International Journal of Sustainable Development Through AI, ML and IoT

Volume 4 | Issue 1 | 2025

ISSN(Online):2584-0827

https://ijsdai.com/index.php/IJSDAI/index

Environmental Ecology and Human Health: A Review of Data-Driven Approaches for Water Quality Evaluation Using Machine Learning

Divpreet Singh
Student, BCA, VIPS-TC, New Delhi
bca_03117702024_divpreet@vipstc.edu.in

Ayushi Kapoor
Student, BCA, VIPS-TC, New Delhi
bca_02417702024_ayushi@vipstc.edu.in

Gagan Jha
Student, BCA, VIPS-TC, New Delhi
bca_03217702024_gagan@vipstc.edu.in

Samadrita Mukherjee
Student, BCA, VIPS-TC, New Delhi
bca_05917702024_samadrita@vipstc.edu.in

Sanjay Prasad Yadav
Student, BCA, VIPS-TC, New Delhi
bca_02717702024_sanjay@vipstc.edu.in

ARTICLE INFO

Received: 01st June 2025
Revised: 17th June 2025
Accepted: 20th July 2025

ABSTRACT

The volume of data related to the aquatic environment has rapidly increased, and machine learning has become an essential tool for data analysis, classification, and prediction. While conventional models used in water-related research tend to be more mechanistic in nature, data-driven machine learning models may be able to solve more complex nonlinear problems efficiently. However, we note that the use of machine learning models and findings has been applied in water environment research to design, monitor, simulate, evaluate, and optimize management systems. ML also contributes to controlling water pollution, improving water quality, and watershed ecosystem security. Machine learning algorithms are well-developed, robust statistical tools that have been applied to many complex problems, including the assessment of different types of water quality in surface water, groundwater, drinking water, sewage, and seawater. we present future uses for machine learning algorithms in aquatic environments.

Introduction

Polluted wastewater generated by rapid economic development is a direct threat to natural

water ecosystems. This has led to the development of a multitude of strategies to combat water pollution. Pollution is a major environmental hazard faced by humanity, and the analysis and assessment of water quality have greatly enhanced the efficiency of pollution control. Different methodologies have emerged globally to assess water quality such as multivariate statistical, fuzzy inference, and water quality index (WQI). [1] Even following accepted protocols, the results of water quality assessments can differ significantly from one experience to the next due to the specific parameter evaluated. Evaluating every water quality parameter is impracticable, given high cost, technical complexity, and variability of water quality parameters. [2] The advancements in machine learning over the last decade have resulted in a degree of optimism among academia that large quantities of data may in the future be procured and analyzed to meet challenging water quality assessment requirements.

Machine learning algorithms, originated from artificial intelligence, [10] analyze data to predict new data. It is one of the most well-known methods for data analysis and processing owing to its precision, flexibility, and extensibility. Machine learning is an approach to model nonlinear relations among grouped factors and to help discover the true mechanisms. [10] The versatility of machine learning, paired with its promise as a tool for environmental research and engineering, enables unprecedented opportunities for analyzing complex environmental issues. Although complicated, the application of machine learning in water quality analysis and assessment could make it more accurate. Drinking water, wastewater, groundwater, surface water, ocean, freshwater, and other kinds of water are complex. [3] Different types of water such as lake, river, and groundwater are different from one another and they have different complexities which make it difficult to conduct research regarding the quality of the water as they need to be treated differently. Based on previous research, these challenges can be addressed by machine learning. Here, we evaluate the advantages and disadvantages of commonly used [23] machine learning algorithms, their application, and performance in a range of water systems (Fig. 1).

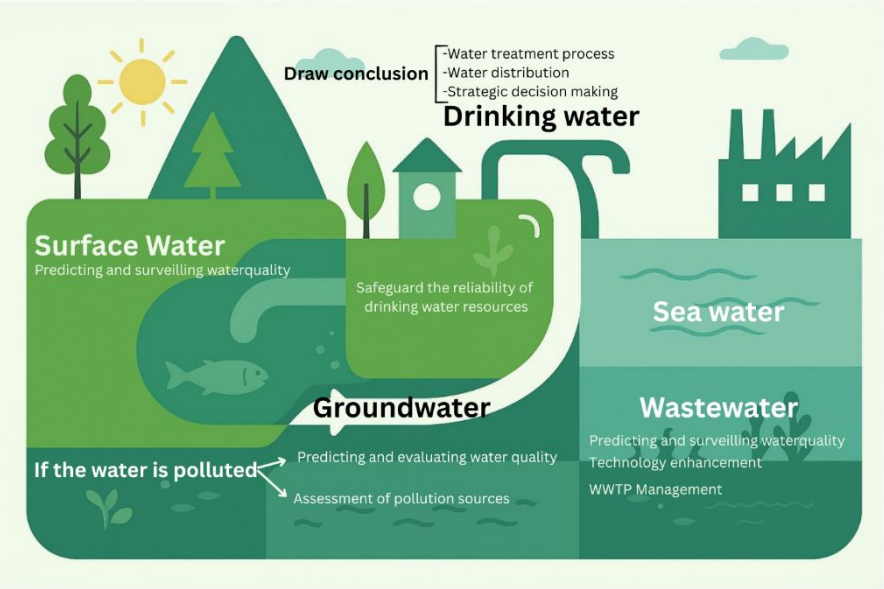

Fig. 1 illustrates the extensive application of machine learning in water systems

Overview

Machine learning is a powerful data analysis technique that has become a popular choice for

identifying patterns or making predictions from large datasets generated by a plethora of scenario-based approaches. [2] What testers must do before practicing machine learning must include: data construction, proper algorithm selection, training model, and model validation. Out of these, one of the most critical is algorithm selection. [5] Two Main Types of Machine Learning Technologies: Supervised and Unsupervised Learning. The key difference between these two classes is the presence of labels in the datasets. With supervised learning, we learn functions to make predictions from labeled training datasets. In the case of training data, each instance contains input and expected output values. Supervised learning algorithms [3] search for the relationships between input and output values, in order to create a prediction model that predicts the outcome given the important input data. Many supervised learning methods have been proposed ranging from linear regression to [7] ANN, DT, SVM, naive Bayes, KNN, RF, etc., which can be used for data classification as well as regression.

