Journal of Applied Chemistry and Biomedical Laboratory Research
https://doi.org/00.0000/0000-000-000

eISSN 3125-1404

Open Access

# Machine Learning Applications in Industrial Pollution Monitoring and Wastewater Management: A Bibliometric Analysis

Cut Dian Nuryanty

*Correspondence:
cutdian.nuryanty@gmail.com

Center for Environmental Pollution Studies, West Java, Indonesia

Received: May 16, 2026
Accepted: June 29, 2026
Published online: June 30, 2026

JACBioLab
JOURNAL OF APPLIED CHEMISTRY AND BIOMEDICAL LABORATORY RESEARCH

## ABSTRACT

**Introduction:** Industrial pollution and wastewater contamination represent escalating global environmental crises, yet a systematic, evidence-based understanding of how machine learning (ML) is being applied to address these challenges remains lacking. This study maps the intellectual landscape and emerging research frontiers in ML-based environmental monitoring and wastewater management through a bibliometric lens. **Materials and Methods:** A total of 171 documents retrieved from the Scopus database (2008–2026) were analyzed using bibliometric methods. VOSviewer was employed to construct co-occurrence keyword networks, overlay maps, and density visualizations. Quantitative analysis of publication trends, document types, top journals, and citation impact. **Results:** Publication output grew exponentially from 2 articles in 2008 to 58 articles in 2025, with 2026 already yielding 34 records as of May. Four thematic clusters were identified: (1) wastewater treatment process optimization using artificial neural networks and deep learning; (2) water quality and pollutant detection; (3) industrial wastewater and environmental monitoring; and (4) emerging contaminants and greenhouse gas modeling. XGBoost, explainable artificial intelligence (XAI), and data-driven modeling represent the most rapidly growing thematic frontiers. **Conclusion:** The field is transitioning from predictive accuracy toward interpretability, real-time deployment, and multi-hazard environmental risk assessment. Critical research gaps include standardized benchmark datasets, cross-domain model transferability, and equitable global collaboration structures. This study provides an evidence-based roadmap for future ML-integrated environmental research.

**Keywords:** Artificial intelligence, Pollution monitoring, Wastewater treatment

Citation:
Nuryanty CD. Machine Learning Applications in Industrial Pollution Monitoring and Wastewater Management: A Bibliometric Analysis. J Appl Chem Biomed Lab Res (JACBioLab). 2026;2(1):18-24. https://doi



## INTRODUCTION

Industrial expansion and rapid urbanization have profoundly intensified the discharge of toxic effluents, micropollutants, and greenhouse gases into aquatic and terrestrial ecosystems. The United Nations Environment Programme estimates that over 80% of global wastewater is discharged untreated into the environment, while industrial pollution accounts for a disproportionate share of surface water contamination in low- and middle-income countries [1]. Conventional wastewater treatment infrastructure, designed primarily for bulk organic matter removal, has proven structurally incapable of addressing the complexity and heterogeneity of modern industrial pollutant mixtures, which increasingly include pharmaceuticals, endocrine disruptors, heavy metals, pesticides, and per- and polyfluoroalkyl substances (PFAS) [2,3]. Traditional physicochemical monitoring methods, though analytically robust, are costly, time-intensive, and unsuited for real-time decision support in dynamic industrial and municipal systems [4]. Temporal data discontinuities hinder proactive pollution response, allowing chronic low-level contamination to accumulate undetected until ecological damage becomes irreversible. Grey water footprint assessments at the river basin level further demonstrate the systemic inadequacy of current accounting frameworks to reflect the true environmental burden of diffuse and point-source pollution [5]. Climate change amplifies these risks by altering hydrological cycles, intensifying storm events, and disrupting the self-purification capacity of receiving water bodies [6].

