Environmental Science Advances

ROYAL SOCIETY OF CHEMISTRY

CRITICAL REVIEW

View Article Online
View Journal

Cite this: DOI: 10.1039/d5va00240k

Artificial intelligence driven bioinformatics for sustainable bioremediation: integrating computational intelligence with ecological restoration

Kashif R. Siddique,[a] Debajyoti Bose, *[b] Riya Bhattacharya,[b] Raul Villamarin Rodriguez[b] and Aritra Ray[c]

Environmental pollution from heavy metals and untreated wastewater poses significant risks to ecosystems and human health, highlighting the urgent need for innovative remediation strategies. Bioremediation employs microorganisms to break down contaminants and presents a sustainable and economical solution. However, conventional techniques such as bioaugmentation and bio-stimulation face challenges due to inefficiencies and the absence of real-time monitoring. This narrative review consolidates the latest developments in AI-driven bioinformatics aimed at enhancing microbial bioremediation, with an emphasis on the degradation of heavy metals and wastewater pollutants. Advanced computational models such as random forest, artificial neural networks, and support vector machines demonstrate high predictive accuracy (R^2 > 0.99) in analysing microbial behaviour and pollutant dynamics and optimizing processes. Bioinformatics tools such as AlphaFold2 and I-TASSER and metagenomic platforms such as QIIME and MG-RAST facilitate accurate identification of microbial communities, genes, and degradation pathways. AI-powered biosensors and advanced deep learning enable the continuous observation of enzymatic activity and the effectiveness of treatments. The combination of AI, metagenomics, and gene editing techniques, such as CRISPR, presents scalable approaches for achieving sustainable bioremediation. The present work emphasizes innovative tools, practical applications such as ANN-RF hybrid models, and prospective pathways, highlighting the significant impact of computational intelligence on ecological restoration.

Received 30th July 2025
Accepted 29th September 2025

DOI: 10.1039/d5va00240k

rsc.li/esadvances

Environmental significance

The integration of artificial intelligence (AI) and bioinformatics in microbial bioremediation, as presented in this work, offers a transformative approach to addressing environmental pollution caused by heavy metals and untreated wastewater. By leveraging advanced computational models such as random forest, artificial neural networks, and support vector machines, alongside bioinformatics tools like AlphaFold2, QIIME, and MG-RAST, this research enhances the efficiency, precision, and scalability of bioremediation processes. These technologies enable real-time monitoring, accurate prediction of microbial behavior, and optimization of pollutant degradation pathways, overcoming limitations of traditional methods like bioaugmentation and bio-stimulation. The application of AI-driven biosensors, metagenomics, and gene-editing techniques, such as CRISPR, facilitates sustainable and cost-effective solutions for ecological restoration. This interdisciplinary approach not only mitigates the adverse impacts of contaminants on ecosystems and human health but also paves the way for innovative, data-driven strategies to achieve long-term environmental sustainability and resilience.

1 Introduction

Bioremediation is a process of breaking down contaminants using microorganisms.[1] Pollution in any form is a major global issue, which adversely affects the environment, humans, animals and plant health.[2] Innovative pollution-reduction strategies are required to solve environmental problems and their harmful impact on human health. Microbial degradation of pollutants and toxins is a promising strategy for environmental remediation.[3,4] Optimization of bioremediation processes is a major task for environmental research. Artificial intelligence or AI is one of the optimum tools to change environmental conditions. AI algorithms have gained popularity in environmental research due to their capability to handle big and complex data, of feature extraction, and of discovering

[a]Department of Analytics, School of Business, Woxsen University, Hyderabad, Telangana, India
[b]Centre of Excellence in Health Technology,Ecosystems & Biodiversity, Woxsen University, Hyderabad, Telangana, India. E-mail: debajyoti1024@gmail.com
[c]College of Engineering, School of Mechanical Engineering, Purdue University, West Lafayette, Indiana, USA

patterns and ability to generate timely ideas to tackle environmental issues.[5,6] Although the potential of AI can only be evaluated using interdisciplinary research elements, this has been developed as a cost effective and viable method for repairing polluted ecosystems. The effects of heavy metal contamination can lead to a variety of health challenges like neurological disorders, reproductive problems, kidney damage and cancer.[7]

According to the World Health Organization in 2021 and United Nations Environment Program, around 80% of the wastewater worldwide is discharged into the untreated environment from economically developing countries without any proper treatment.[8–10] AI-driven optimization of bioremediation using algorithms can be used to evaluate ecological data, estimate pollutant behavior, optimize the process and increase efficiency. Traditional approaches such as bioaugmentation and bio-stimulation have been used for years, yet they still have certain limits. Bioaugmentation is a process of infusing microorganisms to break down the contaminants and enhance the metabolic activity of the existing bacterial community.[11,12] Similarly, bio-stimulation is a process of adding nutrients, oxygen and other chemicals that can improve the pollution degrading capability.[13,14] These traditional bioremediation processes are time-consuming and lack real-time monitoring capabilities, consideration of site heterogeneity, and prediction of microbial behavior.[15] To overcome the limitations of traditional bioremediation approaches, bioinformatics and computational models are being used to predict the microorganism's behavior in response to environmental conditions and pollutions.

Machine learning or ML models like random forests can predict microbial survival based on site parameters and also accelerate the identification of the best parameters. Bioinformatics approaches, such as metagenomics, have shown potential to replace traditional culture methods by accelerating microbial identification for degradation processes and enabling visualization of bacterial communities. Recently computational and bioinformatics techniques have shown potential in various areas of basic and applied sciences. This narrative review synthesizes recent advances in AI-driven bioinformatics for optimizing microbial bioremediation of heavy metals, highlighting key tools, challenges and future directions.

2 AI in microbial selection and optimization

AI leverages digital algorithms to perform complex tasks. Advanced machine learning algorithms like Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), random forests are some promising algorithms.[16,17] Techniques such as clustering and classification help in identifying and characterizing diverse microbial communities for efficient degradation processes. ANNs and ensemble methods can predict microbial community dynamics and explore the relation between environmental conditions and the microbial community for effective bioremediation strategies as presented in Fig. 1. In a recent study by a group, it was observed that random forest has played

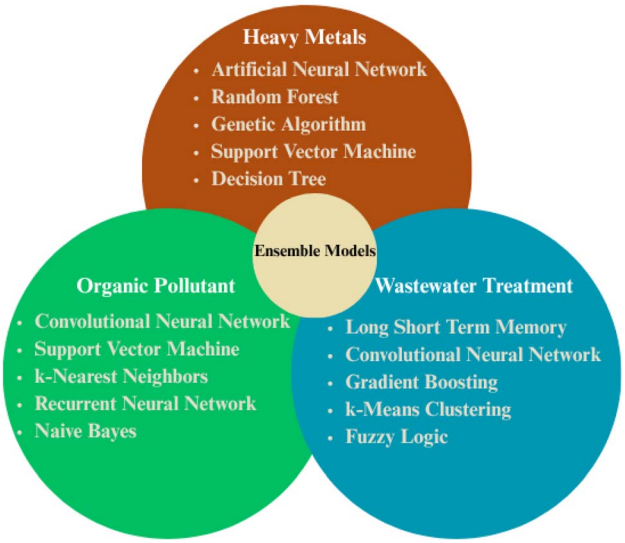

Fig. 1 Machine learning models used in heavy metals, organic pollutants, and wastewater treatment, with ensemble models at the intersection.

