Predicting Water Quality Using Ensemble Machine Learning Models and Remote Sensing Data

Dr. Kaliprasanna Sethy[1],Dr. Megha Mudholkar[2],Dr. Pankaj Mudholkar[3],Pramod Kumar Behera[4] ,Dr B.Karthik[5,]S. Balamuralitharan[6]

[1]Assistant Professor, Department of Civil Engineering, Government College of Engineering, Kalahandi, Bhawanipatna, 766003, Odisha, India;drkps@gcekbpatna.ac.in

[2]Assistant Professor, Department of Computer Engineering, Marwadi University, Rajkot - 360003, Gujarat, India meghakunte2000@gmail.com;https://orcid.org/0000-0003-2016-1525

[3]Associate Professor, Faculty of Computer Applications, Marwadi University, Rajkot - 360003, Gujarat, India mudholkarpankaj@gmail.com;https://orcid.org/0000-0003-1639-0704

[4]Assistant Professor, Odisha University of Technology and Research, Bhubaneswar;pkbeherace@outr.ac.in

[5]Associate professor, Department of EEE,Sona college of technology, Salem 5;Karthik@sonatech.ac.in

[6]Professor, Adjunct Faculty, Department of Mathematics,Saveetha School of Engineering, SIMATS, Saveetha University, Chennai 602105, Tamil Nadu, India.: balamurali.maths@gmail.com

Abstract– *Sustainable management of water resources can hardly be done without water quality prediction. Conventional field-based surveillance approaches are, in most cases, subject to spatio-temporal barriers. A combination of ensemble machine learning (ML) methods and remote sensing data provides a practical and resilient mechanism of forecasting water quality aspects at large spatial scales. In this paper, an ensemble-based predictive model is proposed, which can make use of the remote sensing indicators and the historical water quality data to estimate such important parameters as Biological Oxygen Demand (BOD), Dissolved Oxygen (DO), pH and turbidity. The ensemble modeling method bettered and outweighed single models in accuracy, robustness and generalization and entailed random forest (RF) gradient boosting (GB) and extreme gradient boosting (XGBoost). The technique proposed was tested over satellite images and the in-situ water quality observations of several river basins of India. The findings show that predictive variables in remote sensing like surface temperature, NDVI, and land use are strongly correlated with the water quality indicator hence predictive capacity.*

Keywords– *Water Quality Prediction, Ensemble Machine Learning, Remote Sensing, Random Forest, XGBoost, NDVI, River Basin Management, Environmental Monitoring.*

INTRODUCTION

Clean and safe water is basic to the survival of human beings, health of ecosystems and sustainable economic growth. Nonetheless, rising urban population, industrial effluents, agricultural effluents, and changes in weather have greatly worsened the quality of water in several regions of the globe. Specifically, water bodies like rivers and the open surface water resources close to urban areas as well as industrial centers have become a center of pollution. Like all aspects of environmental management and policy making, surveillance of water quality and knowledge of its spatial-temporal change is particularly important. However, traditional water surveillance approaches mainly depend on the ground-based field work and laboratory tests. These approaches, as accurate as they are, are labour-intensive, costly and lack spatial coverage, and are not suitable to large scale or regular assessments [1].The shortcomings of the conventional water monitoring techniques triggered scientists to search more feasible, inexpensive, and scalable ways to evaluate water safety. In this connection remote sensing has become a potential tool. It has the benefit of providing data of the environment over a broad regional scale with a high time resolution [4]. Land cover temperature (LST), vegetative cover (NDVI) and water body extents (NDWI) are indicators which can be measured through satellite images (e.g. Landsat, Sentinel missions) and are indirectly linked with the water quality. When viewed together with the previous data of water quality, these parameters can yield information about the distributions of pollutants and stress areas in the ecology.Accompanying the development of the remote sensing technology, machine learning (ML) has transformed data-based environmental modeling. The ML models can be used to estimate complex and nonlinear connections between variables, which is a handy tool in environmental predictions tasks where the behavior of a system is dependent on many related factors [7]. In the ML field, ensemble methods, including Random Forests, Gradient Boosting, and XGBoost, have proved to be more effective than other methods because ensemble methods combine several base learners, thus minimizing the variance of the model, enhancing robustness and preventing it in overfitting.A number of researches have shown the possible application of ML in water quality predictions.

Nevertheless, most of these models have been highly dependent on data in-situ which is accounted through limited generalization over other varied regions. On the same note, although there are few studies that have used remote sensing data to map the characteristics of water surfaces, few studies have used a systematic approach of integrating satellite-derived indexes and ensemble ML algorithms to predict the core water quality parameters, such as pH, DO (Dissolved Oxygen), BOD (Biological Oxygen Demand), and turbidity [6].The current hustle and bustle around finding dependable and extensible techniques of estimating the quality of water induces the research. To achieve this goal, we suggest a hybrid model where the multi-source data is used in an ensemble machine learning system by combining satellite remote sensing indicators, precedent water quality data, and other potential factors. The main goal is to further have a better and accurate prediction of important parameters of water quality with this interpretable and high-performing model that can be implemented to different geographic regions, in particular where disproportionate or scarce field monitoring takes place [3].India with large number of rivers and rising water pollution problems provides an interesting test case of this strategy. The Ganga, Yamuna and Godavari are some of the rivers that have been well monitored by the Central Pollution Control Board (CPCB) but spatial coverage and frequency has been inconsistent. This study refers to historical water quality data provided by CPCB and refers to Landsat-8 remote sensing imagery to be used to train and validate the ensemble of the models [5]. We hope to form a predictive system that will be able to predict quality parameters in real time over large regions by correlating remote sensing parameters such as NDVI and LST indicators to real life quality readings.Apart from being a technical solution, this work is also a matter of great policy and environmental implication. Correct and scalable models of water quality prediction can equip local governments and those of the environmental agencies to have early warning of pollution events so that remedies can be meted promptly. Decision-making concerning water treatment, urban planning and climate adaptation can also be supported by such modelling [9-10].