Machine learning (ML) has become a revolutionary methodological paradigm in reaction to these systemic shortcomings. ML algorithms, in contrast to rule-based models, do not require an explicit mechanistic formulation to develop data-driven representations of complex, nonlinear environmental systems [7]. Applications include predicting effluent quality, fault detection in wastewater treatment plants (WWTPs), real-time water quality monitoring [8], chemical fingerprinting of industrial pollution sources [9], non-target screening of contaminants [10], and modeling of greenhouse gas emissions. At operational scales, ensemble techniques like Random Forest and XGBoost, deep learning architectures like long short-term memory (LSTM) networks, and explainable AI (XAI) frameworks [11,12].

Despite this rapid expansion, the field lacks a systematic, quantitative synthesis of its intellectual structure, thematic frontiers, and collaboration dynamics. This study presents a comprehensive bibliometric analysis, mapping thematic clustering, identifying knowledge density hotspots, and delineating the temporal trajectory of emerging frontiers. The novelty lies in its integrative scope, simultaneously analyzing the structural, temporal, and topical dimensions of the literature, and in its critical identification of unresolved methodological and translational challenges that must be addressed to realize the full potential of ML in environmental governance.

## MATERIALS AND METHODS

Bibliometric data were systematically retrieved from the Scopus database on 15 May 2026. The search strategy employed Boolean operators combining "machine learning" OR "artificial intelligence" OR "deep learning" with "wastewater treatment" OR "industrial pollution" OR "environmental monitoring" OR "water quality." Only peer-reviewed articles, reviews, conference papers, and book chapters published in English were eligible. The final dataset comprised 171 documents spanning 2008 to 2026, encompassing 130 original research articles, 28 reviews, 7 book chapters, and 6 conference papers.

Quantitative analysis of publication trends, journal productivity, and citation impact was conducted using VOSviewer (version 1.6.20). Three network visualizations were constructed: (1) a keyword co-occurrence network displaying thematic clusters, (2) a temporal overlay map encoding average year of keyword occurrence by node color, and (3) a density visualization quantifying cumulative co-occurrence frequency per keyword node. A minimum keyword occurrence threshold of three co-occurrences was applied to ensure analytical robustness. This threshold balances inclusivity with statistical reliability, retaining 42 keyword nodes from the original corpus for network construction.

## RESULTS AND DISCUSSION

### Publication Trends

The temporal distribution of 171 documents reveals a pronounced growth trajectory consistent with a field in rapid paradigmatic expansion. Annual output remained minimal prior to 2019 (typically 1–5 publications per year), reflecting nascent adoption of ML in environmental science. A decisive inflection point occurs from 2021 onward, 7 documents in 2021, 13 in both 2022 and 2023, 20 in 2024, and 58 in 2025, representing a 700% increase over the four-year period. The partial 2026 count of 34 publications (through May alone) projects an annualized output of more than 80 articles (Table 1).

This growth reflects qualitative shifts beyond mere quantitative increase. Early publications (2008–2018) were predominantly reactive and descriptive, applying statistical models or artificial neural networks (ANN) to existing monitoring datasets to replicate the behavior of single-site systems [13,14]. Post-2021 literature exhibits a prospective, systems-level orientation: ML is embedded within real-time operational decision frameworks, regulatory compliance monitoring [15], and climate-adaptive water management systems. The exponential trajectory correlates with the democratization of open-source ML tools, the proliferation of environmental sensor networks generating large-volume training datasets, and regulatory recognition that data-driven approaches offer scalable complements to conventional monitoring infrastructure [16].

**Table 1.** Publication Distribution and Key Metrics by Period (2008–2026)

| Period | Publications (n) | Highest-Cited Article (Citations) | Dominant ML Method |
|---|---|---|---|
| 2008–2015 | 11 | Decision support system Manzanares River | ANN, statistical models |
| 2016–2020 | 19 | AI in pollution controls review | ANN, deep learning, LSTM |
| 2021–2023 | 33 | Fault detection WWTP deep learning | RF, XGBoost, LSTM, GPR |
| 2024–2026* | 108 | Carbon emissions in wastewater treatment | XGBoost, XAI, hybrid models |

*2026 data through 15 May 2026. ANN = artificial neural network; RF = random forest; LSTM = long short-term memory; GPR = Gaussian process regression; XGBoost = extreme gradient boosting; XAI = explainable artificial intelligence.