an important role in prediction of bacterial microbiota changes in various contaminated environments. The authors utilized the Web of Science to compile data and collected 6 variables to analyze. Three algorithms were used to predict the effects of microplastics on antibiotic degrading bacteria. The random forest model performed the best with AUC values of 85% and 88%.[18] AI-driven sensors in association with deep learning algorithms are used for real-time monitoring and provide insights into catalytic activities of enzymes.[19] Another investigation showed that biosensors and learning tools can aid in the prediction of long-term treatment outcomes. Researchers investigated the interaction of four commercial dyes by *Bacillus megaterium* H_2 azoreductases using *in silico* analysis (ligand binding site modeling, molecular docking, and molecular dynamics simulation). The obtained binding results suggested stabilization between the complexes.[20] A group of researchers also presented a hypothesis on how nano-biosensors can be used to detect pesticides in the field which can be modified and exploited in the bioremediation process.[21] Random forest models offer high interpretability by revealing feature importance, aiding in understanding key microbial and environmental factors driving bioremediation outcomes. In contrast, artificial neural networks and their hybrids, while highly accurate, are less interpretable due to their complex architectures, necessitating advanced techniques like SHAP or LIME for practical deployment.

One evaluation utilized 22 metagenomic and genomic datasets of microbial communities integrated with AI and ML algorithms to enhance degradation of environmental contaminants and toxins while providing insights into the genetic and functional potential of these communities.[22] Some studies have shown that ML can predict pathways that are involved in degradation by analyzing sequences while functional profiling using AI enables identification of key enzymes and taxa

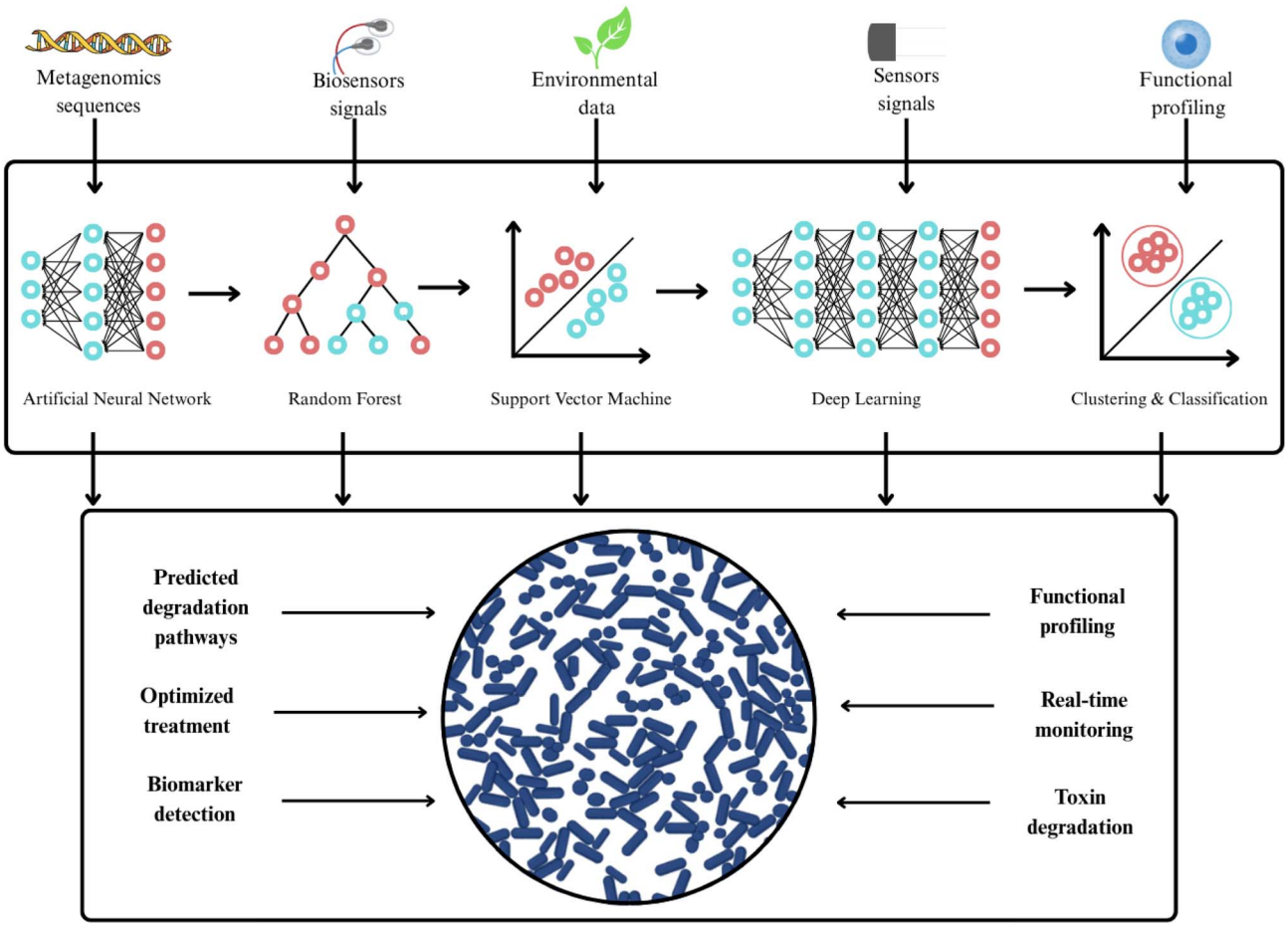

Fig. 2 Integration of machine learning techniques in microbiome analysis for environmental and clinical applications.

responsible for degradation which helps optimization of bioremediation processes.[1,23] Biomarkers like genes, metabolites and proteins can indicate the activity and presence of specific microbial communities capable of degrading pollutants. Algorithms like SVMs, ANNs and RF can identify potential biomarkers by recognizing correlations between the microbial activities and pollutant concentrations as indicated by two independent research groups working at different timelines.[24,25] These integration strategies of ML with microbiome analysis are shown in Fig. 2.

