环境工程学报
CHINESE JOURNAL OF ENVIRONMENTAL ENGINEERING
第 18 卷 第 11 期 2024 年 11 月
Vol. 18, No.11 Nov. 2024
http://www.cjee.ac.cn E-mail: cjee@rcees.ac.cn (010) 62941074
DOI 10.12030/j.cjee.202404012 中图分类号 X524 文献标识码 A

机器学习在地表水水质管理中的应用

王光滔, 赵雯, 江宇静, 刘娟, 朱文磊, 李梅✉
南京大学环境学院, 污染控制与资源化研究国家重点实验室, 地球关键物质循环前沿科学中心, 南京 210023

摘 要 机器学习, 作为人工智能的一个关键子领域, 已在环境领域中发挥着越来越重要的作用, 尤其在处理地表水水质管理中复杂问题时, 机器学习显示出相对传统方法的显著优势。本综述重点探讨了多种机器学习算法在地表水水质管理方面的应用, 分析了溶解氧、生物需氧量、化学需氧量、浊度、温度、pH 等不同水质参数对地表水水质分类、监测及预测结果的影响, 并列举了几种在实际工程应用中最常见的机器学习模型, 如人工神经网络、支持向量机、随机森林及决策树等, 最终归纳总结了几种用于提升输出精度的混合模型在地表水水质管理中的实际应用。综上所述, 实现机器学习对地表水水质准确、高效管理, 不仅取决于选用的水质参数是否能够作为特定算法的数据集, 还依赖于合理使用多种机器学习模型进而增加输出结果的可信度。

关键词 机器学习; 环境工程; 水质管理; 地表水

在过去几十年中, 地表水水质 (water quality, WT) 因各种污染物和废弃物排放而受到了显著影响[1-9]。水质恶化不仅危害了人类健康, 还破坏了水生生物多样性和整个地表水生态系统平衡[10-11]。此外, 气候变化导致的低流量季节水质下降和河流水年平均温度升高, 进一步加剧了地表水体的环境压力[12-14]。为了实现水资源的可持续管理, 识别及应对地表水污染的主要来源和压力, 亟需在国际国内法规和政策及各类框架指令内进行科学的水质管理[15-16]。在众多法规政策文件中, 限制水中特定污染物排放的规定和针对地表水中关键有害物质的界定标准, 二者共同确定了水质评估需要遵循的关键参数上限值[17-18]。水质数据收集通常采用包括现场采样与分析、实验室测试及实时监测传感等多种技术[19-22]。虽然, 高性能传感器在监测水质方面发挥着关键作用, 但较高成本且需定期维护和校准以确保数据的准确性这两大弊端限制了其大规模使用[23-25]。随着技术进步, 关于优化水质管理方法的研究在不断发展。目前在河流水质管理中, 机器学习 (machine learning, ML) 的应用领域主要包括: (a) 水质分类[26-29]; (b) 水质异常检测[30-33]; (c) 水质预测[34-37]。本综述旨在分析上述领域当前面临的挑战, 并深入研究 ML 该如何应对, 进而辅助开发出有效的 ML 框架作为河流/地表水管理决策过程的工具, 最终实现大幅改善河流生态健康并保护生态环境的目标。

水质指数 (water quality index, WQI) 和水质分类标准 (water quality criteria, WQC) 是一类综合反映水质特征的数值。它们由多个参数组成, 每个参数都包含一个 0 到 100 之间的数值 (q_i) 和一个权重 (w_i)。这些数值和权重由专家确定[23, 37], w_i 体现了各参数在水质评估中的重要性[38-40]。水质评估中, 最常用的监测参数包括溶解氧 (dissolved oxygen, DO)、化学需氧量 (chemical oxygen demand, COD)、生化需氧量 (biochemical oxygen demand, BOD)、总溶解固体 (total dissolved solids, TDS)、硝酸盐浓度 (NO_3^-) 和 pH 值。这些参数反映了水体的化学组成和污染水平。物理参数如水温 (water temperature, WT)、浊度、电导率 (electrical conductivity, EC) 以及总固体含量 (total solid, TS), 揭示了水体物理状况和可能的污染源[41-42]。KHULLAR 和 SINGH 以及 SYEED 等[43-44] 给出了关于这些指标的具体定义, 并说明了这些参数是如何影响整体地表水水质的, 对环境监测和管理实践有着重要意义。通过对这些参数的精确测量和监控, 可以更好地理解水体健康状况, 并采取适当措施以保护和改善水质。

其中 WQI 是一个用来综合表示水体状况的无量纲数值[45-46]。WQI 计算公式见式 (1)。

收稿日期: 2024-04-12 **录用日期**: 2024-08-29
基金项目: 江苏省科技计划专项资助项目 (BZ2022006); 国家自然科学基金资助项目 (22176086); 江苏省基础研究计划资助项目 (BK20210189); 江苏省碳达峰碳中和科技创新专项资助项目 (BE202261)
第一作者: 王光滔 (1997—), 男, 硕士研究生, 602023250044@smail.nju.edu.cn ✉**通信作者**: 李梅 (1972—), 女, 博士, 教授, meili@nju.edu.cn

$$WQI = \frac{\sum q_i x w_i}{\sum w_i} \quad (1)$$

这些参数值和相应权重通过 WQI 的计算公式结合起来，以生成一个单一的综合数值，该数值代表了整体水质状况。基于 WQI 结果，可为各个水体建立水质分类标准 (water quality criteria，WQC) 。这些分类提供了为水质快速定级的参考，如 AHMED 等[37] 提供了一个典型分类表 (表 1) 。这些分类代表着不同的水质状况，进而为水资源管理和保护措施提供指导。

表 1 水质分类标准 (WQC) [23]

Table 1 Water quality classification (WQC)[23]

WQI数值	分类
0~25	非常差
25~50	差
50~70	中
70~90	良
90~100	优

1 水质管理中的常见机器学习算法

ML 技术因其高精度、可定制性和处理复杂数据模式的能力，已在环境科学和环境工程学领域得到广泛应用[47-49]。ZHONG 等[50] 分析发现，1990-2020 年间，水 (47.63%) 、空气 (27.32%) 、土壤 (21.02%) 和沉积物 (4.02%) 4 个环境工程代表性领域共发表了 5 855 篇关于 ML 应用技术的研究论文。

ML 在环境科学和环境工程学领域的应用主要集中在 4 个方面：(1) 进行预测；(2) 特征识别；(3) 检测异常；(4) 数据建模[50-54]。ML 应用于 (1) 和 (3) 主要涉及监督学习 (supervised learning，SupVL) [55-56]，如回归和分类，但在一定条件下也可通过非监督学习 (unsupervised learning，UnSupVL) 实现[57-58]。ML 应用于 (2) 和 (4) 通常通过 SupVL 实现，例如应用于 (2) 时采用的线性判别分析 (linear discriminant analysis，LDA)。SupVL 主要用于解决环境问题，例如颗粒物 ($PM_{2.5}$) 预测、水资源可用性评估和废水生化处理系统建模。ZHONG 等[50] 通过 5 个代表性案例阐述了 ML 如何解决复杂的环境问题，并概述了四个主要应用领域，包括预测、特征重要性提取、异常检测和新材料或化学品发现，同时讨论了 ML 在环境科学与工程中应用的挑战和未来机遇。SYEED 等[44] 指出，分析地表水水质对发展中国家尤为重要，过去几十年中多种 ML 模型已被开发以解决水资源管理中的各种挑战。2000—2020 年间，有关 ML 应用于河流研究领域的论文数量从 310 篇增加到 3 444 篇，说明 ML 在该研究领域的重要性显著增加。21 世纪之前，涉及 SupVL 的应用占主导地位，但进入 21 世纪后，涉及 SupVL 和 UnSupVL 的应用数量趋于平衡。

与此同时，神经网络 (neural networks，NN) 和深度学习 (deep learning，DL) 在该领域获得了更多关注，过去二十年中，这两部分的论文发表量占 ML 总论文发表量 15%~21%[59]。根据 ZHU 等[50] 的研究，应用于地表水质评估的 ML 算法包括自举小波神经网络 (bootstrapped wavelet neural network，BWNN) 、人工神经网络 (artificial neural network，ANN) 、自回归综合移动平均 (autoregressive integrated moving average，ARIMA) 、自举人工神经网络 (bootstrapped artificial neural network，BANN) 、长短期记忆网络 (long short-term memory，LSTM) 、nash-sutcliffe 效率 (nash-sutcliffe efficiency，NSE) 、多项式神经网络 (polynomial neural network，PNN) 、级联相关神经网络 (cascade correlation neural network，CCNN) 、tsinghua/temporary deepspeed (TDS) 、深度神经网络 (deep neural network，DNN) 、支持向量回归 (support vector regression，SVR) 、随机森林 (random forest，RF) 、支持向量机 (support vector machine，SVM) 和卷积神经网络 (convolutional neural network，CNN) 。

