10.48047/jocaaa.2025.34.04.53

MACHINE LEARNING-BASED PREDICTION OF DISSOLVED OXYGEN LEVELS IN RIVER ECOSYSTEMS

Dr. P. Babu, U. Satya Narayana, P. Uma Sai Krishna, V. Bharathi

Department of Computer Science and Engineering (AI&ML), Geethanjali Institute of Science and Technology, Nellore, Andhra Pradesh, India.

ABSTRACT

Dissolved oxygen (DO) is a key indicator of river water quality and plays a critical role in sustaining aquatic ecosystems. Adequate DO levels are essential for the respiration and survival of aquatic organisms, making it vital to monitor and predict DO concentrations for effective environmental management and conservation. Traditionally, DO prediction involves collecting water samples from multiple river locations and analyzing them in laboratories. While this method provides valuable data, it has significant drawbacks, including spatial and temporal limitations, high costs, time consumption, and restricted real-time monitoring capabilities. These limitations hinder accurate and continuous assessment of river health. To overcome these challenges, there is a growing interest in leveraging advanced technologies such as machine learning (ML) to enhance DO prediction. ML algorithms can process large volumes of historical and real-time data, uncovering complex, non-linear relationships among various physical and chemical parameters influencing DO levels. This research proposes the development of an ML-based model to predict DO concentrations in river water. By utilizing diverse datasets, the model aims to offer continuous, reliable, and cost-effective predictions that outperform traditional empirical methods. The ability to accurately forecast DO levels not only supports real-time water quality assessment but also strengthens environmental monitoring, conservation planning, and sustainable management of aquatic ecosystems. The outcomes of this research will contribute significantly to modernizing water resource monitoring and ensuring the long-term health of riverine environments.

Keywords: Dissolved oxygen, River ecosystem, Machine learning, Environmental Monitoring, Predictive Modeling.

1.INTRODCUTION

River dissolved oxygen is an important indicator of water quality and ecosystem health, and it is affected by a variety of environmental factors. Insufficient dissolved oxygen can lead to the inhibition of aquatic biological activity or even ecosystem collapse; while too much can cause eutrophication and algal blooms, posing a serious threat to river ecosystems and water quality management [1]. However, traditional prediction models have limited accuracy and applicability because they have difficulty capturing the complex nonlinear relationship between dissolved oxygen and environmental factors. This study aims to combine key water quality parameters with advanced machine learning algorithms to establish a highly accurate dissolved oxygen prediction model to solve the prediction problem and provide a reliable scientific tool for river ecological protection and water quality management [2].

Dissolved oxygen is a key water quality parameter for evaluating the water body rating. There are two main types of dissolved oxygen concentration prediction models: 1. traditional models based on physics, and 2. data-driven models based on artificial intelligence. Traditional prediction models simulate the physical mechanisms in rivers and describe the transfer process of pollutants in water. Physical models can elucidate the hydrodynamic mechanism and show the temporal and spatial trends of pollutants. A number of studies have used physical models to predict water quality and dissolved oxygen. For example, Zehra et al. used the water quality analysis simulation program (WASP) model

10.48047/jocaaa.2025.34.04.53

and the QUAL 2Kw model to systematically analyze the dissolved oxygen and other water quality indicators of the Yamuna River and predict future spatial trends [3]. Zhang et al. used the Mikezl model to simulate the impact of ecological recharge on river hydraulics and discuss the trend of DO under different hydrological conditions. However, the model often requires a large amount of hydrological data during operation, which limits the establishment of models for rivers with fewer data sets [4]. In addition, traditional mechanistic models usually rely on the establishment of hydrodynamic equations, and their internal parameters are complex.

2.LITERATURE SURVEY

Ziyad Sami, et al. [5] developed a reliable prediction model to predict D.O. in the Fei-Tsui reservoir for better water quality monitoring. The proposed model is an artificial neural network (ANN) with one hidden layer. Twenty-nine years of water quality data have been used to validate the accuracy of the proposed model. A different number of neurons have been investigated to optimize the model's accuracy. Statistical indices have been used to examine the reliability of the model.

