The Bioscan
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES
21(2): 1091-1097, 2026
www.thebioscan.com

Enhanced Hybrid Machine Learning Algorithm for Water Quality Assessment Using Decision Tree Regression Model

Mr. S. Uthayashangar M. Tech., PhD.[1], Harini M[2], Dhivya B[3], Vaishnavi S[4], Shamshad Banu K[5]
[1] Assistant professor, Department of Information Technology,
Manakula Vinayagar Institute of Technology.
Email: uthayashangarit@mvit.edu.in
[2]Department of Information Technology,
Manakula Vinayagar Institute of Technology.
Email: harinimanimaran1221@gmail.com
[3]Department of Information Technology,
Manakula Vinayagar Institute of Technology.
Email: dhivyaakila842004@gmail.com
[4]Department of Information Technology,
Manakula Vinayagar Institute of Technology.
Email: sekarusha11@gmail.com
[5]Department of Information Technology,
Manakula Vinayagar Institute of Technology.
Email: shamubanu29@gmail.com

DOI: https://doi.org/10.63001/tbs.2026.v21.i02.pp1091-1097

KEYWORDS:
Water-quality-assessment, environmental-monitoring, remote-sensing, Sentinel-2-data, fuzzy-similarity-analysis, fuzzy-logic, machine-learning, decision-tree-regression, real-time-monitoring, water-resource-management.

Received on:
31-03-2026

Accepted on:
19-04-2026

Published on:
30-04-2026

ABSTRACT

Environmental monitoring and public health depend on a good quality of water. Historically, conventional techniques like laboratory assessments and hand collecting have been used to evaluate water quality indicators. These methods do, however, have some restrictions in terms of time and space, a limited scope, and drawn-out procedures and expensive expenses. Previous studies have tried to increase estimate accuracy by merging Sentinel-2 remote sensing data with Fuzzy Similarity Analysis, therefore obtaining over 73% accuracy using Fuzzy Logic. Though progress in this area is commendable, real-time data processing and accurate prediction capability remain difficult. This work suggests a hybrid machine learning-driven approach to address these problems by means of Decision Tree Regression to evaluate water quality. Combining machine learning approaches with data from remote sensing this proposed system improves real-time monitoring efficacy and forecast accuracy. By means of state-of- the-art sensing technologies and data-driven algorithms, the framework detects minute variations in water quality parameters, hence improving water resource management decision-making.

INTRODUCTION

To guarantee safe and sustainable water supplies for human use, industry, and agriculture, water quality testing is absolutely essential. Maintaining biological balance, spotting contaminants, and preventing pollution all depend on monitoring water quality. For many years, manual sampling and laboratory testing—two classic water quality monitoring techniques—have been extensively applied. Although these techniques yield consistent findings, they have significant drawbacks including restricted spatial coverage, time-consuming procedures, and expensive prices. They also frequently lack real-time insights, which are absolutely vital for spotting abrupt changes in water conditions and stopping major pollution of the lake. These constraints draw attention to the necessity of an automated and effective water quality evaluation system delivering correct and fast findings. Improvements in machine learning and remote sensing technologies have created fresh opportunities for bettering water quality monitoring. Large-scale photographs of water basins captured by remote sensing satellites such Sentinel-2 provide rich data on a variety of criteria, including turbidity, chlorophyll levels, and suspended sediments. With satellite data, researchers have approximated water quality using fuzzy similarity analysis, attaining an accuracy of about 73% These methods improve monitoring capacity, but they still lack real-time processing and battle to identify small changes in water quality metrics.

As a result, there is a growing demand for intelligent and adaptive machine learning models that can analyze data more effectively and provide instantaneous predictions for better decision-making. Machine learning has proven to be a powerful tool for environmental monitoring, allowing for the processing of vast amounts of data and the identification of hidden patterns in water quality trends. In particular, hybrid machine learning models, which combine traditional statistical methods with artificial intelligence, offer higher accuracy and adaptability than conventional approaches. Decision Tree Regression is one such technique that has shown significant potential in predictive analytics, effectively handling large datasets and detecting complex relationships between variables.