在地表水水质管理方面，常用的 ML 模型为用于水质分类、水质预测以及异常检测的树状结构算法。这其中包括决策树 (decision tree，DT) 和 RF、SVM 和 ANN，以及 LSTM，后者也是 ANN 算法中的一种。

机器学习算法是现代计算科学的一个核心分支，其目标是创建能够从数据中自主学习的模型，进而使计算机能够做出智能决策。这些算法的核心在于它们的自适应能力，即在接收到新数据时能够自动调整其行为，而不需要人类开发者进行直接的程序修改[60-61]。由于机器学习中算法种类繁多，在此仅对常见的三种算法进行展开阐述。

1.1 树形算法

DT 是一种 SupVL 算法，在 ML 中通常用于分类和回归。它通过创建树状结构来模拟决策过程[62]。在决策树中，每个决策节点 (或内部节点) 代表一个属性上的测试，每个分支代表测试结果，而每个叶节点 (或终端节点) 代表一个类标或结果。该算法通过分裂数据集的方式，逐步简化问题[26]。该算法由决策节点和叶

节点组成，其结构如图 1 所示。

DT 算法在水质管理中的应用具有数据分类简便的优点，这对于需要多个部门间协调合作的水质管理非常有利。水质数据可能包括数值型、分类型以及时间序列数据。决策树能够高效快速地处理这些不同类型的数据，并有效地捕捉特征间的交互关系，例如，温度和降雨量如何共同影响水体化学需氧量 (COD) 。其次，水质数据往往包含噪声或不完整数据点。决策树不需要复杂的数据预处理，如缩放或归一化，这简化了数据分析流程。然而，在水质管理中，决策树可能会创建过于复杂的模型，以适应训练数据中的特定样本而不是泛化到新的或未见过的数据上，从而造成过拟合的情况，这可能导致对实际应用的预测性能不佳[64]。

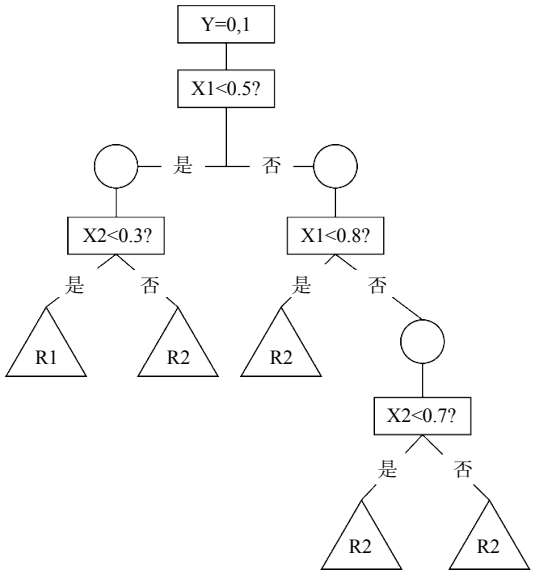


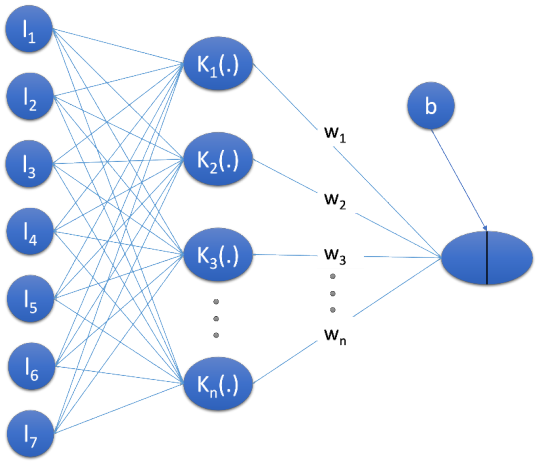
图 1　单型 DT 和 RF 示意图[63]

Fig. 1　Schematic diagram of single-type DT and RF[63]

1.2　支持向量机算法

虽然在过去的几年中 ANN 模型在众多预测任务中被广泛应用，已被证实适用于多种预测场景，但 SVM 模型在许多研究中证实可以在类似任务中提供更高的结果可靠性[50, 65-68]。

SVM 是一种 SupVL 算法，主要用于分类和回归分析。SVM 通过寻找一个最优的超平面来区分不同类别的数据点，使得各类别数据点间的边界最大化。SVM 模型的一般架构，如图 2 所示，通常包括数据输入、特征转换、最优超平面的确定，及最终的分类决策。这个架构体现了 SVM 处理数据的基本步骤，并展示了其如何有效地将数据集划分为不同类别[69]。SVM 擅长分类任务，能够识别和分类不同污染源，具有较强的泛化能力，即使在训练数据较少的情况下，也能准确预测未知数据。此外，SVM 通过使用不同的核函数处理非线性数据，并在处理异常值或噪声时表现出鲁棒性，适用于复杂水质数据模式识别和分析。相比 DT 需要大量数据训练的算法，SVM 在有限数据集上也能取得良好性能，适用于数据资源有限的水质监测项目。



注：K(.)是核函数，而n代表支持向量的个数。

图 2　SVM 架构[70]

Fig. 2　SVM architecture[70]

然而，SVM 水质管理中的应用也面临挑战：首先，选择合适的核函数可能复杂且影响模型性能，需要根据具体的水质数据特点进行调整；其次，SVM 在处理大型数据集时的训练效率较低，这可能限制其在大规模实时水质监测数据分析中的应用；此外，SVM 对数据缺失高度敏感，水质数据常见的缺失或异常值可能会影响模型稳定性和准确性；最后，SVM 模型的参数调节和解释性较为复杂，这可能需要更多的专业知识来确保模型的适当配置和结果解释。

1.3　人工神经网络

ANN 是一种模仿人脑处理信息方式的计算系统[71-73]。它由相互连接的节点或“神经元”组成，通过节点网络传递和处理信息。每个节点可以传递信号给其他神经元，信号在到达某个阈值时，可触发特定输出。人工神经网络 (ANN) 在水质管理中具有显著的应用潜力和挑战。ANN 擅长处理复杂的非线性数据关系，能够处理多种类型数据，并在水质指标实时监测和预测中表现出色。然而，ANN 需要大量高质量的训练数据，且内部机制难以解释，训练过程耗时且需要大量计算资源，对网络结构和超参数设置依赖性高[74]。在水质管理

中，长短期记忆网络 (LSTM) 可以用于时间序列数据的异常值检测，图 3 描述了常用的基于深度学习的算法架构并用于检测异常值的 LSTM。ANN 也可用于预测水质参数变化和污染源识别。综合来看，ANN 在水质管理中能够通过其强大的数据处理能力和模式识别能力提供精确的监测和预测，但其应用也面临数据需求、计算资源消耗和模型解释性等挑战，需结合具体需求和条件，权衡优缺点，选择适当的模型和方法。

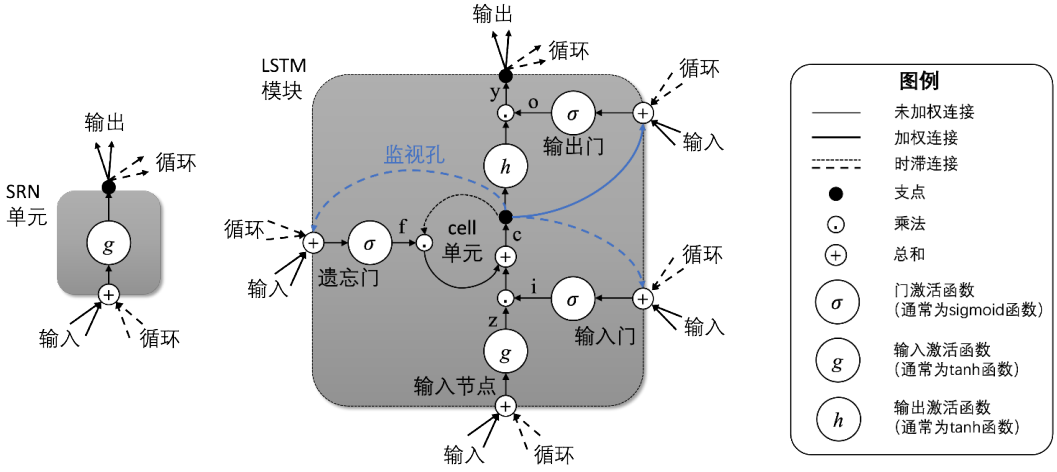


图 3 LSTM 网络架构[75]