Bolick, et al. [6] proposed a Comparison of machine learning algorithms to predict dissolved oxygen in an urban stream. A multiple linear regression model was compared to machine learning algorithms k-nearest neighbor, decision tree, random forest, and gradient boosting. These algorithms were evaluated to understand which best predicted dissolved oxygen (DO) from water temperature, conductivity, turbidity, and water level change at four locations along the urban stream.

Moon, et al. [7] proposed an urban river dissolved oxygen prediction model using machine learning. To predict the optimized WQ, we selected pH, SS, water temperature, total nitrogen (TN), dissolved total phosphorus (DTP), NH3-N, chemical oxygen demand (COD), dissolved total nitrogen (DTN), and NO3-N as the input variables of the AdaBoost model. Dissolved oxygen (DO) was used as the target variable.

Nair, et al. [8] proposed Analysing and Modelling Dissolved Oxygen Concentration Using Deep Learning Architectures. This work employs deep learning algorithms like Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) effective prediction models for forecasting DO levels in river water. The models are developed and validated using the river water quality data collected from eleven sampling stations during the year 2016 to 2020.

Zhou, et al. [9] proposed an interpretable and explainable model that integrates the shapley additive explanations (SHAP) algorithm with the long short-term memory network model (LSTM) is to evaluate the contributions of karst spring discharge, precipitation, water temperature, and specific conductance to DO concentrations in karst spring flow. The hybrid model can predict the temporal fluctuations of DO levels and provide a robust characterization of DO behaviours.

Heddam, et al. [10] proposed the application of long short-term memory (LSTM) deep learning for dissolved oxygen (DO) prediction in rivers. The model was trained and calibrated using three predictors: (i) river water temperature (Tw), (ii) air temperature, and (iii) river discharge (Q). The variables were measured on an hourly time scale and collected from two USGS stations. The LSTM model was compared against genetic programming (GP), the group method of data handling neural network (GMDH), support vector regression (SVR), and Gaussian process regression (GPR) models.

Salas, et al. [11] proposed the Potential of mapping dissolved oxygen in the Little Miami River using Sentinel-2 images and machine learning algorithms. The authors mapped the spatiotemporal changes of DO in the Little Miami River (LMR) using 10-m Sentinel-2 images. We trained two machine learning algorithms – Random Forest (RF) and Support Vector Machine (SVM) – to predict DO concentrations

10.48047/jocaaa.2025.34.04.53

using spectral predictors derived from the satellite images. Moreover, we calculated several metrics, which include Root Mean Squared Error (RMSE), Amount of Variance Explained (AVE), Coefficient of Efficiency (COE), and Normalized Mean Bias (NMB) to assess the performance of the models and accuracy of the DO maps.

Garabaghi, et al. [12] proposed Modeling dissolved oxygen concentration using machine learning techniques with a dimensionality reduction approach. The authors propose an accurate prediction model for DO concentrations. The performance of the Random Forest (RF) and multilayer perceptron (MLP) algorithms was evaluated in generating the regression models. Moreover, the effect of dimensionality reduction of the data by the wrapper feature Selection method on the performance of the models was evaluated.

Ahmed, et al. [13] proposed the development of a dissolved oxygen forecast model using a hybrid machine-learning algorithm with hydro-meteorological variables. This work aims to forecast dissolved oxygen (DO) concentration using a multivariate adaptive regression spline (MARS) hybrid model coupled with maximum overlap discrete wavelet transformation (MODWT) as a feature decomposition approach for Surma River water using a set of water quality hydro-meteorological variables.

Heddam, et al. [14] proposed a novel hybrid model based on signal processing decomposition, extreme learning machine and parallel chaos search for forecasting DO several days in advance. The correlation between DO data at several times lags were calculated using the autocorrelation function (ACF) and the partial autocorrelation function (PACF). The DO concentration time series were decomposed using the empirical wavelet transform technique (EWT), and the multiresolution analysis (MRA) components were then provided.

Khabusi, et al. [15] proposed A Deep Learning Approach to Predict Dissolved Oxygen in Aquaculture. This study aimed at designing a prediction model for DO in aquatic environments. To achieve the objective, time series data consisting of 70374 records and 15 attributes from Mumford Cove in Connecticut, USA collected for over 5 years was preprocessed and used to train long-short term memory (LSTM) recurrent neural network (RNN) for DO prediction.