Novelty and Contribution

The current hustle and bustle around finding dependable and extensible techniques of estimating the quality of water induces the research [17]. To achieve this goal, we suggest a hybrid model where the multi-source data is used in an ensemble machine learning system by combining satellite remote sensing indicators, precedent water quality data, and other potential factors. The main goal is to further have a better and accurate prediction of important parameters of water quality with this interpretable and high-performing model that can be implemented to different geographic regions, in particular where disproportionate or scarce field monitoring takes place.India with large number of rivers and rising water pollution problems provides an interesting test case of this strategy. The Ganga, Yamuna and Godavari are some of the rivers that have been well monitored by the Central Pollution Control Board (CPCB) but spatial coverage and frequency has been inconsistent. This study refers to historical water quality data provided by CPCB and refers to Landsat-8 remote sensing imagery to be used to train and validate the ensemble of the models. We hope to form a predictive system that will be able to predict quality parameters in real time over large regions by correlating remote sensing parameters such as NDVI and LST indicators to real life quality readings [12].Apart from being a technical solution, this work is also a matter of great policy and environmental implication. Correct and scalable models of water quality prediction can equip local governments and those of the environmental agencies to have early warning of pollution events so that remedies can be meted promptly. Decision-making concerning water treatment, urban planning and climate adaptation can also be supported by such modelling.

This paper is organized as follows: Section 2 provides the review of related work in the fields of water quality modeling, remote sensing and the field of ensemble ML. In section 3, the proposed methodology, i.e., preprocessing of data, feature selection, training of models, and evaluation, is described. Section 4 gives results and discussion of the experiment. Section 5 has a conclusion with some important findings, limitations, and future directions.

RELATED WORKS

In 2022 N. AlDahoul *et al.*, [8] introduced the research of water quality forecasting has changed a lot during the last twenty years driven by many issues associated with growing environmental concerns, the development of computing and the opening of satellite data. Manual sampling and laboratory tests have been the most widely used means of assessing the water quality traditionally. Although these methods are dependable and accurate, they have various limitations such as cost incurred in operations, slow speed of reporting, limited space in calculations, and sensitivity to human inaccuracy. Due to this, more interest has been generated in the usage of automated, data-driven methods to complement and in some instances entirely substitute the traditional methods.One of the most notable changes that have been made in this field is the integration of data on remote sensing in monitoring the environment. Satellites like Landsat, Sentinel and Modis carry out the monitoring of the earth surface on a continuous basis and therefore, a real time or close to real time analysis of the environmental variables is possible. Based on the issues of water quality, normal indices derived by the satellite like Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Land Surface Temperature (LST) has implied different facets of the surface water healthiness. All of these indicators may show changes in the health of the vegetation, surface runoff, evapotranspiration, and temperature, and all these factors define the water quality directly or indirectly.