Unsupervised learning is a pattern recognition problem, and it works without any labels. Unlabeled training datasets are used. Unsupervised learning [23] uses dimensionality reduction and clustering to group the training data. The number of categories is unclear as well as what they exactly are. It is widely used for classification and association mining under the umbrella of unsupervised learning. Commonly used methods in unsupervised machine learning are principal component analysis (PCA) and K-means. [2] Reinforcement learning is a machine learning method that lets machines derive appropriate reactions to unanswered problems. [12] It is the least common class of machine learning to be employed in the topic of the aquatic environment compared to the first two classes.

3. Application of machine learning for different water environments

Machine learning for water treatment and management systems [1] (taken from Gassman et al.), including real-time monitoring, prediction, pollutant source tracking, concentration estimation, resource allocation, and [23]technology optimization

3.1. Applications in surface water

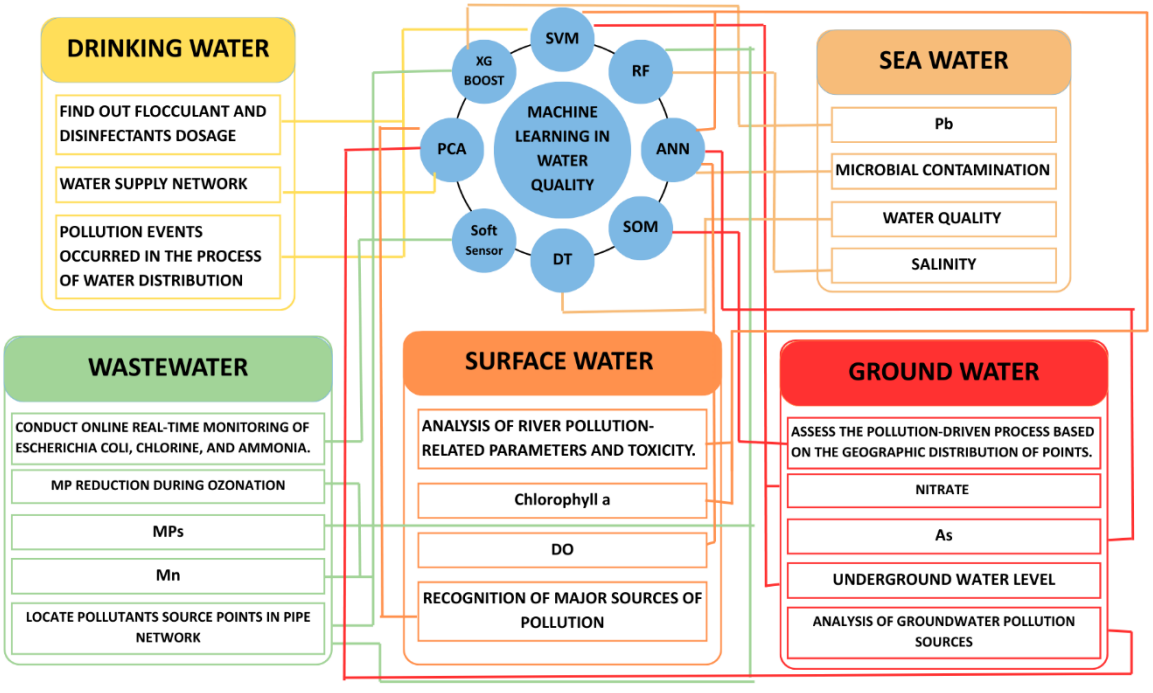

Fig. 2 showcases the diverse applications of machine learning algorithms in different aspects of water treatment and management. The figure highlights commonly used techniques such as support vector machines (SVM), random forests (RF), artificial neural networks (ANN), self-organizing maps (SOM), decision trees (DT), principal component analysis (PCA), and extreme gradient boosting (XGBoost), in addressing key parameters like dissolved oxygen (DO) levels and micropollutant (MP) detection

Wastewater composed of human-originated effluent from municipal and industrial activities is one of the principal sources of urban water quality degradation. Machine learning has been increasingly applied in surface water quality research. [7] Many approaches are present to predict surface water quality and analyze it; Table 1 presents some. The models for machine learning have been optimized, and their prediction accuracy improved.

Getting data is an essential step in building machine learning models. Water quality monitoring data, either [18] compliance-based or at intervals, can provide benchmarks for water system management. Traditional monitoring methods are still heavily relied on by environmental authorities. Conventional means for monitoring in situ are impractical. [3] Remote sensing technologies not only offer the potential for timely and extensive monitoring of water quality but also expose the transport and dispersion of such challenging and elusive contaminants to conventional techniques. Sagan et al., for instance, found that machine learning with experiment-based optimizations allows sophisticated, real-time monitoring sensor data and satellite data. [18] Compared to conventional models, the accuracies of PLS, SVR, and DNN models were more impressive. Several water quality characteristics, such as [7] pathogen concentration, cannot be assessed directly by remote sensing when either no optically active remote sensors currently exist or high-resolution hyperspectral data are lacking. But these variables can be indirectly approximated from other available variables which are measurable. [5] Wu et al. used a convolutional neural network (CNN) in to differentiate between clean and unclean water images. The attentional neural network was applied to a water surface image dataset and proven to work well. [2] With CNNs, we use the reflectance image when we input it directly, so feature engineering and parameter tuning are not needed. A sparse matrix and performance degradation may also be caused if the data is missing, incorrect, or integrated into the supercomputer in a destroyed form, whether due to equipment error or human error.