3 Bioinformatics and metagenomics

Bioinformatics aids in using microarray data by enhancing the structural characterization of proteins.[26] In a recent study[27] a comprehensive protein library was constructed that supports heavy metal bio-removal which was modeled through AlphaFold2. Bioinformatics can help in *in silico* studies and analysis of the data and can also enhance bioremediation using the databases for gene identification and microbial degradation pathways of compounds Table 2. Tools such as I-TASSER, Phyre2 and SWISS-MODEL can also be used for protein structural prediction which is often the initial step for detection of active sites to determine enzymatic function. Other tools like CASTp are used for automated detection of active sites. Bioinformatics tools such as PathPred and the University of Minnesota Pathway Prediction System (UMPPS) are freely available tools that provide users with a variety of biochemical reactions ultimately leading to modification of pathways.[4] Tools like BLAST are alignment tools applied for the identification of resemblances among sequences of both protein and structures based on the hypothesis that homologous sequences are expected to function similarly. Other tools like the genome scale metabolic model (GSMM) and constraint based reconstruction and analysis (COBRA) use genetic information from databases for construction of metabolic pathways, while KEGG, BioCyc and others provide all the probable metabolic pathways.

Complete wide range analysis of bacterial communities will play an important role in identification of new genes and metabolic pathways for bioremediation. Bioinformatics approaches have increased our capability to detect pollution sources and screen fluctuations in microorganisms throughout the process. An overview of successful studies where AI can be successfully integrated into bioremediation models is shown in Table 1. Additionally, metagenomics is a critical tool in bioremediation and provides an insightful grasp of the structural and functional properties of the bacterial communities that are engaged in bioremediation processes as presented in Table 2.

Table 1 AI application in different bioremediation approaches, leveraging the capabilities of AI algorithms, to improve efficiency, accuracy, and sustainability

AI application	Bioremediation approach	Ref.
Heavy metal removal	Constructed wetlands	16 and 28
Microbial selection	Bioaugmentation	7 and 16
Pollution monitoring	Data driven monitoring	29 and 30
Nutrient supplementation	Biostimulation	31 and 32
Real-time adaptive control	Dynamic bioremediation	33
Wastewater treatment	Optimized constructed wetlands	31
Nanotechnology integration	Enhanced bioremediation with nanoparticles	7

Table 2 Tools used in metagenomic analysis enabling the study of microbial communities in polluted environments to identify and understand the microorganisms involved in pollutant degradation[34,35]

Name of the software	Application areas in bioremediation
MetagenomeSeq	Evaluation of the abundance of 16S rRNA genes in meta-profiling
UCLUST	A clustering tool, which utilizes USEARCH to allocate sequences to clusters
Mothur	Used in the quality analysis of reads for taxonomic classification
NGSQC toolkit	Method of performing quality control analysis in a direct environment
RDP (Ribosomal Database Project)	Biodiversity analysis, sequence arrangement, alignment, trimming, and taxonomic classification of sequences
Pfam	A large collection of families and domains expressed using profile HMMs and multiple sequence alignments
Prodigal	Identification of translation initiation sites in prokaryotic genes
CAMERA	A server for a metagenomic database containing sequences from environmental samples collected during the GOS
envDB	Prokaryotic taxa environmental distribution database and a tool server
myPhyloDB	A tool used for the purpose of storage and metagenomic analysis
FUNGIpath	A database used for metagenomic and pathological studies of fungi
PyNAST	Aligned sequences of representative OTUs
Meta MIS	Analysis of microbial interaction
FOAM	Created to screen environmental metagenomic sequencing datasets and to offer a novel functional ontology specialized in categorizing gene functions pertinent to environmental microbes using HMMs

This comprises the direst isolation of the genome from the sources to identify composition without any process of isolation and cultivation in the laboratory. Variability in sequencing technologies and biased datasets can limit model generalizability across diverse contaminated sites. Models trained on specific datasets may not perform well under unseen environmental conditions due to site-specific factors like pH or microbial diversity.

The workflow of metagenomics during the extraction of the DNA from the sources and construction of a genomic library, analysis and sequencing of the data to target genes for further application is shown in Fig. 3. The online accessibility of Metagenomic Analysis Shell for Microbial Communities (SmashCommunity), Meta Genome Analyzer (MEGAN), Metagenomic Rapid Annotation using Subsystems Technology (MG-RAST) and Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) equips scientists with cutting-edge approaches for metagenomic based studies. Metagenomics is a strong tool for exposing diverse bacterial species and their importance, but it also has some limitations such as a lack of ideal guidelines for each specific bacterial species. There are several issues with the quality of metagenomics because of longer reads, large data and numerous error models. Similar datasets can be interpreted differently by distinct sequencing technologies. Amplicon based sequencing suffers from biased amplification and is unable to amplify unknown regions. High quality datasets are crucial for training AI and ML algorithms; incomplete and biased data can lead to inaccurate predictions.

4 Real world applications

Real-time applications in evaluations have involved hybrid techniques by merging two ML techniques including ANN-RF and SVM-RF.[9,10] Four-layer ANNs, four-layer RNNs, typical adaptive neuro-FISs and typical CNNs are the four most common neural network structures used in water treatment and monitoring as presented in Fig. 4. ANNs are the most preferred in comparison with other algorithms such as Genetic Algorithms (GAs) and SVMs. ML models predict the output percentage of the absorbate, absorption capacity, effluent concentrations, and water quality parameters.