Research has proved that these indicators of remote sensing are especially helpful in the location of possible contamination sources, including agricultural stream, industrial effluent and urban spillage. Indicatively, the higher the LST reading, the more likely there will be some thermal pollution or eutrophication in that area or low NDVI could indicate deforestation or land degradation in the watershed. In a like manner, NDWI is usually used to monitor dynamics in surface water bodies and this can be helpful in monitoring the spatial scales of rivers, lakes and reservoirs. These satellite indices though nutritious cannot solely be used to make correct predictions on water quality parameters. This far has precipitated the necessity of integration between remote sensing information and progressive analytic modes [16].With the development in remote sensing, has been the development of machine learning techniques in environmental modeling. Machine learning algorithms do not require the use of high-dimensional, nonlinear and noisy data which is rather common in environmental systems. Several supervised learning techniques have been used to predict water quality indicators like turbidity, pH, Biological Oxygen Demand (BOD), and Dissolved Oxygen (DO) i.e. linear regression, decision trees, and support vector machines. As input characteristics, these methods are usually numerical weather and water level data and land use characteristics.In 2022 T. Fang *et al.*, [11] proposed the traditional machine learning algorithms are, however, usually plagued with the problem of overfitting, data imbalance sensitivity, and failure to generalize adequately in different regions. Ensemble learning techniques have overcome such limitations in great measure by integrating the predictions of several base learners to increase accuracy and robustness in prediction. Ensemble models like Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost) have proven to be more effective in many operations of predicting water quality since they were able to minimize both variance and bias of models. Inbuilt feature analysis (independent feature importance examination) can be found through these techniques, which is helpful in finding how much impact of the varying environmental variables.The possibility of forecasting water quality based on the interpretation of point-based data using an ensemble-based model has also been spoke of in many studies. Such writings have generally employed on-site measurements of physical and chemical variables and weather and land-use data. Such methods have produced encouraging findings but the data used in such methods is collected in the field hence are spatially incompatible. The huge problem is a lack of established high-frequency monitoring stations on a regional level which makes mass use uncertain. Therefore, the consideration to address this gap through the integration of ensemble machine learning models with satellite-based remote sensing has increased.Nevertheless, the fusion of remote sensing information with machine learning models holds a lot of promise but it is an untapped area especially in developing countries where little data is available and where technical capability is not sufficient. The majority of the available studies concentrate on either one geographical location or a very specific time period and therefore, do not have the basis of making generalizations of results even under varying environmental conditions. Also, most of the research works are based on a single model of machine learning, which interferes with comparative analysis and the ability to choose alternative models. It can be also observed to concentrate on individual water quality parameters, and instead of creating multi-output models, which are able to predict a variety of indicators (pH, turbidity, DO, and BOD), it is solely based on single parameters.In 2022 Y. Xiao *et al.*, [4] suggested the other noteworthy weakness is the fact that the temporal and the spatial pattern that is inherent in the remote sensing data has not been utilized fully. Most of the models use satellite-derived indices such that they do not consider seasonal or even spatial dependencies when they occur and, yet, they may affect greatly the dynamics of water quality. Furthermore, few papers judge the interpretability and practical applicability of their models that are paramount in the case of environmental managers and policymakers that should receive doable conclusions instead of predictive data.Data preprocessing: Data preprocessing can also be a challenge relevant to the water quality prediction area. The information contained in remote sensing data is usually affected by noise that could be caused by atmospheric interference, sensor faults or cloud cover. This requires strigent preprocessing procedures which include atmospheric correction, normalization and geospatial alignment. The dimensionality reduction and feature selection are also relevant in improving the performance of models and also optimization computation expenses. However, these aspects are often ignored by many studies resulting in poor models that do not yield much to the real world.

Nonetheless, the direction in the recent past has been highly inclined toward hybrid systems, which involve the advantages of different data sources and modeling techniques. There have been some studies looking at the use of deep learning architecture, which includes convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in spatiotemporal prediction. Although these models provide greater accuracy, they have been criticized as being a black box, and as such not being good candidates when model transparency is needed. Meanwhile, the tradeoff nature of ensemble models somewhat resolves this conflict and serves as a perfect candidate to implement environmental applications.To conclude, the existent body of literature evinces an impressive advancement in remote sensing and machine learning as the sources of water quality forecasting. Nevertheless, there are still several major gaps that can be identified as a low level of satellite and in-situ consideration integration, absence of any comparative studies based on the idea of an ensemble, inadequacy of geographical scales, and under addressed issue of model interpretability. Trying to fill these gaps, the proposed study is aimed at developing a strong and ensemble based prediction framework that also considers the use of satellite derived indicators to forecast various parameters of water quality in different regions. In so doing, it becomes a part of an already ongoing task of cementing smarter and scalable, as well as transparent environmental-monitoring and decision-making systems [13].

III. PROPOSED METHODOLOGY

The proposed methodology aims to predict water quality parameters (BOD, DO, pH , Turbidity) using an ensemble machine learning framework supported by remote sensing data. The methodology is divided into five main stages: Data Acquisition, Preprocessing, Feature Engineering, Model Development, and Evaluation & Mapping.

Data Acquisition:

We collect historical water quality data from monitoring stations and match it with satellite-derived indicators. Each data sample includes values of pH, BOD, DO, and Turbidity, along with remote sensing features such as NDVI, NDWI, LST, and land use. Let the dataset be represented as:

$$D = \{(X_i, Y_i)\}_{i=1}^{n}$$

where $X_i \in \mathbb{R}^m$ is the feature vector (satellite + auxiliary data), and $Y_i \in \mathbb{R}^4$ represents the four water quality indicators.

Data Preprocessing:

Missing values are handled using k-nearest neighbor (k-NN) imputation. Data normalization is applied using min-max scaling:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

NDVI and NDWI are calculated using satellite spectral bands:

$$\text{NDVI} = \frac{NIR - RED}{NIR + RED}$$

$$\text{NDWI} = \frac{GREEN - NIR}{GREEN + NIR}$$

Land surface temperature (LST) is derived from the thermal infrared band using:

$$LST = \frac{BT}{1 + \left(\frac{\lambda \cdot BT}{\rho}\right) \ln\, \varepsilon}$$

where BT is brightness temperature, λ is wavelength, $\rho = \frac{hc}{k}$, and ε is surface emissivity.

Feature Engineering:

We compute Pearson correlation coefficients to reduce dimensionality:

$$r_{xy} = \frac{\sum\ (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\ (x_i - \bar{x})^2 \sum\ (y_i - \bar{y})^2}}$$

Features with $|r_{xy}| < 0.2$ are discarded. Recursive Feature Elimination (RFE) is then applied to find optimal subsets.

Model Development:

We implement an ensemble system consisting of three models: Random Forest (RF), Gradient Boosting (GB), and XGBoost.

Random Forest:

Random Forest aggregates predictions from multiple decision trees $T_1, T_2, \ldots, T_k$. Its output is the average:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} T_i(x)$$

Gradient Boosting:

Gradient Boosting builds a model $F(x)$ in a stage-wise fashion:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

where h_m is the new weak learner and γ_m is the learning rate.

