富与完善,显著提高了监测的时效性与空间精度。Long等[21]开发了一种基于多任务图机器学习的时空建模框架,显著提升街道级$PM_{2.5}$污染的监测精度,空间分辨率达30~200 m,时间分辨率为1 h;Colléaux等[22]提出了一种基于支持向量回归的低成本传感器数据校正方法,通过量化温度、湿度及气体交叉敏感性对传感器读数的影响,显著提升了CO、O_3和CO_2的监测精度。

(2)空气污染源识别。随着工业化和城市化的快速发展,对及时准确识别空气污染提出了重大挑战,机器学习因其强大的模式识别能力,被广泛应用于污染源识别中。Choi等[23]采用RF、SVM和ANN等算法对韩国5类工业排放的27种空气污染物进行分类以追踪污染源特征,RF模型的分类性能最优,准确率达到96.9%,且氯化氢和乙醛被识别为关键判别污染物;Rutherford等[24]基于卷积神经网络(convolutional neural network,CNN)结合激发-发射矩阵荧光光谱来识别吸入颗粒物,对木材燃烧、香烟烟雾和柴油废气3种实验室污染源产生的$PM_{2.5}$是否存在进行识别,总体准确率为89.0%。

(3)空气污染预测。该方法主要包括基于物理化学机理的模拟驱动方法和基于数学统计原理的数据驱动方法2类。WRF-Chem和CMAQ是常见的模拟驱动方法,通过耦合气象参数、化学反应、排放源特征和输运扩散机制来模拟污染物的生成、迁移过程,具有较强的物理解释性[2]。但其建模较复杂、计算资源消耗大,对数据完整性和准确性依赖较强。机器学习属于数据驱动方法,在挖掘污染物浓度与气象条件、排放源特征等变量间的内在规律方面展现出显著优势。Wang等[13]基于带外部输入的非线性自回归和时间卷积网络(temporal convolution network,TCN)模型对昆明高原地区臭氧浓度进行预测,TCN模型的预测精度更高,MAPE为5.91%,R^2=0.901;张容硕等[25]基于light gradient boosting machine(LightGBM)模型预测郑州市大气中$PM_{2.5}$

和O_3浓度,对$PM_{2.5}$和O_3浓度预测的准确率分别为80.8%和52.5%。

2.2 机器学习在水污染中的应用

随着工业化进程的加快以及农业面源污染的加剧,水体中氮、磷、有机污染物和重金属含量超标问题频发,导致水质恶化、水华现象频发。水质参数种类众多、污染过程复杂以及时空分布差异显著,这给数据采集、污染源识别和污染趋势预测等方面带来巨大挑战。机器学习凭借其强大的非线性建模能力和多变量综合分析优势[26],成为水污染研究的重要方法。

(1)水质监测。水的化学、物理、生物和微生物综合表征被称为水质。传统水质监测主要依赖地面监测站,由于站点分布不均,以及人力、物力限制,难以全面获取大范围内的水质空间分布数据,遥感弥补了这方面的不足[27],融合遥感数据与机器学习的水质监测手段得到广泛应用。吴奇等[28]基于Landsat 8遥感影像和反向传播神经网络,构建了总氮和氨氮的反演模型,揭示了太子河干流2014—2019年总氮和氨氮浓度的时空分布特征;Cao等[29]基于多元线性回归和ANN模型,结合Landsat 8遥感影像数据,构建了滇池水体pH、溶解氧、氨氮和高锰酸盐指数的反演模型,得到2013—2018年滇池水质参数的时空分布规律。

(2)水污染源识别。准确识别污染源是控制和治理污染的有效措施,机器学习通过分析大量水质数据来提取关键特征,有助于提高污染源识别的准确性和效率。Elsayed等[30]采用24种机器学习算法对加拿大安大略省南部农业流域地表水中的营养物质浓度进行分类,集合袋装树是硝酸盐浓度分类的最佳算法,准确率为90.9%,加权KNN是总磷浓度分类的最佳算法,准确率为87.0%。集合子空间判别算法对可溶性活性磷和总溶解磷浓度的总体分类准确率最高,分别为79.2%和77.9%;Liu等[31]采用包括RF、DT和梯度提升决策树(gradient boosting decision tree,GBDT)在内多种机器学习算法,对核主成分与原始水质参数进行回归分析以解析污染源特征,GBDT模型的拟合性能最优,R^2=0.988,MAE为0.05。