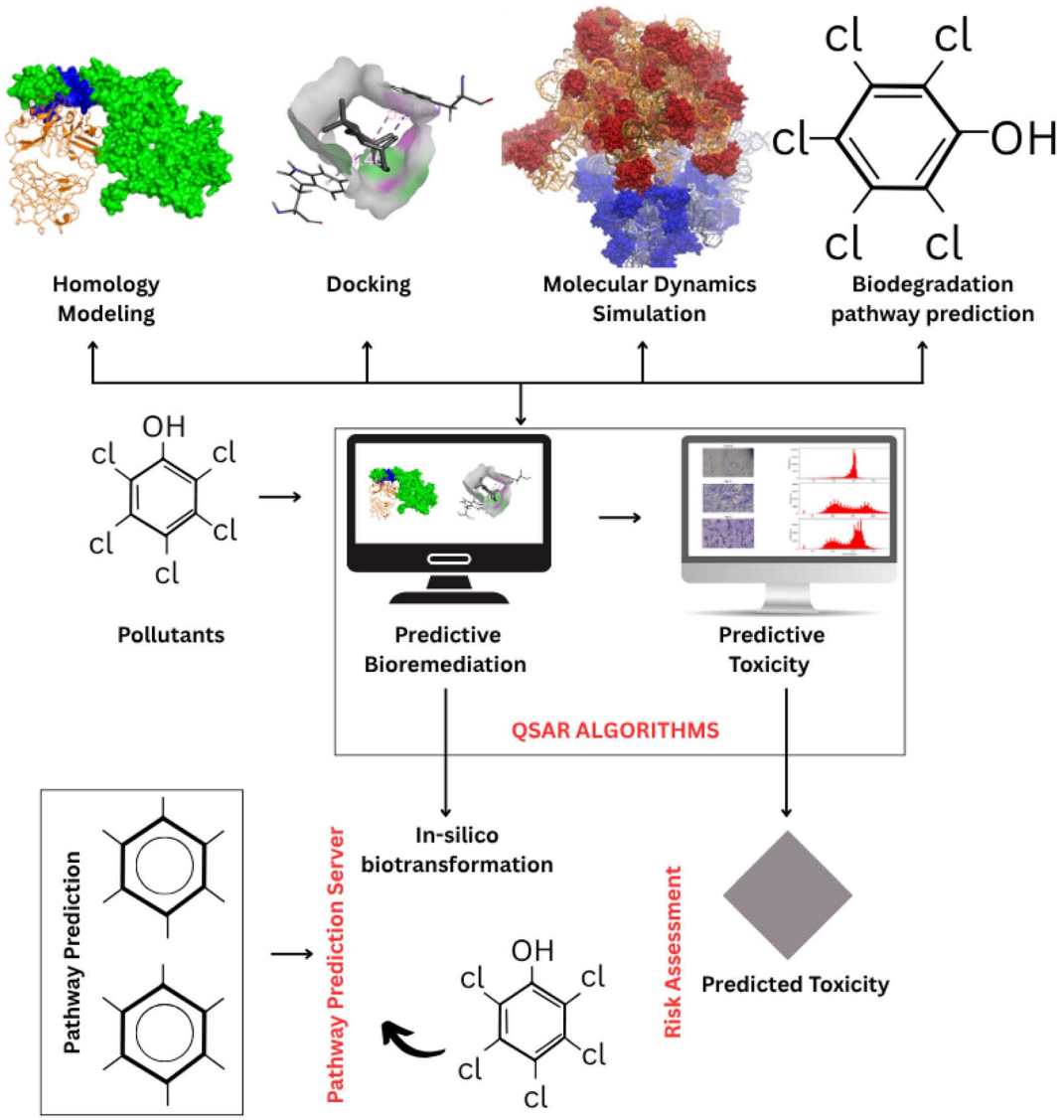

Fig. 3 Graphical illustration of different computational techniques used in predictive bioremediation.

Other models with notable success are ANFISs, SVMs and RF, generally achieving R^2 values of more than 0.9 and in some cases greater than 0.99. Another process known as membrane filtration which refers to separation of contaminants using barriers or filters can be optimized using ML techniques. Like the previous process, ANNs are the most dominant models among RNNs, SVMs and ANNs. ML techniques like Recurrent Neural Networks (RNNs), Random Forest (RF), Support Vector Machines (SVMs), Artificial Neural Networks (ANNs) and Fuzzy Interference Systems (FISs) are shown in Table 3.

One study reported the use of three different models for the prediction of copper removal in an absorption process using clay as the primary adsorbent.[2] These three ML models include ANNs, SVMs and RF developed using open-source software for programming language *R*. The dataset was divided into two parts for training and testing with a ratio of 8 : 2 providing 80% for the training and 20% for the testing. The ANN consisted of a four layered model having one input and output layer and two hidden layers, each neuron relying on the linear output function. The RF model utilizes 76 samples to develop decision trees and an SVM model was developed using a linear kernel. RF and ANN models showed the best performance in terms of accuracy achieving greater than 0.99, while SVMs achieved 0.93.

Studies on water quality management use many models such as ANNs, ANFISs, RNNs, and RF. The ANFIS has outperformed typical ANNs and SVMs and in some cases it was outperformed by hybrid models. ANN and ANFIS models achieved R^2 values greater than 0.999 with both models forecasting the water level.[46,47] Although the studies compared different models, they lack comparisons to simpler baseline models (*e.g.*, linear regression or traditional statistical methods) to contextualize AI model performance.

Baseline comparisons would clarify whether complex AI models provide significant improvements over simpler approaches, justifying their computational cost. In other studies authors compared both ANFIS and ANN models for

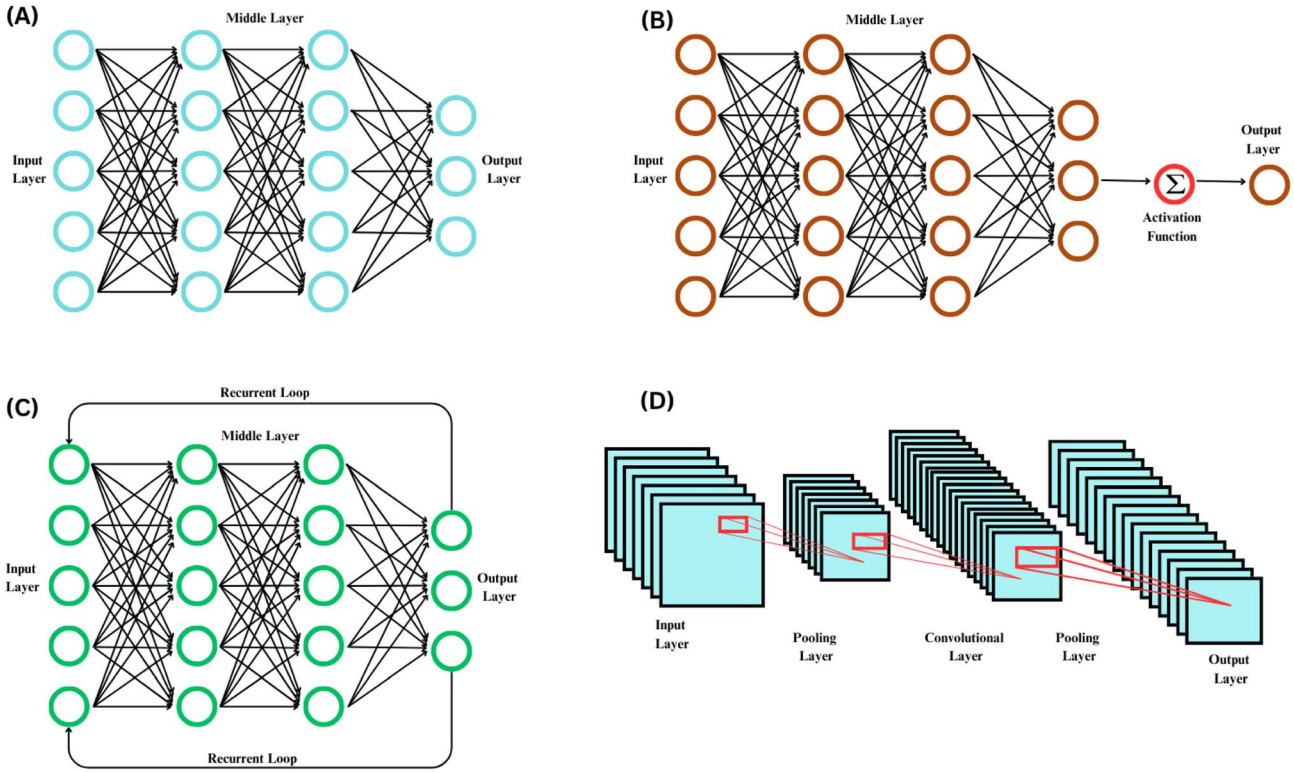

Fig. 4 Illustration of different neural network architectures used in machine learning and deep learning applications for bioremediation studies. (A) Four-layered ANN, (B) four-layered ANFIS, (C) four-layered RNN and (D) four-layered CNN.