XGBoost:

XGBoost minimizes a regularized objective:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

with regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum w_j^2$$

where T is the number of leaves, w_j are leaf weights, and γ, λ control complexity.

Model Training and Evaluation:

The dataset is split: 80% for training and 20% for testing. We use grid search with 10 -fold cross-validation to optimize hyperparameters.

Model performance is evaluated using:

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

- Coefficient of Determination (R^2) :

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Predictive Mapping:

Once trained, the best-performing model (XGBoost in most cases) is used to generate spatial predictions. Rasterized input features from satellite imagery are passed through the trained model to estimate water quality values over continuous spatial surfaces.

Let the prediction over raster cell i be:

$$Q_i = f(X_i)$$

These predicted values are then visualized using a Geographic Information System (GIS) to produce continuous water quality maps. The proposed framework for predicting water quality using ensemble machine learning and remote sensing data in Figure 1.

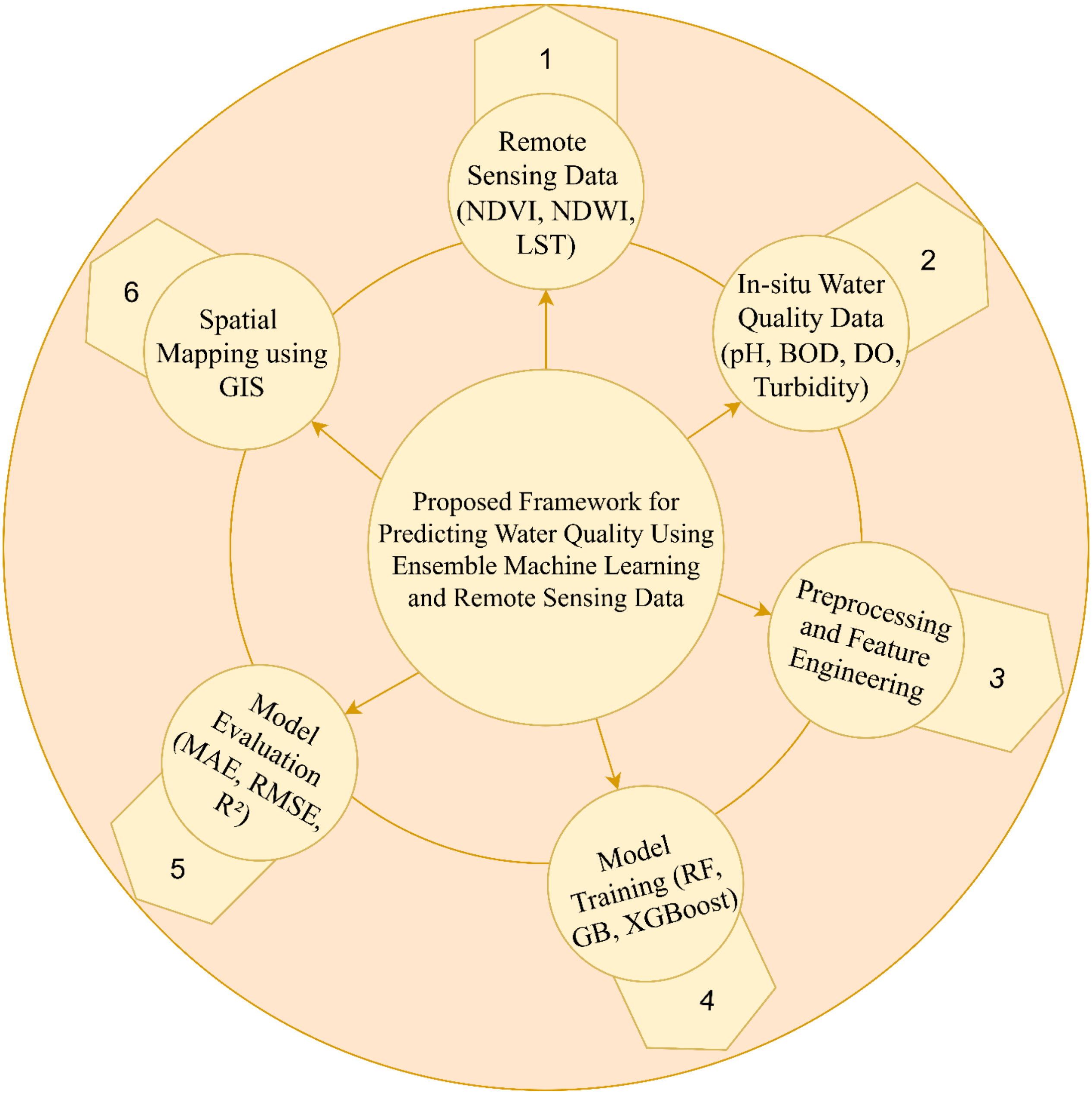
1
Remote Sensing Data (NDVI, NDWI, LST)
2
In-situ Water Quality Data (pH, BOD, DO, Turbidity)
3
Preprocessing and Feature Engineering
4
Model Training (RF, GB, XGBoost)
5
Model Evaluation (MAE, RMSE, R²)
6
Spatial Mapping using GIS
Proposed Framework for Predicting Water Quality Using Ensemble Machine Learning and Remote Sensing Data