Siddik, et al. [16] proposed the Application of machine learning approaches in predicting estuarine dissolved oxygen (DO) under a limited data environment. The application of machine learning (ML) approaches to predict estuarine dissolved oxygen (DO) from a set of environmental covariates including nutrients remains unexplored due to nutrient data unavailability.

Adedeji, et al. [17] proposed Predicting in-stream water quality constituents at the watershed scale using machine learning. The authors implemented five ML algorithms—Support Vector Machines, Random Forest (RF), eXtreme Gradient Boost (XGB), ensemble RF-XGB, and Artificial Neural Network (ANN) —and demonstrated our modeling framework in an inland stream—Bullfrog Creek, located near Tampa, Florida.

Kim, et al. [18] proposed Machine learning predictions of chlorophyll-a in the Han river basin, Korea. This work developed a model to predict concentrations of chlorophyll-a ([Chl-a]) as a proxy for algal population with data from multiple monitoring stations in the Han river basin, by using machine-learning predictive models, then analyzed the relationship between [Chl-a] and the input variables of the optimized model.

Yan, et al. [19] proposed a new framework to predict long-term water quality by using Bayesian-optimised machine learning methods and key pollution indicators collected from monitoring stations in

10.48047/jocaaa.2025.34.04.53

the Pearl River Estuary, Guangdong, China. The optimised stacked generalisation (SG-op) model achieved the best performance with the highest accuracy (0.992) and Kappa coefficient (0.987).

3.PROPOSED METHODOLOGY

Fig. 1 showcases a comprehensive data analysis and machine learning workflow for water quality testing data. It encompasses data loading, cleaning, visualization, correlation analysis, linear regression, random forest regression, and evaluation. The goal is to gain insights into the dataset and develop predictive models for dissolved oxygen levels, which can be valuable for water quality assessment and environmental monitoring.

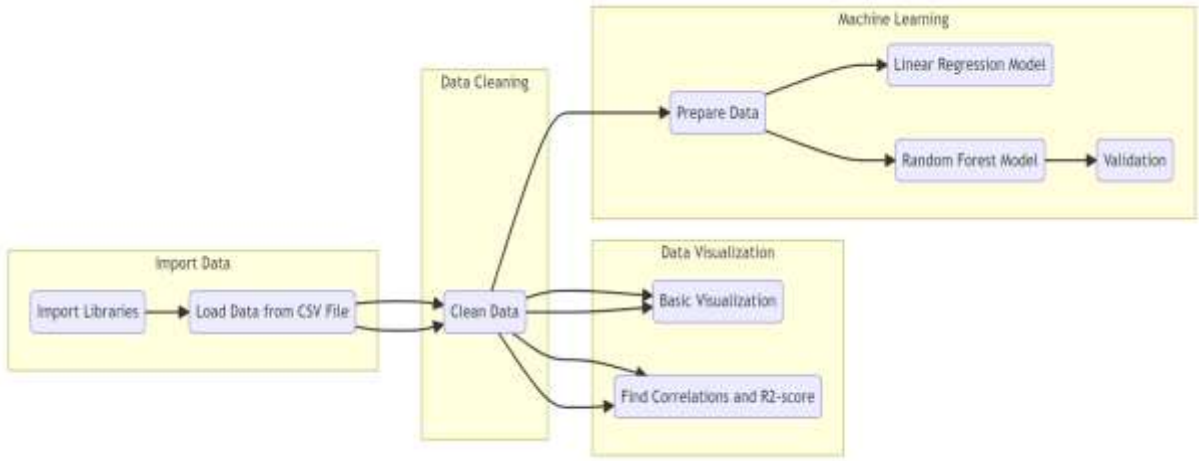

Fig. 1: Block diagram of proposed system architecture.