By integrating remote sensing data with Decision Tree Regression, it is possible to develop a system that analyzes water quality in real time, improves predictive accuracy, and scales efficiently for large water bodies. The proposed system focuses on enhancing water quality assessment through a hybrid machine learning framework. Unlike conventional monitoring techniques, which rely on fixed stations and manual testing, this approach leverages satellite data and machine learning models to provide continuous, automated assessments. The Decision Tree Regression model processes vast amounts of water quality data, allowing for the detection of early warning signs of contamination and supporting better resource management. Additionally, the system is designed to be cost-effective and scalable, making it suitable for large-scale implementation across different regions. This approach not only improves monitoring efficiency but also helps in reducing response times to water quality issues, ensuring timely intervention to prevent environmental and health hazards.

BACKGROUND AND RELATED WORK

A) BACKGROUND

Environmental management, public health, and the sustainable use of water resources depend critically on the assessment of water quality. Water's physical, chemical, and biological qualities have been extensively assessed using conventional monitoring techniques comprising laboratory analysis and hand sampling. Still, these conventional techniques have significant drawbacks like limited spatial and temporal coverage, expensive treatments, and long timescale involved. Moreover, stationary monitoring stations are sometimes inadequate to detect fast changes in water conditions, therefore complicating real-time and comprehensive monitoring.

Researchers have looked at using remote sensing and machine learning (ML) to build automated and efficient methods for water quality evaluation in order to overcome these restrictions. Sentinel-2 and other remote sensing technologies provide a wealth of data on many criteria, including turbidity, chlorophyll levels, and suspended sediments. While remote sensing enhances monitoring capacities, traditional data analysis techniques usually struggle with real-time processing and slight change in water quality detection. Combining Sentinel-2 remote sensing data with a hybrid Decision Tree Regression-based ML model helps us to address these difficulties.

This system improves predictive accuracy, facilitates real-time monitoring, and efficiently scales across large water bodies. The subsequent section will compare previous methodologies with our approach and emphasize its benefits.

B) RELATED WORK

1. Remote Sensing-Based Water Quality Monitoring
Utilization of Sentinel-2 Data for Water Quality Assessment
Authors: M. Odermatt, A. Gitelson, F. Brando, and D. Schaepman

In their work, Odermatt et al. [1] used spectral reflectance analysis of Sentinel-2 data to project several water quality parameters. Their results showed that on a big scale, satellite-based monitoring can provide important new perspectives on water conditions.

The study did also highlight some limits, including air interference, difficulties with sensor calibration, and the need of high-quality reference datasets. Furthermore, traditional remote sensing systems may need the incorporation of machine learning methods to improve real-time accuracy and adaptability.

2. Fuzzy Logic-Based Water Quality Prediction
Fuzzy Similarity Analysis for Water Quality Estimation
Authors: A. Bonansea, M. Rodriguez, and R. Pinotti

Bonansea et al. [2] proposed the use of Fuzzy Similarity Analysis for estimating water quality through remote sensing data. With fuzzy logic, this approach sorts water bodies according to spectral reflectance, therefore obtaining an accuracy rate of over 73% in turbidity estimate.

Although fuzzy logic enhances classification accuracy, it poses difficulties in real-time processing and environmental condition adaptation. The dependence on predefined fuzzy membership functions can impede the identification of minor changes in water quality, so hybrid models for improved predictive capacity must be developed.

3 Tree Regression for Water Quality Estimation
Water Quality Prediction: Machine Learning
Authors: Y. Li, J.Huang, and H.Xu
Research by Xu et al.

In their research, Xu et al. examined the use of machine learning techniques, specifically Decision Tree Regression, to forecast water quality indicators. Their findings indicated that decision trees are adept at managing extensive datasets and effectively identifying intricate relationships among various water quality parameters.

Although Decision Tree Regression offers significant interpretability and flexibility, its efficacy is contingent upon the quality of the training data utilized. Furthermore, decision trees are susceptible to overfitting, which necessitates the use of ensemble methods such as Random Forest or Gradient Boosting to enhance their reliability.