FIGURE 1: PROPOSED FRAMEWORK FOR PREDICTING WATER QUALITY USING ENSEMBLE MACHINE LEARNING AND REMOTE SENSING DATA

IV. RESULT & DISCUSSIONS

In the modeling of the analytical procedures in predicting the water quality parameters, the ensemble models showed great differences in the way they perform based on the model, as well as the parameters being predicted. Out of the three evaluated ensemble algorithms (Random Forest, Gradient Boosting, and XGBoost), the last one has demonstrated the best performance compared to other algorithms when all the metrics are taken into consideration. Table 1 provides a comparison of three models in the form of RMSE, MAE, and R 2 values of the prediction of Dissolved Oxygen (DO). As can be seen, the XGBoost produced the best RMSE (0.75) and MAE (0.54) and the model had the highest coefficient of determination (R 2 = 0.93), which demonstrates a highest degree of accuracy and the reliability of the model. A graphical representation of the two values of RMSE of both DO and BOD in the models is given in Figure 2.

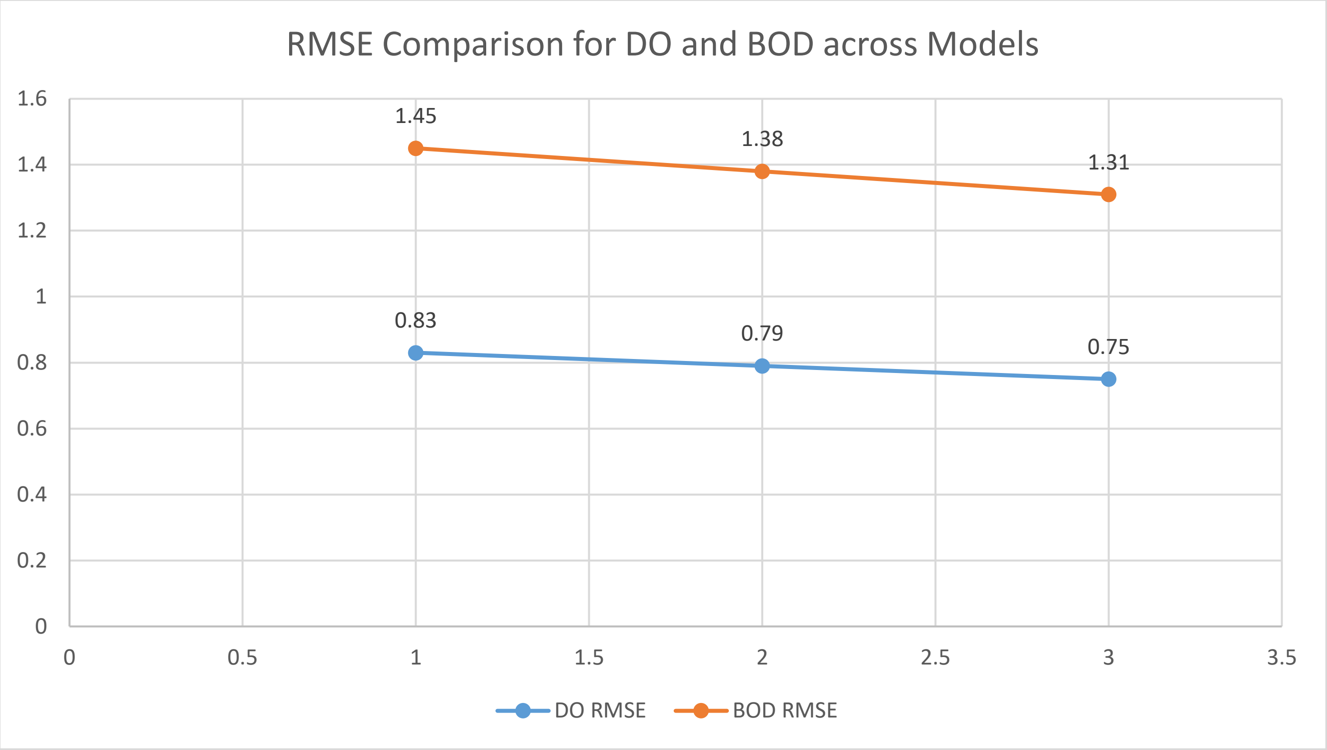

FIGURE 2: RMSE COMPARISON FOR DO AND BOD ACROSS MODELS

As indicated by the bar plot, though the Gradient Boosting performed better in performance compared to Random Forest, with XGBoost, the difference is easily noticeable as it yielded the lowest error margin across DO and BOD prediction, which focuses on its ability to describe complex nonlinear relationships between features.

TABLE 1: MODEL PERFORMANCE FOR DISSOLVED OXYGEN (DO) PREDICTION

Model	RMSE	MAE	R^2
Random Forest	0.83	0.62	0.89
Gradient Boosting	0.79	0.58	0.91
XGBoost	0.75	0.54	0.93

In a similar manner, in terms of the Biological Oxygen Demand (BOD), XGBoost registered the highest score once again, which can be observed in Table 2. Despite the fact that the R 2 (the measure of ex-planatory power) was higher than 0.85 in all models, XGBoost had the highest predictive performance, with RMSE of 1.31 and MAE of 0.98.