Fig. 3 LSTM architecture

综上所述，在水质管理中，常见的机器学习算法都有其特定的优缺点。例如，SVM 擅长处理高维和非线性数据，但对大数据集和参数选择敏感，训练时间较长。ANN 能建模复杂关系，适应大量数据特征，但计算成本高，且网络结构选择困难。决策树易于理解，对数据预处理要求低，但容易过拟合。随机森林在多变量处理上表现优秀，具有较好的泛化能力，但在大数据集上训练成本高，结果难以解释。KNN 实现简单，对非线性问题有效，但对大数据集计算成本高，且对异常值敏感。选择适当的算法需要考虑数据特性、问题复杂度和资源可用性，通常通过对比测试多种算法来确定最佳方案。未来水质管理中的机器学习算法发展将聚焦于集成学习和多模型融合以提高准确性和鲁棒性，深度学习的广泛应用以处理复杂的时空数据，以及自动化机器学习 (auto-ml) 以高效挑选最佳模型和参数。同时，将增强模型的解释性和可解释性，确保专业人士和决策者能够理解和信任预测结果。此外，融合传统的过程基模型和数据驱动的机器学习方法将提升模型准确性和泛化能力。

2 机器学习在河流/地表水水质分类、预测与检测中的应用

机器学习在河流和地表水水质分类、预测与检测中的应用正在迅速发展，为环境监测和保护提供了强大的工具。机器学习算法可以通过分析历史水质数据，自动分类水样中的污染物类型，帮助快速、准确地确定水体污染源，提高环境监测效率；通过收集和处理大量水质参数数据，识别出水质变化的模式，预测未来的水质状况，例如，时间序列分析和回归模型可用来预测污染物浓度的变化趋势，帮助相关部门提前采取应对措施；此外，非监督学习算法如聚类分析和孤立森林算法能够识别水质数据中的异常值，检测潜在污染事件，提供早期预警。通过结合物联网技术，机器学习还可实时分析来自传感器网络的大规模水质数据，提供即时水质监控和预警系统。这些技术的应用不仅提高了水质监测的效率和准确性，还为水资源管理和环境保护政策的制定提供了科学依据。

2.1 机器学习用于水质分类

城市和农村地区的人类活动被认为是河流水质恶化的主要原因[76-77]。为了更好地管理水质，引入 WQI 指数对于维护人类和生态环境的健康至关重要[78-80]。在选定的研究文献中，最常用于水质评估的参数包括 DO、BOD、NO_3^-、pH 值和 EC[39, 81-83]。其次，COD、TS、TDS、磷酸盐 (PO_4^{2-}) [39]、浊度、大肠杆菌群 (fecal coliform，FC) 、总大肠杆菌群 (total coliform，TC)、总大肠菌群细菌 (total coliform bacteria，TCB)、盐度 (Salinity, Sal.)、悬浮固体 (suspended solids，SS)、总有机碳 (total organic carbon，TOC) 和氨氮

(ammonia nitrogen，AN) 等参数也被广泛使用[26, 84-86]。这些研究应用了多种 ML 算法来处理和分析水质数据，以提高评估准确性和效率。这些 ML 算法的应用情况列在表 2 中，包括各种用于分类、预测和模式识别的先进技术。通过这些算法，我们可以从复杂的水质数据集中提取有价值的信息，进而为水资源管理和保护决策提供科学依据。

表 2　用于水质分类的常用 ML 算法

Table 2　Common ML algorithms for WQC

应用领域	算法类型
水质分类 (确定WQI和WQC)	神经网络 (人工神经网络、前馈神经网络) [83, 85-86] 随机森林[39, 83] 多项式逻辑回归[83] 支持向量机[83, 86] 袋装树模型[83] 决策树[39, 86] K-近邻[85] 随机树[39] 减少误差修剪树[39] 混合模型：12种混合算法，作为独立算法与袋式分类、CV参数选择和可随机过滤分类、自适应回归-支持向量机、自适应神经模糊推理系统的组合[39, 84-85]。

对于多元线性回归 (multiple linear regression，MLR) 和 RF 算法，HASSAN 等[83] 开发了一款软件应用程序，实时预测印度 3 个不同水质等级 (良好、较差和不适合饮用) 的水质并处理在处理缺失数据。在使用 MLR、RF、增强树 (boosted trees，BT) 、NN 和 SVM 分类模型中，性能表现分别为 99.83%、98.99%、98.99%、98.65% 和 96.98%。NN、BT 和 SVM 这些高级机器学习模型在对 pH、EC、DO、TC 和 BOD 这些变量的重要性评估中表现突出。ANN、DT 和 SVM 在输入和输出变量之间的非线性和复杂关系方面表现出色，因此被广泛采用。SHAMSUDDIN 等[86] 同样使用高级模型对 2022 年兰加特河流域的水质进行多类别分类研究。他们使用了 ANN、DT 和 SVM 模型将 III 级和 II 级的水质按照供水/渔业适用进行了准确地分类。这 3 种模型的性能均超过了 85%，其中 SVM 的宏观准确度和精度分别为 96.35% 和 91.97%，ANN 为 95.62% 和 92.06%，DT 为 94.71% 和 89.22%。此外，他们发现 SVM 在处理大数据集和预测 WQI 方面的效率超过了 ANN，而且在处理小数据集时的适用性被核函数的优化和改进所增强。线性回归被证实是最适合水质分类的函数，SILLBERG 等[84] 采用了 AR-SVM 模型并综合 11 个水质参数对河流水质进行分类。其中，最主要的水质参数包括氨氮 (NH_3-N) 、TCB、FCB、BOD、DO 和 Sal。6 个主要水质参数的准确度和精确度分别为 0.94 和 0.84。AR-SVM 模型在 16 个数据集中的 15 个 (93.75%) 中取得了良好的一致性，并与传统水质指数计算结果具有良好的对应性。AL-ADHAILEH 和 ALSAADE[85] 使用 ANFIS、KNN 和 FFNN 预测了印度不同水体的水质指数。ANFIS 在确定 WQI 时显示出较高的效率和准确性，回归系数高达 96.17%。FFNN 模型在分类 WQC 时表现出卓越的稳定性，准确率和精确度分别达到 100% 和 99.96%，而 KNN 的准确率和精确度分别为 80.63% 和 82.50% (表 3) 。BUI 等[39] 在评估伊朗河的月度水质指数时，使用了 4 个独立模型和 12 个混合数据挖掘模型。这些模型的性能受到不同水质参数组合影响。在 16 个已验证的算法中，所有模型都表现良好，但袋集回归树 (bagging-aided regression tree，BA-RT) 预测 WQI 的能力最强 (R^2=0.941) ，而组合验证-回归树 (cross-validation pruned regression tree，CVPS-REPT) 预测 WQI 的能力最低 (R^2=0.853) 。基于树的混合模型在稳健性和灵活性方面优于独立模型，尤其是融合袋集算法后。

以上案例表明多种机器学习算法被应用于水质评估和分类，其中 MLR、RF、AT、NN 和 SVM 等方法在预测不同水质等级时显示出较高准确度。在大数据集处理和 WQI 预测方面，SVM 表现优越。ANFIS、KNN、FFNN 等则在特定应用中表现出高效率和准确性。这些研究表明，集成学习方法和基于树的模型，尤其是融合复杂算法的模型，可能在未来的水质数据处理与分类中成为主流趋势。

表 3 不同研究人员使用机器学习对水质分类和预测的结果[83-86]

Table 3 Results of water quality classification and prediction using machine learning by different researchers[83-86]

团队	分类模型	准确率/%	精确率/%	召回率/%
HASSAN[83]	神经网络	98.65	–	–
	随机森林	98.99	–	–
	多项式罗辑回归	99.83	–	–
	支持向量机	96.98	–	–
	增强树	98.99	–	–
SHAMSUDDIN[86]	人工神经网络	95.62	92.06	77.39
	决策树	94.71	89.22	76.35
	支持向量机	96.35	91.97	84.89
SILLBERG[84]	人工智能-支持向量机	94.00	84.00	97.00
AL-ADHAILEH和ALSAADE[85]	K-近邻	80.63	82.50	86.84
	前馈神经网络	100	99.96	100

综上所述，机器学习在水质分类中的应用带来了多方面的优势，如能够处理和分析大量复杂数据，精准识别水质类别和变化趋势，并实现实时监测和预警，从而提高水质管理的效率和响应速度。此外，它还能减少人力成本并提高预测的准确性。然而，包括对高质量数据的依赖性，模型可能的黑盒性质导致解释困难，过拟合风险，以及对先进计算资源和专业知识的高需求也造成了其广泛使用的局限性。其次，pH、DO 等动态环境因素的变化也可能影响模型的稳定性和准确性。尽管机器学习在水质分类中展现了处理复杂数据和实现实时监测的显著优势，但其对高质量数据的依赖性和动态环境因素的影响 (pH、DO、TDS 等影响水质) 需进一步优化，在将机器学习技术应用至河流/地表水水质分类时应提前对输出结果准确性进行预估，以确保在实际应用中的稳定性和准确性。