(3)水质预测。机器学习在水质预测中展现出高度自动化和高效性,通过多变量综合分析能够精准识别水质变化与不同土地利用之间复杂的非线性关系,具有强大的自适应能力,能够应对不同水体环境和多种污染因素。Ejaz等[17]采用6种机器学习算法对污染水体的水质指数进行预测分析,梯度提升模型的精确度最优,R^2和RMSE分别为0.850、8.670;Wang等[32]提出一种融合长短期记忆网络(long short-term memory,LSTM),门控循环单元与贝叶斯优化算法的混合深度学习模型,对平原河网区域的水质预测,比LSTM模型平均RMSE低18.1%。

(4)估算水污染修复效果。水污染修复采用膜分离、活性炭吸附、高级氧化等技术,不同技术对特定污染物的处理效果及适用条件存在显著差异。利用机器学习模型分析污染物特性和工艺参数,能够精准预测处理效率并优化运行条件,从而实现更高的污染物去除率以及降低化学药剂消耗[33]。Ren等[34]利用极端梯度提升(extreme gradient boosting,XGBoost)模型筛选渗透性反应屏障的反应材料,提高零价铁基反应材料的选择效率和实用性,用于地下水修复;Du等[35]通过构建三维CNN代理模型并结合非支配排序遗传算法,在保证预测精度(R^2≈0.990)的同时,计算效率较传统方法提升约99.8%,能够有效识别成本与修复效率之间的权衡关系,为地下水修复提供了新的思路。

2.3 机器学习在土壤污染中的应用

随着工业化进程的推进与农业生产强度的增加,土壤中重金属和有机污染物的累积问题日益严重,导致土壤环境质量下降。土壤污染具有污染物种类多样、成因复杂和时空分布不均等特点,使得传统监测与评估方法在大范围、高精度应用中面临诸多挑战。机器学习因具备非线性建模、多源数据融合及特征挖掘等优势,成为土壤污染研究的重要方法[15]。

(1)土壤污染物监测。传统实验室分析方法虽具有较高精度,但存在耗时长、成本高、难以实现大范围快速监测等局限。高光谱遥感技术因具备高分辨率和大范围覆盖等优势,结合机器学习的反演模型,成为土壤污染监测的重要手段。Zhang[36]提出了一种结合CNN、注意力机制、图神经网络的高光谱反演模型SpecMet,在低浓度汞的反演中精度提升20.3%;Zhang等[37]提出了一种融合谱段上下文信息的自注意力深度学习网络模型,在保持原始光谱信息的同时,显著提升反演精度,在土壤重金属污染监测展现应用潜力。

(2)土壤污染源识别。土壤污染源识别有同位素示踪法、多元统计法和空间插值法等方法,机器学习能够优化传统方法的局限性,准确高效地识别土壤污染的特征及来源,为有效控制和管理土壤污染提供支持。Zheng等[38]构建了一个结合自组织映射、正矩阵因子分解模型和GBDT算法的土壤重金属污染溯源方法,对中国南方某工业城市272个土壤样本进行分析,识别出工业与交通排放、污水灌溉及自然成因是3类主要污染来源;Zhou和Li[39]提出了一种具有创新性的t分布-随机邻域嵌入-RF-Shapley加法解释模型,对黄河三角洲土壤中潜在有毒元素(potentially toxic elements,PTEs)的分布特征及来源进行研究,钠、有机质和磷是影响PTEs的主要因素。

(3)土壤污染预测。机器学习在土壤污染预测时展现出强大的应用潜力,通过挖掘环境变量与重金属浓度之间的复杂非线性关系,显著提升了预测模型的精度和解释性。Keçeci等[40]构建了DT、LR、RF和XGBoost模型对土壤中镉浓度进行预测,XGBoost模型表现最优(R^2=0.981);Ma等[41]将反距离加权和普通克里金法相结合,构建整合空间区域化指数(spatial regionalization indices,SRIs),并引入RF算法构建RF-SRIs模型,从而显著提升了对铜、锌等8种重金属的预测精度(R^2>0.960),有效解决了传统空间预测方法在分层异质性区域的局限性。

(4)估算土壤污染修复效果。土壤污染修复通常依赖生物炭、纳米材料、聚合物吸附剂等材料,不同材料对污染物的去除效果差异显著。利用机器学习技术筛选修复材料和优化设计,能达到提高修复效果、降低成本投入、缩短修复周期的目的[7]。Du等[42]采用RF模型预测生物炭对土壤中对镉的固定效率,影响土壤中镉的固定效率因素生物炭性质高于实验条件和土壤性质,为确定不同地区土壤镉污染修复的最佳生物炭特性提供科学决策;Zhang等[43]采用SVM、ANN、RF、XGBoost模型预测土壤气相抽提(soil vapor extraction,SVE)对有机污染土壤的修复效率,XGBoost模型的预测精度最高(R^2=0.973),时间、污染物类型和温度是影响SVE修复效率的3个最重要特征。