**Keyword Co-occurrence Network**

Keyword co-occurrence network analysis using VOSviewer identified four distinct thematic clusters within the corpus (Figure 1). Each cluster represents a coherent intellectual community with shared methodological vocabularies, yet substantial inter-cluster linkages reveal a field characterized by cross-domain methodological fertilization rather than disciplinary fragmentation.

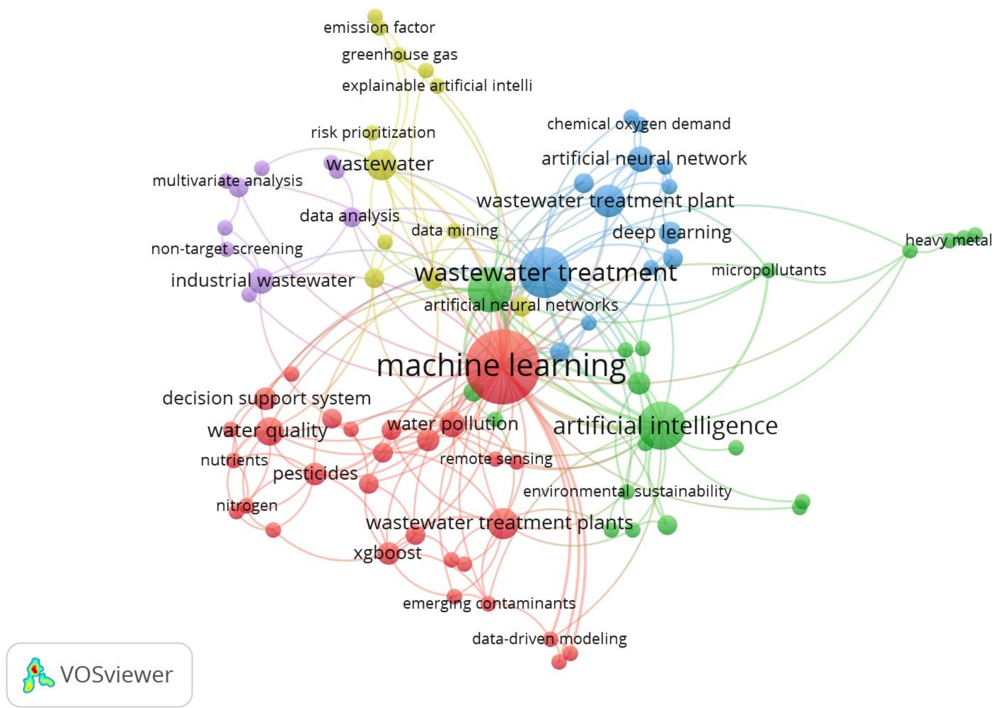


**Figure 1.** Keyword co-occurrence network map

Cluster 1 (red) centers on "machine learning," "artificial intelligence," "water quality," and "water pollution," interconnected with "XGBoost," "remote sensing," and "wastewater treatment plants." With 41 and 17 co-occurrences respectively, "machine learning" and "artificial intelligence" represent the field's principal organizing constructs. The presence of XGBoost as a distinct high-frequency node signals methodological maturation, gradient boosting algorithms have displaced the generic "neural network" framing in many recent studies due to their superior interpretability and computational efficiency. Applications in this cluster range from satellite-based surface water quality mapping to multi-step ahead effluent parameter prediction.