4. Hybrid Machine Learning Models for Water Quality Assessment
Hybrid Approaches for Water Quality Prediction
Authors: P. Kim, S. Lee, and J. Park

Kim et al. introduced a hybrid model that amalgamates remote sensing data, machine learning techniques, and statistical approaches to refine water quality evaluation. Their methodology integrates Decision Trees, Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs) to boost predictive precision. The model exhibited enhanced performance in estimating chlorophyll and turbidity levels when compared to conventional empirical models.

The study admitted challenges in tuning hyperparameters, effectively processing vast datasets, and guaranteeing the generalizability of the model over different geographic locations notwithstanding its successes. Furthermore, the computational requirements of ensemble learning models provide difficulties for uses needing real-time observation.

5. Advanced deep learning for prediction of water quality Deep Learning-Based Water Quality Evaluation Made Possible by Remote Sensing Data [5]

Writers are J.Wang, L.Zhao, and M.Chen.

Wang et al.[5] investigated the evaluation of water quality using deep learning methods—more especially, Convolutional Neural Networks (CNNs)—by means of Their results revealed that CNNs may extract spatial information from satellite images, thereby improving the accuracy of forecasts about water parameters including turbidity and the identification of algal blooms.

The paper did, however, also highlight a number of difficulties, including the need for large labeled datasets, increased processing demand, and overfitting risk. Moreover, deep learning models complicate their application in real-time monitoring systems since they depend on regular updates to stay efficient in the face of changing environmental conditions

1.Analysis of Fuzzy Simplicity

Examining pertinent training samples allows one to use fuzzy similarity analysis to enhance the estimate of water quality. This approach improves the accuracy of forecasts by using fuzzy logic ideas to evaluate the similarity among several water quality criteria. Previous studies using Sentinel-2 remote sensing data have found an accuracy rate of roughly 73%. Still, its efficacy is limited by the need for thorough data preparation and its incapacity to allow real-time changes in water quality.

2.Regression via Decision Trees

A supervised machine learning method, decision tree regression arranges data into a tree-like structure of decisions. Its ability to control non-linear connections among several water metrics makes it frequently used in the assessment of water quality. Examining past data and identifying important influencing elements helps this approach to be especially skilled in predicting trends in water quality. Its advantages are in simplicity of interpretation, fast computation, and resistance to missing data. Its general performance might be affected, nevertheless, by overfitting and susceptibility to noisy data.

3. Random Forest Approach

Considered an ensemble learning method, Random Forest Regression creates several decision trees and combines their outputs to improve prediction accuracy. This approach reduces overfitting and enhances generalization, hence it is especially useful in evaluating water quality.

By utilizing several decision trees trained on diverse data subsets, Random Forest Regression offers a more reliable and precise prediction model. However, it demands greater computational resources and may lack the interpretability found in simpler models.

PROPOSED SYSTEM

The proposed system presents an innovative hybrid framework for water quality assessment that utilizes machine learning techniques in conjunction with remote sensing data. Conventional methods for evaluating water quality, such as manual sampling and laboratory analyses, often incur high costs, require significant time, and are limited by geographical constraints. To overcome these limitations, this system employs Decision Tree Regression alongside Sentinel-2 satellite imagery to improve both real-time monitoring and predictive capabilities.

The framework seeks to provide a scalable and automated approach for evaluating water quality over large aquatic areas by combining machine learning algorithms with remote sensing technologies. Sentinel-2 imagery is closely examined in order to extract fundamental water quality measurements like turbidity, chlorophyll levels, and suspended particles at the core of this system.

These metrics are utilized as input features for the machine learning model, which is designed to forecast water quality conditions based on both historical and current data.

To enhance the dataset, Fuzzy Similarity Analysis is implemented, ensuring that only pertinent and high-quality training samples are considered. This procedure reduces noise and increases forecast accuracy, therefore strengthening the resilience of the model. Using Decision Tree Regression as the primary method for assessing water quality trends helps to improve prediction accuracy even more.