3.1 Random Forest Regressor

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

10.48047/jocaaa.2025.34.04.53

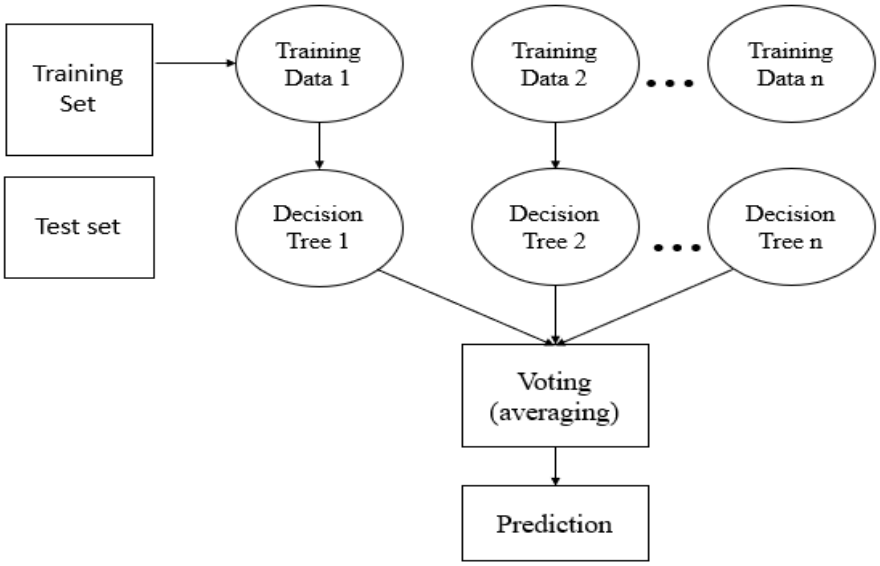

Fig. 2: Working of Random Forest regressor.

The Random Forest algorithm operates through a series of structured steps. Initially, it selects n random samples from the original dataset containing k records, often using bootstrapping (sampling with replacement). For each of these samples, an individual decision tree is constructed independently. Once all trees are built, each tree generates its own prediction. In the case of classification tasks, the final output is determined by majority voting, where the most frequent class among the trees is selected. For regression tasks, the final prediction is obtained by averaging the outputs of all the decision trees. This ensemble approach helps improve accuracy and reduce overfitting compared to using a single decision tree.

4. RESULTS AND DICUSSION

Water quality is a crucial aspect of environmental management, and it is essential to measure various physical, chemical, and biological parameters to monitor it effectively. The dataset of 200 rows contains measurements of six critical water quality parameters widely used in water quality monitoring and analysis. The dataset provides a representative snapshot of water quality and can be used for various research, education, and decision-making purposes.

Figure 3 presents a heatmap illustrating the correlation matrix of numerical features related to river water quality, including pH, temperature (°C), turbidity (NTU), dissolved oxygen (mg/L), and conductivity (µS/cm). The color gradient from blue to red indicates the strength and direction of the correlation, with red representing strong positive correlation and blue indicating negative correlation. The matrix reveals that dissolved oxygen has a strong positive correlation with conductivity (0.76) and pH (0.71), meaning that increases in these variables are generally associated with higher dissolved oxygen levels. There is a weaker, but still positive correlation with temperature (0.25). In contrast, turbidity is negatively correlated with dissolved oxygen (–0.28), suggesting that higher turbidity tends to reduce oxygen levels. The matrix helps identify which features may have a significant influence on dissolved oxygen and are therefore useful predictors in modeling efforts.