Cluster 2 (blue) consolidates wastewater treatment process optimization, with core nodes including "wastewater treatment plant," "deep learning," "artificial neural networks," and "chemical oxygen demand." This is the operationally most mature ML application domain, predictive models for effluent quality parameters and fault detection systems have progressed to full-scale implementation. LSTM-based fault detection in a full-scale WWTP provides early warning 2–4 hours before conventional alarm thresholds, a system-level advance with direct operational impact. The co-occurrence of "deep learning" and "artificial neural networks" as neighboring but distinct nodes reflects an ongoing methodological transition from classical feedforward networks toward recurrent architectures capable of capturing temporal dynamics in effluent quality data.

Cluster 3 (yellow-green) encompasses industrial wastewater characterization and environmental source tracking, anchored by "industrial wastewater," "non-target screening," "multivariate analysis," and "data mining." This cluster's methodological sophistication, deploying support vector classification on high-resolution mass spectrometry data, achieves 92–100% cross-validated accuracy in identifying wastewater across five environmental source matrices. Shu *et al.* [17] extended this approach to domestic-industrial wastewater separation using fluorescence excitation-emission matrices combined with ML classification, demonstrating the cluster's capacity for operationally deployable source discrimination tools. Peripheral "greenhouse gas" and "emission factor" nodes indicate conceptual expansion toward atmospheric co-monitoring, aligned with carbon accounting imperatives in the wastewater sector [18].

Cluster 4 (purple) groups emerging and niche topics: "pesticides," "nutrients," "nitrogen," "decision support systems," and "risk prioritization." Though smaller in node count, this cluster is intellectually significant because it represents the convergence of agrochemical pollution monitoring and evidence-based environmental decision-making. Fischer *et al.* [19] demonstrated decision support frameworks for contaminants of emerging concern in drinking water catchments, integrating multi-criteria analysis with physicochemical property data, an approach that ML could substantially amplify through data-driven prioritization. The connections of this cluster to the central ML node confirm that these application domains are adopting data-driven approaches, though at earlier stages of methodological maturity than Cluster 2.

**Temporal Overlay Analysis**

The temporal overlay visualization assigns color values to keywords based on their average year of co-occurrence, enabling identification of chronological knowledge stratification (Figure 2). Keywords rendered in cooler colors (blue-purple, representing 2016–2019) include "wastewater treatment," "artificial neural networks," "decision support system," and "water quality" reflecting established research domains with publication histories predating the current ML acceleration. Keywords in warmer colors (yellow-green, 2023–2026) include "XGBoost," "explainable artificial intelligence," "data-driven modeling," and "remote sensing" representing the field's current cutting edge.