2.4 机器学习在固体废弃物中的应用

固体废弃物的综合治理与再生利用对实现可持续发展具有重要意义。随着城市化和工业化的快速推进,固体废弃物的种类日益多样且数量持续增长,给环境治理和社会经济发展带来巨大压力。传统固体废弃物管理方式多依赖人工分类、经验判断和简单统计预测,存在效率低、误差大等缺点,难以满足城市精细化及智能化管理的需求。机器学习凭借其强大的数据挖掘和模式识别能力,成为推动固体废弃物管理转型升级的重要驱动力[44]。

(1)固体废弃物监测。固体废弃物的准确监测是实现科学管理和资源化利用的前提。传统的人工巡检方式不仅耗时耗力,还难以在大范围区域内实现高效和全面的覆盖,容易造成数据缺失与监测盲区。随着遥感和机器学习技术的快速发展,基于高光谱影像的自动化监测方法成为固体废弃物分布研究的重要方向。Zeng等[45]利用多尺度CNN对高光谱图像数据的像素进行分类,以提取垃圾区域的位置和大小,为大范围环境中垃圾分布的高效监测与治理提供了科学手段。

(2)固体废弃物分类。目前,擅长图像识别任务的深度学习在固体废弃物分类中应用逐渐增加,为构建智能化废物分类回收系统提供了有力支撑,从而降低成本和处理时间[46]。Abu-Qdais等[10]基于CNN模型结合图像处理技术,开发了固体废弃物智能分类系统,比RF、SVM等传统机器学习模型识别的准确度高;Hossen等[47]基于一种新型的垃圾分类深度神经网络(garbage classifier deep neural network,GCDN-Net)对不同类别的废弃物进行分类,GCDN-Net对垃圾与非垃圾图像区分的准确率为95.8%,精确率为95.8%,召回率为95.8%。

(3)固体废弃物产生量预测。在城市生活垃圾管理领域,精准预测固体废物产生量对于制定科学的管理策略至关重要。近年来,多种机器学习和深度学习模型被应用于这一研究,且取得了显著成效。Lin等[48]构建集成一维CNN、LSTM和Attention的模型,来预测城市固体废物产生量,预测精度达到95.3%;Fatovatikhah等[44]提出了一种LSTM-SVM的混合模型对各类洪水废物产生量进行预测,LSTM-SVM模型的平均RMSE为0.478,平均MAPE为0.257。

2.5 机器学习在噪声中的应用

近年来噪声问题持续引起全社会的广泛关注。交通运输、工业生产以及建筑施工等多重因素叠加,使得噪声污染呈现范围广、持续时间长以及影响群体大的特点,对人类身心健康和城市发展构成

挑战。机器学习对环境参数的自适应学习能力强,能有效提升监管效率[49]。

(1)噪声监管。多数国家已经制定噪声法规,规定不同区域和时段的最大允许噪声水平,但噪声监管费力、耗时且低效,机器学习能提高监管效率,降低监管所需资源。Carrasco等[9]利用深度学习构建噪声识别系统,显著减少人工审核量(约97.0%),提升了城市环境噪声监管效率;Mann等[50]开发了一种随机效应广义线性模型,能够量化解释变量对交通噪声污染的影响,有助于确定资源的优先级和制定噪声控制策略。

(2)噪声估算。机器学习方法凭借其强大的数据处理与模式识别能力,展现出优于传统模型的潜力,能更精准地估算交通及环境噪声水平。Fallah-Shorshani等[51]对比了土地利用回归、XGBoost以及3个声学模型(CadnaA、Harmonoise和FHWA-TNM2.5)在交通噪声评估中的表现,XGBoost模型在次干道与住宅道路上最准确,CadnaA模型在高速公路和辅助道路上表现更好;Liu等[14]采用RF和土地利用回归模型对加拿大5个城市的环境噪声水平进行估算,RF模型的整体性能更优,R^2=0.580,RMSE为4.44 dB。

以上研究表明,RF、SVM和ANN等算法被广泛应用于环境科学与工程领域,在数据处理与分析方面展现出较高准确性和效率,不仅提高了环境信息的处理能力,还有助于优化环境治理决策。在环境科学与工程领域,常见机器学习算法的优缺点总结见表1。

3 机器学习在环境科学与工程中面临的问题和挑战

(1)缺乏高品质数据。数据不仅是机器学习的知识来源,还是进行特征选择的重要依据[52]。数据质量对机器学习预测结果的影响至关重要,如果训练数据存在偏差,模型预测结果通常也会反映出这种偏差。受诸多因素的影响,获取高质量数据存在困难。同时,多源数据异质性显著,增加了数据融合的复杂性。此外,数据共享和使用受限于隐私保护。例如,由于与环境污染暴露相关的病例数据涉及个人隐私[53],进一步加重数据孤岛问题,制约了模型的可靠性与泛化能力。