10.48047/jocaaa.2025.34.04.53

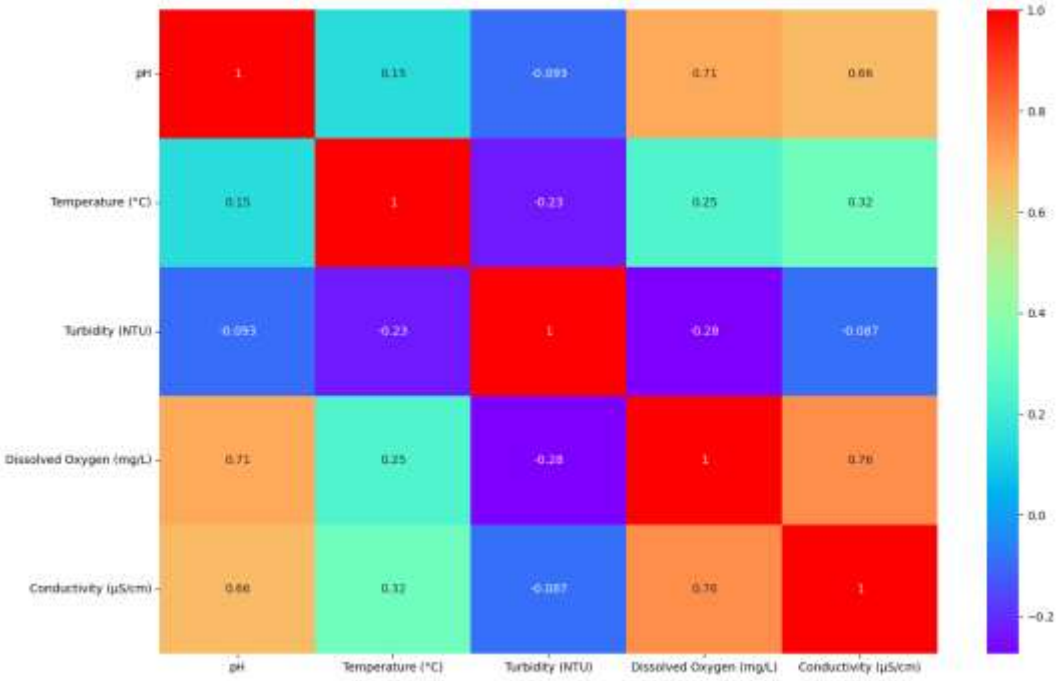

Fig. 3: Visualizing the correlation matrix of numerical columns.

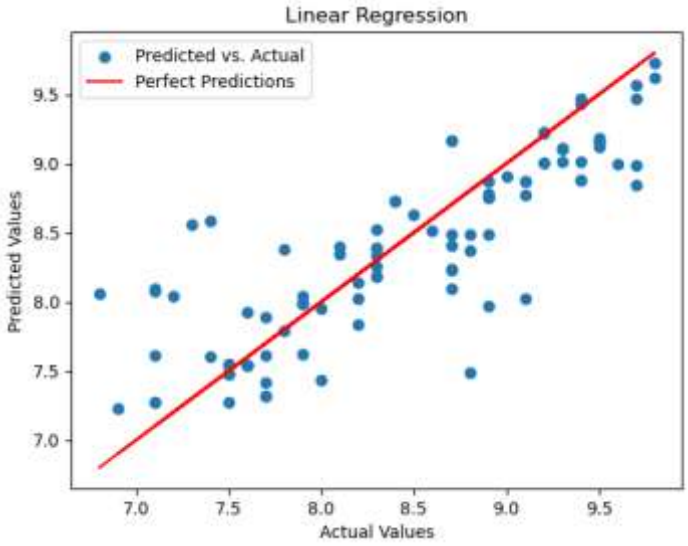

Fig. 4: Best fit line obtained using existing Linear Regression.

10.48047/jocaaa.2025.34.04.53

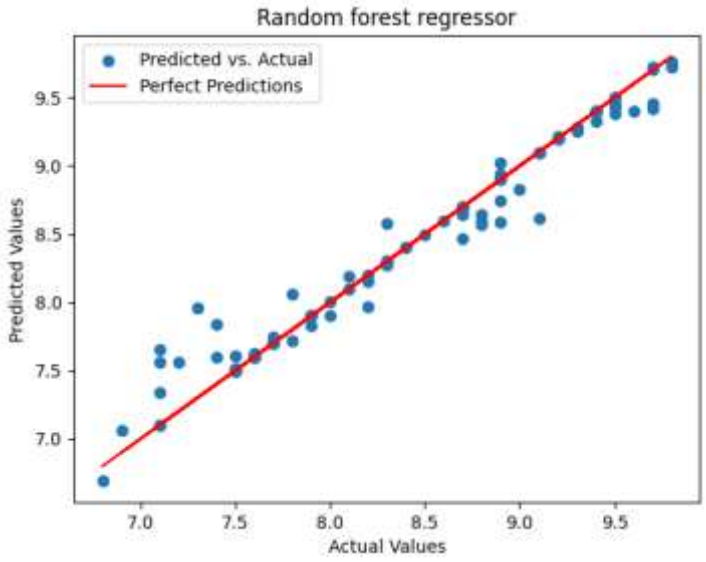

Fig. 5: Best fit line obtained using proposed random forest regressor.

