International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Comparative Analysis of Machine Learning Models for Water Quality Prediction Using Remote Sensing Data

Gowshik Sabarish A M, Mohamed Harshath S, Madheshwaran S J, Felix Robin S
Department of Civil Engineering, Kumaraguru College of Technology, Coimbatore-641049, Tamil Nadu, India.

Nishant S
Assistant Professor, Department of Civil Engineering, Kumaraguru College of Technology, Coimbatore-641049, Tamil Nadu, India.

Abstract: Water quality monitoring is an essential process for ecological balance and sustainable environmental management. However, the traditional field-based sampling and testing, though highly accurate, are time consuming and spatially limited. This explores the integrated work of remote sensing and machine learning techniques for water quality monitoring. Water samples were collected form Ukkadam Periyakulam Lake, Coimbatore and the corresponding Sentinel-2 spectral data were extracted from Google Earth Engine. Water quality parameters including turbidity, dissolved oxygen (DO), pH, total suspended solids (TSS), biochemical oxygen demand (BOD), and chemical oxygen demand (COD) were predicted using multiple machine learning models such as random forest (RF), extreme gradient boosting (XGBoost) and support vector regression (SVR). The results indicated that turbidity was predicted with highest accuracy with random forest achieving a test R^2 of 0.84, indicating strong relation between the observed and predicted values. Dissolved oxygen showed moderate prediction with SVR performing best generalisation among the three models. pH and TSS showed reasonable trends with acceptable accuracy, while BOD and COD has showed relatively poor results due to their weak spectral dependency. Overall, the results from this study highlights the effective performance of machine learning models integrated with remote sensing on optically active water quality parameters.

Keywords: Remote Sensing, Machine Learning, Water Quality Parameters, Spectral data.

1. INTRODUCTION

Water quality assessment is an important factor that is to be considered while thinking of ecological balance, ecosystem health, and sustainable practices (Zhang et al., 2025) (Awasthi et al., 2025). Major physiochemical parameters such as turbidity, dissolved oxygen (DO), potential of hydrogen (pH), total suspended solids (TSS), biochemical oxygen demand (BOD), and chemical oxygen demand (COD) are considered widely for evaluating the conditions of water quality (Xiao et al., 2022), as they reflect the physical, chemical and biological process taking place within the water body. Continuous monitoring of these parameters is a must as they are highly dynamic and can get influenced by multiple environmental factors, in order to manage the water quality effectively. Conventional approach of in-situ sampling and testing requires more time and is labour-intensive. Utilizing the modern machine learning and remote sensing techniques has made possible to overcome these limitations with ease and support the monitoring with spatial and temporal analysis, which is also a limitation in the traditional sampling method (Li et al., 2022) (Tian et al., 2022).

Many earlier studies have shown tremendous results of predicting water quality parameters by integrating remote sensing data with machine learning methods. Multispectral sensors such as Sentinel-2 provide valuable spectral information regarding the reflectance from the water surface, which can be used to predict water quality parameters, especially the optically active ones (Chowdhury et al., 2025; Dawn et al., 2025a). Integrating them with machine learning models will further enhance the capability of predictive modelling in water quality assessment. Machine learning models have the ability to capture the complex and non-linear relationship between the spectral information and water quality parameter, which is hard in traditional statistical methods. Machine learning algorithms such as random forest (RF), extreme gradient boosting (XGBoost) and support vector regression (SVR) have proven their strong predictive capabilities consistently when compared to the rest. Random forest has been widely recognised for its robustness and handling large dataset (Breiman, 2001), XGBoost is known for its iterative boosting and avoiding overfitting by incorporating regularizations (T. Chen & Guestrin, 2016; Kim et al., 2014) and SVR is has the capability to handle the complex relations especially when the dataset is limited (Smola & Schölkopf, 2004) .

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A number of studies have used these machine learning algorithms in particular for assessment of water quality predictions. RF has successfully achieved high accuracy in parameters predictions like turbidity and suspended matter under stable environmental conditions (Liu et al., 2021; Silveira Kupssinskü et al., 2020). Additionally, RF has also been used in water quality classification and multi parameter analysis in environmental systems (Jena et al., 2023; Prasad et al., 2021). On the other hand, XGBoost has been used in dissolved oxygen prediction and key parameters and is has shown greater predictive ability, especially when combined with feature selection and optimization techniques (Tiyasha et al., 2021; Xiao et al., 2022). It has also shown better performance than other models when its use has been extended to satellite-based and multi-source modelling (Dawn et al., 2025b; Zhao et al., 2024). SVR has been applied in predicting parameter with non-linear relationship with spectral data such, particularly the ones with smaller dataset (Kim et al., 2014). Furthermore, SVR has shown improved performance when integrated with hybrid and ensemble approaches in complex environments (W. Chen et al., 2026a; Choudhary et al., 2025). Despite the advancements made in water quality assessment through remote sensing and machine learning techniques, few challenges remain to stay (Gao et al., 2024; Singh et al., 2026). Turbidity and total suspended solids have been consistently predicted with higher accuracies due to the nature of exhibiting strong relations with spectral data; they are optically active parameters which can be assessed effectively. In contrast, non-optically active parameters like BOD and COD are difficult to estimate due to complex biochemical influence and poor relation with spectral data. Though many studies have attempted to overcome this limitations by incorporating various complex variables, the results still vary across datasets. (Dawn et al., 2025b).

Based on the observations from the previous studies, this study focuses on evaluating multiple machine learning models for key water quality parameters using Sentinel-2 satellite data. The study further analyses parameter wise model performance to understand the characteristics of optical and non-optical parameters. By integrating remote sensing with machine learning, this study aims to contribute towards the development of efficient and scalable data-driven water quality monitoring (Choudhary et al., 2025; Yee Wong et al., 2023).

2. MATERIALS AND METHODS

Study Area

The area selected for this study is Ukkadam Periyakulam Lake, located in Coimbatore district of Tamil Nadu, India. The coordinates of the lake ranges approximately between latitudes 10.98°N–11.00°N and longitudes 76.95°E–76.97°E. The lake experiences a tropical climate ranging from moderate to high temperatures and seasonal rainfall driven by both south-west and north-east monsoons. The average temperature this region experiences ranges between 25°C and 32°C and the annual rainfall is approximately between 600 to 700 mm.

It is an urban freshwater lake that forms a part of Noyyal river basin, contributing majorly towards the groundwater recharge. The primary inflow is from the storm water channels and the urban drainage systems. Due to its urban setting, the lake undergoes anthropogenic pressures such as solid waste disposal, urban runoff, release of untreated waste water and nutrients loading from its surroundings, contributing to eutrophication and deterioration of the quality of water resulting in increased concentration of toxic compounds.