estimation of the Water Quality Index (WQI). Both the models showed relative success by achieving accuracy greater than 0.99.[48,49] Federated learning technology and edge cloud-based servers are also used to develop an automated system for prediction and monitoring as presented in Fig. 5.[50] High R^2 values (*e.g.*, >0.99) suggest potential overfitting, especially for complex models like ANNs and hybrid ANN-RF. Overfitting occurs when models fit training data too closely, failing to generalize to new data. The use of an 80 : 20 train-test split is a basic validation approach, but it may not detect overfitting in

Table 3 Instances of successful implementation of ML models used in bioremediation, water treatment and monitoring

Techniques	Applications	Water treatment applications	Ref.
Random forest	Regression, classification and SVMs	Adsorption percent removal and dissolved oxygen modeling	36
Support vector machines/ regressions	SVMs, regression, classification/ pattern analysis	Membrane process parameter, biological oxygen demand (BOD) and chemical oxygen demand (COD) modelling	37 and 38
Fuzzy inference system	Regression, classification and stochastic algorithms	Disinfection by-product modeling	39
Artificial neural network	Regression, classification and SVMs	Adsorption process modeling and dissolved oxygen concentration modeling	40
k-Nearest neighbor	Classification and SVMs	Aquaponics growth stage classification	41 and 42
Recurrent neural network/long short-term memory	Regression, classification and SVMs	Membrane process parameter and dissolved oxygen concentration modeling	43
Adaptive neuro-fuzzy inference systems	Regression, classification and SVMs	Membrane process parameter, biological oxygen demand (BOD) and chemical oxygen demand (COD) modelling	44
Convolutional neural network	Regression, classification and SVMs	Adsorption process modeling	45

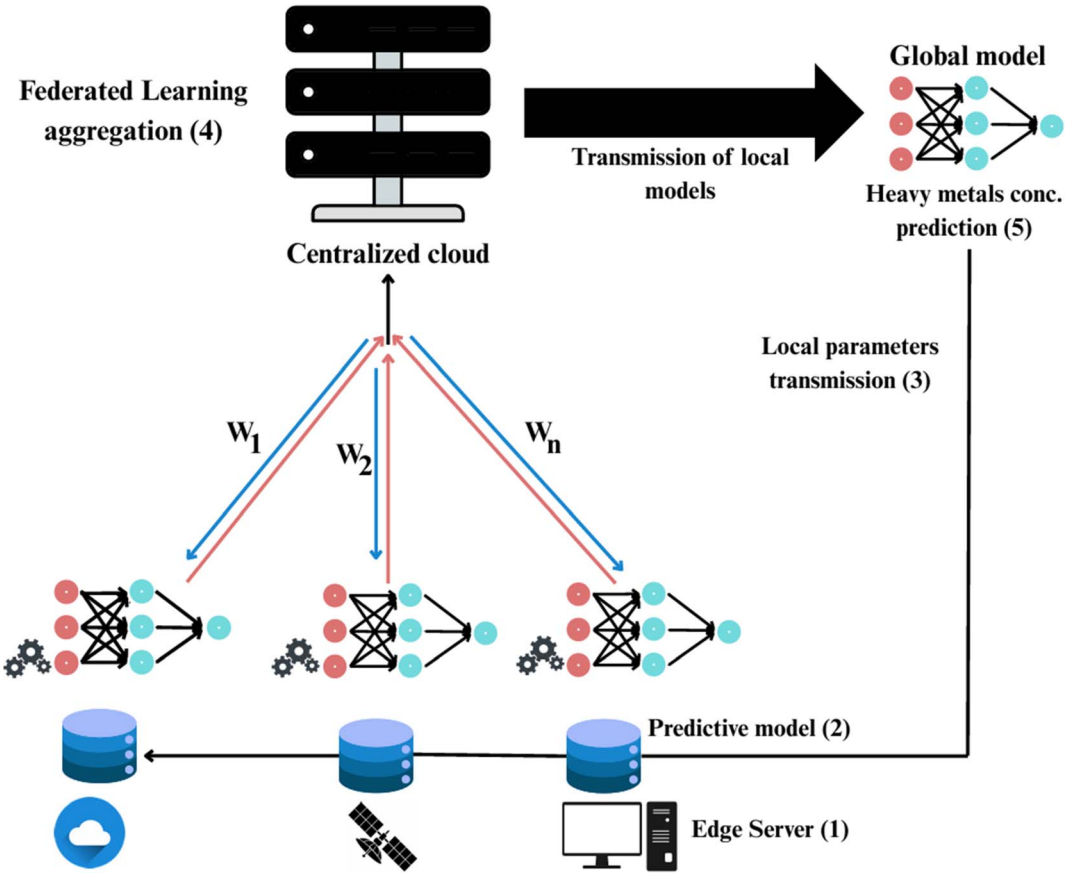

Fig. 5 Federated learning system for predicting heavy metal concentrations using decentralized edge models aggregated into a global model.

small or non-diverse datasets which is common in bioremediation studies. Techniques like *k*-fold cross-validation or regularization could mitigate overfitting.

To illustrate the practical impact of AI-driven bioremediation, two case studies demonstrate successful implementations at meaningful scales. First, a municipal wastewater treatment facility in Hyderabad, India, employed an ANN-RF hybrid model to optimize microbial degradation of heavy metals (*e.g.*, cadmium and lead) in effluent from industrial discharges. Using metagenomic data analyzed *via* QIIME (Table 2), the model identified key microbial taxa and predicted optimal bioaugmentation strategies, achieving a 95% reduction in heavy metal concentrations across 10 000 m^3 of wastewater daily. Real-time biosensors, integrated with deep learning algorithms (Fig. 1), enabled continuous monitoring, ensuring stable performance over six months. Second, a field-scale bioremediation project at a chromium-contaminated mining site in Odisha, India, utilized RF models to predict microbial survival under varying pH and temperature conditions. By integrating metagenomic insights from MG-RAST (Table 2), the project enhanced bio-stimulation, reducing chromium levels by 90% across a 5 hectare site over one year. These cases highlight the scalability of AI-driven approaches, leveraging models like ANN-RF and RF (Table 3) to address large-scale environmental challenges, though site-specific recalibration remains critical for sustained success.