(2)模型可解释性与可移植性挑战。尽管机器学习算法的数据挖掘能力不断提升,但其在环境领域的应用仍面临模型自身带来的挑战,主要体现在可解释性和可移植性2个方面。一方面,机器学习普遍存在“黑箱”问题,即可解释性不足。具体而言,模型中间过程或内部机制难以被理解,使得对预测结果的信任度和接受度降低[54]。一般情况下,模型准确度的提升往往伴随可解释性的下降[55]。因此,如何平衡模型的准确性与可解释性仍是一个核心挑战。另一方面,环境问题存在的强地域依赖性限制了模型的可移植性。在特定区域训练得到的特征-响应关系,会因地理环境差异而失效,使得预测结果缺乏通用性[56]。例如,某些模型在温带平原河流富营养化治理中表现良好,但在高寒地区河流中可能效果不佳,因为这些河流的水温常年偏

表1 环境科学与工程领域常见机器学习算法的特点

Table 1 characteristics of common machine learning algorithms in environmental science and engineering

机器学习算法	实现原理	优点	缺点
线性回归	基于最小二乘法,假设输入特征与输出变量存在线性关系,适合处理线性可分的数据[40]	计算简单,训练速度快,可解释性强	不能拟合非线性数据,抗噪能力弱
神经网络	受生物神经元结构启发,通过多层非线性变换逼近复杂函数,适用于高维、复杂数据特征提取[4,23]	具有大规模处理信息、分布式储存、自组织、自学习及自适应能力,具有良好的容错性及可扩展性[36]	需要大量有代表性数据,训练时间长,模型解释性差,易过拟合
RF	通过集成多棵决策树来提到模型的预测精度和泛化能力,可处理高维数据[18]	易于实现、抗过拟合、性能优良,对缺失值不敏感,具有很强的鲁棒性	结构复杂,计算成本较高,可解释性较差,对小样本噪声敏感[41]
SVM	基于结构风险最小化,通过核函数将数据映射到高维空间实现非线性分类,适合小样本、高维数据[17,43]	核函数灵活,算法简单,泛化能力强	对参数和核函数选择敏感,大规模数据训练慢,对缺失值敏感
最近邻算法	基于距离度量(如欧氏距离)的惰性学习方法,通过假设相似输入具有相似输出,适用于低密度、非线性数据[23]	实现简单,适合局部模式识别	计算复杂度高,不适合大数据集,当数据不均衡时,预测偏差较大
梯度提升树	通过迭代构建多个弱学习器,每一步优化前一步的残差,适合处理特征交互和非线性关系[38]	预测精度高,泛化能力强,自动特征处理,灵活性高,有助于特征重要性评估	训练速度较慢,参数调节复杂,不适合高维稀疏数据,占用内存大

低、光照强度高、微生物活性受低温抑制,与温带地区季节性藻类爆发的影响因素有所不同。

4 展 望

4.1 创建数据与模型共享平台

由于环境数据具有多源异构、时空尺度差异大等特点,构建的标准化共享数据平台成为解决问题的关键突破口[57]。平台允许科研人员访问和利用现有数据,避免因重复收集和处理数据造成的资源和时间上的浪费,大量节省模型训练时间和提高数据品质。此外,模型共享平台还有利于模型在跨区域、多尺度场景中的快速复现与应用迭代,进而研究出泛化能力更强的机器学习模型,提高研究的可靠性和透明度。例如,谷歌旗下的人工智能公司DeepMind与欧洲生物信息研究所合作免费公开超过2亿个蛋白质结构预测结果,帮助科研人员解析未知功能的蛋白质结构,促进了医疗、生物等相关领域的发展。

4.2 探索集成学习模型

面对环境领域的复杂性和不确定性问题,集成学习模型通常比单一机器学习模型的准确性和鲁棒性更强[58]。集成学习通过结合多个学习器,克服单一模型的局限性,降低过拟合风险,提高了模型的泛化能力[8]。集成学习模型凭借其在大规模数据处理、非线性问题解析以及多变量优化任务中的卓越性能,已成为解决复杂环境问题的新方法。针对不同的环境问题,没有通用的机器学习模型,集成学习能依据不同场景需求,有针对性地选择互补性算法和模型,显著提高模型的预测精度。