2.2 机器学习用于水质预测

水质监测和预测有助于提升水质控制和调节、优化灌溉质量及策略、提高水产养殖效率、改善饮用水处理方法以及制定水污染预防策略[87-89]，在水资源管理中扮演着至关重要的角色。根据 AL-ADHAILEH 和 ALSAADE、Khullar 和 Singh 的研究[43, 85]，多种 ML 算法已被成功应用于这些目标场景。表 4 中总结了几种用于水质预测的 ML 算法，其中 ANN、深度神经网络 (DNN) 和 SVM 因其高效性而被频繁使用[90]。在地表水水质预测中，常用的参数包括 DO、水温 (WT) 、pH 值、SS、NO_3^-、TDS、EC、浊度、BOD 和 COD。根据 SYEED 等[44] 的研究，有时也会根据数据可用性、河流类型和地点使用其他参数，如粪大肠菌群、氯化物、硫酸盐及有机和无机污染物 (图 4) 。

表 4 水质预测中常见的 ML 算法

Table 4 Common ML algorithms for water quality prediction

应用领域	算法类型
水质预测及评估	决策树回归[91]
	基于决策树的混合模型：先进数据处理-随机森林和先进数据处理-极致梯度提升[92]
	基于DNN的混合模型：双LSTM模型[43]
	神经网络及其变体 (反向传播神经网络、一般回归神经网络、循环神经网络、深度神经网络及其变体 (卷积神经网络、长短期记忆以及二者组合) [38, 43, 68, 70, 93, 94-95]
	分组数据处理方法[68]
	支持向量机[68, 70]
	树外回归[95]
	支持向量回归[43, 91, 95]

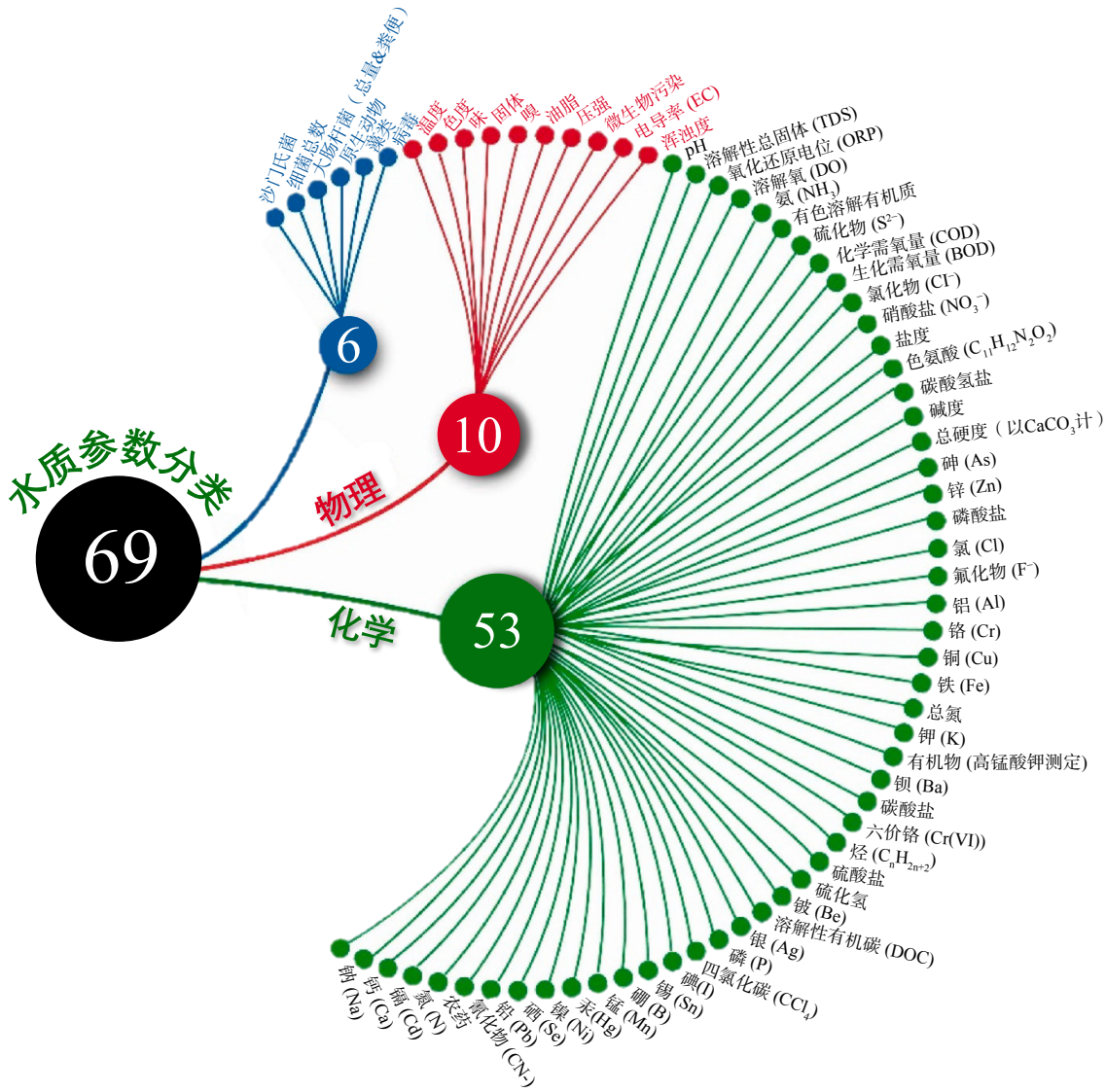


图 4　按照自然因素 (如物理、化学、生物和细菌学) 对 69 个水质参数进行分类

Fig. 4　Sixty-nine water quality parameters were categorized according to natural factors (e.g., physical, chemical, biological, and bacteriological)

水质参数之间相互关联意味着单个参数的变化可能会影响其他参数值，因此评估每个参数的重要性及其相互关系显得尤为重要[96-97]。例如，ZHU 等[50] 研究发现溶解氧是全球关注的主要水质指标之一，它与 pH 值、温度和氮氧化物 (NO_x) 浓度等输入参数有显著相关性。除了预测特定参数的浓度或基于多参数的总体水质，还可应用 ML 算法进行其他类型的预测，如月径流预测[98] 或水位预测[93]。

为了对未来的水状况进行建模，常见的方法是从已发表的文章或公共服务/监测站收集大量数据，并创建数据库，作为数据集输入待开发的 ML 模型。然而，有时会出现数据不完整或缺失的问题，这可以通过添加其他类型的数据 (如更普遍的水文数据) [99] 或使用额外的数据后处理工具 (如多元贝叶斯不确定性处理器 (MBUP)) [100] 来解决。开发多个模型并比较，以提高研究的有效性，进而挑选出精度最高、误差最小的模型。例如，ANN 和 SVM 算法已成功预测了中国河流的总氮和总磷浓度。SVM 因其优化参数少并避免过度训练的能力，在泛化效果方面更为优越。同时，遗传算法和试错分析法也常被用于优化模型参数[70]。考虑到不同模型的局限性，混合模型可能优于单一模型。虽然 LIU 等[97] 指出深度学习 LSTM 模型的良好预测能力，但 KHULLAR 和 SINGH[43] 指出，单一 CNN 和 LSTM 模型往往存在复杂度高和预测准确率低的问题，而改进的 CNN-LSTM 模型则能够克服这些缺点 (表 5) 。此外，结合先进数据去噪技术 (如 CEEMDAN)

表 5 使用机器学习得到 COD 的预测值与实际值的比较[43]

Table 5 Comparison of predicted and actual values of COD obtained using machine learning[43]

方法	均方误差	均方根误差	平均绝对误差	平均绝对百分比误差
支持向量回归	0.491	0.711	0.596	54.28
人工神经网络	0.566	0.652	0.521	54.21
随机森林	0.487	0.568	0.614	53.24
逻辑回归	0.401	0.480	0.556	51.29
长短期记忆网络	0.328	0.401	0.358	46.82
融合卷积神经网络的长短期记忆	0.218	0.268	0.214	34.22
深度学习-双向长短期记忆网络	0.015	0.117	0.115	20.32