Task	Algorithms Used	Sample Size	Input Data	Performance	Reference
Predicting Oxygen Levels (DO)	BWNN, ANN, ARIMA, BANN	340	Oxygen levels in the water	BWNN performed the best with 95% accuracy. BANN was just behind with 90%, and ARIMA and ANN were a bit less accurate, with ARIMA at 80% and ANN at 75%.	
Predicting Oxygen Levels (DO)	LSTM	240	Oxygen in the water	LSTM was successful at 70% of the sites, with decent results overall.	
Predicting Oxygen Levels (DO)	PNN	1700	Water quality info like temperature, pH, and nutrients	PNN achieved about 73% accuracy, showing strong correlation with water conditions.	
Predicting Oxygen Levels (DO)	CCNN	210	Oxygen levels, pH, temperature, and other water quality factors	Achieved 80% accuracy with a slight margin of error.	
Predicting Biological Oxygen Demand (BOD)	DNN, SVR, RF	29000	Sea conditions, temperature, and oxygen levels	DNN did about 20% better than other models in terms of accuracy.	
Predicting EC, $HCO3^-$, $SO4^{2-}$	SVM, ANN	Data from 1960s	pH levels, temperature, and general water chemistry	SVM had a small advantage over ANN, with SVM at 85% and ANN at 80%.	
Predicting Total Nitrogen (TN) & Phosphorus (TP)	SVM, ANN	680	River flow, rainfall, oxygen, TN, TP	SVM performed better with 88% accuracy than ANN at 82% accuracy.	
Predicting Water Quality	DT, RF, DCF, and others	31000	Oxygen, ammonia, pH, temperature, and more	Decision Trees and Random Forest hit a great accuracy of 85% and 90%, leading the pack.	
Predicting Nutrients (TRP, $NO3^-N$, TP, $NH4^-N$)	RF	20000	Water temperature, flow, chlorophyll levels, and pH	Reduced error by 50% compared to simpler models, with solid performance.	

Predicting Chlorophyll-a	SVM, ANN	340	Chlorophyll-a levels, temperature, wind speed	SVM showed a slight edge over ANN, with SVM at 88% and ANN at 83%.	
Predicting Algal Blooms	ANFIS	880	Water data like COD, BOD, chlorophyll, nutrients, temperature	ANFIS showed the best results, working well in both predicting and classifying blooms.	
Optimizing Hyperparameters	SVR	230	Chlorophyll-a, turbidity, oxygen levels, pollution	Achieved an accuracy of 76% for chlorophyll and 80% for suspended solids.	
Detecting Water Pollution	Attention Neural Network	920	Water body images	Clean water detection had 67% accuracy, and polluted water had 71%.	
Detecting Water Pollution	CNN, SVM, RF	95	Satellite images and water quality levels	CNN led with 94% accuracy, SVM was at 90%, and RF reached 85%.	
Assessing Heavy Metal Pollution	PCA	40	Metal content (Cu, Pb, Zn, etc.)	PCA successfully identified polluted areas with around 83% accuracy.	
Selecting Water Quality Indicators (WQI)	PCA	220	Oxygen levels, pH, TDS, BOD, nitrate, chloride, temperature	PCA helped identify key water quality parameters like DO and pH, contributing to WQI calculation.	

Table 1 . MACHINE LEARNING IN WATER QUALITY MONITORING

Abbreviations:

- **DO**: Dissolved Oxygen, **BWNN**: Bootstrapped Wavelet Neural Network, **ANN**: Artificial Neural Network, **ARIMA**: Autoregressive Integrated Moving Average, **BANN**: Bootstrapped Artificial Neural Network , **LSTM**: Long Short-Term Memory **PNN**: Polynomial Neural Network **BOD**: Biological Oxygen Demand **COD**: Chemical Oxygen Demand **EC**: Electrical Conductivity , **CCNN**: Cascade Correlation Neural Network **TDS**: Total Dissolved Solids **RMSE**: Root Mean Square Error **DNN**: Deep Neural Network **SVR**: Support Vector Regression **RF**: Random Forest **SVM**: Support Vector Machine **TP**: Total Phosphorus **TN**: Total Nitrogen **TRP**: Total Reactive Phosphorus **TOC**: Total Organic Carbon **TSS**: Total Suspended Solids **DTP**: Dissolved Total Phosphorus **BGA-PC**: Blue-Green Algae Phycocyanin

FDOM: Fluorescent Dissolved Organic Matter **CNN**: Convolutional Neural Network **PCA**: Principal Component Analysis **WQI**: Water Quality Index

There are many different ways of handling the Data cleaning, it includes using of averages and medians, other alternative method is using combination of machine learning and matrix completion to supplement the raw data. We cannot directly use the set of data for Data cleaning. For this, [18] Ma et al suggested an approach in which the DNN (Deep Neural Network) and deep Matrix Factorization (deep MF) is combined together to predict the [24] Biological Oxygen Demand (BOD) . Things not only done by suggesting but also have to be checked with a real-life case. So they went with the verification for this method and verified the validity and reliability of this method using the New York Harbor waters as a case study. Data cleaning enhances the quality of the data and thus enhances the accuracy of machine learning model applications.

In general there are two aspects to which the prediction accuracy is related for the machine learning applications i.e., the model selection and the quality of the training dataset. In machine learning there are two [3] algorithms:- Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) which is used for classification and regression, but they differ in their approach and strength. Both have provided a great excellence performance in predicting the water quality components. There are some cases, where [16] SVM's prediction accuracy is higher and also show higher generalization ability than ANN. The reason is that the optimization of model parameters in neural network is quite unstable, which in return affected the accuracy of the ANN by nonlinear disturbances. SVM becomes more effective than ANN when it comes to minimizing the generalization error as [22] SVM uses an upper bound on the generalization error instead of minimizing the training error. Since, the river system has dynamically changed with more complexity over the period of time , therefore one of the most effective way to manage rivers is the monitoring of water quality in real time, on in the absence of monitoring conditions make predictions based on the other data. Researchers have approved that long short-term memory(LSTM) Networks and [23] bootstrapped wavelet neural networks (BWNN) are fully sufficient to handle fluctuating and nonseasonal time-series water-quality data. [6] Autoregressive integrated moving average (ARIMA) model is from one of the traditional statistical theories, which can be applied to the time series prediction, but basically they are part of linear models. [26] ARIMA is less superior than the BWNN model as it is influenced by the self-adaption during the learning process of the ANN and the time-frequency properties of the wavelet basis functions. On the other hand, LSTM model (type of recurrent neural network structure), learns directly from time-series data. LSTM and BWNN can recognize the nonlinear relationship between the variables and their respective predicted variables more accurately, and then transfer the vital information from the past to the future.