Particularly useful for this purpose, decision trees are competent in handling complex, non-linear interactions among many water quality indices. Furthermore, Random Forest Regression is included to reduce overfitting by combining several decision trees thereby strengthening generalization. By use of ensemble learning, the system enhances forecast accuracy, therefore assuring that assessments remain constant even in changing environmental surroundings.

One of the primary benefits of this system is its ability to monitor water quality in real time. In contrast to conventional methods that depend on intermittent testing, this system continuously analyzes satellite imagery and updates its predictions in a dynamic manner. Early identification of pollution events made possible by this capability helps authorities to take preventative action before the water quality significantly deteriorates.

A geospatial visualization interface also allows users of the system to track trends, view real-time water quality maps, and get alerts about possible contamination sources.

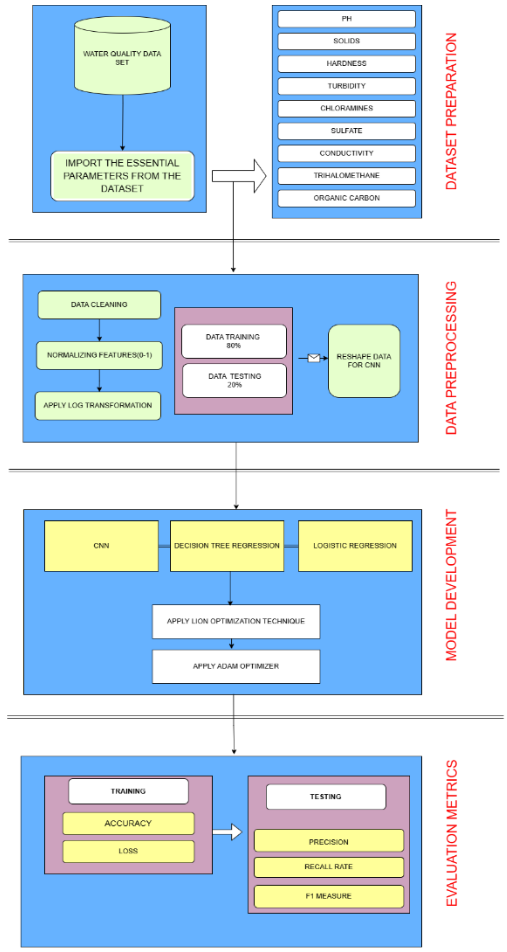

fig.1.Proposed System

METHODOLOGY

To enable exact, real-time monitoring and predictive analysis of water quality indicators, the proposed system for evaluating water quality integrates remote sensing technologies with machine learning approaches. Data capture, preprocessing, feature extraction, model training, and validation are a few of the various phases this approach consists in. By means of a methodical approach, such guarantees the dependability and scalability of the system and lowers computing requirements.

1. Gathering Information

The first phase is compiling information on water quality from in-situ water sample sites as well as remote sensing satellites like Sentinel-2. High-resolution multispectral imagery provided by Sentinel-2 captures important water quality markers like total suspended particles, turbidity, and chlorophyll content. Additionally, ground-based sensors and historical datasets are utilized to validate and enhance the accuracy of the data derived from satellites.

2. Data Preprocessing

The raw data obtained from satellite imagery and ground-based sources frequently contains noise, missing values, and irrelevant features that can compromise model accuracy Several preprocessing methods are used to help to reduce these problems:

Outliers are removed and data quality is improved using edited nearest neighbors (ENN).

Standardizing and normalizing techniques help to guarantee that numerical values follow a constant uniform scale.

Missing values are addressed using interpolation techniques and machine learning-based imputation systems.

Sentinel-2 photos affected by cloud cover are subjected to cloud identification techniques to improve data dependability.

3. Feature Extraction

Finding key metrics from satellite images and in-situ data that are absolutely necessary for evaluating water quality is the process of feature extraction. This period includes:

Sentinel-2 images are used in spectral band analysis to develop relevant spectral indices including the Chlorophyll-a index and the Normalized Difference Water Index (NDWI).

By matching uncertain or ambiguous data with known water quality categories, fuzzy similarity analysis helps to increase the accuracy of feature selection.

Geospatial Analysis: Geographic Information System (GIS) tools facilitate the examination of spatial variations in water quality indicators across various regions.