和极梯度提升以及射频技术的混合模型，已被证明在短期水质预测中非常有效[96]。

机器学习在水质预测中提供了高度自动化和效率，通过复杂模式识别能力有效预测水质变化，并且具有强适应性和多变量分析能力。然而，其性能高度依赖于数据质量，且复杂模型的透明度较低可能限制其在需要明确决策过程场合中的应用。此外，过拟合问题和对高级技术及资源的需求也是实施机器学习水质预测时需要考虑的重要因素。因此，虽然机器学习为水质预测带来了显著优势，但在应用时也需谨慎处理这些潜在的局限性。未来水质预测和评估的趋势可能倾向于使用集成学习和混合模型，这些模型结合了多种算法来提高预测精度和稳定性[101-103]。特别是，深度学习模型 (如 LSTM 和 Bi-LSTM) 和先进数据处理技术 (如 CEEMDAN) 在水质参数预测方面显示出极大潜力[104-108]。综上所述，结合 LSTM、Bi-LSTM 和 CEEMDAN 技术的混合模型，通过深入研究和纳入不同水质参数之间的关系，能够显著提高水质预测的精确度和可靠性，这为未来水质管理的智能化和高效化提供了坚实技术基础。

2.3 机器学习用于水质异常值检测

在水资源管理中，识别供水数据中的异常情况，如缺失值、异常模式或数据规格不一致的过程，被称为异常检测[31, 109-110]。这一过程涉及到应用 ML 模型来识别和处理这些不规则数据。根据训练数据的不同，ML 模型可以分为两大类：需要基于标注数据集进行训练和校准的 SupVL 模型，以及不依赖于标注数据集的非监督学习 UnSupVL 模型。由于 SupVL 模型通常需要较大的标注数据集，因此更适用于数据丰富且已经标注的场景中。而在数据标注成本高或者难以获得足够标注数据的情况下，UnSupVL 模型可作为一个有效的替代选择。RUSSO 等[30] 的研究指出 UnSupVL 模型特别适合在从未标注的数据中识别异常模式。

表 6 中列出了几种用于异常检测的 ML 算法。这些算法的选择和应用取决于具体的数据特性和检测目标，包括处理缺失值、识别异常数据模式或处理数据不一致性等任务。应用上述算法将提高数据的准确性和可靠性，从而优化整体水资源管理和决策过程。

表 6 用于异常值检测的常见 ML 算法

Table 6 Common ML algorithms for outlier detection

应用领域	算法类型
水质异常检测	逻辑回归[110]
	支持向量机[110]
	长短期记忆[110-111]
	人工神经网络[110-111]
	深度神经网络[110]
	循环神经网络[110]
	线性判别分析[110]
	带有极端学习机的卷积神经网络[111]
	序列到序列[111]
	卷积门控循环单元[111]
	贝叶斯自回归和隔离森林[109]

MUHAREMI 等[110] 比较了 ML 算法与逻辑回归在水质 (WQ) 数据预测准确性方面，以及不同模型在水质数据中的表现。为此，他们应用了 SVM、ANN、DNN、RNN、LSTM 和 LDA 等 ML 算法，并与逻辑回归进行对比。实验结果按 F1 分数 (一种准确性指标) 排序，发现 SVM 模型表现最佳 (F1 分数为 0.989 1)，其次是 DNN (0.948 5)、LSTM (0.902 3)、RNN (0.834 5)、逻辑回归 (0.602 7)、ANN (0.576 8) 和 LDA (0.082 0)。除 SVM、逻辑回归和 ANN 外，其他模型在处理

不平衡数据集时表现出较大的脆弱性。

神经网络算法模仿了人脑的结构，包括 ANN (由众多相互连接的神经元组成的网络) 、DNN (具有多个隐藏层的有效模型) 和 RNN (具有多个循环和隐藏层的模型) 。其中，RNN 利用递归循环，即输出状态反馈到每个节点的输入状态。LSTM 模型在进行准确预测的同时，能有效地学习有用信息并遗忘无用信息。而 LDA 尽管在独立测量方面表现出色，但传统识别技术为该模型带来了挑战。MIAU 和 HUNG[111] 重点比较了 ANN、卷积神经网络 (CNN) 、LSTM、Seq2seq 和卷积门控循环单元 (Conv-GRU) 模型在台湾淡水河流域水位预测中的表现。性能指标包括均方根误差 (RMSE) 、平均绝对误差 (MAE) 和平均绝对百分比误差 (MAPE) ，其中 Conv-GRU 模型表现最佳 (RMSE 为 0.774，MAE 为 0.567，MAPE 为 30.684) ，其次是 LSTM (RMSE 为 1.032，MAE 为 0.620，MAPE 为 31.035) 和 CNN (RMSE 为 1.144，MAE 为 0.745，MAPE 为 37.154) 。这表明 Conv-GRU 模型的实际值与预测值之间的误差极小，在预测河流水位时效果最好；LSTM 和 CNN 在预测河流水位时误差稍大，但小于 ANN 和 Seq2seq。CNN 能够发现局部趋势，并观察到相同的模式在不同地方重复出现，因此取得了良好的预测结果。只有综合 CNN 和 GRU 的模型在预测性能上优于其他四个模型，因为它们是时间序列建模器，能及早显示异常行为[112-114]。Seq2seq 在多步时间序列预测基础上提供了逐序列预测，而 LSTM 和 ANN 的预测结果显示了其对复杂数据预测的能力不足[110] (表 7)。

表 7　使用不同模型对台北市不同桥站水位的预测值与实际值的对比[110-111]

Table 7　Comparison of predicted and actual water levels at different bridge stations in Taipei City using different models[110-111]

算法及样本量	均方根误差	平均绝对百分比误差	平均绝对误差
人工神经网络-180×8×5×6*	1.527	55.215	0.984
卷积神经网络-180×24×4×6	1.144	37.154	0.745
长短期记忆网络-180×32×3×6	1.032	31.035	0.62
序列到序列-180×32×5×6	1.431	49.903	0.933
卷积门控循环单元-180×8×5×6	0.774*	30.684*	0.567*

*注：表示共有180组数据，每个数据矩阵大小为8×5，共进行6次数据输入。

LIU 等[97] 结合贝叶斯自回归 (BAR) 模型和隔离森林 (IF) 算法，对美国西弗吉尼亚州波托马克河的水质数据进行了异常检测。作为质量参数的评价指标包括 RMSE、MAE 和均方误差 (MSE) ，以浊度 (TURB) 、比电导率 (SC) 和 DO。误差指标值分别为 RMSE (TURB 为 0.169 4，SC 为 0.083 1，DO 为 0.033 2) 、MAE (TURB 为 0.108 6，SC 为 0.045 3，DO 为 0.028 2) 和 MSE (TURB 为 0.028 7，SC 为 0.006 9，DO 为 0.001 1) 。这两个算法在异常检测方面都取得了卓越的结果，并证明了在水质监测和应急响应预警方面的有效性。

3　结论与展望

近年来，在环境工程特别是河流与地表水质管理领域中，ML 技术已成为关键工具。研究中广泛应用的算法包括 DT、RF、ANN、DNN、SVM、LSTM、Conv-GRU，以及基于 DNN 的多模型融合策略[115-119]。这些算法不仅能高效地进行水质分类、预测和异常检测，还可通过多模型融合策略克服单模型的限制，提高整体分析精度。深度学习 (DL) 技术的运用也促进了更高效、成本更低的水资源管理，有助于实现生态可持续性。然而，为了充分利用这些系统的潜力并验证它们在实际应用中的效能，需要进行更深入的研究和实践探索。未来研究可能聚焦于优化算法性能、融合多种数据源，并开发更有效的计算方法处理大规模数据集。长期的环境监测与评估对确保所采纳技术的可持续性和对生态系统的正面影响也至关重要。此外，提高算法的解释性和透明度，以增强其在政策制定和公众沟通中的有效性，也是未来研究的一个重点方向。

基于以上讨论，本文认为在未来几年内，ML 在辅助水质管理方面应遵循以下原则。

1) 提高模型的预测精度：虽然多种 ML 模型的比较分析已被广泛认可并实施以确定最合适的模型，但输

入参数对预测精度的显著影响不容忽视。因此，精选合适水质参数 (water quality parameters，WQPs) 至关重要。尽管 ML 算法在不同应用场景的泛化可能面临挑战，泛化能力对于跨领域的实际应用仍然是关键。在模型开发和评估过程中，收集各种气候和水文条件下的地表水水质参数，并将更多维度的数据 (如水文学、形态学、地质学等) 纳入训练数据集，可以有效提高模型的输出精度[34, 95]。