The accuracy of the prediction of machine learning models is dependent on the features which are used to train the models. If there are redundant variables present, then it will tend to the reduce the inverse power and accuracy of the model, and also increase the complexity. The most widely concerned surface water quality parameters is the dissolved Oxygen (DO) [24] , which tells about the aquatic ecosystem's status and how much it is suitable for the aquatic organisms to sustain. In the Danube River, the linear polynomial neural network (PNN) was used to get a insight of DO concentration. The most important features which were affecting the prediction accuracy which include:- among 17 water quality parameters,

temperature, pH, BOD, and phosphorus concentration. [26] St. John's River, USA, has prediction of DO concentration among the five input features (chloride, NOx, total dissolved solids, pH, and water temperature), and among all these five inputs pH and NOx are strongly correlated with DO which can in turn affect the prediction accuracy. These outcomes are very familiar with those obtained by the [24] Chen et al, who stated that the Input parameters do affect the prediction performance of the model. Apart from other regular water parameters, there is another concern in surface water quality prediction which is eutrophication. Formulated on the adaptive neuro-fuzzy inference system (ANFIS) model, [16] Ly et al. found that there are certain reasons like combined interaction of the nutrients, organic matter, and environmental elements which is responsible for algal blooms. Part et al prediction the concentration of chlorophyll-a [2] in two reservoirs in the U.S., by using the meteorological data and weekly water quality data and he found that SVM and ANN both had near about similar prediction accuracies. The prediction accuracy get highly improved with the addition of meteorological factors. More factors like regional hydrological and socioeconomic factors can be added to the machine learning model so that the results may provide much stronger support for the comprehensive management of the regional water environment.

Moreover, the machine learning model's performance can also depend on its architecture. We should thoroughly go through and analyze the logical structure of algorithms as it is also a crucial part in the successful application of machine learning. When the [16] PNN gets compared with the other traditional neural network models, it has an advantage over them. The advantage of PNN in determining the key model parameters which are discussed above is that the number of hidden neurons and layers of PNN is directly determined by the data which saves the time for trial. The lower root mean square error (RMSE) of [7] DNN model which is 19.20%-25.16% as compared to the traditional machine learning model. The lower RMSE value of [16] DNN indicates better predictive performance. The reason behind the lower RMSE value is that there are multiple layers between the input and output layers of DNN, and also it uses more advanced activation functions than ANN. It helps in reducing the difficulty of training as it is more favourable to model convergence than the sigmoid used by traditional ANN. [1] LSTM works relatively well based on time series, when it come to predict water quality over the time.

3.2. Applications in Groundwater

It is crucial to ensure the safety of ground water for public health, as it is an important source of drinking water. Machine learning has potential applications in groundwater analysis, which includes assessing the quality of groundwater and predicting pollution sources.

Multivariate statistical techniques, including [24] Principal Component Analysis (PCA) and cluster analysis, have been employed for groundwater evaluation for a long time . Techniques like Decision Trees(DT), Random Forests(RF), Support Vector Machines(SVM) and Artificial Neural Networks(ANN) are also utilized to determine the quality of groundwater. Comparing different ML algorithms is what studies of the same tend to emphasize in order to determine models that are applicable for specific problems. For example, [23] Jeihouni et al. compared five data mining algorithms (such as typical DT, RF, chi-square automatic interaction detector, and Iterative Dichotomizer 3) to determine significant parameters influencing groundwater in semi-arid areas and estimate high-quality zones in Tabriz, Iran. [24] Lee et al. applied a self-organizing map (SOM) and fuzzy c-means clustering to classify

urban groundwater in Seoul, South Korea, in terms of pollution levels and spatial distribution. Geographic Information System (GIS) software is often combined with ML to produce pollution maps and determine contaminated areas more accurately.

The intricate hydrogeological conditions of groundwater, unlike surface water, render quality prediction difficult. However, ML has been applied to evaluate large scale data and predictions of future quality. [23] Agrawal et al. integrated PSO with SVM to estimate and predict groundwater quality index (WQI), proving the feasibility of the method. Individual pollutants such as nitrate and arsenic have been successfully modeled: [25] Arabgol et al. estimated nitrate concentration with SVM, whereas Sajedi Hosseini et al. combined the use of Boosted Regression Trees, Multivariate Discriminant Analysis, and SVM to forecast nitrate pollution risk in Iran's Lennart Plain. Ransom et al. applied ML to predict nitrate levels throughout the U.S., demonstrating its feasibility. Cho et al employed ANN to forecast risks of arsenic contamination in Cambodia, Laos, and Thailand. [23] Groundwater levels have also been predicted by models like ANFIS, DNN, and SVM, with DNN reporting the best seasonal prediction accuracy. Yadav et al. employed ensemble modeling to forecast levels in Indian cities with an accuracy of 85%.

Ensuring the safety of groundwater is imperative, therefore, determining the causes of pollution may be found to be useful. [24] PCA and cluster analysis are common in recent studies. Celestino et al. used PCA with subsequent K-means cluster analysis to distinguish natural and anthropogenic geochemical alterations. Chen et al. also utilized PCA and multivariate statistical methods to identify factors influencing groundwater quality.