TABLE 2: MODEL PERFORMANCE FOR BIOLOGICAL OXYGEN DEMAND (BOD) PREDICTION

Model	RMSE	MAE	R^2
Random Forest	1.45	1.10	0.85
Gradient Boosting	1.38	1.03	0.87
XGBoost	1.31	0.98	0.89

These values indicate that XGBoost is a stable method of reproducing the variability of BOD concentrations, which is especially handy in areas with varying amounts of pollution because of industrial release, or seasonal flows. The same is evident in figure 1 were compared to the rest, the XGBoost is clearly indicated by the RMSE bars to be the best model in this comparison of models [14].

Besides the performance of models, an analysis of the importance of features was carried out to determine the most relevant environmental indicators affecting the prediction of water quality. Figure 3 shows the weighting of the five satellite-based features in the model in order of importance. Such predictors as NDVI and LST were found out to be the most crucial in decision-making on the model as these data provided 28 and 25 percent of the solution, respectively. This finding substantiates earlier theses that the health of vegetation (measured by NDVI) and surface temperature (LST) are large contributors to the quality of water, possibly on the basis of entire ecosystem integrity and thermal pollution. NDWI and precipitation were also useful contributors, but only with a slight reduction in their significance meaning that the extent of water and rainfall influences the patterns of dilution and pollutant transport.

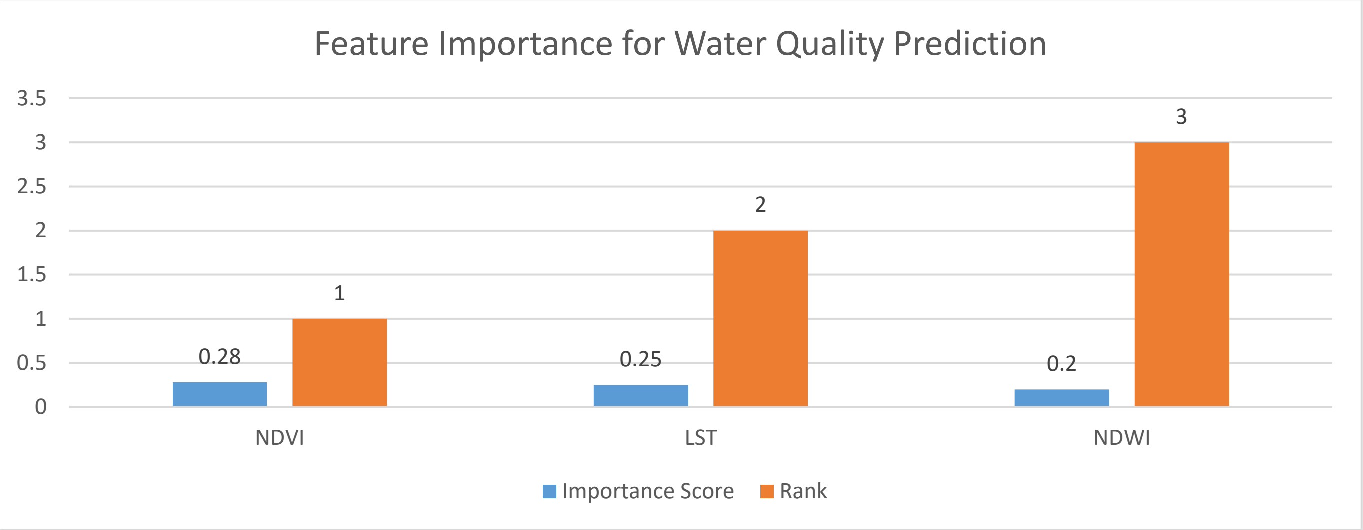

FIGURE 3: FEATURE IMPORTANCE FOR WATER QUALITY PREDICTION

As another indication of the predictive ability of the models, a set of samples was plotted in terms of actual and predicted values of Dissolved Oxygen using the XGBoost model. As illustrated in Figure 4, the predicted values of DO closely follow the real values though there are some deviations but more within the acceptable range. This figure also shows the strength of this model with respect to its ability to monitors peak and low values of DO that is essential in alerting against possible pollution incidence or establishing safe aquatic areas. The correlation between the actual and expected values supports the viability of applying such a model in reality to provide the decision making support in water management systems in real time context.