4.3 跨学科技术融合

机器学习与传统技术相结合,能够充分利用二者之间的优势。将物理先验知识作为约束条件与数据驱动相结合的模型,能够有效提升机器学习模型的泛化能力和物理一致性。物理信息神经网络(physics-informed neural networks,PINNs)是一种结合物理规律的深度学习模型,通过将物理方程嵌入神经网络的训练过程,使得模型既能拟合实测数据,又能严格遵循已知的物理规律[59]。例如,PINNs在解决水泥水化过程中的碳足迹问题时,用于分析添加补充胶凝材料,不仅减少了水泥水化过程中的碳排放,还解决了传统机器学习对于复杂物理-化学过程理解的局限性[60]。

机器学习凭借其卓越特性在空气污染、水污染、土壤污染、固体废弃物、噪声污染等方面取得一系列进展,突破传统研究方法的技术壁垒,极大地推动了环境领域相关研究的发展。目前机器学习仍存在一些问题和挑战,如何创建数据与模型共享平台、探索集成学习方法及跨学科技术融合是今后研究的一些重点。本文旨在为机器学习在环境科学与工程中的应用提供一个梳理和例子,也为相关环境问题的治理提供一些借鉴和启发。

©The editorial department of Journal of Capital Normal University (Natural Science Edition) 2025. This is an open access article under the CC BY-NC-ND 4.0 license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

参 考 文 献

[1] LUKAS P, MELESSE A M, KENEA T T. Predicting reservoir sedimentation using multilayer perceptron - Artificial neural network model with measured and forecasted hydrometeorological data in Gibe-III reservoir, Omo-Gibe River basin, Ethiopia [J]. Journal of Environmental Management, 2024, 359: 121018. DOI:10.1016/j.jenvman.2024.121018.

[2] MA X X, LIU H N, PENG Z. Improving WRF-Chem PM2.5 predictions by combining data assimilation and deep-learning-based bias correction [J]. Environment International, 2025, 195: 109199. DOI: 10.1016/j.envint.2024.109199.

[3] HUANG K, LIDBURY B A, THOMAS N, et al. Machine learning and multi-omics in precision medicine for ME/CFS [J]. Journal of Translational Medicine, 2025, 23(1): 68. DOI:10.1186/s12967-024-05915-z.

[4] ZHOU G L, WANG S R. Enhancing credit risk decision-making in supply chain finance with interpretable machine learning model [J]. IEEE Access, 2025, 13: 14239-14251. DOI:10.1109/ACCESS.2025.3530433.

[5] LI C L, LI F D, LIU C, et al. Deep learning-based geological map generation using geological routes [J]. Remote Sensing of Environment, 2024, 309: 114214. DOI:10.1016/j.rse.2024.114214.

[6] ZHU M Y, WANG J W, YANG X, et al. A review of the application of machine learning in water quality evaluation [J]. Eco-Environment & Health, 2022, 1(2): 107-116. DOI:10.1016/j.eehl.2022.06.001.

[7] YIN R X, LI X, NING Y T, et al. Machine learning unveils the role of biochar application in enhancing tea yield by mitigating soil acidification in tea plantations

[J]. Science of the total environment, 2025, 965: 178597-178597. DOI:10.1016/j.scitotenv.2025.178597.

[8] LIN S B, KIM J W, HUA C B, et al. Comparing artificial and deep neural network models for prediction of coagulant amount and settled water turbidity: lessons learned from big data in water treatment operations [J]. Journal of Water Process Engineering, 2023, 54: 103949. DOI:10.1016/j.jwpe.2023.103949.

[9] CARRASCO V, ARENAS J P, HUIJSE P, et al. Application of deep learning to enforce environmental noise regulation in an urban setting[J]. Sustainability, 2023, 15(4): 3528. DOI:10.3390/su15043528.

[10] ABU-QDAIS H, SHATNAWI N, AL-ALAMIE E. Intelligent system for solid waste classification using combination of image processing and machine learning models [J]. Journal of Experimental & Theoretical Artificial Intelligence, 2024, 37(6): 1063-1074. DOI: 10.1080/0952813X.2024.2323043.

[11] VAPNIK V N. An overview of statistical learning theory [J]. IEEE Transactions on Neural Networks, 1999, 10(5): 988-999. DOI:10.1109/72.788640.

[12] KHANDELWAL K, NANDA S, DALAI A K. Machine learning to predict the production of bio-oil, biogas, and biochar by pyrolysis of biomass: a review [J]. Environmental Chemistry Letters, 2024, 22(6): 2669-2698. DOI:10.1007/s10311-024-01767-7.

[13] WANG Q Y, LIU H Y, LI Y J, et al. Predicting plateau atmospheric ozone concentrations by a machine learning approach: a case study of a typical city on the southwestern plateau of China [J]. Environmental Pollution, 2024, 363(1): 125071. DOI: 10.1016/j.envpol.2024.125071.