The deteriorating water quality conditions, particularly the presence of elevated nutrient levels and toxic compounds, make this lake an important case study for evaluating water quality from an aquaculture perspective. Parameters such as dissolved oxygen and pH are critical for aquatic organism health, and their assessment provides insights into the suitability of such urban water bodies for aquaculture-related applications.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

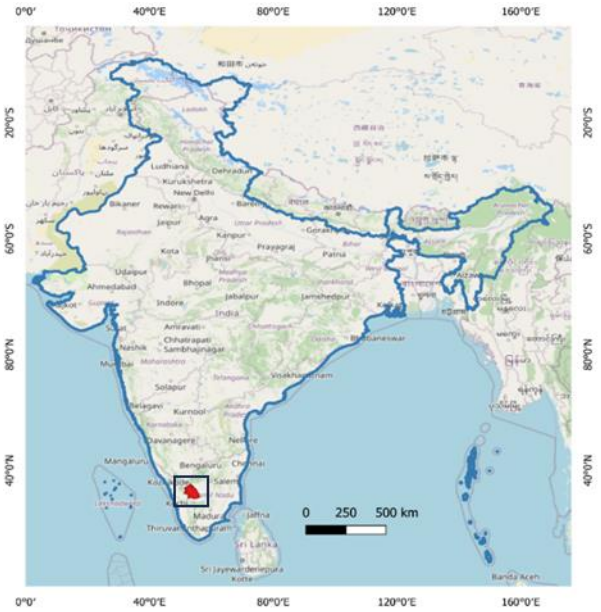


Fig 1: Location of Coimbatore district in India.

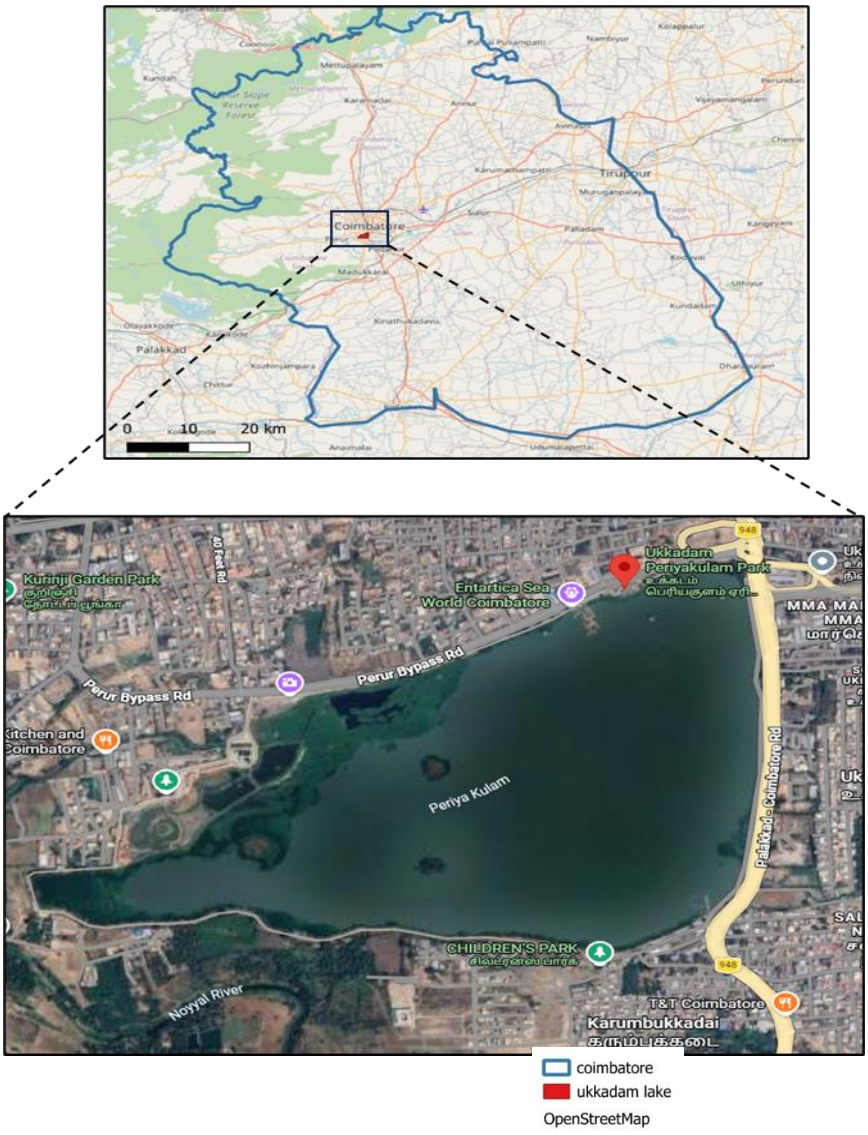


Fig 2: Location of study area

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Water Sampling and Laboratory testing

A total of 104 water samples were collected from multiple locations including the inlet regions and mid lake regions across the Ukkadam Periyakulam lake to analyze the spatial variation of several water quality parameters. The sampling was carried out in three different phases to ensure the temporal variability. In the first phase, 26 samples were collected on December 3, 2025, in the second phase, 26 samples were collected on January 27, 2026, and in third and final phase, 52 samples were collected on February 6, 2026. On all the dates, the samples were collected between the time frame of 10:30 AM to 11:30 AM, in order to maintain the consistency of overpass time of the Sentinel 2 satellites. This synchronisation ensures the correspondence of in-situ sample collection and spectral data derived from satellites, considering the revisit interval of 5 days.

All the samples were collected at the depth of 10 cm (approx..) from the water surface. The samples were collected in clean water bottles and were properly labelled during the collection process. Within half an hour from collection, the samples were stored in freezers in order to maintain the integrity and prevent biological changes. Laboratory analysis was conducted in the Environmental laboratory of Kumaraguru College of Technology, Coimbatore. The water quality parameters including turbidity, potential of hydrogen (pH), dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), and total suspended solids (TSS) were tested and analysed using the standard procedures as described in American Public Health Association (APHA) methods (Gikas et al., 2023). Turbidity was determined by Nephelometric method (method 2130 B in APHA), pH by Electrometric method (method 4500-H^+B in APHA), DO Winkler method (method 500-O C in APHA), COD by Dichromate closed reflux method (method 5220 D in APHA), TSS by Gravimetric method (method 2540 D in APHA) and BOD was calculated using COD values by the ratio BOD/COD=0.6. For aquaculture systems, the BOD/COD ratio usually ranges between 0.2 and 0.5, representing optimal biodegradability and stable water conditions. Here in the study, 0.6 is adopted, indicating higher biodegradability, which is characteristic of wastewater influenced urban water bodies.

The measured water quality data were then used along with the spectral data acquired from satellites to develop machine learning models for monitoring and managing water quality for aquaculture conditions.