2) 集成多源数据：河流水质受诸多因素影响，包括气候变化、工业排放和农业活动等。综合运用遥感卫星、地面监测站和社会经济数据库等多源数据，可以实现更全面的水质评估。这种数据集成有助于识别影响水质的关键因素，并提高预测模型的准确性。例如，结合气象数据和水质参数可以更好地理解和预测极端天气事件对水质的影响[99]。

3) 提高模型的可解释性：现代 ML 模型，特别是深度学习模型，虽然预测精度高，但通常缺乏透明度，限制了其在更广泛领域如环境管理中的应用。采用可解释的机器学习框架可以提供预测结果的同时解释这些预测的产生过程，如 LIME 通过局部扰动数据分析模型预测的变化，而 SHAP 值计算每个特征对预测结果的贡献，这种透明度对于获得政策制定者和公众的信任至关重要。

4) 实时监测与预测系统的开发：随着物联网 (IoT) 技术的发展，实时水质监测成为可能。结合 IoT 传感器的实时数据和 ML 模型，可以开发出能即时识别和预测水质问题的系统，这对及时响应污染事件和制定有效的水质管理策略至关重要。

5) 强调跨学科合作的重要性：水质管理是一个多学科领域，涉及环境科学、工程学、计算机科学等。有效的 ML 解决方案需要这些领域专家的知识和技能。例如，环境科学家可提供关于水体系统的关键见解，而数据科学家可以贡献先进的分析技术。这种跨学科合作有助于开发更准确、更实用的 ML 应用。

6) 注重可持续性和伦理性的考量：在开发和部署 ML 模型时，需要考虑其长期环境影响和社会伦理问题。例如，数据收集过程中需要确保数据渠道公开透明，搜集流程符合相关法律法规，并且数据处理应遵循公平和透明的原则。

7) 政策和标准的制定：随着 ML 技术在水质管理中的应用日益普及，相关政策和标准制定变得尤为重要。这些政策和标准应涵盖数据收集、处理、模型开发和应用等各个方面，以确保技术安全、有效和公正应用。

参 考 文 献

[1] TUNG T M, YASEEN Z M. A survey on river water quality modelling using artificial intelligence models: 2000–2020[J]. Journal of Hydrology, 2020, 585: 124670.

[2] JIANG Y, TIAN S, LI H, et al. Harnessing microbial electrosynthesis for a sustainable future[J]. The Innovation Materials, 2023, 1(1): 100008.

[3] CAO Y, BAO Q, MIAO Y, et al. Biomimetic attempts in electrochemiluminescence[J]. The Innovation Materials, 2023, 1(3): 100034.

[4] CHEN G, WANG Q, CHU X. Accelerated spread of Fukushima's waste water by ocean circulation[J]. The Innovation, 2021, 2(2): 100119.

[5] BASU N B, VAN METER K J, BYRNES D K, et al. Managing nitrogen legacies to accelerate water quality improvement[J]. Nature Geoscience, 2022, 15(2): 97-105.

[6] YU J, ZHAO L, LIANG X Z, et al. The mediatory role of water quality on the association between extreme precipitation events and infectious diarrhea in the Yangtze River Basin, China[J]. Fundamental Research, 2023, 4.3: 495-504.

[7] ZAMORA-LEDEZMA C, NEGRETE-BOLAGAY D, FIGUEROA F, et al. Heavy metal water pollution: A fresh look about hazards, novel and conventional remediation methods[J]. Environmental Technology & Innovation, 2021, 22: 101504.

[8] 董素涵, 刘萌硕, 蔡闻琪, 等. 磺胺甲噁唑淡水水生生物水质基准与生态风险评估[J]. 环境科学学报, 2023, 43(5): 496-504.

[9] 张秋英, 李兆, 王健祺, 等. 南水北调东线湖泊硫酸盐污染现状与成因分析—以东平湖为例[J]. 环境科学学报, 2023, 43(7): 48-55.

[10] AMOATEY P, BAAWAIN M S. Effects of pollution on freshwater aquatic organisms[J]. Water Environment Research, 2019, 91(10): 1272-1287.

[11] YAN Z, ZHENG X, FAN J, et al. China national water quality criteria for the protection of freshwater life: Ammonia[J]. Chemosphere, 2020, 251: 126379.

[12] JEPPESEN E, BEKLIOĞLU M, ÖZKAN K, et al. Salinization increase due to climate change will have substantial negative effects on inland waters: A call for multifaceted research at the local and global scale[J]. The Innovation, 2020, 1(2): 100030.

[13] YUAN W, LIU Q, SONG S, et al. A climate-water quality assessment framework for quantifying the contributions of climate change and human activities to water quality variations[J]. Journal of Environmental Management, 2023, 333: 117441.

[14] RYBERG K R, CHANAT J G. Climate extremes as drivers of surface-water-quality trends in the United States[J]. Science of the Total Environment, 2022, 809: 152165.

[15] GREENHALGH S, SAMARASINGHE O. Sustainably managing freshwater resources[J]. Ecology and Society, 2018, 23(2): 44.

[16] JAM K, NOROOZI A, MOSAVI S H. A holistic view of sustainability in water resources management in the European Union: challenges and threats[J]. Environment, Development and Sustainability, 2023, 26(8): 1-34.

[17] Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 Establishing a framework for Community Action in the Field of Water Policy[J]. Council Decision of, 2000, 327: 1-23.

[18] BERTHET A, VINCENT A, FLEURY P. Water quality issues and agriculture: An international review of innovative policy schemes[J]. Land Use Policy, 2021, 109: 105654.

[19] BEHMEL S, DAMOUR M, LUDWIG R, et al. Water quality monitoring strategies—A review and future perspectives[J]. Science of the Total Environment, 2016, 571: 1312-1329.

[20] MEYER A M, KLEIN C, FüNFROCKEN E, et al. Real-time monitoring of water quality to identify pollution pathways in small and middle scale rivers[J]. Science of the Total Environment, 2019, 651: 2323-2333.

[21] ALTENBURGER R, AIT-AISSA S, ANTCZAK P, et al. Future water quality monitoring—Adapting tools to deal with mixtures of pollutants in water resource management[J]. Science of the Total Environment, 2015, 512: 540-551.

[22] CZYCZULA RUDJORD Z, REID M J, SCHWERMER C U, et al. Laboratory development of an AI system for the real-time monitoring of water quality and detection of anomalies arising from chemical contamination[J]. Water, 2022, 14(16): 2588.

[23] AHMED A N, OTHMAN F B, AFAN H A, et al. Machine learning methods for better water quality prediction[J]. Journal of Hydrology, 2019, 578: 124084.

[24] PARK J, KIM K T, LEE W H. Recent advances in information and communications technology (ICT) and sensor technology for monitoring water quality[J]. Water, 2020, 12(2): 510.

[25] KRUSE P. Review on water quality sensors[J]. Journal of Physics D: Applied Physics, 2018, 51(20): 203002.

[26] ABUZIR S Y, ABUZIR Y S. Machine learning for water quality classification[J]. Water Quality Research Journal, 2022, 57(3): 152-164.

[27] NASIR N, KANSAL A, ALSHALTONE O, et al. Water quality classification using machine learning algorithms[J]. Journal of Water Process Engineering, 2022, 48: 102920.

[28] XIN L, MOU T. Research on the Application of Multimodal-Based Machine Learning Algorithms to Water Quality Classification[J]. Wireless Communications and Mobile Computing, 2022, 2022: 9555790.

[29] VENKATA VARA PRASAD D, SENTHIL KUMAR P, VENKATARAMANA L Y, et al. Automating water quality analysis using ML and auto ML techniques[J]. Environmental Research, 2021, 202: 111720.

[30] RUSSO S, BESMER M D, BLUMENSAAT F, et al. The value of human data annotation for machine learning based anomaly detection in environmental systems[J]. Water Research, 2021, 206: 117695.

[31] DOGO E M, NWULU N I, TWALA B, et al. A survey of machine learning methods applied to anomaly detection on drinking-water quality data[J]. Urban Water Journal, 2019, 16(3): 235-248.

[32] SOUZA A P, OLIVEIRA B A, ANDRADE M L, et al. Integrating remote sensing and machine learning to detect turbidity anomalies in hydroelectric reservoirs[J]. Science of the Total Environment, 2023, 902: 165964.

[33] NASSIF A B, TALIB M A, NASIR Q, et al. Machine learning for anomaly detection: a systematic review[J]. Ieee Access, 2021, 9: 78658-78700.

[34] KHULLAR S, SINGH N. Machine learning techniques in river water quality modelling: a research travelogue[J]. Water Supply, 2021, 21(1): 1-13.

[35] WAGLE N, ACHARYA T D, LEE D H. Comprehensive review on application of machine learning algorithms for water quality parameter estimation using remote sensing data[J]. Sensors & Materials, 2020, 32(11 Pt. 4): 3879-3892.