In a study, authors utilized a conventional response surface methodology and an ANN model for enhancing fluoranthene degradation by *Mycobacterium litorale*. The study used optimized $CaCl_2$, KH_2PO_4 and NH_4NO_3. The designed model maximized fluoranthene degradation. It was obtained that the model could efficiently simulate the degradation process and the output received from the ANN model was reliable, precise and reproducible. The authors claimed that the degradation on the 3rd day was better with 51.28% in comparison to that obtained with an unoptimized degradation method which was only 26.37% after 7 days.[51]

5 Challenges and prospects

To fully realize the promises of multi-omics integration in systems biology, researchers will need to address some challenges which include documentation and integration of data collection and analytical methods and establishment of databases on metabolites and pathways. CRISPR and gene editing are some novel technologies that offer the possibility of optimizing bacterial metabolism with high accuracy. The overall goal is moving towards a more efficient, cost-effective, and sustainable solution for bioremediation. For training AI and ML, there is a requirement of high quality and comprehensive datasets as incomplete and biased data will lead to incorrect predictions. Database availability is another major concern for

AI and ML; existing databases may not cover full diversity of microbial communities and their metabolites.

6 Practical limitations of deploying AI-driven bioremediation systems

Deployment of AI driven bioremediation systems especially involvement of gene editing tools like CRISPR must comply with environmental regulations. For instance, regulations like EPA guidelines or the European Union's genetic modification directives have strict controls on releasing genetically modified organisms (GMOs) to the environment to prevent ecological imbalances. AI systems relying on the metagenomic data involve handling large datasets which may raise concerns. Regulatory frameworks like the General Data Protection Regulation (GDPR) in Europe could complicate data management. Also, we noticed a lack of ideal guidelines for metagenomic analysis (*e.g.*, variability in sequencing technologies). This absence of standardized protocols for AI model validation and deployment in bioremediation can hinder regulatory approval, as agencies may require consistent, reproducible methodologies. Use of advanced models like ANNs, RF and SVMs generally requires significant computational power but in developing countries access to high performance computational infrastructure or cloud-based servers may be limited, increasing the costs and reducing the scalability. Implementing AI-driven systems, including biosensors and real-time monitoring tools, requires substantial investment in hardware, software, and maintenance. Small-scale or rural bioremediation projects may lack funding, as the manuscript notes that no funding was received for this work, reflecting a broader challenge in resource allocation. Apart from these, there are other challenges that are related to data availability and quality, such as the need for high-quality datasets for training AI models. In real-world settings, collecting comprehensive, unbiased data from diverse contaminated sites is resource-intensive, requiring specialized equipment and trained personnel, which may not be feasible in low-resource environments. Operating AI-driven systems, such as those integrating metagenomics or machine learning models, demands expertise in bioinformatics, data science, and environmental microbiology. The manuscript's reliance on complex tools like AlphaFold2 and I-TASSER suggests a steep learning curve, which may not be practical in regions with limited access to trained professionals. Real-time monitoring systems, such as AI-powered biosensors, require ongoing maintenance and troubleshooting. Without onsite expertise, system failures or misinterpretations of AI outputs could compromise bioremediation efficiency. Deploying these systems in non-ideal settings necessitates training local personnel, which can be time-consuming and costly. The manuscript's call for interdisciplinary collaboration underscores the need for knowledge transfer, but this is challenging in areas with limited educational infrastructure.

Fundamental limitations of AI approaches in environmental contexts pose significant challenges to their widespread adoption in bioremediation. Data scarcity, due to the high cost and complexity of collecting diverse metagenomic and environmental datasets, restricts the training of robust AI models like ANNs and RF, particularly for rare pollutants or understudied microbial communities. Environmental heterogeneity, characterized by variations in site-specific factors such as pH, temperature, and pollutant profiles, limits the generalizability of models trained on specific datasets, as seen in applications like heavy metal removal (Table 1). Furthermore, scaling laboratory results, such as those from AlphaFold2 or QIIME analyses, to field applications is hindered by uncontrolled environmental variables and the need for real-time model recalibration, necessitating robust validation strategies and adaptive algorithms to ensure reliable performance across diverse ecological settings.

Expansion of coverage is required for a more complete understanding of microbial metabolism.[52] A further major difficulty is that distinct sequencing technologies interpret the same datasets differently, and missing environmental context limits application of supervised ML. There is still a need for wet-lab validation because of the limited biodegradation database availability. Researchers can combine more sensitive tools to be used for these challenging tasks, such as advanced structural elucidation and optimization. Enhancing multi-omics and genetic engineering tools can help develop a more sustainable bioremediation strategy. Further refinements in metagenomics and bioinformatics will help in precise ecological interpretations. These omics approaches can anticipate microbial metabolism at contaminated sites; therefore, utilizing multi-omics approaches can lead to new hypothesis and theories. Another area of investigation would be an experimental approach from a multidisciplinary perspective to achieve better prediction and application.

Generated data might not be utilized in different ML techniques to test effectiveness. Another major challenge is management of several hazardous organic compounds which can be reduced by the development of new monitoring techniques. Most of the literature on metagenomics focuses on a limited range of enzymes, mostly esterases and oxidases, while the others also play an essential role in biodegradation of pollutants and toxins.[6] Closed loop systems should emphasize the development of closed systems where AI/ML models' predictions are iteratively validated through experimental testing and refinement creating a feedback cycle that will enhance both predictive accuracy and practical applicability.

7 Outlook

In hindsight, bioinformatics tools like AlphaFold, I-TASSER and SWISS-MODEL are used for protein modeling, which is an important step in metagenomics. Recent developments in sequencing technologies have removed the obstacles and opened the door for metagenomics revealing novel information about microbial diversity. Metagenomics can analyze the genetic material of microbial communities to solve problems and discover new enzymes, genes and metabolic pathways. Metagenomics tools like QIIME, UPARSE, and MOTHUR facilitate bioinformatics analysis with MG-RAST and MetaPhlAn2

providing phylogenetic analysis and functional insights. These web-based tools are crucial for advancing omics. AI and ML have shown potential to solve complex issues faced during the bioremediation process. ML algorithms like ANNs, CNNs, RNNs and RF have achieved an accuracy of more than 0.99. ML models have optimized, predicted, modeled and automated some applications of bioremediation and its strategies. To advance AI driven bioremediation researchers should prioritize standardized protocols that integrate metagenomics with ML frameworks for accurate and reproducible analysis. There is a need for robust publicly available datasets representing diverse contaminated environment data to enhance generalizability. Enhancements of bioinformatics pipelines such as QIIME and MG-RAST are required to manage noise in the datasets and integration of protein structure prediction tools is needed to accelerate novel discoveries.

ML algorithms have shown better results in prediction, modelling and optimization of water treatment processes to degrade pollutants and toxins. Though many studies reviewed with success, there are sets of challenges and limitations which include data management, reproducibility and transparency that must be addressed. Interdisciplinary collaborations integrating bioinformatics, metagenomics and machine learning will be pivotal in overcoming these challenges. By bridging the gaps between these interdisciplinary collaborations future research can be transformative in bioremediation ultimately contributing to more effective bioremediation strategies and environmental solutions.