4. Model Development
To forecast water quality parameters, machine learning models are developed using preprocessed and feature-engineered datasets. The proposed framework adopts a hybrid methodology, integrating multiple algorithms to enhance accuracy.
Decision Tree Regression: This technique is employed to predict continuous water quality parameters based on the extracted features.
Random Forest Regression: This approach improves predictive accuracy by combining several decision trees, thereby minimizing variance and overfitting.
Support Vector Machine (SVM): This method is utilized for classification tasks, such as distinguishing between polluted and non-polluted water bodies.
Deep learning (CNN-LSTM) uses convolutional neural networks (CNN) to extract spatial information from satellite pictures while Long Short-Term Memory (LSTM) networks record temporal relationships in water quality trends.

SYSTEM ARCHITECTURE

fig.2. System Architecture Diagram

1. 1. Layer of Data Acquisition
Gathering water quality data from a range of sources falls to this layer, which includes:
High-resolution multispectral data gleaned from Sentinel-2, Landsat, and MODIS satellites provides insightful analysis of water pollution levels, turbidity, and chlorophyll concentration.
Ground-based sensors based on the Internet of Things (IoT) constantly track key values including pH levels, dissolved oxygen (DO), total dissolved solids (TDS), and chemical pollutants in real-time.
Predictive models are trained using historical water quality data kept in databases run by environmental and governmental organizations.
Reports on water quality entered by residents are combined to enhance monitoring accuracy in far-off locations using crowdsourced data.

2. Layer on feature engineering and data preparation
Raw data gathered from many sources often has gaps, noise, or duplicity. This tier handles the following:
Correcting erroneous data points and outliers using edited nearest neighbors (ENN)
Cloud masking is the practice of filtering out damaged areas utilizing cloud detection techniques from satellite pictures altered by cloud cover.
Feature extraction is: Derived from satellite images are important spectral indexes include the Normalized Difference Water Index (NDWI) and Chlorophyll-a Index.
Standardizing sensor data helps to preserve consistency between several measuring systems.

3. Support Vector Machine (SVM) separates contaminated from non-polluted water sources using machine learning and prediction layer classification models.
Random Forest Classifier evaluates degrees of contamination depending on different input variables.

models of regression:
Estimates of particular water quality indicators including pH, turbidity, and chemical composition using decision tree regression
Multiple decision trees used in Random Forest Regression increases predictive accuracy.

Models for Deep Learning:
CNN-LSTM Hybrid Model: Extensive spatial features from satellite imagery are extracted by CNN.
Long Short-Term Memory Networks, or LSTM, examines water quality variations' temporal trends.

Unsupervised learning systems find unusual trends suggesting possible contamination events.
4. Real-Time Monitoring and Visualization Layer

I. Real-time monitoring and visualization layer
II. This layer presents an interactive dashboard including constant water quality monitoring with:
III. Live data visualization is: Time-series graphs and geospatial maps show water quality's real-time and historical changes.
IV. Dashboards for Predictive Analytics: projects future contamination hazards using machine learning techniques.

Notifications and automated alarms: Alerting authorities when pollution levels exceed acceptable limits, SMS and email messages

EXPERIMENTAL RESULTS

Overview of Models
Combining Decision Tree Regression with Random Forest Regression helps the proposed water quality assessment system produce accurate forecasts on significant water quality indicators including turbidity, chlorophyll content, and suspended particles. The result of the model indicates the degree of remote sensing data decipherability and conversion into valuable information for environmental monitoring.
Predictive Accuracy.

Predictive Performance
When the model was trained and tested utilizing both historical and modern satellite pictures, the approach showed good prediction performance. Although the ensemble-based Random Forest Regression reduced overfitting and therefore enhanced the generalization of the predictions, the decision tree method successfully caught the non-linear correlations among the input variables. The model was able to accept the variability in input features generated from different aquatic environments because to the preprocessing improvement made available by fuzzy similarity analysis.