[14] LIU Y, GOUDREAU S, OIAMO T, et al. Comparison of land use regression and random forests models on estimating noise levels in five Canadian cities [J]. Environmental Pollution, 2020, 256: 113367. DOI: 10.1016/j.envpol.2019.113367.

[15] 傅嘉辉,陈云,王丽娜,等.机器学习在土壤污染识别中的应用研究进展[J].湖南师范大学自然科学学报, 2024, 47(5): 86-94. DOI: 10.7612/j. issn. 2096-5281.2024.05.010.
FU J H, CHEN Y, WANG L N, et al. Research progress in the application of machine learning in soil contamination identification [J]. Journal of Natural Science of Hunan Normal University, 2024, 47(5): 86-94.

[16] FERREIRA B, SILVA R G, ITEN M. Earth observation satellite imagery information based decision support using machine learning [J]. Remote Sensing, 2022, 14(15): 3776. DOI:10.3390/rs14153776.

[17] EJAZ U, KHAN S M, JEHANGIR S, et al. Monitoring the Industrial waste polluted stream - integrated analytics and machine learning for water quality index assessment [J]. Journal of Cleaner Production, 2024, 450: 141877. DOI:10.1016/j.jclepro.2024.141877.

[18] WEI P, HAO S, SHI Y, et al. Combining google traffic map with deep learning model to predict street-level traffic-related air pollutants in a complex urban environment [J]. Environment International, 2024, 191: 108992. DOI:10.1016/j.envint.2024.108992.

[19] MIZIKOVSKY D, NAVAL SANCHEZ M, NEFZGER C M, et al. Organization of gene programs revealed by unsupervised analysis of diverse gene-trait associations [J]. Nucleic Acids Research, 2022, 50(15): e87. DOI: 10.1093/nar/gkac413.

[20] LATIF S D, AHMED A N. A review of deep learning and machine learning techniques for hydrological inflow forecasting [J]. Environment, Development and Sustainability, 2023, 25(11): 12189-12216. DOI: 10.1007/s10668-023-03131-1.

[21] LONG Q, MA J, GUO C, et al. High-resolution spatio-temporal estimation of street-level air pollution using mobile monitoring and machine learning [J]. Journal of Environmental Management, 2025, 377: 124642. DOI: 10.1016/j.jenvman.2025.124642.

[22] COLLÉAUX Y, WILLAUME C, MOHANDES B, et al. Air pollution monitoring using cost-effective devices enhanced by machine learning [J]. Sensors, 2025, 25(5): 1423. DOI:10.3390/s25051423.

[23] CHOI Y, KANG B, KIM D. Utilizing machine learning-based classification models for tracking air pollution sources: a case study in Korea [J]. Aerosol and Air Quality Research, 2024, 24(7): 230222. DOI: 10.4209/aaqr.230222.

[24] RUTHERFORD J W, DAWSON-ELLI N, MANICONE A M, et al. Excitation emission matrix fluorescence spectroscopy for combustion generated particulate matter source identification [J]. Atmospheric Environment, 2020, 220: 117065. DOI: 10.1016/j.atmosenv.2019.117065.

[25] 张容硕,谢沛远,陈宏飞,等.基于机器学习的郑州市大气$PM_{2.5}$与O_3浓度预测方法及气象因子的影响分析[J].环境科学研究, 2024, 37(3): 469-478. DOI: 10.13198/j.issn.1001-6929.2023.12.03.
ZHANG R S, XIE P Y, CHEN H F, et al. Prediction method and meteorological factors impact analysis of atmospheric $PM_{2.5}$ and O_3 concentration in Zhengzhou

City based on machine learning [J]. Research of Environmental Sciences, 2024, 37(3): 469-478.

[26] JING Z M, ZHANG Y, LIU X L, et al. Identifying human activities causing water pollution based on microbial community sequencing and source classifier machine learning [J]. Environment International, 2025, 195: 109240. DOI: 10.1016/j.envint.2024.109240.

[27] 梁文广,陈伟,王金东,等.利用Sentinel-2卫星影像反演骆马湖叶绿素A浓度[J].测绘通报,2024(10):1-6. DOI:10.13474/j.cnki.11-2246.2024.1001.
LIANG W G, CHEN W, WANG J D, et al. Inversion of chlorophyll-A concentration in Luoma lake using Sentinel-2 satellite imagery [J]. Bulletin of Surveying and Mapping, 2024(10): 1-6.