Consent for publication

All authors agree to this submission.

Conflicts of interest

The authors declare no known competing interests.

Data availability

Data will be made available from the corresponding author upon reasonable request.

References

1 S. Haque, N. Srivastava, D. B. Pal, M. F. Alkhanani, A. H. Almalki, M. Y. Areeshi, *et al.*, Functional microbiome strategies for the bioremediation of petroleum-hydrocarbon and heavy metal contaminated soils: A review, *Sci. Total Environ.*, 2022, **833**, 155222.

2 S. K. Bhagat, K. Pyrgaki, S. Q. Salih, T. Tiyasha, U. Beyaztas, S. Shahid, *et al.*, Prediction of copper ions adsorption by attapulgite adsorbent using tuned-artificial intelligence model, *Chemosphere*, 2021, **276**, 130162.

3 R. Bhattacharya, D. Bose, P. Ganti, A. Rizvi, G. Halder and A. Sarkar, Bioelectricity production and bioremediation potential of Withania somnifera in plant microbial fuel cells with food wastes as enrichment, *Energy Nexus*, 2024, **15**, 100314.

4 L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, *et al.*, Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0, *Nat. Protoc.*, 2019, **14**(3), 639–702.

5 J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, *et al.*, Recent advances in convolutional neural networks, *Pattern Recognit.*, 2018, **77**, 354–377.

6 C. Firincă, L. G. Zamfir, M. Constantin, I. Răut, M. L. Jecu, M. Doni, *et al.*, Innovative Approaches and Evolving Strategies in Heavy Metal Bioremediation: Current Limitations and Future Opportunities, *J. Xenobiot.*, 2025, **15**(3), 63.

7 R. Patowary, A. Devi and A. K. Mukherjee, Advanced bioremediation by an amalgamation of nanotechnology and modern artificial intelligence for efficient restoration of crude petroleum oil-contaminated sites: a prospective study, *Environ. Sci. Pollut. Res.*, 2023, **30**(30), 74459–74484.

8 World Health Organization: WHO, Antimicrobial resistance, World Health Organization: WHO, 2023, 2023, November 21, internet, **https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance**.

9 G. Alam, I. Ihsanullah, M. Naushad and M. Sillanpää, Applications of artificial intelligence in water treatment for optimization and automation of adsorption processes: Recent advances and prospects, *Chem. Eng. J.*, 2022, **427**, 130011.

10 Y. Xie, Y. Chen, Q. Lian, H. Yin, J. Peng, M. Sheng, *et al.*, Enhancing Real-Time Prediction of Effluent Water Quality of Wastewater Treatment Plant Based on Improved Feedforward Neural Network Coupled with Optimization Algorithm, *Water*, 2022, **14**(7), 1053.

11 D. Bose, R. Bhattacharya, M. Gopinath, P. Vijay and B. Krishnakumar, Bioelectricity production and bioremediation from sugarcane industry wastewater using microbial fuel cells with activated carbon cathodes, *Results Eng.*, 2023, **18**, 101052.

12 E. Raper, T. Stephenson, D. R. Anderson, R. Fisher and A. Soares, Industrial wastewater treatment through bioaugmentation, *Process Saf. Environ. Prot.*, 2018, **118**, 178–187.

13 R. Bhattacharya, D. Bose, J. Yadav, B. Sharma, E. Sangli, A. Patel, *et al.*, Bioremediation and bioelectricity from Himalayan rock soil in sediment-microbial fuel cell using carbon rich substrates, *Fuel*, 2023, **341**, 127019.

14 N. Nivetha, B. Srivarshine, B. Sowmya, M. Rajendiran, P. Saravanan, R. Rajeshkannan, *et al.*, A comprehensive review on bio-stimulation and bio-enhancement towards remediation of heavy metals degeneration, *Chemosphere*, 2023, **312**, 137099.

15 D. Bose, M. Santra, R. V. S. P. Sanka and B. Krishnakumar, Bioremediation analysis of sediment- microbial fuel cells for energy recovery from microbial activity in soil, *Int. J. Energy Res.*, 2021, **45**(4), 6436–6445.

16 P. K. Gupta, B. Yadav, A. Kumar and S. K. Himanshu, Machine learning and artificial intelligence application in

constructed wetlands for industrial effluent treatment: advances and challenges in assessment and bioremediation modeling, in *Bioremediation for Environmental Sustainability*, Elsevier, 2021, pp. 403–414.

17 P. K. Gupta, B. Yadav, A. Kumar and S. K. Himanshu, Machine learning and artificial intelligence application in constructed wetlands for industrial effluent treatment: advances and challenges in assessment and bioremediation modeling, in *Bioremediation for Environmental Sustainability*, Elsevier, 2021, pp. 403–414.

18 J. Wang, C. Peng and X. Liu, Prediction of bacterial microbiota changes in various microplastics-contaminated environments based on machine learning, *J. Environ. Chem. Eng.*, 2025, **13**(5), 117461.

19 Y. Yang, A. Jerger, S. Feng, Z. Wang, C. Brasfield, M. S. Cheung, *et al.*, Improved enzyme functional annotation prediction using contrastive learning with structural inference, *Commun. Biol.*, 2024, **7**(1), 1690.

20 H. A. Oyewusi, R. A. Wahab, K. A. Akinyede, G. M. Albadrani, M. Q. Al-Ghadi, M. M. Abdel-Daim, *et al.*, Bioinformatics analysis and molecular dynamics simulations of azoreductases (AzrBmH2) from Bacillus megaterium H2 for the decolorization of commercial dyes, *Environ. Sci. Eur.*, 2024, **36**(1), 31.

21 S. Srinivasan, D. Raajasubramanian, N. Ashokkumar, V. Vinothkumar, N. Paramaguru, P. Selvaraj, *et al.*, Nanobiosensors based on on-site detection approaches for rapid pesticide sensing in the agricultural arena: A systematic review of the current status and perspectives, *Biotechnol. Bioeng.*, 2024, **121**(9), 2585–2603.

22 M. S. Ayilara and O. O. Babalola, Bioremediation of environmental wastes: the role of microorganisms, *Front. Agron.*, 2023, **5**, 1–15.

23 J. Wijaya, J. Park, Y. Yang, S. I. Siddiqui and S. Oh, A metagenome-derived artificial intelligence modeling framework advances the predictive diagnosis and interpretation of petroleum-polluted groundwater, *J. Hazard. Mater.*, 2024, **472**, 134513.

24 S. Vikram, L. D. Guerrero, T. P. Makhalanyane, P. T. Le, M. Seely and D. A. Cowan, Metagenomic analysis provides insights into functional capacity in a hyperarid desert soil niche community, *Environ. Microbiol.*, 2016, **18**(6), 1875–1888.

25 S. Zhong, K. Zhang, M. Bagheri, J. G. Burken, A. Gu, B. Li, *et al.*, Machine Learning: New Ideas and Tools in Environmental Science and Engineering, *Environ. Sci. Technol.*, 2021, 1–14.