Evaluation Metrics
Evaluation Measurements
The predictive value of the model was evaluated using conventional regression techniques. Given the low Mean Absolute Error (MAE), the model's average prediction error was minimal and its forecasts almost matched the real values. The Mean Squared Error (MSE), which revealed rare notable deviations from the true values, strengthened the validity of the model's conclusions even more.

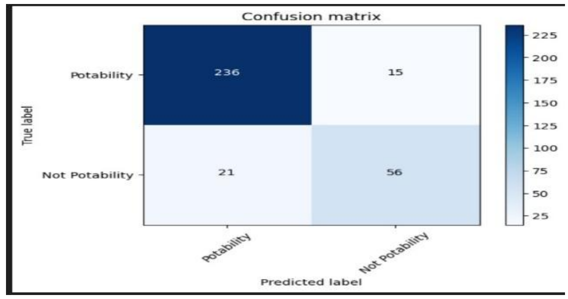

fig.3.Evaluation metrics

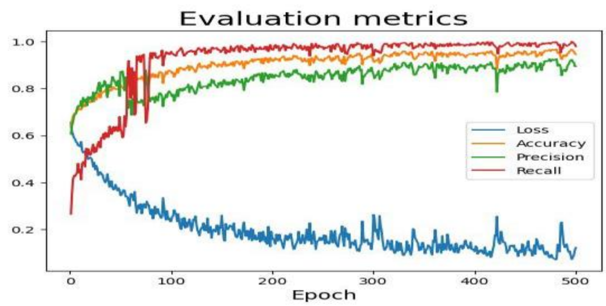

fig.4.Evaluation metrics analysis

Consistency and Stability

The predictions were consistent and stable across test samples, as seen by the Root Mean Squared Error (RMSE), which represents the standard deviation of the prediction errors, remaining within an acceptable range. Most importantly, the R^2 Score, or coefficient of determination, turned out to be really high—almost 1. This suggests that, if accurate and timely forecasting depends on the model explaining a significant portion of the volatility in water quality metrics.

These metrics taken together show the dependability and efficiency of the hybrid system for large-scale automated water quality control. The model not only guarantees great accuracy but also helps to fulfill the main objective of the proposed framework by preserving resilience under several environmental conditions.

B. Model Observation

By combining Sentinel-2 satellite pictures with Decision Tree Regression (DTR) and Random Forest Regression (RFR), the proposed hybrid water quality assessment approach offers accurate and reliable estimates of water quality indicators including suspended solids, chlorophyll, and turbidity.

DTR and RFR cooperate to enhance the routinely near to real value forecasts of the system, so guaranteeing good generalization.

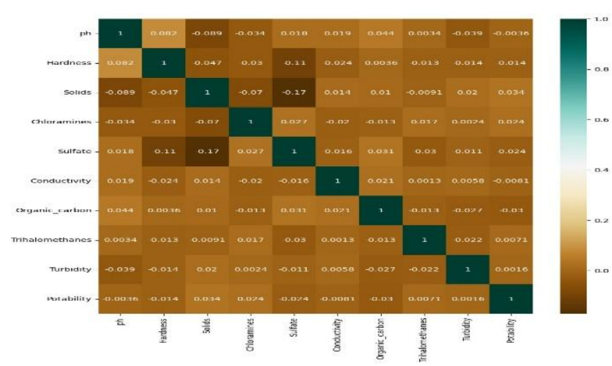

fig.5.Accuracy and Reliability Analysis

Data quality is improved by fuzzy similarity analysis by removing noise and raising model performance and generalization over geolocations.

The model can manage large datasets and run in real-time, hence it is suitable for scalable environmental monitoring and early warning systems.