Decision trees in data mining are commonly employed to predict groundwater quality dynamics. The models' algorithms can learn to associate input-output variables with understandable rules. [24] RF, for instance, performs well because it has high accuracy and generalization power—attaining 97.1% [1] accuracy on continuous datasets and providing realistic insights for groundwater planning. [9] Ensemble models, particularly those based on boosting techniques, improve prediction by aggregating weak learners. Even though different models are combined to create models with less variance, overfitting is still an issue.

In conclusion, machine learning offers robust tools for groundwater quality analysis, prediction, and management. Methods from statistical analysis and decision trees to neural networks and hybrid models have all played a role in more accurate assessments. However, model interpretability, data availability, and proper model choice are still essential for sustainable groundwater management.

3.3. Applications in Drinking water

Machine learning has applications in drinking water treatment and management systems, including from source water management to processes of treatment, distribution networks, and decision assistance. Drinking water typically comes from water available at surface or in ground. Employment of ML in assessing and predicting source water quality can assist with early detection and control of contaminants.

In a research paper by [2] Bouamar et al. , the viability of using multisensor-based Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for dynamic water quality monitoring was evaluated as far back as 2007. Both models achieved satisfactory

classification accuracy for the two water quality categories tested, and [18] SVM was shown to be more stable than ANN. Subsequently, Wu et al. developed an adaptive frequency analysis technique, employing water quality records from four Norwegian cities. Their method provided useful information for risk alerting at early stages, water quality management, and strategic planning. Other contributions include [20] Liu et al., who used Long Short-Term Memory (LSTM) networks and Deep Neural Networks (DNN) to create a time-series model that could predict water quality six months ahead with significant accuracy. Arnon et al. also used SVM to create a system that could predict pollution events under unknown conditions using ultraviolet absorption measurements. The model provided high detection accuracy on four datasets and kept error rates low.

Despite studies having been conducted with emphasis on chemical and physical parameters, microbial indicators, specifically [25] Escherichia coli (E. coli), have not been considered frequently. ML can also be utilized in the estimation of coagulant and disinfectant quantities in water treatment plants. Owing to its ease of operation and stability in performance, [3] SVM is also frequently applied to flocculation and disinfection process design. [16] Wang et al. , for example, established an SVM-based model of chemical dosing control for controlling residual free chlorine based on predictive residual free chlorine levels. [20] SVM outperformed the classical proportional-integral-derivative (PID) feedback controllers in effectiveness.

Prioritizing the quality of drinking water supply has motivated research in monitoring the correct operation of urban water infrastructure, fault detection, and disaster anticipation. Because of the complexity of the systems, water from water treatment plants, meeting the required standards, can also be re-contaminated during transport. This can be monitored by biological stability indicators and controlled by disinfection techniques.

Cluster analysis [12] has been found to be effective in identifying water quality differences between networked systems. Tian et al., for instance, applied clustering to evaluate the contribution of mixed-source water to aluminum (Al) residue concentrations in large-scale urban water supplies, including seasonal patterns and Al transport behavior. [20] Brester et al. obtained precise water quality assessment through casting methods with a Random Forest (RF) algorithm.

In the meantime, pipe burst failures can contribute to significant loss and contamination of water during transport. While deep learning algorithms may be able to predict where pipe bursts are likely to happen, the results nevertheless contain high degrees of uncertainty. For this, [26] Rayaroth et al. presented a bagging classifier based on random decision trees optimized with a shuffled frog-leaping algorithm successfully detecting leak locations in distribution pipes with low numbers of sensors deployed at optimal positions. The designed pipeline lifespan is yet another crucial aspect of planning water supply systems. [23] Almheiri et al.suggested a meta-learning model combining a neural network, and found that residual chlorine concentration has a significant impact on pipe lifespan.

Furthermore, water distribution system pollution incidents have been predicted using [22] SVM models. Park et al. also quantitatively assessed the impacts of disasters on water distribution systems by combining PCA, AHP, RF, and XGBoost models. Nonetheless, real-time data acquisition is still a problem, which hinders the application of the method in practice.

Water production capacity today becomes a determining factor in balancing regional development and population growth. Zhang et al. developed a [19] hybrid statistical model combining ANN and genetic algorithms for predicting the operation of water treatment plants. It can predict changes in production rates under different scenarios of fluctuating parameters, which enable operators to rapidly adjust treatment system settings. Concurrently, [18] Cardoso et al. created an automated monitoring system for urban water networks based on time-series clustering. Their findings indicated that water demand peaked between 3 a.m. and 6 a.m. during summer seasons, probably because of irrigation of public green areas. To enhance short-term water demand prediction, Guo et al. employed a [22] Gated Recurrent Unit (GRU) network with an interval setup of 15 minutes and successfully predicted water consumption for the next 15 minutes as well as the next 24 hours. Ghiassi et al. experimented with three models—Dynamic Artificial Neural Network (DAN2), time-delay focused neural networks, and K-Nearest [19] Neighbors (KNN)—for daily, weekly, and monthly water demand forecasting in Tehran. Among these, [22] DAN2 provided the most accurate predictions, with accuracy levels of 96%, 99%, and 98% for the daily, weekly, and monthly predictions, respectively.

In conclusion, ML methods like ANN and SVM are now universally applied in drinking water applications, particularly when dealing with high-dimensional datasets. Their short training times of a few seconds make them useful for real-time dynamic monitoring of drinking water quality and safety. While [17] ANN's accuracy has improved significantly because of advances in training methodologies, its susceptibility to noise remains a concern. SVM, by contrast, is inherently more noise-resistant, which has prompted growing interest in combining both approaches to leverage their respective strengths.

3.4. Applications in Wastewater

Machine learning (ML) contributes greatly to wastewater treatment through facilitating water quality monitoring and prediction, optimizing treatment technology, and optimizing the operation and management of wastewater treatment plants (WWTPs). [20] Industrial and domestic wastewater contains a diverse variety of contaminants, thus requiring a quality assessment prior to treatment . Rosen et al. came up with a method that integrates multiresolution analysis and principal component analysis (PCA) and provides more sensitivity in the monitoring of sewage indicators over various scales than PCA alone.