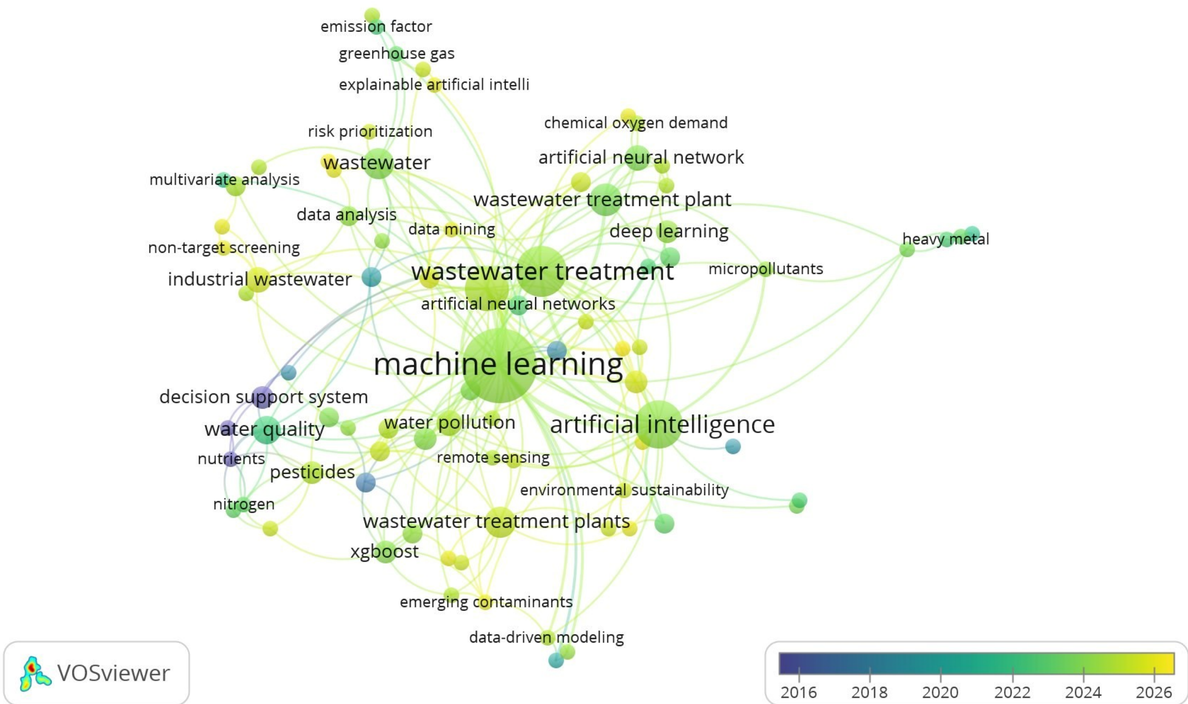


**Figure 2.** Temporal overlay map of keyword co-occurrence network

The emergence of XAI as a high-recency node warrants critical analysis. Methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) address a fundamental limitation of black-box ML: the inability to provide mechanistically interpretable outputs amenable to regulatory scrutiny and operational trust by engineers. Soleymani Hasani *et al.* [20] demonstrated that XAI-based analysis of lake water quality models revealed nitrogen and phosphorus as dominant predictors, knowledge directly actionable by lake managers, highlighting XAI's potential to bridge the gap between ML sophistication and environmental management practice. Similarly, volatile organic compound (VOC) monitoring in wastewater systems now integrates e-nose sensor arrays with ML classification, with XAI applied to identify sensor contributions and guide instrument calibration.

The overlay analysis also shows "data-driven modeling" and "emerging contaminants" as clearly recent nodes. The expanding application of ML to PFAS monitoring, recently demonstrated by Dong *et al.* [21] through systematic ML-based database mining across 40,000+ PFAS records to model environmental fate and treatment efficiency, exemplifies the frontier where data availability and ML capability converge to address previously intractable contaminant management challenges. Similarly, carbon emission accounting in wastewater treatment systems, driven by decarbonization regulatory imperatives, is developing as a high-velocity research topic where ML provides the predictive architecture for real-time carbon footprint monitoring.

**Density Map**

The density visualization (Figure 3) supports the cluster analysis results and also shows gradients of knowledge concentration with spatial accuracy. The densest concentration of the brightest yellow hotspots, signifying the highest density of the field, is focused at the intersection of "machine learning", "wastewater treatment" and "artificial intelligence". This suggests that the field's knowledge core is treatment-process optimization. Secondary density peaks at "water quality" and "deep learning" nodes represent sub-nuclear activity zones, reflecting the maturity of these sub-domains