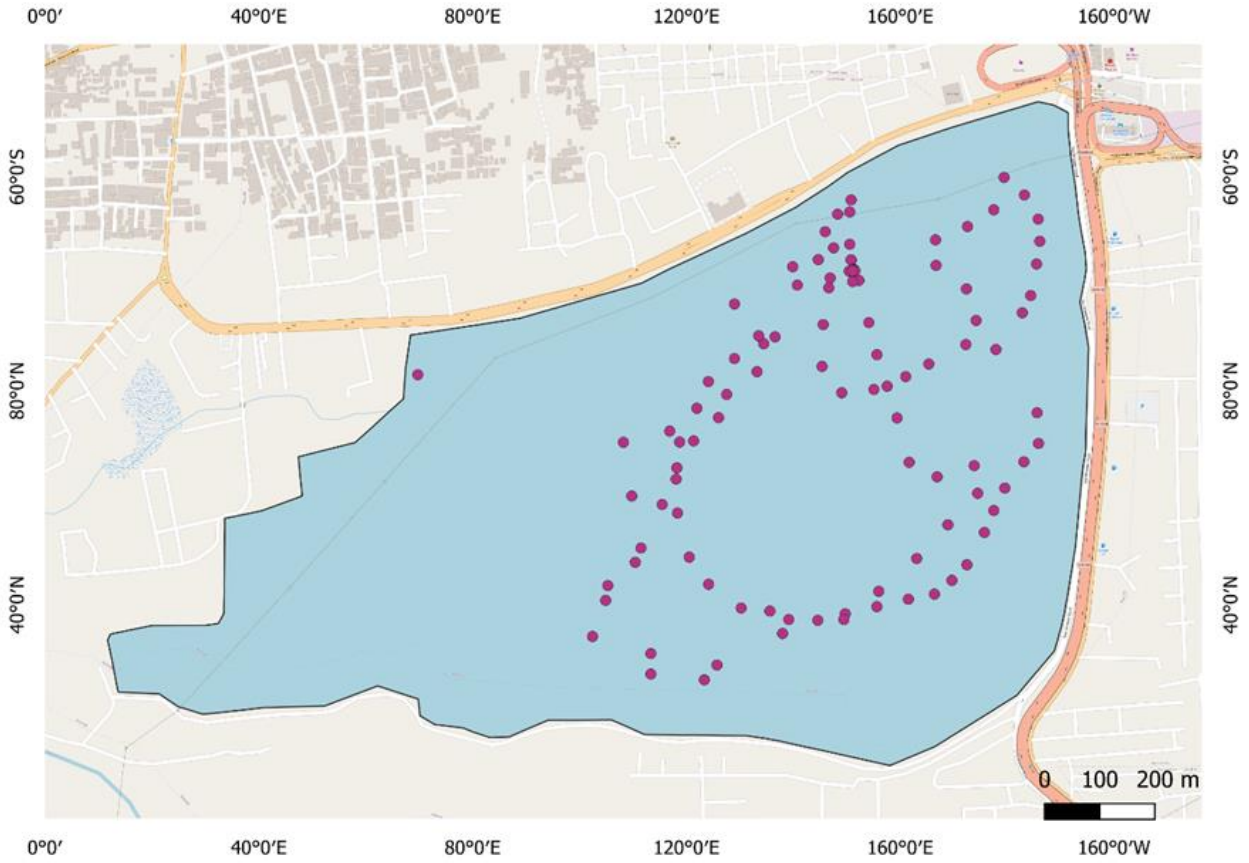


Fig 3: Location of sampling points

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Satellite Data and Preprocessing

Satellite data used in this study were obtained from Sentinel 2 mission (Deng et al., 2024), which operated and managed by the Europe Space Agency. Surface reflectance imagery from Sentinel 2 level 2A product (COPERNICUS/S2_SR_HARMONIZED) was used in the study; it helps in providing the atmospherically corrected multispectral data across all the 13 spectral bands provided by the Sentinel 2 satellites. The bands acquired from the satellite includes B1 (coastal aerosol), B2 (blue bands), B3 (green bands), B4 (red bands), B5 (red edge 1), B6 (red edge 2), B7(red edge 3), B8 (near infrared (NIR)), B8A (narrow NIR), B9 (water vapour), B11 (shortwave infrared 1 (SWIR)), and B12 (SWIR 2).

Satellite data processing and extraction were performed using Google Earth Engine (GEE) (França et al., 2026; Salas et al., 2025). Sampling locations (latitude and longitude) and date were provided as the input information to extract the multispectral band data. For each sampling point, a temporal window of ± 2 to ± 3 days was used for selecting the multispectral images. The images were filtered based on spatial location and sorted by cloud cover percentage in order to obtain the least cloudy values possible. As a result, spectral bands were obtained by applying scaling factors on the digital number to convert them to surface reflectance. The scaling factor used on digital numbers is 0.0001 and 0.001 was used on atmospheric parameters such as Aerosol Optical Thickness (AOT) and Water Vapour (WVP). A buffer region of 20 m was used around each sampling point to reduce spatial noise.

In addition to the spectral bands, temperature data was extracted using ERA5-Land Hourly dataset. Air temperature at the height of 2 m corresponding to the time of 11:00 AM was extracted for each sampling point. They were converted from Kelvin to degree Celsius and included to the dataset as additional information.

The usage of GEE ensured efficient spectral data extraction from the satellites, saving time by reducing the manual work of downloading the spectral images and performing manual atmospheric corrections. This approach ensured improved temporal alignments and reduced avoidable errors. The final dataset now consists of the lab measured values of the parameters, satellite-derived spectral bands, atmospheric parameters, and temperature data, making it ready to use for machine learning models to predict the concentration of the various parameters and analyze the quality variations, with particular relevance to aquaculture applications.

Feature Selection

Feature selection was used in this study to improvise the model performance by identifying the relevant input variables and reducing the redundancy in the dataset (Padilla-Mendoza et al., 2023; Zhu et al., 2023). Feature selection includes selection of useful bands and band ratios for model development. Spectral bands of Sentinel-2 such as B2 (blue), B3 (green), B4 (red), B8 (near infrared), B11 and B2 (shortwave infrared), and the band ratios such as green-to-red and blue-to-green have been used in this study. The feature selection aimed to retain only the most important variables, thereby reducing the noise and improvising the model generalisation, ensuring efficient and reliable mode development.

Table 1: Feature Engineering used in this study

Type	Feature used
Spectral Bands	B2, B3, B4, B8, B11, B12
Spectral indices	NDWI = (B3 – B8) / (B3 + B8) MNDWI = (B3 – B11) / (B3 + B11)
Band ratios	B3/B2, B4/B3, B8/B4
Transformed features	log(B3), log(B8)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Model Development

The key objective of this study is to develop predictive machine learning models for estimating water quality parameters relevant to aquaculture using spectral data obtained from the satellite. The input variables consisted of the surface reflectance values obtained from multi spectral images and the output variables included the predicted values of turbidity, dissolved oxygen, pH, total suspended solids, biochemical oxygen demand, chemical oxygen demand, and ammonia. In this study, a regression-based approach was adopted, because the models were to predict the numerical values rather than the categories. Different machine learning algorithms were trained and analysed to identify the best suitable model for each parameter. Three primary machine learning algorithms were used: Support Vector Regression (SVR), Random Forest Regression (RF), and Extreme Gradient Boosting Regression (XGBoost). Since the relation between the spectral data and the water quality parameters are complex and non-linear and changes across the different parameters, multiple machine learning models were used to identify the best suitable model for each parameter, rather than relying on one single model for all parameters (Chowdhury et al., 2025).

Support Vector Regression (SVR)

Known for its effectiveness in modelling non-linear relationships, SVR is one of the widely used machine learning algorithm. Using kernel functions, SVR transforms the input data into a higher-dimensional feature space, where an optimal hyperspace is constructed to reduce the prediction errors.