Large-scale data requires real time online monitoring. [22] A soft sensor based on a black-box model, for example, was proposed for real-time monitoring of E. coli , and it was shown that E. coli concentrations significantly increase following heavy rains—most likely due to suspended sewer sediments being mobilized by urban storm runoff Through the fusion of soft sensors and [24] artificial neural networks (ANN), scientists have developed systems able to continuously monitor chlorine and ammonia levels, also providing solutions towards mitigating WWTP high operation costs and technical complexities [19] Qin et al. created a sensor system with a boosting-iterative predictor weighting-partial least squares (Boosting-IPW-PLS) method combined with UV spectrometry and turbidimetry. The multiple sensor system, which detected chemical oxygen demand (COD) and total suspended solids, successfully reduced irrelevant variables by giving lower weights to them. The predictive model obtained was well representative of real water quality data, supported by a high correlation coefficient.

ML can be utilized to examine past data to optimize wastewater treatment systems. [6] Fang et al. employed SVM with an adaptive genetic algorithm to model anaerobic, anoxic, and oxic treatment conditions. Their research was designed to minimize anoxic tank capacities, thus conserving space. ML has also optimized tertiary treatments like reverse osmosis (RO), nanofiltration (NF), ozonation, and adsorption. [17] Cha et al., for instance, used a random forest (RF) model to predict the removal of micropollutants (MPs) in ozonation and attained enhanced removal efficiencies. High-resolution fluorescence excitation-emission matrix (EEM) data has also enhanced ML models' precision by enabling enhanced analysis of the nonlinear dependencies between organic matter and oxidizer exposure. For membrane-based treatment, MP removal prediction is vital in selecting appropriate membranes. [7] [14] Teychene et al. employed decision trees (DT) to identify the mechanisms by which RO and NF systems remove MPs, and concluded that particle size exclusion, electrostatic repulsion, and adsorption were the predominant mechanisms. [21] XGBoost was also employed to predict MP removal efficiency in RO and NF systems. Sigmund et al. built two neural-network-based models for the selection of the most efficient adsorbents for various pollutants. These research studies highlight the enormity of what ML can contribute to advanced treatments of wastewater, particularly when used to treat emerging pollutants.

Predicting and understanding intricate environmental circumstances are offered by machine learning.

ANN is capable of solving complex nonlinear environmental challenges, especially when it comes to removing pollutants [8].

An ANN model that predicts COD and BOD contents in treated wastewater was developed by [13] Bayat Varkeshi et al.. Currently, the majority of water quality prediction models aim to predict concentrations of a particular pollutant. Abdi et al. employed CatBoost, following an evaluation of tetracycline (TC) photodegradation rates under different conditions, to precisely predict TC removal when utilizing metal-organic frameworks. [17] Baek et al. built three models—RF, SVM, and ANN—to forecast the removal of five MPs, with RF yielding the best results.

Biological indicators may also be simulated using ML. Bayesian methods, such as naive Bayes and [15] semi-naive Bayesian networks, have been used to predict pathogen removal efficiency and study reduction-relationship, operating conditions-relationship, and monitoring parameter-relationship. RF was used by Roguet et al. to predict Clostridiales and [17] Bacteroidales levels in wastewater. RF also assists in creating methods for tracing sources of fecal contamination and hence preventing waterborne disease transmission.

The operation of [11] WWTPs relies on numerous parameters, and management and maintenance of WWTPs can have an issue on costing. On that note, machine learning can be utilized to make sense in investigating the possibilities of cutting cost, along with enhancing operation. Gomez-Munoz et al. utilized Bayes' theorem to quantify various [17] WWTP cost elements, facilitating improved construction, regulatory, and operational decision-making. Harmful pollutants in sewer systems may interfere with plant operations, but models such as [8] XGBoost and RF can assist in their detection as well as sources. Even though flow-measuring sensors exist in sewer pipes, the likelihood of inaccuracy due to contamination, corrosion, and extreme turbidity increases with time. This may result in inaccurate measurements. Deep learning optimizes measurement accuracy by utilizing

sensors. These flow-measurement sensors are usually placed within sewage pipes. [21] Instability in measurement due to contamination, corrosion, and high turbidity, though, can result in erroneous measurements. Deep learning has the potential to optimize the accuracy of measurement in different scenarios by upgrading current sensors. [8] Ji et al. trained a model using past data on recognized sensor failures to detect faults, adjust the treatment process accordingly, and ensure the continuous running of the WWTP.

In conclusion, Machine learning improves wastewater treatment by making real-time water quality monitoring, pollutant prediction, and process optimization possible.

Models such as ANN, SVM, RF, and XGBoost sense toxic substances including E. coli, [17] COD, BOD, and micropollutants, enhancing the detection of pollution at an early stage and improving treatment processes like ozonation and membrane filtration. Soft sensors and deep learning enhance hazard monitoring accuracy and sensor reliability, but [21] ML also facilitates the choice of treatment chemicals, the identification of sources of pollution, equipment failure prediction, and operational cost management. In summary, ML makes wastewater treatment systems more efficient and safer.

3.5 Application of machine learning in marine environments

Nowadays [13] seawater pollution is becoming a serious threat as it is affecting the ecosystem of the Earth. Introducing technology, with the help of machine learning monitoring seawater pollutants would become easier. In short, we can resolve these issues by providing different solutions. Bhagat et al, established a lead-protection algorithm by using XGBoost, and by using the historical monitoring data from the [4] Bramble and Deception Bay stations in Australia he trained the model. As a result, it worked successfully and found that the trained model performed very well in selecting the input parameters and predicting the water quality. [15] Gon calves et al. introduced a waste mapping program which was based on radio frequency (RF) and an automatic unmanned aerial vehicle (UAV) system so that it could monitor coastal plastic waste automatically. To predict the concentration of coastal microbial pollution in beach water, an ensemble machine learning approach with a two- layered learning structure was proposed. For the improvement of the accuracy of [22] antibiotic resistance gene (ARG) prediction in beach water, [6] LSTM-CNN model was implemented by Jang et al. and predicted a single ARG successfully. Using machine learning classification algorithms, Mancia et al. found differentially expressed genes in dolphins which was exposed to marine pollutants.