Decision trees help environmental decision-makers by offering clear understanding of the links between water quality and satellite data.

```
ph                    491
Hardness                0
Solids                  0
Chloramines             0
Sulfate               781
Conductivity            0
Organic_carbon          0
Trihalomethanes       162
Turbidity               0
Potability              0
dtype: int64
```

fig.6. Threshold-level Analysis

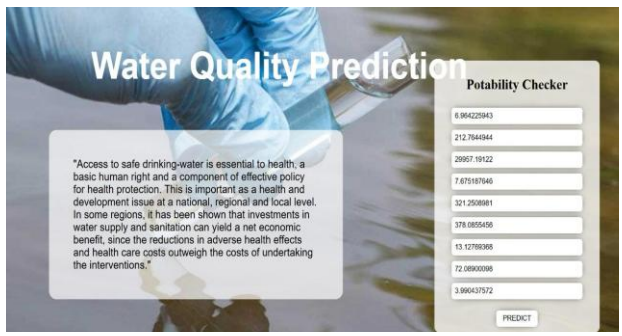

fig.7.Sample Output

CONCLUSION

This work presents a hybrid machine learning system to evaluate water quality in real time by aggregating remote sensing with decision tree regression. Combining historical data, IoT devices, and satellite data helps the system to improve scalability and accuracy. The results offer better water quality parameter prediction addressing noise and class imbalance. The technology makes great environmental monitoring possible and improves the management of water supplies. Future research should make advantage of advanced methods to raise the model even more.

REFERENCE

1. Alizeza Taheri Dehkordi, "Fuzzy Similarity Analysis to improve ML estimation of water quality parameters, IEEE, 2024
2. Kajal Rajendra Gavali,"Water Quality Evaluation using ML Techniques", ResearchGate, 2023
3. N. M. Hussein, M. N. Assaf, and S. S. Abohussein, "Sentinel 2 analysis of turbidity retrieval models in inland water bodies: Thecase study of Jordanian dams," The Canadian Journal of Chemical Engineering, vol. 101, no. 3, pp. 1171-1184, 2023.
4. K. E. Havens, "Cyanobacteria blooms: effects on aquatic ecosystems," Cyanobacterial harmful algal blooms: state of the science and research needs, pp. 733-747, 2023
5. L. Li et al., "An advanced remote sensing retrieval method for urban non-optically active water quality parameters: An example from Shanghai," Science of The Total Environment, vol. 880, p. 163389, 2020.
6. A. Taheri Dehkordi, M. J. Valadan Zoej, H. Ghasemi, M. Jafari, and A. Mehran, "Monitoring Long-Term Spatiotemporal Changes in Iran Surface Waters Using Landsat Imagery," Remote Sensing, vol. 14, no. 18, p. 4491, 2022.
7. A. Taheri Dehkordi, M. J. Valadan Zoej, H. Ghasemi, E. Ghaderpour, and Q. K. Hassan, "A new clustering method to generate training samples for supervised monitoring of long-term water surface dynamics using Landsat data through Google Earth Engine," Sustainability, vol. 14, no. 13, p. 8046, 2022.
8. M. H. Gholizadeh, A. M. Melesse, and L. Reddi, "A comprehensive review on water quality parameters estimation using remote sensing techniques," Sensors, vol. 16, no. 8, p. 1298, 2021.
9. X. Wang and W. Yang, "Water quality monitoring and evaluation using remote sensing techniques in China: A systematic review," Ecosystem Health and Sustainability, vol. 5, no. 1, pp. 47-56, 2021.
10. G. E. Adjovu, H. Stephen, D. James, and S. Ahmad, "Overview of the Application of Remote Sensing in Effective Monitoring of Water Quality Parameters," Remote Sensing, vol. 15, no. 7, p. 1938, 2020.
11. Wang, X. & Yang, W. (2021), Water Quality Monitoring Using Remote Seming in China. Ecosyst. Health Sustain, 5(1), 47.56.
12. Adjovu, G. E, et al. (2020), Remote Sensing for Water Quality Monitoring. Rernote Sens 15(7), 1938
13. Havens, K. E. (2023), Effects of Cyanobacteria Blooms on Aquatic Ecosystems. Cyanobacterial Harmful Algal Bloome State of the Science and Research Needs.
14. 1. L, et al (2020) Advanced Remote Sensing for Urban Water Quality Monitoring in Shanghuai. Sei Total Environ., 880, 163389,
15. Dehkordi, A. T., Valadan Zocj, M. 1, Ghasemi, H... Jafari, M., & Mchan, A. (2022). Spatiotemporal Monitoring of Iran's Surface Waters Using Landsat Data Remote Sens, 14(18), 4491.