In this study, Radial Based Function (RBF) kernel was used due to its special ability if capturing complex non-linear patterns in water quality data. In order to improve the performance of SVR, feature scaling was applied to the input data, prior to the model training phase. As a result, SVR model showed a strong predictive capability especially for parameters like dissolved oxygen and pH; these parameters are known to contain prominent non-linear relationships (Kim et al., 2014; Smola & Schölkopf, 2004).

Random Forest Regression (RF)

Random Forest is a prominent ensemble machine learning algorithm that creates decision trees during the training process and use them to generate the final predictions. As its name suggests, it creates multiple trees (forest) and each tree is trained on a random subset of data and features, finally combining them all and providing the average of the trees as the final result. This ensures the robustness and generalization of the cases that involve complex dataset.

In this study, RF performed effectively in turbidity prediction due to its ability to handle non-linear and noisy data (Breiman, 2001; Zheng et al., 2024).

Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced gradient boosting machine learning algorithm, that builds sequential models, where each model tries to improve and correct the errors made by its predecessor. This iterative method ensures the improved prediction achieving the highest accuracy possible. XGBoost uses parallel processing and L1 (Lasso) and L2 (Ridge) regularisation to prevent the model from producing over fitting results and ensure proper generalization. In this study, XGBoost outperformed other models during the prediction and analysis of total suspended solids (TSS) (T. Chen & Guestrin, 2016; Xiao et al., 2022).

All machine learning model predictions for the water quality parameters were implemented using Python in a cloud based Jupyter Notebook platform called Google Colab. All the necessary libraries such as scikit-learn, Numpy, etc., were used for the model development and analysis. Two types of input configurations were considered in this study, namely satellite band-only inputs and hybrid inputs combining spectral bands with auxiliary parameters. The performance of these approaches was evaluated during the model assessment phase.

Linear Regression Model

Linear regression models create a direct linear relationship between the input variable and target variables, which makes it useful for understanding the fundamental relation between the spectral information and the water quality parameters. The main limitation in these models is that they limited to capture the non-linear relation between the spectral data and the water quality parameters. Previous studies have also stated about the underperformance of linear regression models while predicting complex parameters (Gao et al., 2024).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

In this study, linear regression models were developed to serve as a baseline for comparing their performance with the machine learning models

Model Evaluation

The data was split into training set (70%) and testing set (30%), to ensure the representative distribution of the samples, prior to model training in order to prevent data leakage and hide the testing set during the training phase.

To ensure further robustness of the model, 5-fold cross validation approach was used to the dataset. In this approach, the dataset was divided into 5 folds, ensuring 70% (training split) of each fold serve as training set and the remaining 30% was for testing. This method of validation ensured that every part of the dataset undergoes training and testing and the prediction results remain consistent throughout the process. Additionally, this approach helps in detecting overfitting and increase the reliability of the model.

The prediction results of the machine learning models were evaluated using the statistical metrics to assess their accuracy and reliability. Since a single metric cannot describe the performance of the of a model completely, three primary metrics were used to evaluate different aspects of prediction error and model reliability (Campos et al., 2026).

Coefficient of Determination (R^2)

The coefficient of determination (R^2) is a statistical evaluation metric that provides the amount of variance in the predicted data given by the machine learning model. This metric says how well the predicted results match with the actual values.

The R^2 value is calculated by using the formula:

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

Where:

- y_i = observed (actual) value
- $\hat{y}_i$ = predicted value
- $\bar{y}$ = mean of observed values
- n = total number of observations

R^2 values range between 0 to 1, where values closer to 1 usually indicates that the model is performing better. This explains that the model was able to explain a larger portion of variability in the data. This study utilizes R^2 to evaluate the overall goodness of fit of the models and how well the model captures the relationship between spectral data and the water quality parameters.

Root Mean Square Error (RMSE)

Root Mean Square (RMSE) calculates the average magnitude of prediction errors by squaring the differences between the predicted values and the actual values, then calculating the square root of the average of the squared differences.

The RMSE value is calculated by using the formula:

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

Where:

- y_i = observed (actual) value
- $\hat{y}_i$ = predicted value
- n = total number of observations

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

In RMSE calculation, higher weightage is allocated to the large errors due to the squaring of differences. This helps in identifying the models that cause significant deviations in predictions. In this study, RMSE was used to quantify the overall accuracy and penalize the models that produce large errors.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) provides the magnitude of prediction errors by calculating the average absolute of the difference between the predicted values and the actual values.

MAE is calculated by using the formula:

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

Where:

- y_i = observed (actual) value
- $\hat{y}_i$ = predicted value
- n = total number of observations

Unlike RMSE, MAE does not allocate any weightage to the large error values. It treats every error equally, making MAE a useful metric to understand the average error. In this study, MAE was used alongside RMSE to provide a clear understanding of model performance, ensuring the evaluation is not overly influenced by extreme values.

Water Quality Index (WQI)

The water quality index approach was adopted to provide a simplified version of the overall water quality by combining multiple parameter predictions into a single numerical value. In this study, Weighted Arithmetic Water Quality Index (WAWQI) method was adopted due to its simplicity and flexibility, unlike other complex index formulas (Lencha et al., 2021; Uddin et al., 2023). WAWQI has the advantage of allowing explicit assignment of weights to individual parameters according to their environmental and fish health impacts.

Alternative WQI approaches like Entropy Weighted WQI (EW WQI), Canadian Council of Ministers of the Environment WQI (CCME WQI), Oregon WQI, and Irish Water Quality Index (IEWQI) were considered but not adopted. The CCME WQI was primarily designed for extensive temporal datasets and long-term monitoring, making it less suitable for predicted values. The Oregon WQI follows a fixed parameter structure and does not allow modification of parameter importance, limiting its applicability for domain-specific contexts. The EW WQI method determines parameter weights based on statistical variability, which may not reflect the ecological importance of critical parameters such as dissolved oxygen in aquaculture environments. Similarly, the IEWQI is primarily developed for coastal and transitional water quality assessment under regulatory frameworks, which may limit its adaptability for aquaculture-specific conditions and machine learning-based predictive modelling (Baharudin et al., 2021; Gikas et al., 2023).

The WQI was calculated using the predicted values of the aquaculture relevant key water quality parameters provided from the developed machine learning models. Each parameter was assigned a weight considering their impacts on fish health and aquatic ecosystem.

The overall WQI was calculated using the WAWQI formula:

$$WQI = \frac{\sum(W_i x Q_i)}{\sum W_i}$$

Where:

- W_i = weightage
- Q_i = quality rating

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Since the sum of the assigned weights is unity, the equation can we written as:

$$WQI = \sum(W_i \times Q_i)$$

The quality rating Q_i, was calculated using parameter specific formulas based on the impacts of the parameter. Parameters like turbidity, BOD, COD, and TSS were calculated using the formula:

$$Q_i = \frac{C_i}{S_i} \times 100$$

Where:

- C_i = is the predicted value
- S_i = corresponding standard values

This formula was used in order to state that higher concentrations indicate deterioration in water quality.