Moreover, many other researchers have deeply focused on developing the surveillance technologies for algal blooms which can lead to severe contamination. Further, XGBoost model was trained by adding more efficient spectral characteristics of different water types and algal blooms by [11] Ghatkar et al. helping the model to distinguish between the algae and identify the algae that cause algal blooms. [4] Du et al. used a method called hierarchical cluster analysis water quality evaluation based on the which was based on the [9] Mahalanobis distance, for the evaluation of the water quality along with the North Yellow and Bohai Seas. Coming to the conclusion, there are many different machine learning methods which can identify the types of seawater pollutants, can distinguish between the different species present, determine the concentration and distribution of pollutants, and at the end provide a useful data of the status of the marine organisms. Just by combining the human mind and technology we can make the best results which has a positive effect on the ecosystem of earth and our's life also.

In order to protect the marine ecosystem, seawater quality monitoring plays a very crucial role. Many researchers have applied the machine learning in monitoring of seawater quality. In the year 2001, a researcher named Alshehri et al. proposed a near-shore water quality prediction model which was based on KNN. [13] Further Sheng et al. combined BPNN, SVR, and LSTM models to establish a water quality prediction method, and this method provide great results by improving the water quality prediction accuracy. Another researcher, Zohu et al. proposed a water quality prediction method on an already improved grey regression analysis algorithm and [14] LSTM on the basis of the multivariate correlation and the time series characteristics provided in water quality information. Du et al. collected data by a geosynchronous ocean color imagery and from 1240 water quality sample points along with the coast of [15] Zhejiang, analyzes it using a water quality assessment method with a geographic neural network weighted regression model. As per reports [23], by 2050 75% of the world's population will face a freshwater crisis. In the areas where water shortages at its extreme, desalinated seawater is an important source of freshwater for them. Some seawater desalination difficulties will remain with the low efficiency and reliability of desalination systems is being the major obstacles. Alshehri et al. decided to improve the seawater treatment performance of water treatment plants, so he used a [6] CNN model and transfer the learning in order to classify salt particles with different concentrations in water. [21] Chawla et al. used the machine learning algorithms like linear regression, RF, SVM and LSTM, in such a way that he predicted the salinity and development trend of the Salton sea which makes easier for the long-term management of seawater salinity and seawater desalination.

In the previous literature, the single water quality prediction models has been thoroughly described and recently the integrated model has come into action.

We have discussed different-different mechanisms which works upon the different input features and also provide different predictive performances. The integration model was suggested by [21] Sheng et al. When we talk about the prediction model, the best fits data first or the one which may have a higher chance in predicting the prediction is selected before making the prediction. [21] This is a model selection algorithm based on input features, on the other hand XGBoost method which was proposed by Bhagat et al. can screen input features and can select more than 1 (5-9) features out of 21 features to be combined with ANN and the other application methods. During the training stage, the information gathered by the model will not be lost. [11] High accuracy and fast speed is the main advantages of XGBoost model which makes it a promising modeling algorithm. But when it comes to feature selection algorithm, then it will depend on the sample size.

Concluding remarks

As we have discussed so many methods and algorithms which is used by the machine learning models to solve the water environmental issues . Basically, it acts as a powerful tool as we can predict and get different insights from this such as water quality, optimize water resource allocation, manage water resource shortages, etc. Apart from this, there are several challenges which remain in fully applying machine learning approaches in this field to evaluate water quality:

(1)Machine learning works on the large amounts of high quality data . Providing sufficient data with high accuracy in water treatment and management systems is quite a difficult task as it leads to high cost consumption and there might have limitations in technology also.

(2)There are complex conditions in real water treatment and management systems, so the currently introduced algorithms can be applied to some specific systems which obstruct the wide application of machine learning approaches.

(3)To apply these machine learning algorithms practically would require researchers who must have the professional background knowledge.

To overcome these challenges the following changes should be considered:

(1)More improved and advanced sensors, including soft sensors, should be introduced and apply them in water quality monitoring to gather the sufficient and accurate data . All these would enhance the efficiency of the machine learning models .

(2)As per the water treatment and management systems, the algorithm's feasibility and reliability should be improved and more effective algorithms and models should be developed.

(3)More advanced knowledge about the technology used in machine learning should be provided and train them. Develop advanced machine learning technique and use them in engineering practises.

References

[1] K. Shaukat, F. Tahir, U. Iqbal, and S. Amjad, "A Comparative Study of Numerical Analysis Packages," *Int. J. Comput. Theory Eng.*, vol. 10, no. 3, pp. 67–72, 2018, doi: 10.7763/IJCTE.2018.V10.1201.

[2] C. Gupta, "Coronamask: A Face Mask Detector for Real-Time Data," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5624–5630, Aug. 2020, doi: 10.30534/IJATCSE/2020/212942020.

[3] C. Gupta* and P. N. S. Gill, "Machine Learning Techniques and Extreme Learning Machine for Early Breast Cancer Prediction," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 4, pp. 163–167, Feb. 2020, doi: 10.35940/IJITEE.D1411.029420.

[4] M. T. Verma* and D. N. Singh Gill, "Email Spams via Text Mining using Machine Learning Techniques," *Int. J. Innov. Technol. Explor. Eng.*, vol. 4, no. 9, pp. 2535–2539, Feb. 2020, doi: 10.35940/IJITEE.D1915.029420.

[5] T. Verma and N. S. Gill, "Machine Learning Techniques for Better Data Driven Decisions Revisited," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 4, pp. 460–464, Apr. 2020, doi: 10.35940/IJEAT.D6766.049420.