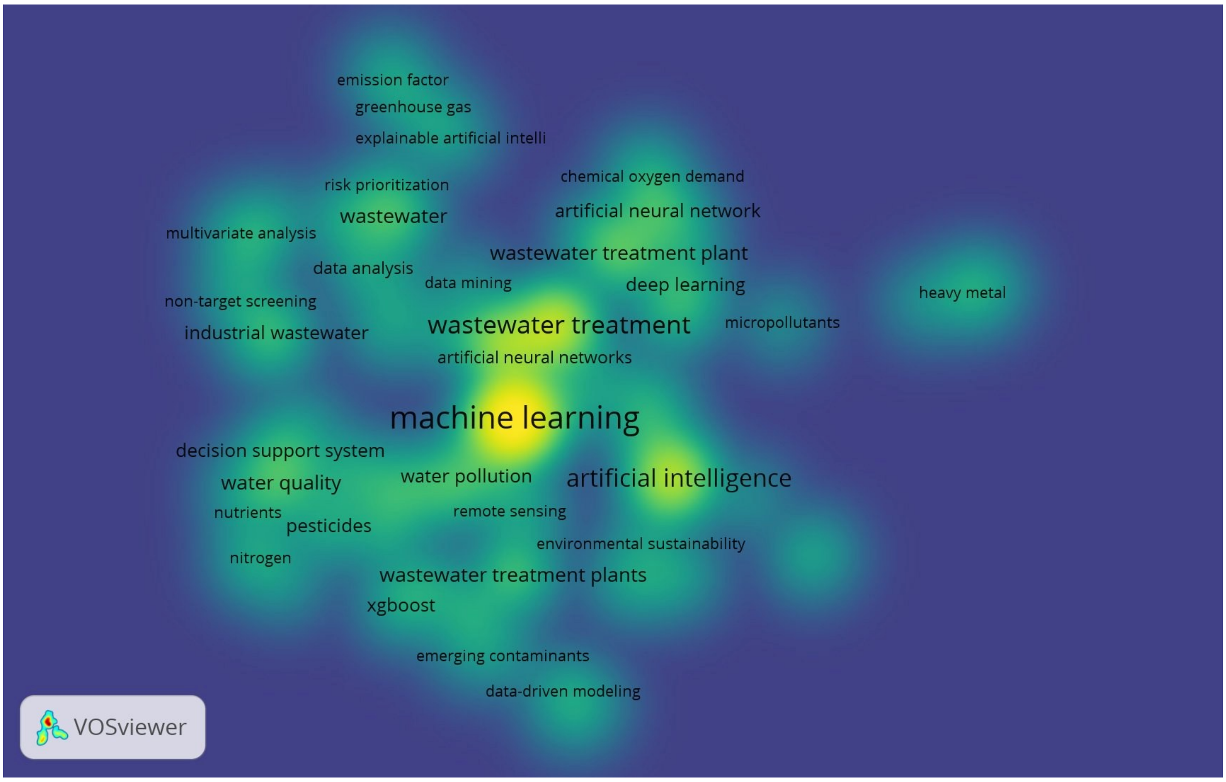


**Figure 3.** Keyword co-occurrence density map

Crucially, the density gradient toward peripheral nodes "micropollutants," "heavy metals," "emerging contaminants," and "PFAS" is steep, indicating that, despite their recognized environmental significance, these contaminant classes remain systematically underserved by ML methodologies relative to conventional effluent parameters. This structural gap is particularly salient for PFAS, where complicated mixture toxicity, environmental persistence, and regulatory uncertainty make a compelling justification for ML-based predictive risk assessment [22]. In contrast, antibiotic sensing in wastewater, now nearing operational feasibility via surface-enhanced Raman scattering combined with ML signal interpretation [23], shows that novel sensor-ML integration is technically feasible for emerging contaminants, emphasizing the analytical rather than fundamental nature of the gap.

## CONCLUSIONS

The bibliometric research presented here shows that ML applications for industrial pollution monitoring and wastewater management have grown exponentially from 2021 to 2026, driven by breakthroughs in ensemble learning, deep learning architectures, and the growing integration of XAI frameworks. The four connected topic clusters are the optimization of the field wastewater treatment process, the real-time monitoring of water quality, the characterization of industrial sources, and the management of emergent contaminants, each with distinct methodological characteristics and operational maturity levels. Three priority actions are warranted to unlock the transformative potential of the field in environmental governance: (1) creation of open, standardised benchmark datasets to evaluate environmental ML models and facilitate systematic cross-study comparison, (2) deliberate reorganisation of international collaboration networks to inclusively integrate institutions from high-pollution, low-resource settings where ML-based monitoring tools are most urgently needed, and (3) codification of uncertain These structural adjustments will ensure that the quantitative growth demonstrated here translates into qualitative achievements in safeguarding global aquatic and terrestrial habitats.