[36] GUPTA D, MISHRA V K. Development of entropy-river water quality index for predicting water quality classification through machine learning approach[J]. Stochastic Environmental Research and Risk Assessment, 2023, 37(11): 4249-4271.

[37] AHMED U, MUMTAZ R, ANWAR H, et al. Efficient water quality prediction using supervised machine learning[J]. Water, 2019, 11(11): 2210.

[38] ANTANASIJEVIĆ D, POCAJT V, POVRENOVIĆ D, et al. Modelling of dissolved oxygen content using artificial neural networks: Danube River, North Serbia, case study[J]. Environmental Science and Pollution Research, 2013, 20: 9006-9013.

[39] BUI D T, KHOSRAVI K, TIEFENBACHER J, et al. Improving prediction of water quality indices using novel hybrid machine-learning algorithms[J]. Science of the Total Environment, 2020, 721: 137612.

[40] NONG X, SHAO D, ZHONG H, et al. Evaluation of water quality in the South-to-North Water Diversion Project of China using the water quality index (WQI) method[J]. Water Research, 2020, 178: 115781.

[41] HAMAIDI-CHERGUI F, BRAHIM ERRAHMANI M. Water quality and physicochemical parameters of outgoing waters in a pharmaceutical plant[J]. Applied Water Science, 2019, 9(7): 165.

[42] QURESHI S S, CHANNA A, MEMON S A, et al. Assessment of physicochemical characteristics in groundwater quality parameters[J]. Environmental Technology & Innovation, 2021, 24: 101877.

[43] KHULLAR S, SINGH N. Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation[J]. Environmental Science and Pollution Research, 2022, 29(9): 12875-12889.

[44] SYEED M M, HOSSAIN M S, KARIM M R, et al. Surface water quality profiling using the water quality index, pollution index and statistical methods: A critical review[J]. Environmental and Sustainability Indicators, 2023, 100247.

[45] SUTADIAN A D, MUTTIL N, YILMAZ A G, et al. Development of river water quality indices—a review[J]. Environmental Monitoring and Assessment, 2016, 188: 1-29.

[46] GORDE S, JADHAV M. Assessment of water quality parameters: a review[J]. Journal of International Environmental Application and Science, 2013, 3(6): 2029-2035.

[47] ZHONG S, ZHANG K, BAGHERI M, et al. Machine learning: new ideas and tools in environmental science and engineering[J]. Environmental Science & Technology, 2021, 55(19): 12741-12754.

[48] ZHU J J, YANG M, REN Z J. Machine learning in environmental research: Common pitfalls and best practices[J]. Environmental Science & Technology, 2023, 57(46): 17671-17689.

[49] RAZAVI S, HANNAH D M, ELSHORBAGY A, et al. Coevolution of machine learning and process-based modelling to revolutionize Earth and

environmental sciences: A perspective[J]. Hydrological Processes, 2022, 36(6): e14596.
[50] ZHU M, WANG J, YANG X, et al. A review of the application of machine learning in water quality evaluation[J]. Eco-Environment & Health, 2022, 2: 107-116.
[51] ZHANG Y, WRIGHT M A, SAAR K L, et al. Machine learning-aided protein identification from multidimensional signatures[J]. Lab on a Chip, 2021, 21(15): 2922-2931.
[52] ZHAO Y, DENG G, ZHANG L, et al. Based investigate of beehive sound to detect air pollutants by machine learning[J]. Ecological informatics, 2021, 61: 101246.
[53] LIN S, FANG X, FANG G, et al. Ultrasensitive detection and distinction of pollutants based on SERS assisted by machine learning algorithms[J]. Sensors and Actuators B: Chemical, 2023, 384: 133651.
[54] CHOWDHURY M A F, ABDULLAH M, AZAD M A K, et al. Environmental, social and governance (ESG) rating prediction using machine learning approaches[J]. Annals of Operations Research, 2023, https://doi.org/10.1007/s10479-023-05633-7.
[55] RANI V, NABI S T, KUMAR M, et al. Self-supervised learning: A succinct review[J]. Archives of Computational Methods in Engineering, 2023, 30: 2761-2775.
[56] LIU X, ZHANG F, HOU Z, et al. Self-Supervised learning: Generative or contrastive[J]. Ieee Transactions on Knowledge and Data Engineering, 2023, 35(1): 857-876.
[57] MOZELLI A, TAHERINEJAD N, JANTSCH A. A study on confidence: An unsupervised multiagent machine learning experiment[J]. Ieee Design & Test, 2022, 39(3): 54-62.
[58] HE Z, QUAN C, WANG S, et al. A comparative study of unsupervised deep learning methods for MRI reconstruction[J]. Investigative Magnetic Resonance Imaging, 2020, 24(4): 179-195.
[59] HO L, GOETHALS P. Machine learning applications in river research: Trends, opportunities and challenges[J]. Methods in Ecology and Evolution, 2022, 13(11): 2603-2621.
[60] CIABURRO G. Machine fault detection methods based on machine learning algorithms: A review[J]. Mathematical Biosciences and Engineering, 2022, 19(11): 11453-11490.
[61] ELBASI E, ZAKI C, TOPCU A E, et al. Crop prediction model using machine learning algorithms[J]. Applied Sciences, 2023, 13(16): 9288.
[62] ARIEF V N, DELACY I H, BASFORD K E, et al. Application of a dendrogram seriation algorithm to extract pattern from plant breeding data[J]. Euphytica, 2017, 213(4): 85.
[63] SONG Y-Y, YING L. Decision tree methods: applications for classification and prediction[J]. Shanghai archives of psychiatry, 2015, 27(2): 130.
[64] QUINLAN J R. C4. 5: Programs for machine learning[M]. Morgan Kaufmann, 1993.
[65] MYLES A J, FEUDALE R N, LIU Y, et al. An introduction to decision tree modeling[J]. Journal of Chemometrics: A Journal of the Chemometrics Society, 2004, 18(6): 275-285.
[66] RIGATTI S J. Random forest[J]. Journal of Insurance Medicine, 2017, 47(1): 31-39.
[67] PARK Y, CHO K H, PARK J, et al. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea[J]. Science of the Total Environment, 2015, 502: 31-41.
[68] HAGHIABI A H, NASROLAHI A H, PARSAIE A. Water quality prediction using machine learning methods[J]. Water Quality Research Journal, 2018, 53(1): 3-13.
[69] NOBLE W S. What is a support vector machine?[J]. Nature Biotechnology, 2006, 24(12): 1565-1567.
[70] LIU M, LU J. Support vector machine-an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river?[J]. Environmental Science and Pollution Research, 2014, 21: 11036-11053.
[71] KUJAWA S, NIEDBAŁA G. Artificial neural networks in agriculture[J]. Agriculture, 2021, 11(6): 497.
[72] ZADOR A M. A critique of pure learning and what artificial neural networks can learn from animal brains[J]. Nature Communications, 2019, 10(1): 3770.
[73] HASSON U, NASTASE S A, GOLDSTEIN A. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks[J]. Neuron, 2020, 105(3): 416-434.
[74] ZOU J, HAN Y, SO S-S. Overview of artificial neural networks[J]. Artificial neural networks: methods and applications, 2009, 14-22.
[75] GREFF K, SRIVASTAVA R K, KOUTNiK J, et al. LSTM: A search space odyssey[J]. IEEE transactions on neural networks and learning systems, 2016, 28(10): 2222-2232.
[76] TYAGI S, SHARMA B, SINGH P, et al. Water quality assessment in terms of water quality index[J]. American Journal of water resources, 2013, 1(3): 34-38.
[77] JI L, LI Y, ZHANG G, et al. Anthropogenic disturbances have contributed to degradation of river water quality in arid areas[J]. Water, 2021, 13(22): 3305.
[78] WANG X, ZHANG F, DING J. Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China[J]. Scientific reports, 2017, 7(1): 12858.
[79] CHOU J S, HO C C, HOANG H S. Determining quality of water in reservoir using machine learning[J]. Ecological informatics, 2018, 44: 57-75.
[80] CHIDIAC S, EL NAJJAR P, OUAINI N, et al. A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives[J]. Reviews in Environmental Science and Bio/Technology, 2023, 22(2): 349-395.
[81] NAJAFZADEH M, HOMAEI F, FARHADI H. Reliability assessment of water quality index based on guidelines of national sanitation foundation in natural streams: Integration of remote sensing and data-driven models[J]. Artificial Intelligence Review, 2021, 54(6): 4619-4651.
[82] SHAH M I, ABUNAMA T, JAVED M F, et al. Modeling surface water quality using the adaptive neuro-fuzzy inference system aided by input optimization[J]. Sustainability, 2021, 13(8): 4576.
[83] HASSAN M M, HASSAN M M, AKTER L, et al. Efficient prediction of water quality index (WQI) using machine learning algorithms[J]. Human-Centric Intelligent Systems, 2021, 1(3-4): 86-97.
[84] SILLBERG C V, KULLAVANIJAYA P, CHAVALPARIT O. Water quality classification by integration of attribute-realization and support vector machine