26 D. Chettri, A. K. Verma, M. Chirania and A. K. Verma, Metagenomic approaches in bioremediation of environmental pollutants, *Environ. Pollut.*, 2024, **363**, 125297.

27 T. Uttarotai, N. Mukjang, N. Chaisoung, W. Pathom-Aree, J. Pekkoh, C. Pumas, *et al.*, Putative Protein Discovery from Microalgal Genomes as a Synthetic Biology Protein Library for Heavy Metal Bio-Removal, *Biology*, 2022, **11**(8), 1226.

28 P. P. Biswas, W. H. Chen, S. S. Lam, Y. K. Park, J. S. Chang and A. T. Hoang, A comprehensive study of artificial neural network for sensitivity analysis and hazardous elements sorption predictions *via* bone char for wastewater treatment, *J. Hazard. Mater.*, 2024, **465**, 133154.

29 A. K. Wani, F. Rahayu, I. Ben Amor, M. Quadir, M. Murianingrum, P. Parnidi, *et al.*, Environmental resilience through artificial intelligence: innovations in monitoring and management, *Environ. Sci. Pollut. Res.*, 2024, **31**(12), 18379–18395.

30 H. Tao, Z. S. Al-Khafaji, C. Qi, M. Zounemat-Kermani, O. Kisi, T. Tiyasha, *et al.*, Artificial intelligence models for suspended river sediment prediction: state-of-the art, modeling framework appraisal, and proposed future research directions, *Eng. Appl. Comput. Fluid Mech.*, 2021, **15**(1), 1585–1612.

31 S. Sahu, A. Kaur, G. Singh and A. S. Kumar, Harnessing the potential of microalgae-bacteria interaction for eco-friendly wastewater treatment: A review on new strategies involving machine learning and artificial intelligence, *J. Environ. Manage.*, 2023, **346**, 119004.

32 S. E. Bibri, J. Krogstie, A. Kaboli and A. Alahi, Smarter eco-cities and their leading-edge artificial intelligence of things solutions for environmental sustainability: A comprehensive systematic review, *Environmental Science and Ecotechnology*, 2024, **19**, 100330.

33 N. B. Chang, G. Mohiuddin, A. J. Crawford, K. Bai and K. R. Jin, Diagnosis of the artificial intelligence-based predictions of flow regime in a constructed wetland for stormwater pollution control, *Ecol. Inform.*, 2015, **28**, 42–60.

34 R. N. Bharagava, D. Purchase, G. Saxena and S. I. Mulla, Applications of Metagenomics in Microbial Bioremediation of Pollutants, in *Microbial Diversity in the Genomic Era*, Elsevier, 2019, pp. 459–477.

35 P. Arya, Ravindra. Metagenomics based approach to reveal the secrets of unculturable microbial diversity from aquatic environment, in *Recent Advancements in Microbial Diversity*, Elsevier, 2020, pp. 537–559.

36 Y. Liu, Y. Wang and J. Zhang, *New Machine Learning Algorithm*: Random Forest, 2012, pp. 246–252, DOI: 10.1007/978-3-642-34062-8_32.

37 C. Cortes and V. Vapnik, Support-vector networks, *Mach. Learn.*, 1995, **20**(3), 273–297.

38 K. S. Chua, Efficient computations for large least square support vector machine classifiers, *Pattern Recognit. Lett.*, 2003, **24**(1–3), 75–80.

39 C. Moraga, E. Trillas and S. Guadarrama, Multiple-Valued Logic and Artificial Intelligence Fundamentals of Fuzzy Control Revisited, in *Artificial Intelligence in Logic Design*, Springer Netherlands, Dordrecht, 2004, pp. 9–37.

40 R. E. Uhrig, Introduction to artificial neural networks, in *Proceedings of IECON '95 - 21st Annual Conference on IEEE Industrial Electronics*, IEEE, pp. , pp. 33–37.

41 H. Samet, K-Nearest Neighbor Finding Using MaxNearestDist, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008, **30**(2), 243–252.

42 L. Jiang, Z. Cai, D. Wang and S. Jiang, Survey of Improving K-Nearest-Neighbor for Classification, in *Fourth International*

Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), IEEE, 2007, pp. 679–683.

43 R. DiPietro and G. D. Hager, Deep learning: RNNs and LSTM, in *Handbook of Medical Image Computing and Computer Assisted Intervention*, Elsevier, 2020, pp. 503–519.

44 J. Farhoudi, S. M. Hosseini and M. Sedghi-Asl, Application of neuro-fuzzy model to estimate the characteristics of local scour downstream of stilling basins, *J. Hydroinform.*, 2010, **12**(2), 201–211.

45 J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, *et al.*, Recent advances in convolutional neural networks, *Pattern Recognit.*, 2018, **77**, 354–377.

46 M. Al-Yaari, T. H. H. Aldhyani and S. Rushd, Prediction of Arsenic Removal from Contaminated Water Using Artificial Neural Network Model, *Appl. Sci.*, 2022, **12**(3), 999.

47 H. Mazaheri, M. Ghaedi, M. H. Ahmadi Azqhandi and A. Asfaram, Application of machine/statistical learning, artificial intelligence and statistical experimental design for the modeling and optimization of methylene blue and Cd(ii) removal from a binary aqueous solution by natural walnut carbon, *Phys. Chem. Chem. Phys.*, 2017, **19**(18), 11299–11317.

48 M. S. Mazloom, F. Rezaei, A. Hemmati-Sarapardeh, M. M. Husein, S. Zendehboudi and A. Bemani, Artificial Intelligence Based Methods for Asphaltenes Adsorption by Nanocomposites: Application of Group Method of Data Handling, Least Squares Support Vector Machine, and Artificial Neural Networks, *Nanomaterials*, 2020, **10**(5), 890.

49 Y. Mesellem, A. A. El Hadj, M. Laidi, S. Hanini and M. Hentabli, Computational intelligence techniques for modeling of dynamic adsorption of organic pollutants on activated carbon, *Neural Comput. Appl.*, 2021, **33**(19), 12493–12512.

50 Z. M. Yaseen, The next generation of soil and water bodies heavy metals prediction and detection: New expert system based Edge Cloud Server and Federated Learning technology, *Environ. Pollut.*, 2022, **313**, 120081.

51 D. R. Dudhagara, R. K. Rajpara, J. K. Bhatt, H. B. Gosai and B. P. Dave, Bioengineering for polycyclic aromatic hydrocarbon degradation by Mycobacterium litorale: Statistical and artificial neural network (ANN) approach, *Chemom. Intell. Lab. Syst.*, 2016, **159**, 155–163.

52 S. Khanna and A. Kumar, Bioinformatics Toward Improving Bioremediation, in *Biotechnological Innovations for Environmental Bioremediation*, Springer Nature Singapore, Singapore, 2022, pp. 631–669.