	Actual	Predicted
361	8.9	8.900
73	7.9	7.910
374	8.3	8.300
155	6.9	7.063
104	7.9	7.835
..	...	...
347	7.5	7.500
86	9.0	8.821
75	8.0	8.006
438	8.7	8.700
15	6.8	6.693

[100 rows x 2 columns]

Fig. 6: Actual and predicted values side by side.

Figure 4 shows a scatter plot comparing actual dissolved oxygen values to those predicted by a linear regression model. The red line represents the ideal case where predicted values perfectly match actual values. Most of the blue data points are scattered along the red line, indicating a general positive trend and some predictive power in the linear regression model. However, there are noticeable deviations, especially at the mid and lower ranges of DO levels, suggesting that the linear model does not fully capture the complex dependencies among the variables. This results in a moderate fit, with some predictions overestimating or underestimating the actual values.

Figure 5 presents a similar scatter plot but uses the random forest regressor for prediction. The distribution of blue points is much tighter around the red perfect-fit line, indicating a stronger agreement between predicted and actual values. This demonstrates that the random forest model is significantly more accurate than linear regression in this context. The improved performance can be attributed to the model's ability to handle non-linear relationships and interactions among input features such as pH, temperature, and turbidity, which linear regression fails to model effectively.

Figure 6 provides a tabular comparison of actual and predicted dissolved oxygen values across 100 samples. The table shows close agreement between values, with many entries matching exactly or

10.48047/jocaaa.2025.34.04.53

deviating only slightly by a few hundredths. For instance, in one row the actual DO value is 8.9, while the model predicts it exactly as 8.900. In another row, the actual value is 7.9 and the predicted value is 7.910, showing minimal error. This consistency across a large number of samples reinforces the accuracy and reliability of the random forest model for predicting dissolved oxygen levels in river water.

Table. 1: Performance comparison of quality metrics obtained using linear regressor (LR) model and random forest regressor (RFR) model.

Model	MSE	MAE	R^2-score
LR model	0.18	0.30	0.56
RFR model	0.0245	0.081	0.959

Table 1 provides a performance comparison between the Linear Regression (LR) model and the Random Forest Regressor (RFR) model based on three key evaluation metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2-score. The LR model yields an MSE of 0.18 and an MAE of 0.30, indicating a relatively higher level of prediction error. Its R^2-score of 0.56 suggests that the linear model explains only 56% of the variance in the dissolved oxygen data, reflecting moderate predictive capability. In contrast, the RFR model significantly outperforms the LR model across all metrics, with a much lower MSE of 0.0245 and MAE of 0.081, indicating highly accurate predictions with minimal error. Most notably, the RFR achieves an R^2-score of 0.959, meaning it accounts for 95.9% of the variance in the data. This demonstrates that the random forest model provides a far more robust and reliable prediction of dissolved oxygen levels compared to the linear regression approach.

5.CONCLUSION

In conclusion, this study presents a thorough analysis and application of machine learning techniques on a water quality dataset to predict dissolved oxygen levels—an essential indicator of aquatic ecosystem health. The process began with meticulous data preprocessing, including loading, cleaning, and handling missing or duplicate values to ensure the reliability of the dataset. Preliminary visualizations offered an overview of variable distributions and interdependencies, while a detailed correlation heatmap highlighted strong relationships among water quality parameters. Linear regression analysis helped uncover how pH and other features relate to dissolved oxygen levels. In the modeling phase, the dataset was divided into input features and the target variable, and two regression models—Linear Regression and Random Forest Regressor—were trained and evaluated. Visual assessments through scatter plots of predicted versus actual values revealed that the Random Forest model provided significantly more accurate predictions. The generation of a predictions DataFrame enabled a clear comparison of actual and predicted values, further emphasizing model effectiveness. Overall, this work demonstrates the potential of machine learning in water quality monitoring, offering a scalable, accurate, and efficient alternative to traditional methods. These insights can contribute meaningfully to environmental conservation efforts and the sustainable management of freshwater resources.

REFERENCES

1. Hutchins MG, Qu Y, Charlton MB. Successful modelling of river dissolved oxygen dynamics requires knowledge of stream channel environments. J Hydrol. 2021;603:126991.