for the Chao Phraya River[J]. Journal of Ecological Engineering, 2021, 22(9): 70-86.
[85] HMOUD AL-ADHAILEH M, WASELALLAH ALSAADE F. Modelling and prediction of water quality by using artificial intelligence[J]. Sustainability, 2021, 13(8): 4259.
[86] SHAMSUDDIN I I S, OTHMAN Z, SANI N S. Water quality index classification based on machine learning: A case from the Langat River Basin model[J]. Water, 2022, 14(19): 2939.
[87] ZHANG M, HUANG Y, XIE D, et al. Machine learning constructs color features to accelerate development of long-term continuous water quality monitoring[J]. Journal of Hazardous Materials, 2024, 461: 132612.
[88] CHEN P, WANG B, WU Y, et al. Urban river water quality monitoring based on self-optimizing machine learning method using multi-source remote sensing data[J]. Ecological Indicators, 2023, 146: 109750.
[89] LI Y, WANG X, ZHAO Z, et al. Lagoon water quality monitoring based on digital image analysis and machine learning estimators[J]. Water Research, 2020, 172: 115471.
[90] KRISHNAN S R, NALLAKARUPPAN M, CHENGODEN R, et al. Smart water resource management using Artificial Intelligence—A review[J]. Sustainability, 2022, 14(20): 13384.
[91] NOURAKI A, ALAVI M, GOLABI M, et al. Prediction of water quality parameters using machine learning models: a case study of the Karun River, Iran[J]. Environmental Science and Pollution Research, 2021, 28(40): 57060-57072.
[92] ZHAO Y, YU T, HU B, et al. Retrieval of water quality parameters based on near-surface remote sensing and machine learning algorithm[J]. Remote Sensing, 2022, 14(21): 5305.
[93] SAMANTARAY S, DAS S S, SAHOO A, et al. Monthly runoff prediction at Baitarani river basin by support vector machine based on Salp swarm algorithm[J]. Ain Shams Engineering Journal, 2022, 13(5): 101732.
[94] BAEK S-S, PYO J, CHUN J A. Prediction of water level and water quality using a CNN-LSTM combined deep learning approach[J]. Water, 2020, 12(12): 3399.
[95] ASADOLLAH S B H S, SHARAFATI A, MOTTA D, et al. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models[J]. Journal of Environmental Chemical Engineering, 2021, 9(1): 104599.
[96] LU H, MA X. Hybrid decision tree-based machine learning models for short-term water quality prediction[J]. Chemosphere[J], 2020, 249: 126169.
[97] LIU P, WANG J, SANGAIAH A K, et al. Analysis and prediction of water quality using LSTM deep neural networks in IoT environment[J]. Sustainability, 2019, 11(7): 2058.
[98] EL BILALI A, TALEB A. Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment[J]. Journal of the Saudi Society of Agricultural Sciences, 2020, 19(7): 439-451.
[99] ZHI W, FENG D, TSAI W-P, et al. From hydrometeorology to river water quality: can a deep learning model predict dissolved oxygen at the continental scale?[J]. Environmental Science & Technology, 2021, 55(4): 2357-2368.
[100] ZHOU Y. Real-time probabilistic forecasting of river water quality under data missing situation: Deep learning plus post-processing techniques[J]. Journal of Hydrology, 2020, 589: 125164.
[101] ZHENG Z, DING H, WENG Z, et al. Research on a multiparameter water quality prediction method based on a hybrid model[J]. Ecological informatics, 2023, 76: 102125.
[102] HAQ K P R A, HARIGOVINDAN V P. Water quality prediction for smart aquaculture using hybrid deep learning models[J]. Ieee Access, 2022, 10: 60078-60098.
[103] ZHANG Y, LI C, JIANG Y, et al. A hybrid model combining mode decomposition and deep learning algorithms for detecting TP in urban sewer networks[J]. Applied Energy, 2023, 333: 120600.
[104] YANG H, LIU S. A prediction model of aquaculture water quality based on multiscale decomposition[J]. Mathematical Biosciences and Engineering, 2021, 18(6): 7561-7579.
[105] JIANGE J, LIQIN Z, SENJUN H, et al. Water quality prediction based on IGRA-ISSA-LSTM model[J]. Water, Air, & Soil Pollution, 2023, 234(3): 172.
[106] SONG C, YAO L. A hybrid model for water quality parameter prediction based on CEEMDAN-IALO-LSTM ensemble learning[J]. Environmental Earth Sciences, 2022, 81(9): 262.
[107] YANG Z, ZOU L, XIA J, et al. Inner dynamic detection and prediction of water quality based on CEEMDAN and GA-SVM models[J]. Remote Sensing, 2022, 14(7): 1714.
[108] DONG L, ZHANG J. Predicting polycyclic aromatic hydrocarbons in surface water by a multiscale feature extraction-based deep learning approach[J]. Science of the Total Environment, 2021, 799: 149509.
[109] LIU J, WANG P, JIANG D, et al. An integrated data-driven framework for surface water quality anomaly detection and early warning[J]. Journal of Cleaner Production, 2020, 251: 119145.
[110] MUHAREMI F, LOGOFĂTU D, LEON F. Machine learning approaches for anomaly detection of water quality on a real-world data set[J]. Journal of Information and Telecommunication, 2019, 3(3): 294-307.
[111] MIAU S, HUNG W-H. River flooding forecasting and anomaly detection based on deep learning[J]. Ieee Access, 2020, 8: 198384-198402.
[112] PRASAD D V V, VENKATARAMANA L Y, KUMAR P S, et al. Analysis and prediction of water quality using deep learning and auto deep learning techniques[J]. Science of the Total Environment, 2022, 821: 153311.
[113] ZHU G, LIN J, FANG H, et al. A flocculation tensor to monitor water quality using a deep learning model[J]. Environmental Chemistry Letters, 2022, 20(6): 3405-3414.
[114] WAI K P, CHIA M Y, KOO C H, et al. Applications of deep learning in water quality management: A state-of-the-art review[J]. Journal of Hydrology, 2022, 613: 128332.
[115] CHEN H, ZHANG C, YU H, et al. Application of machine learning to evaluating and remediating models for energy and environmental engineering[J]. Applied Energy, 2022, 320: 119286.

[116] LAKHOUIT A, SHABAN M, ALATAWI A, et al. Machine-learning approaches in geo-environmental engineering: Exploring smart solid waste management[J]. Journal of Environmental Management, 2023, 330: 117174.
[117] COJBASIC S, DMITRASINOVIC S, KOSTIC M, et al. Application of machine learning in river water quality management: a review[J]. Water Science & Technology, 2023, 88(9): 2297-2308.
[118] JAFARI I, LUO R, LIM F Y, et al. Machine-learning-assisted prediction and optimized kinetic modelling of residual chlorine decay for enhanced water quality management[J]. Chemosphere, 2023, 341: 140011.
[119] DING F, ZHANG W, CAO S, et al. Optimization of water quality index models using machine learning approaches[J]. Water Research, 2023, 243: 120337.

(责任编辑: 金曙光)

Application of machine learning to surface water quality management

WANG Guangtao, ZHAO Wen, JIANG Yujing, LIU Juan, ZHU Wenlei, LI Mei*

The Frontiers Science Center for Critical Earth Material Cycling, State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing 210023, China
*Corresponding author, E-mail: meili@nju.edu.cn

Abstract Machine learning, a key subfield of artificial intelligence, has been playing an increasingly important role in the environmental field. When dealing with complex problems in surface water quality management, it shows significant advantages over traditional methods. This review focused on the applications of various machine learning algorithms in surface water quality management. It analyzed the effects of different water quality parameters, such as dissolved oxygen, biological oxygen demand, chemical oxygen demand, turbidity, temperature, pH, etc., for surface water quality classification, monitoring, and prediction. This review also provided an in-depth discussion of several machine learning models that were commonly used in real-world engineering applications, such as artificial neural networks, support vector machines, random forests, decision trees, and deep learning. In addition, this review explored the application of hybrid models for improving output accuracy in surface water quality management. In summary, the realization of machine learning for accurate and efficient management of surface water quality not only depends on suitability of selected parameters for specific algorithms but also relies on reasonable use of multiple machine learning models to increase the credibility of the output results.

Keywords machine learning; environmental engineering; water quality management; surface water