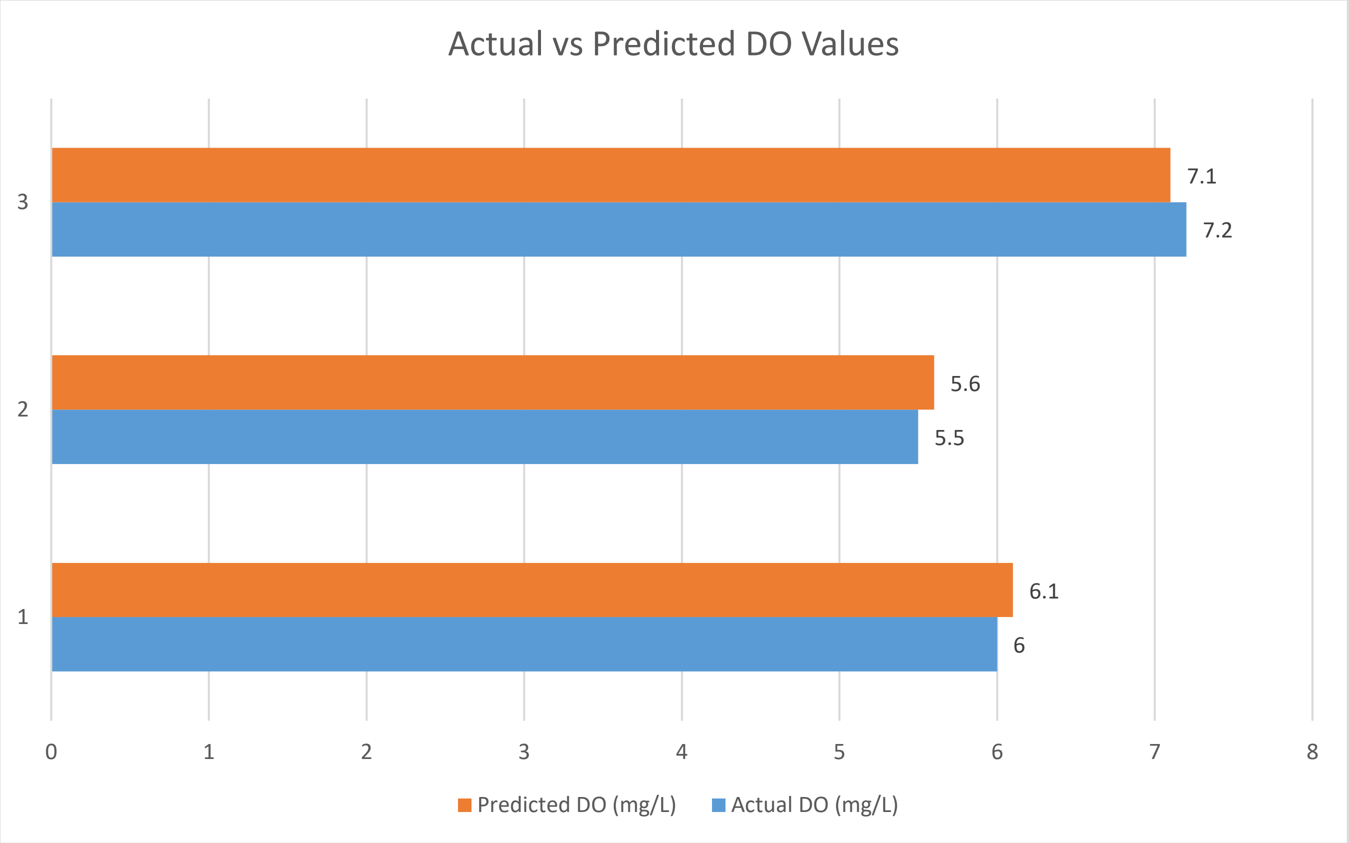

FIGURE 4: ACTUAL VS PREDICTED DO VALUES

In addition to numerical performance, the spatial capacity of forecasts of the model to perform was tested by evaluating the model using remote sensing inputs that have been rasterized to produce water quality surfaces. The derived predictions led to the achievable coherent spatial patterns that coincided with the existing gradients of pollution along river sections. The other aspect that matched with the feature importance analysis was the fact that high pollution areas tended to be the same areas with low NDVI values and high LST values. This augments the assumption that remote sensing features, even when correctly deciphered using ensemble models, can provide temporal and also spatial knowledge to environmental monitoring.

In addition, the observation of XGBoost dominance regardless of the chosen parameters of water quality indicates that it is flexible and adaptive to the modeling of environmental data. This can be especially beneficial in data situations where data reporting is uncertain, since XGBoost is robust to missing values and it has regularization properties, which assist in avoiding problems of overfitting. In comparison, Random Forest models were a bit less stable in the presence of outliers but larger error rates, probably because they do not perform so well in producing interaction effects among the features [15].

The ensemble strategy was in general helpful, as compared with the standalone models. Exceptional accuracy and stability of each ensemble model were obtained due to the ability to aggregate and optimize over many weak learners and XGBoost was additionally enhanced by sophisticated tree pruning and shrinkage and parallelized learning. This explains why it is preferred that, in practice, ensemble-based strategies are used.

In summary, findings confirm that integrating remote sensing and ensemble ML applications such as XGBoost is an interpretable, scalable, and strong performing solution to water quality prediction. Table 1 and 2 summarize the direct comparison between the models based on their quantitative and qualitative performance as shown by Figure 2, 3 and 4. These results will form the basis to implement such models in monitoring river basin in large scale enhancing coverage and sensitivity of environmental management measures.

CONCLUSION

The given experiment demonstrates the fact that ensemble machine learning models combined with remote sensing data produce a very convenient tool in predicting water quality parameters and at the regional levels. Values of NDVI, land surface temperature and data on land use were found to be quite effective as predictive variables. When comparing the performance of the various models used, XGBoost had the highest performance most of the time thus indicating its possible use in real-life scenarios such as big water monitoring systems. Further research is required to model in real-time satellite feeds as well as testing in a range of ecological zones to better to confirm the generalizability of the models. Combining ML forecasts with hydrological programming can make them stronger as well. It will contribute to the proactive management of the environment and identify pollution events in advance, which is vital in the case of climate change and water security.