[28] 吴奇,宫福征,白伟桦,等.太子河干流总氮与氨氮水质参数反演及时空变化研究[J].生态与农村环境学报, 2024, 40(8): 1017-1028. DOI: 10.19741/j. issn. 1673-4831.2023.0366.
WU Q, GONG F Z, BAI W H, et al. Study on the inversion and spatio-temporal variations of total nitrogen and ammonia nitrogen water quality parameters in the mainstream of Taizi River [J]. Journal of Ecology and Rural Environment, 2024, 40(8): 1017-1028.

[29] CAO J J, WEN X P, LUO D Y, et al. Study on water quality inversion model of dianchi lake based on landsat 8 data[J]. Journal of Spectroscopy, 2022: 3341713. DOI: 10.1155/2022/3341713.

[30] ELSAYED A, RIXON S, LEVISON J, et al. Application of classification machine learning algorithms for characterizing nutrient transport in a clay plain agricultural watershed [J]. Journal of Environmental Management, 2023, 345: 118924. DOI: 10.1016/j.jenvman.2023.118924.

[31] LIU Y, ZHENG N R, YANG S H, et al. Identification of water pollution sources in the Daluxi River using kernel principal component analysis and gradient boosting decision tree [J]. Environmental Earth Sciences, 2025, 84(9): 236. DOI: 10.1007/s12665-025-12241-0.

[32] WANG K F, LIU L, BEN X C, et al. Hybrid deep learning based prediction for water quality of plain watershed [J]. Environmental Research, 2024, 262: 119911. DOI: 10.1016/j.envres.2024.119911.

[33] LYU H F, XU Z M, ZHONG J, et al. Machine learning-driven prediction of phosphorus adsorption capacity of biochar: insights for adsorbent design and process optimization [J]. Journal of Environmental Management, 2024, 369: 122405. DOI: 10.1016/j.jenvman.2024.122405.

[34] REN Y M, CUI M C, ZHOU Y Y, et al. Zero-valent iron based materials selection for permeable reactive barrier using machine learning [J]. Journal of Hazardous Materials, 2023, 453: 131349. DOI: 10.1016/j.jhazmat.2023.131349.

[35] DU J W, SHI X Q, MO S X, et al. Deep learning based optimization under uncertainty for surfactant-enhanced DNAPL remediation in highly heterogeneous aquifers [J]. Journal of Hydrology, 2022, 608: 127639. DOI: 10.1016/j.jhydrol.2022.127639.

[36] ZHANG L. A hyperspectral metal concentration inversion method using attention mechanism and graph neural network [J]. Ecological Informatics, 2024, 83: 102792. DOI: 10.1016/j.ecoinf.2024.102792.

[37] ZHANG T Y, FU Q, TIAN R Q, et al. A spectrum contextual self-attention deep learning network for hyperspectral inversion of soil metals [J]. Ecological Indicators, 2023, 152: 110351. DOI: 10.1016/j.ecolind.2023.110351.

[38] ZHENG J T, WANG P, SHI H Y, et al. Quantitative source apportionment and driver identification of soil heavy metals using advanced machine learning techniques [J]. Science of the Total Environment, 2023, 873: 162371. DOI: 10.1016/j.scitotenv.2023.162371.

[39] ZHOU M G, LI Y H. Spatial distribution and source identification of potentially toxic elements in Yellow River delta soils, China: an interpretable machine-learning approach [J]. Science of The Total Environment, 2024, 912: 169092. DOI: 10.1016/j.scitotenv.2023.169092.

[40] KEÇECI M, GÖKMEN F, USUL M, et al. Prediction of cadmium content using machine learning methods [J]. Environmental Earth Sciences, 2024, 83(12): 362. DOI: 10.1007/s12665-024-11672-5.

[41] MA X D, GUAN D X, ZHANG C S, et al. Improved mapping of heavy metals in agricultural soils using machine learning augmented with spatial regionalization indices [J]. Journal of Hazardous Materials, 2024, 478: 135407. DOI: 10.1016/j.jhazmat.2024.135407.

[42] DU Z L, SUN X, ZHENG S N, et al. Optimal biochar selection for cadmium pollution remediation in Chinese agricultural soils via optimized machine learning [J]. Journal of Hazardous Materials, 2024, 476: 135065. DOI: 10.1016/j.jhazmat.2024.135065.

[43] ZHANG S, ZHAO J T, ZHU L Z. Predicting removal efficiency of organic pollutants by soil vapor extraction based on an optimized machine learning method [J]. Science of the Total Environment, 2024, 927: 172438.

DOI:10.1016/j.scitotenv.2024.172438.

[44] FATOVATIKHAH F, AHMEDY I, NOOR R M. Waste prediction approach using hybrid long short-term memory with support vector machine[J]. International Journal of Computational Intelligence Systems, 2024, 17(1):103. DOI:10.1007/s44196-024-00485-w.