In case of dissolved oxygen (DO), which has a positive influence on water quality, inverse formula approach was used to ensure that lower DO indicates poor water quality. The formula used was:

$$Q_{DO} = \frac{S_{DO} - C_{DO}}{S_{DO}} \times 100$$

For pH, which affects the aquatic life through both acidic and alkaline deviations, the quality rating was calculated based on its deviation from neutral conditions:

$$Q_{pH} = \frac{|C_{pH} - 7|}{S_{pH} - 7} \times 100$$

All quality ratings were constrained to be positive values in order to maintain consistency in the index calculations. The calculated WQI values were then classified into different categories to facilitate interpretation of water quality conditions. All the standard permissible limits (S_i) are as per the Food and Agricultural Organisation (FAO). These guidelines are widely adopted in aquaculture research and practice.

Spatial Mapping

Spatial mapping was done to visualise the distribution of the water quality parameters across the study area. The predicted results provided by the machine learning models were used to generate the spatial variability of the parameters. Geographic Information Systems (GIS) was used to interpolate the predicted values. IDW interpolation method was adopted in this study to estimate the values of the unsampled locations, who's values were based on the nearby sampling points. This approach allows continuous representation of water quality conditions and helps identify the patterns and potential pollution hotspots (W. Chen et al., 2026b; Hridoy et al., 2025).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

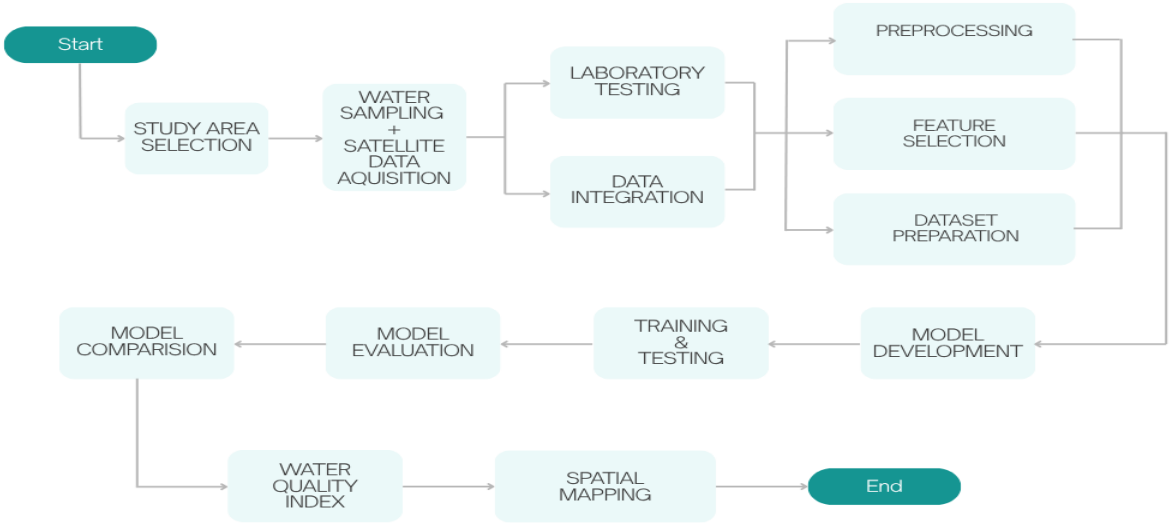


Fig 4: Methodology flowchart

3. Results and Discussion

Table 2: Performance comparison of machine learning models for water quality parameters

Parameter	Dataset	Model	R^2	MAE	RMSE
Turbidity	Training	RF	0.976	1.127	1.471
		XGB	0.975	1.117	1.514
		SVR	0.853	3.133	3.671
	Testing	RF	0.845	2.476	3.57
		XGB	0.728	3.36	4.735
		SVR	0.818	2.988	3.874
DO	Training	RF	0.769	0.92	1.434
		XGB	0.844	0.657	1.017
		SVR	0.646	1.107	1.774
	Testing	RF	0.604	1.022	1.745
		XGB	0.644	1.043	1.654
		SVR	0.777	0.876	1.309
pH	Training	RF	0.953	0.061	0.093
		XGB	0.631	0.179	0.26
		SVR	0.767	0.112	0.207
	Testing	RF	0.551	0.161	0.265
		XGB	0.503	0.194	0.279
		SVR	0.699	0.123	0.217
TSS	Training	RF	0.933	123.699	171.701
		XGB	0.942	77.178	158.807

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

		SVR	0.301	332.379	555.991
	Testing	RF	0.537	225.693	419.696
		XGB	0.62	281.418	379.869
		SVR	0.312	328.949	511.29
BOD	Training	RF	0.205	44.102	115.121
		XGB	0.979	5.046	18.727
		SVR	0.023	50.464	127.623
	Testing	RF	-0.137	113.355	301.799
		XGB	-0.1195	117.47	299.468
		SVR	-0.108	110.981	298.007
COD	Training	RF	0.197	73.81	192.923
		XGB	0.978	8.47	31.5426
		SVR	0.023	84.108	212.708
	Testing	RF	-0.142	189.338	503.928
		XGB	-0.14	198.978	503.772
		SVR	-0.108	185	496.665

The performance of machine learning models random forest (RF), extreme gradient boosting (XGBoost), and support vector regression (SVR) were evaluated and analysed for predicting water quality parameters using Sentinel-2 derived spectral bands. The evaluation was carried out by using coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) for both training and testing datasets.

Among all the parameters, turbidity has shown the highest prediction accuracy, proving its strong relationship with spectral reflectance and confirming its optically active nature. Both RF and XGBoost has similar as well as high training R^2 of 0.976 and 0.975 respectively. However, RF has performed the best with test R^2 of 0.845 and test RMSE of 3.57, followed by SVR, and XGBoost showed comparatively lower generalisation. The prediction performance of DO was moderate when compared to turbidity. XGBoost has shown the highest training dataset R^2 of 0.844. However, in the testing dataset, SVR outperformed other models with the highest R^2 of 0.777 and lowest RMSE of 1.309, indicating better generalisation. SVR was followed by XGBoost and RF. The prediction of pH shows relatively strong performance across models. RF has achieved the highest training R^2 of 0.953 and SVR demonstrated better testing performance with R^2 of 0.699 and lowest RMSE. This indicates that SVR provides better generalisation. TSS showed strong training performances, but moderate testing accuracy. With testing R^2 as 0.62, XGBoost has outperformed RF and SVR. SVR performed poorly with lower accuracy in both training and testing datasets. In contrast, both BOD and COD showed poor performance across all models, although XGBoost achieved high training accuracy for both the parameters. The testing performance was negative for all models, indicating poor generalisation and non-optical nature.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