[6] B. Qian, R. Wei, Y. Wu, and C. Ma, "An Interactive Multi-task Learning Model for Aspect-Based Sentiment Analysis," *2023 8th Int. Conf. Comput. Commun. Syst. ICCCS 2023*, pp. 1075–1081, 2023, doi: 10.1109/ICCCS57501.2023.10151000.

[7] K. S. Eljil, F. Nait-Abdesselam, E. Hamouda, and M. Hamdi, "Enhancing Sentiment Analysis on Social Media with Novel Preprocessing Techniques," *J. Adv. Inf. Technol.*, vol. 14, no. 6, pp. 1206–1213, 2023, doi: 10.12720/JAIT.14.6.1206-1213.

[8] S. Kumar and M. T. Uddin Haider, "Sentiment Analysis of Political Party Twitter

Data using Ensemble Learning Classifier," *2023 14th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2023*, 2023, doi: 10.1109/ICCCNT56998.2023.10308149.

[9] C. Jin, M. A. Bin Ahmadon, and S. Yamaguchi, "A Proposal of Sentiment Analysis Approach for Comments Based on Semi-supervised Learning and Topic Analysis," *2024 12th Int. Conf. Inf. Educ. Technol. ICIET 2024*, pp. 343–347, 2024, doi: 10.1109/ICIET60671.2024.10542760.

[10] Z. Zhu, Y. Ma, T. Chen, and T. Wang, "Domain-dependent Dictionary Construction for Sentiment Analysis," *2023 6th Int. Conf. Big Data Artif. Intell. BDAI 2023*, pp. 181–186, 2023, doi: 10.1109/BDAI59165.2023.10257048.

[11] Sangeeta and N. S. Gill, "Review of Factors Affecting Efficiency of Twitter Data Sentiment Analysis," *Int. J. Comput. Theory Eng.*, vol. 12, no. 2, pp. 53–58, Apr. 2020, doi: 10.7763/IJCTE.2020.V12.1263.

[12] S. Rani, N. S. Gill, and P. Gulia, "Analyzing impact of number of features on efficiency of hybrid model of lexicon and stack based ensemble classifier for twitter sentiment analysis using WEKA tool," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 22, no. 2, pp. 1041–1051, May 2021, doi: 10.11591/IJEECS.V22.I2.PP1041-1051.

[13] A. Sagu, "Machine Learning Decision Tree Classifier and Logistics Regression Model," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1.4, pp. 163–166, Sep. 2020, doi: 10.30534/IJATCSE/2020/2491.42020.

[14] S. Rani, N. S. Gill, and P. Gulia, "Survey of Tools and Techniques for Sentiment Analysis of Social Networking Data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, pp. 222–232, 2021, doi: 10.14569/IJACSA.2021.0120430.

[15] Sangeeta, P. Gulia, and N. S. Gill, "Comprehensive Analysis of Flow Incorporated Neural Network based Lightweight Video Compression Architecture," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 3, pp. 503–508, Jul. 2021, doi: 10.14569/IJACSA.2021.0120360.

[16] Sangeeta, P. Gulia, and N. S. Gill, "Flow incorporated neural network based lightweight video compression architecture," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 2, pp. 939–946, May 2022, doi: 10.11591/IJEECS.V26.I2.PP939-946.

[17] C. Gupta, N. S. Gill, and P. Gulia, "SSDT: Distance Tracking Model Based on Deep Learning," *Int. J. Electr. Comput. Eng. Syst.*, vol. 13, no. 5, pp. 339–348, Jul. 2022, doi: 10.32985/IJECES.13.5.2.

[18] S. Yadav, P. Gulia, and N. S. Gill, "Flow-MotionNet: A neural network based video compression architecture," *Multimed. Tools Appl.*, vol. 81, no. 29, pp. 42783–42804, Dec. 2022, doi: 10.1007/S11042-022-13480-0/METRICS.

[19] A. Saini, N. S. Gill, and P. Gulia, "Gray scale image denoising technique using regression based residual learning," *Multimed. Tools Appl.*, vol. 83, no. 2, pp. 3547–3566, Jan. 2024, doi: 10.1007/S11042-023-15603-7/METRICS.

[20] Bharti, N. S. Gill, and P. Gulia, "Exploring machine learning techniques for fake profile detection in online social networks," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 3, pp. 2962–2971, Jun. 2023, doi: 10.11591/ijece.v13i3.pp2962-2971.

[21] S. Yadav *et al.*, "Video Object Detection from Compressed Formats for Modern Lightweight Consumer Electronics," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1,

pp. 4507–4514, Feb. 2024, doi: 10.1109/TCE.2023.3325480.

[22] C. Gupta *et al.*, "A Real-Time 3-Dimensional Object Detection Based Human Action Recognition Model," *IEEE Open J. Comput. Soc.*, vol. 5, pp. 14–26, 2024, doi: 10.1109/OJCS.2023.3334528.

[23] S. Yadav *et al.*, "Design of a Lightweight Compressed Video Stream-Based Patient Activity Monitoring System," *Comput. Mater. Contin.*, vol. 78, no. 1, pp. 1253–1274, Jan. 2024, doi: 10.32604/CMC.2023.042869.

[24] Nisha, N. S. Gill, and P. Gulia, "A review on machine learning based intrusion detection system for internet of things enabled environment," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 2, pp. 1890–1898, Apr. 2024, doi: 10.11591/ijece.v14i2.pp1890-1898.

[25] C. Gupta, N. S. Gill, P. Gulia, and J. M. Chatterjee, "A novel finetuned YOLOv6 transfer learning model for real-time object detection," *J. Real-Time Image Process.*, vol. 20, no. 3, pp. 1–19, Jun. 2023, doi: 10.1007/S11554-023-01299-3/METRICS.

[26] S. Joshi *et al.*, "Object detection and classification from compressed video streams," *Expert Syst.*, vol. 42, no. 1, p. e13382, Jan. 2023, doi: 10.1111/EXSY.13382;JOURNAL:JOURNAL:14680394;PAGE:STRING:ARTICLE/CHAPTER.