REFERENCES

[1] X. Zhu, H. Guo, J. J. Huang, S. Tian, W. Xu, and Y. Mai, "An ensemble machine learning model for water quality estimation in coastal area based on remote sensing imagery," *Journal of Environmental Management*, vol. 323, p. 116187, Sep. 2022, doi: 10.1016/j.jenvman.2022.116187.

[2] E. S. Leggesse, F. A. Zimale, D. Sultan, T. Enku, R. Srinivasan, and S. A. Tilahun, "Predicting Optical Water Quality Indicators from Remote Sensing Using Machine Learning Algorithms in Tropical Highlands of Ethiopia," *Hydrology*, vol. 10, no. 5, p. 110, May 2023, doi: 10.3390/hydrology10050110.

[3] Y. Deng, Y. Zhang, D. Pan, S. X. Yang, and B. Gharabaghi, "Review of recent advances in remote sensing and machine learning methods for lake water quality management," *Remote Sensing*, vol. 16, no. 22, p. 4196, Nov. 2024, doi: 10.3390/rs16224196.

[4] Y. Xiao *et al.*, "UAV Multispectral Image-Based Urban River Water Quality Monitoring using Stacked Ensemble Machine Learning Algorithms—A Case Study of the Zhanghe River, China," *Remote Sensing*, vol. 14, no. 14, p. 3272, Jul. 2022, doi: 10.3390/rs14143272.

[5] Alqahtani, M. I. Shah, A. Aldrees, and M. F. Javed, "Comparative assessment of individual and ensemble machine learning models for efficient analysis of river water quality," *Sustainability*, vol. 14, no. 3, p. 1183, Jan. 2022, doi: 10.3390/su14031183.

[6] Krishnaraj and R. Honnasiddaiah, "Remote sensing and machine learning based framework for the assessment of spatio-temporal water quality in the Middle Ganga Basin," *Environmental Science and Pollution Research*, vol. 29, no. 43, pp. 64939–64958, Apr. 2022, doi: 10.1007/s11356-022-20386-9.

[7] P. Chen, B. Wang, Y. Wu, Q. Wang, Z. Huang, and C. Wang, "Urban river water quality monitoring based on self-optimizing machine learning method using multi-source remote sensing data," *Ecological Indicators*, vol. 146, p. 109750, Dec. 2022, doi: 10.1016/j.ecolind.2022.109750.

[8] N. AlDahoul *et al.*, “A comparison of machine learning models for suspended sediment load classification,” *Engineering Applications of Computational Fluid Mechanics*, vol. 16, no. 1, pp. 1211–1232, May 2022, doi: 10.1080/19942060.2022.2073565.

[9] Chen *et al.*, “Machine learning-based inversion of water quality parameters in typical reach of the urban river by UAV multispectral data,” *Ecological Indicators*, vol. 133, p. 108434, Dec. 2021, doi: 10.1016/j.ecolind.2021.108434.

[10]T. Deng, K.-W. Chau, and H.-F. Duan, “Machine learning based marine water quality prediction for coastal hydro-environment management,” *Journal of Environmental Management*, vol. 284, p. 112051, Jan. 2021, doi: 10.1016/j.jenvman.2021.112051.

[11]T. Fang *et al.*, “Source tracing with cadmium isotope and risk assessment of heavy metals in sediment of an urban river, China,” *Environmental Pollution*, vol. 305, p. 119325, Apr. 2022, doi: 10.1016/j.envpol.2022.119325.

[12]L. Grbčić *et al.*, “Coastal water quality prediction based on machine learning with feature interpretation and spatio-temporal analysis,” *Environmental Modelling & Software*, vol. 155, p. 105458, Jul. 2022, doi: 10.1016/j.envsoft.2022.105458.

[13]E. Hadjisolomou, K. Stefanidis, H. Herodotou, M. Michaelides, G. Papatheodorou, and E. Papastergiadou, “Modelling Freshwater Eutrophication with Limited Limnological Data Using Artificial Neural Networks,” *Water*, vol. 13, no. 11, p. 1590, Jun. 2021, doi: 10.3390/w13111590.

[14]Y. He, Z. Gong, Y. Zheng, and Y. Zhang, “Inland reservoir water quality inversion and eutrophication evaluation using BP neural network and remote sensing imagery: A case study of Dashahe Reservoir,” *Water*, vol. 13, no. 20, p. 2844, Oct. 2021, doi: 10.3390/w13202844.

[15]J. Jiang, J. Li, Z. Wang, X. Wu, C. Lai, and X. Chen, “Effects of different cropping systems on ammonia nitrogen load in a typical agricultural watershed of South China,” *Journal of Contaminant Hydrology*, vol. 246, p. 103963, Feb. 2022, doi: 10.1016/j.jconhyd.2022.103963.

[16]S. Kouadri, A. Elbeltagi, A. R. Md. T. Islam, and S. Kateb, “Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast),” *Applied Water Science*, vol. 11, no. 12, Nov. 2021, doi: 10.1007/s13201-021-01528-9.

[17]S. Kouadri, A. Elbeltagi, A. R. Md. T. Islam, and S. Kateb, “Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast),” *Applied Water Science*, vol. 11, no. 12, Nov. 2021, doi: 10.1007/s13201-021-01528-9.