[45] ZENG D, ZHANG S, CHEN F S, et al. Multi-scale CNN based garbage detection of airborne hyperspectral data[J]. IEEE Access, 2019, 7: 104514-104527. DOI: 10.1109/ACCESS.2019.2932117.

[46] XIA W J, JIANG Y P, CHEN X H, et al. Application of machine learning algorithms in municipal solid waste management: A mini review [J]. Waste Management & Research, 2022, 40 (6) : 609-624. DOI: 10.1177/0734242X211033716.

[47] HOSSEN M M, ASHRAF A, HASAN M, et al. GCDN-Net: garbage classifier deep neural network for recyclable urban waste management [J]. Waste Management, 2024, 174: 439-450. DOI: 10.1016/j.wasman.2023.12.014.

[48] LIN K S, ZHAO Y C, TIAN L, et al. Estimation of municipal solid waste amount based on one-dimension convolutional neural network and long short-term memory with attention mechanism model: a case study of Shanghai [J]. Science of the Total Environment, 2021, 791: 148088. DOI: 10.1016/j. scitotenv. 2021. 148088.

[49] GIWA S O, NWAOKOCHA C N, OSIFEKO O M, et al. Appraising machine learning algorithms in predicting noise level and emissions from gasoline-powered household backup generators [J]. International Journal of Environmental Science and Technology, 2024, 22(5): 3071-3088. DOI:10.1007/s13762-024-05987-w.

[50] MANN S, SINGH G. Random effect generalized linear model-based predictive modelling of traffic noise [J]. Environmental Monitoring and Assessment, 2024, 196(2):168. DOI:10.1007/s10661-023-12285-4.

[51] FALLAH-SHORSHANI M, YIN X Z, MCCONNELL R, et al. Estimating traffic noise over a large urban area: An evaluation of methods [J]. Environment International, 2022, 170: 107583. DOI: 10.1016/j.envint. 2022.107583.

[52] 程浩淼,杨富康,张健,等.基于机器学习的河湖水质评估的研究进展[J].环境科学,2025,46(6):3349-3360. DOI:10. 13227/j.hjkx.202406089.

CHENG H M, YANG F K, ZHANG J, et al. Research progress on river and lake water quality assessment based on machine learning [J]. Environmental Science, 2025, 46(6): 3349-3360.

[53] LOVRIC M, GAJSKI G, FERNáNDEZ-AGÜERA J, et al. Evidence-driven indoor air quality improvement: an innovative and interdisciplinary approach to improving indoor air quality [J]. BioFactors, 2025, 51(1): e2126. DOI:10.1002/biof.2126.

[54] HU B B, DAI Y D, ZHOU H, et al. Using artificial intelligence to rapidly identify microplastics pollution and predict microplastics environmental behaviors [J]. Journal of Hazardous Materials, 2024, 474: 134865. DOI:10.1016/j.jhazmat.2024.134865.

[55] FUNG P L, SAVADKOOHI M, ZAIDAN M A, et al. Constructing transferable and interpretable machine learning models for black carbon concentrations [J]. Environment International, 2024, 184: 108449. DOI: 10.1016/j.envint.2024.108449.

[56] LOTFY A A, ABUARAB M E, FARAG E, et al. Forecasting blue and green water footprint of wheat based on single, hybrid, and stacking ensemble machine learning algorithms under diverse agro-climatic conditions in Nile Delta, Egypt [J]. Remote Sensing, 2024, 16(22): 4224. DOI:10.3390/rs16224224.

[57] HE Y Q, CHEN M, WEN Y N, et al. An open online simulation strategy for hydrological ensemble forecasting [J]. Environment Modelling & Software, 2024, 174: 105975. DOI:10.1016/j.envsoft.2024.105975.

[58] ZHAO Y B, YU T, HU B L, et al. Retrieval of water quality parameters based on near-surface remote sensing and machine learning algorithm [J]. Remote Sensing, 2022, 14(21): 5305. DOI:10.3390/rs14215305.

[59] WESSELKAMP M, MOSER N, KALWEIT M, et al. Process-informed neural networks: a hybrid modelling approach to improve predictive performance and inference of neural networks in ecology and beyond [J]. Ecology Letters, 2024, 27(11): e70012. DOI: 10.1111/ele.70012.

[60] RAHMAN M A, LU Y. EcoBlendNet: a physics-informed neural network for optimizing supplementary material replacement to reduce the carbon footprint during cement hydration [J]. Journal of Cleaner Production, 2024, 464: 142777. DOI: 10.1016/j.jclepro.2024.142777.

(责任编辑:李拓宇)