**Funding:** This research received no external funding.

**Acknowledgments:** The author acknowledges the Center for Environmental Pollution Studies, West Java, Indonesia, for institutional support.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## REFERENCES

1. Ye Z, Yang J, Zhong N, Tu X, Jia J, Wang J. Tackling environmental challenges in pollution controls using artificial intelligence: A review. Sci Total Environ [Internet]. 2020;699. Available from: https://www.scopus.com/pages/publications/85074240649?origin=resultslist
2. Sarkar DJ, Das Sarkar S, Das BK, Sahoo BK, Das A, Nag SK, et al. Occurrence, fate and removal of microplastics as heavy metal vector in natural wastewater treatment wetland system. Water Res [Internet]. 2021;192. Available from: https://www.scopus.com/pages/publications/85099860065?origin=resultslist
3. Mamandipoor B, Majd M, Sheikhalishahi S, Modena C, Osmani V. Monitoring and detecting faults in wastewater treatment plants using deep learning. Environ Monit Assess [Internet]. 2020;192(2). Available from: https://www.scopus.com/pages/publications/85078450728?origin=resultslist
4. Liu Z, Xu Z, Zhu X, Yin L, Yin Z, Li X, et al. Calculation of carbon emissions in wastewater treatment and its neutralization measures: A review. Sci Total Environ [Internet]. 2024;912. Available from: https://www.scopus.com/pages/publications/85180536764?origin=resultslist
5. Pellicer-Martínez F, Martínez-Paz JM. Grey water footprint assessment at the river basin level: Accounting method and case study in the Segura River Basin, Spain. Ecol Indic [Internet]. 2016;60:1173–83. Available from: https://www.scopus.com/pages/publications/84942798932?origin=resultslist
6. Whelan MJ, Linstead C, Worrall F, Ormerod SJ, Durance I, Johnson AC, et al. Is water quality in British rivers "better than at any time since the end of the Industrial Revolution"? Sci Total Environ [Internet]. 2022;843. Available from: https://www.scopus.com/pages/publications/85133229221?origin=resultslist
7. Nagpal M, Siddique MA, Sharma K, Sharma N, Mittal A. Optimizing wastewater treatment through artificial intelligence: recent advances and future prospects. Water Sci Technol [Internet]. 2024;90(3):731–57. Available from: https://www.scopus.com/pages/publications/85201245056?origin=resultslist
8. Safder U, Kim J, Pak G, Rhee G, You K. Investigating Machine Learning Applications for Effective Real-Time Water Quality Parameter Monitoring in Full-Scale Wastewater Treatment Plants. Water (Switzerland) [Internet]. 2022;14(19). Available from: https://www.scopus.com/pages/publications/85139824784?origin=resultslist
9. Dávila-Santiago E, Shi C, Mahadwar G, Medeghini B, Insinga L, Hutchinson R, et al. Machine Learning Applications for Chemical Fingerprinting and Environmental Source Tracking Using Non-target Chemical Data. Environ Sci Technol [Internet]. 2022;56(7):4080–90. Available from: https://www.scopus.com/pages/publications/85127435242?origin=resultslist
10. Lim YM, Swamy V, Ramakrishnan N, Chan ES, Kesuma HP. Volatile organic compounds (VOCs) in wastewater: Recent advances in detection and quantification. Microchem J [Internet]. 2023;195. Available from: https://www.scopus.com/pages/publications/85174735993?origin=resultslist
11. Jiang Y, Li C, Song H, Wang W. Deep learning model based on urban multi-source data for predicting heavy

metals (Cu, Zn, Ni, Cr) in industrial sewer networks. J Hazard Mater [Internet]. 2022;432. Available from: https://www.scopus.com/pages/publications/85126682237?origin=resultslist