2. Xu C, Luo P, Wu P, Song C, Chen X. Detection of periodicity, aperiodicity, and corresponding driving factors of river dissolved oxygen based on high-frequency measurements. J Hydrol. 2022;609:127711.

3. Zehra R, Singh SP, Verma J, Kulshreshtha A. Spatio-temporal investigation of physico-chemical water quality parameters based on comparative assessment of QUAL 2Kw and WASP

10.48047/jocaaa.2025.34.04.53

model for the upper reaches of Yamuna River stretching from Paonta Sahib, Sirmaur district to Cullackpur, North Delhi districts of North India. Environ Monit Assess. 2023;195(4):480. pmid:36930328

4. Zhang X. Simulation study on the impact of South–North water transfer central line recharge on the water environment of Bai River. Water. 2023;15(10):1871.
5. Ziyad Sami, Balahaha Fadi, et al. "Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan." Scientific Reports 12.1 (2022): 3649.
6. Bolick, Madeleine M., et al. "Comparison of machine learning algorithms to predict dissolved oxygen in an urban stream." Environmental Science and Pollution Research (2023): 1-22.
7. Moon, J., Lee, J., Lee, S., & Yun, H. (2022). Urban river dissolved oxygen prediction model using machine learning. Water, 14(12), 1899.
8. Nair, Jitha P., and M. S. Vijaya. "Analysing And Modelling Dissolved Oxygen Concentration Using Deep Learning Architectures." International Journal of Mechanical Engineering 7 (2022): 12-22.
9. Zhou, Renjie, and Yanyan Zhang. "Predicting and explaining karst spring dissolved oxygen using interpretable deep learning approach." Hydrological Processes 37.8 (2023): e14948.
10. Heddam, Salim, et al. "Predicting dissolved oxygen concentration in river using new advanced machines learning: Long-short term memory (lstm) deep learning." Computers in Earth and Environmental Sciences. Elsevier, 2022. 1-20.
11. Salas, Eric Ariel L., et al. "Potential of mapping dissolved oxygen in the Little Miami River using Sentinel-2 images and machine learning algorithms." Remote Sensing Applications: Society and Environment 26 (2022): 100759.
12. Garabaghi, Farid Hassanbaki, Semra Benzer, and Recep Benzer. "Modeling dissolved oxygen concentration using machine learning techniques with dimensionality reduction approach." Environmental Monitoring and Assessment 195.7 (2023): 879.
13. Ahmed, Abul Abrar Masrur, et al. "The development of dissolved oxygen forecast model using hybrid machine learning algorithm with hydro-meteorological variables." Environmental Science and Pollution Research 30.3 (2023): 7851-7873.
14. Heddam, Salim. "Parallel Chaos Search Based Incremental Extreme Learning Machine Based Empirical Wavelet Transform: A New Hybrid Machine Learning Model for River Dissolved Oxygen Forecasting." Computational Intelligence for Water and Environmental Sciences. Singapore: Springer Nature Singapore, 2022. 355-376.
15. Khabusi, Simon Peter, and Yo-Ping Huang. "A Deep Learning Approach to Predict Dissolved Oxygen in Aquaculture." 2022 International Conference on Advanced Robotics and Intelligent Systems (ARIS). IEEE, 2022.
16. Siddik, Mohammad Abu Zafer. "Application of machine learning approaches in predicting estuarine dissolved oxygen (DO) under a limited data environment." Water Quality Research Journal 57.3 (2022): 140-151.
17. Adedeji, Itunu C., Ebrahim Ahmadisharaf, and Yanshuo Sun. "Predicting in-stream water quality constituents at the watershed scale using machine learning." Journal of Contaminant Hydrology 251 (2022): 104078.
18. Kim, Kyung-Min, and Johng-Hwa Ahn. "Machine learning predictions of chlorophyll-a in the Hanriver basin, Korea." Journal of Environmental Management 318 (2022): 115636.
19. Yan, Tao, Annan Zhou, and Shui-Long Shen. "Prediction of long-term water quality using machine learning enhanced by Bayesian optimisation." Environmental Pollution 318 (2023): 120870.