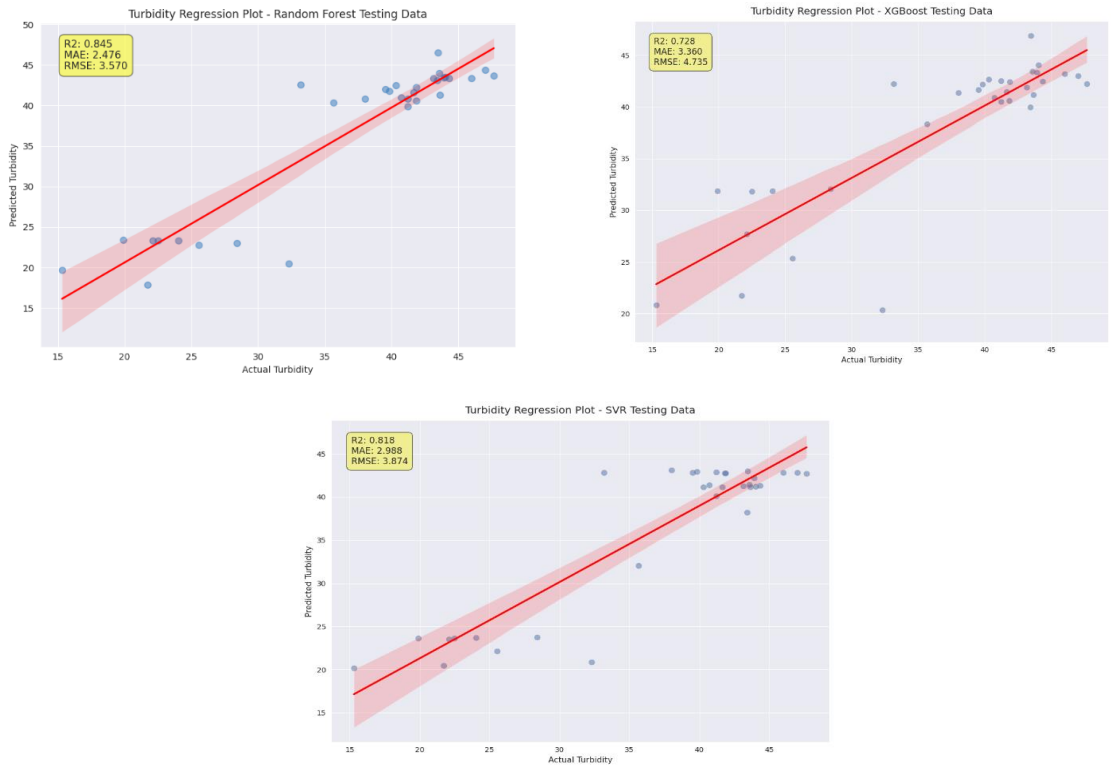


Fig 5: Scatter plots of actual vs predicted turbidity

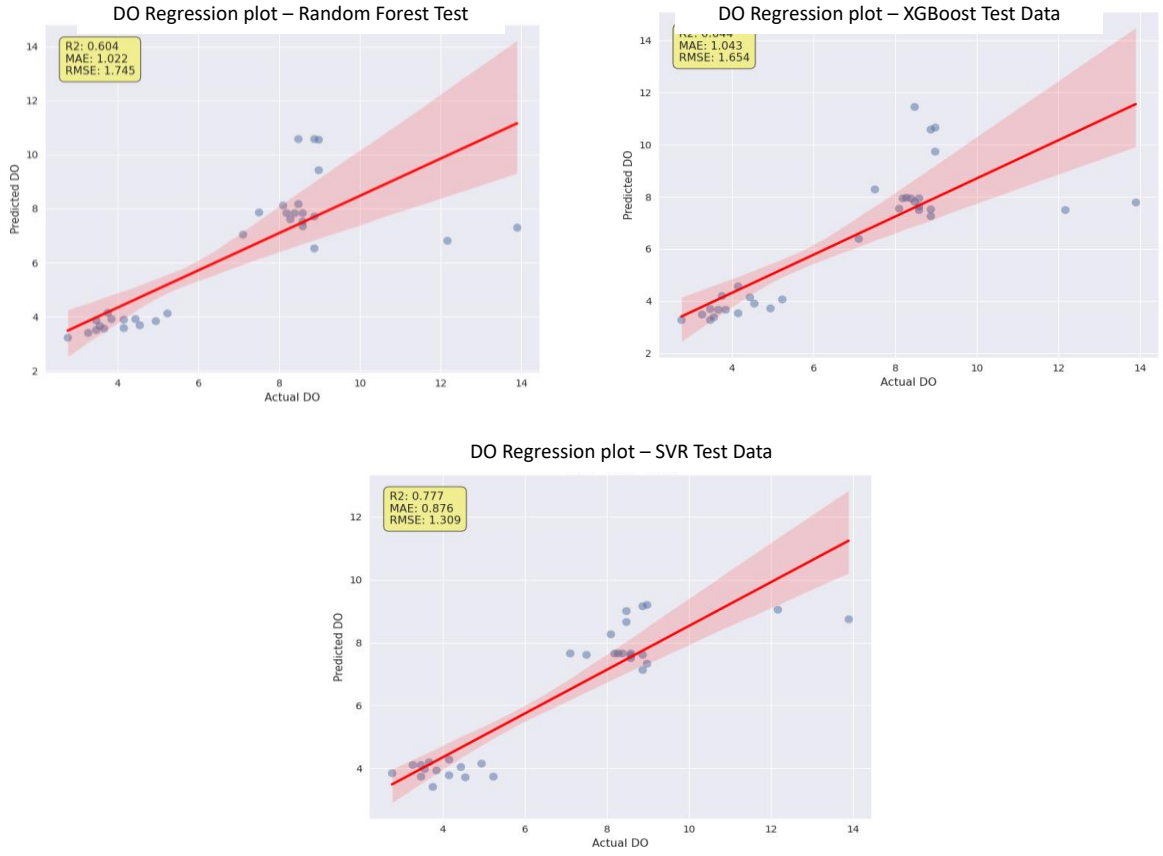


Fig 6: Scatter plot of actual vs predicted DO

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

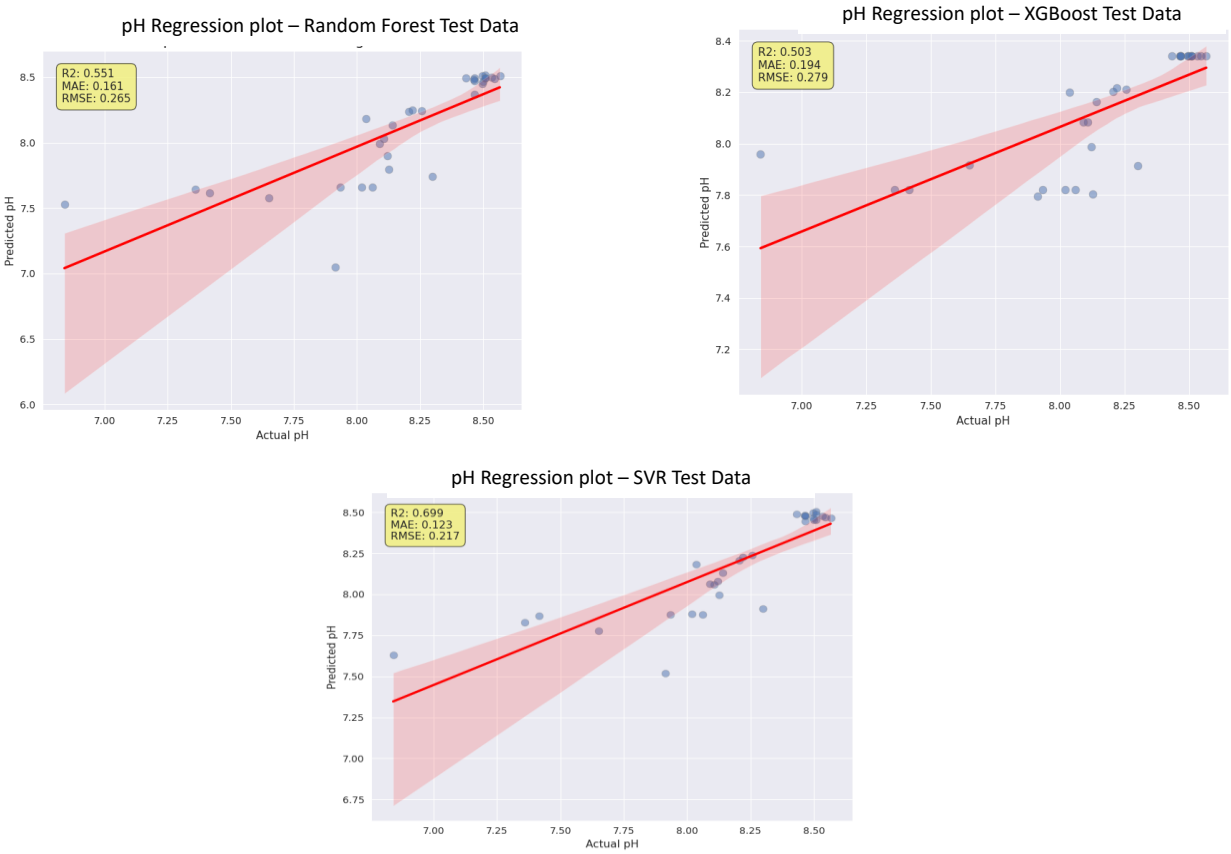


Fig 7: Scatter plots of actual vs predicted pH

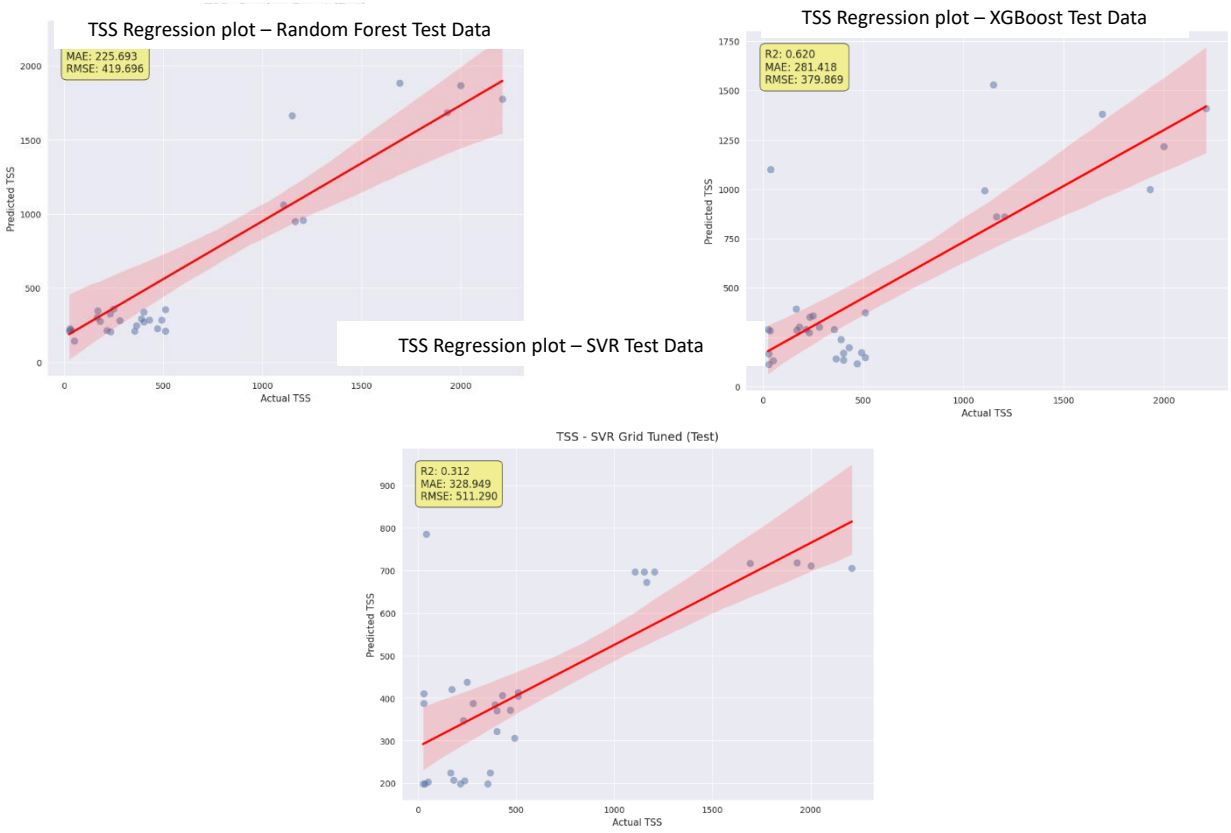


Fig 8: Scatter plots of actual vs predicted TSS

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

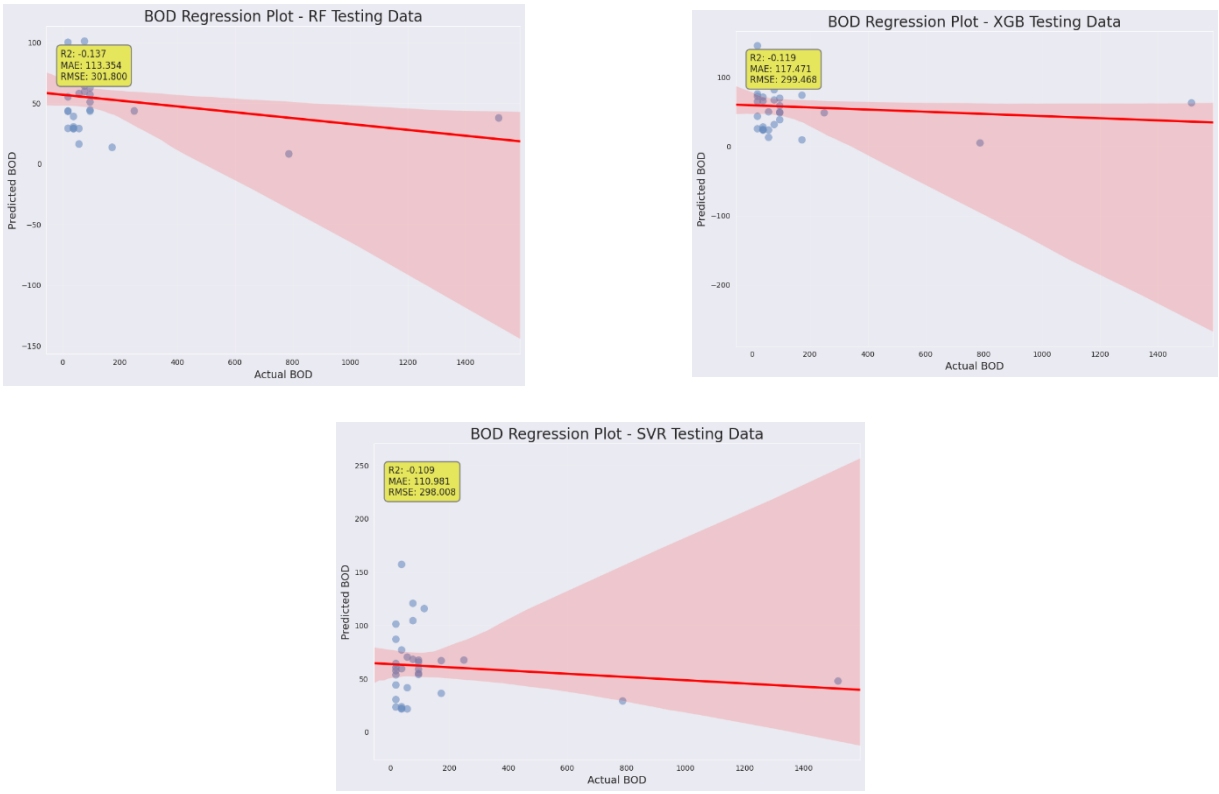


Fig 9: Scatter plots of actual vs predicted BOD

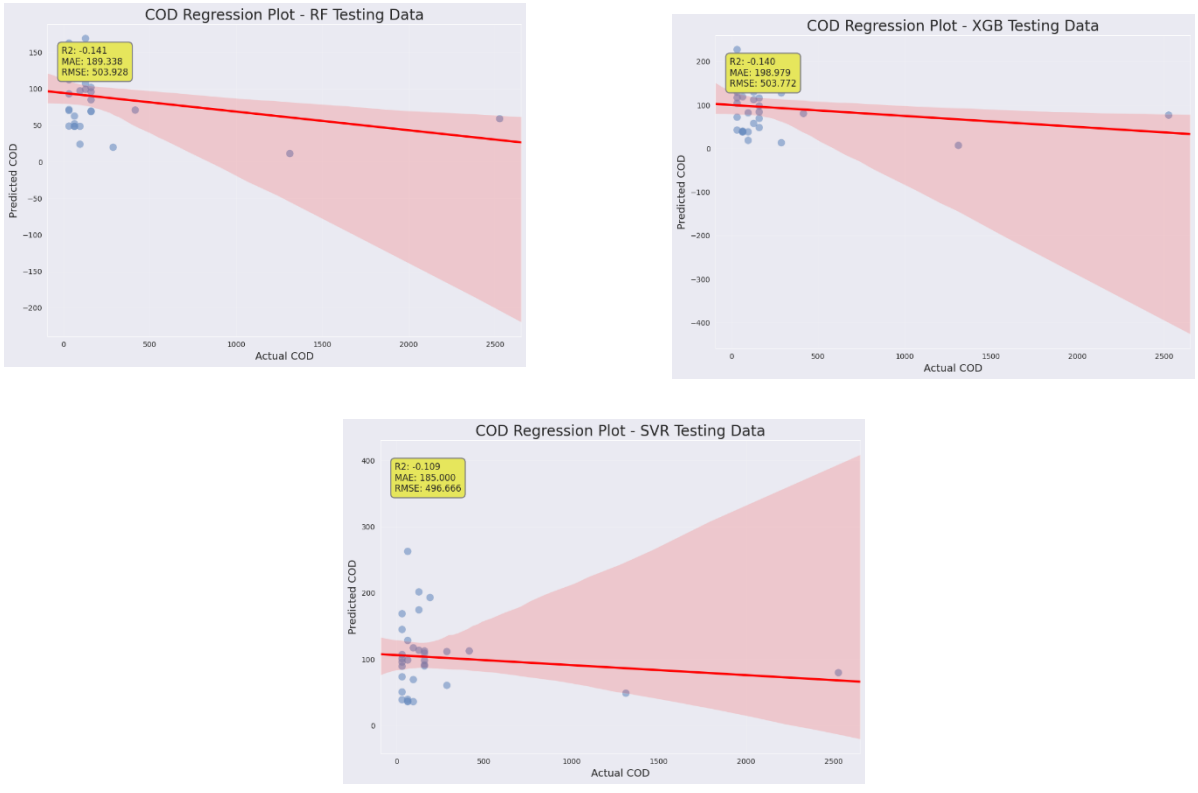


Fig 10: Scatter plots of actual vs predicted COD

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The scatter plots of predicted versus observed values indicate that predictions for turbidity, DO, pH, and TSS are generally distributed close to the 1:1 line, confirming good agreement. However, larger deviations are observed for BOD and COD.