12. Zhang H, Ren X, Chen S, Xie G, Hu Y, Gao D, et al. Deep optimization of water quality index and positive matrix factorization models for water quality evaluation and pollution source apportionment using a random forest model. Environ Pollut [Internet]. 2024;347. Available from: https://www.scopus.com/pages/publications/85188459124?origin=resultslist
13. Zeinolabedini M, Najafzadeh M. Comparative study of different wavelet-based neural network models to predict sewage sludge quantity in wastewater treatment plant. Environ Monit Assess [Internet]. 2019;191(3). Available from: https://www.scopus.com/pages/publications/85061639922?origin=resultslist
14. Paredes J, Andreu J, Solera A. A decision support system for water quality issues in the Manzanares River (Madrid, Spain). Sci Total Environ [Internet]. 2010;408(12):2576–89. Available from: https://www.scopus.com/pages/publications/77951754088?origin=resultslist
15. Hammond P, Suttie M, Lewis VT, Smith AP, Singer AC. Detection of untreated sewage discharges to watercourses using machine learning. npj Clean Water [Internet]. 2021;4(1). Available from: https://www.scopus.com/pages/publications/85102436130?origin=resultslist
16. Wani AK, Rahayu F, Ben Amor I, Quadir M, Murianingrum M, Parnidi P, et al. Environmental resilience through artificial intelligence: innovations in monitoring and management. Environ Sci Pollut Res [Internet]. 2024;31(12):18379–95. Available from: https://www.scopus.com/pages/publications/85187160185?origin=resultslist
17. Shu Y, Kong F, He Y, Chen L, Liu H, Zan F, et al. Machine learning-assisted source tracing in domestic-industrial wastewater: A fluorescence information-based approach. Water Res [Internet]. 2025;268. Available from: https://www.scopus.com/pages/publications/85206333849?origin=resultslist
18. Marinelli E, Radini S, Foglia A, Lancioni N, Piasentin A, Eusebi AL, et al. Validation of an evidence-based methodology to support regional carbon footprint assessment and decarbonisation of wastewater treatment service in Italy. Water Res [Internet]. 2021;207. Available from: https://www.scopus.com/pages/publications/85119113843?origin=resultslist
19. Fischer A, ter Laak T, Bronders J, Desmet N, Christoffels E, van Wezel A, et al. Decision support for water quality management of contaminants of emerging concern. J Environ Manage [Internet]. 2017;193:360–72. Available from: https://www.scopus.com/pages/publications/85013650629?origin=resultslist
20. Soleymani Hasani S, Arias ME, Nguyen HQ, Tarabih OM, Welch Z, Zhang Q. Leveraging explainable machine learning for enhanced management of lake water quality. J Environ Manage [Internet]. 2024;370. Available from: https://www.scopus.com/pages/publications/85206174383?origin=resultslist
21. Dong J, Kim S, Young SD, Li C, Jin Z, Lee D, et al. Machine learning for monitoring per- and polyfluoroalkyl substance (PFAS) in California's wastewater treatment plants: An assessment of occurrence and fate. J Hazard Mater [Internet]. 2025;492. Available from: https://www.scopus.com/pages/publications/105001806267?origin=resultslist
22. Jiang L, Yao J, Ren G, Sheng N, Guo Y, Dai J, et al. Comprehensive profiles of per- and polyfluoroalkyl substances in Chinese and African municipal wastewater treatment plants: New implications for removal efficiency. Sci Total Environ [Internet]. 2023;857. Available from: https://www.scopus.com/pages/publications/85140361895?origin=resultslist
23. Huang YH, Wei H, Santiago PJ, Thrift WJ, Ragan R, Jiang S. Sensing Antibiotics in Wastewater Using Surface-Enhanced Raman Scattering. Environ Sci Technol [Internet]. 2023;57(12):4880–91. Available from: https://www.scopus.com/pages/publications/85151044177?origin=resultslist