The comparison between machine learning models and regression-based approaches further highlights the superiority of machine learning techniques.

Table 3: Linear regression model performance

Regression models showed significantly lower performance compared to machine learning models. For example, turbidity achieved an R^2 of 0.787, while DO and TSS showed weak correlations with R^2 values of 0.226 and 0.215 respectively. BOD and COD exhibited negligible predictive capability with negative R^2 values, confirming the limitations of linear regression in modelling complex environmental relationships.

Parameter	Equation	R2	RMSE
Turbidity	Turbidity = 43.466850 - (60.177839 * B2 (blue)) + (28.437788 * B3 (green)) + (22.675374 * B4 (red)) - (35.594486 * B8 (NIR)) + (91.815632 * B11 (SWIR 1)) - (84.077381 * B12 (SWIR 2))	0.787	4.190
DO	DO = 3.449174 + (0.538437 * B2 (blue)) + (317.589192 * B3 (green)) - (345.811647 * B4 (red)) - (152.504728 * B8 (NIR)) + (223.886541 * B11 (SWIR 1)) - (35.820433 * B12 (SWIR 2))	0.226	2.437
pH	pH = 7.866422 - (15.804186 * B2 (blue)) - (1.366983 * B3 (green)) + (22.499176 * B4 (red)) + (5.929357 * B8 (NIR)) - (2.661643 * B11 (SWIR 1)) - (8.793555 * B12 (SWIR 2))	0.636	0.238
TSS	TSS = 1041.948262 + (13692.625796 * B2 (blue)) + (12438.513876 * B3 (green)) - (30479.548602 * B4 (red)) - (12728.752530 * B8 (NIR)) + (15732.816301 * B11 (SWIR 1)) - (1120.590834 * B12 (SWIR 2))	0.215	546.473
BOD	BOD = 19.868537 - (2028.443688 * B2 (blue)) + (1192.749123 * B3 (green)) + (761.431484 * B4 (red)) + (2382.718278 * B8 (NIR)) - (3322.665410 * B11 (SWIR 1)) + (1102.957708 * B12 (SWIR 2))	-0.002	283.263
COD	COD = 33.114229 - (3380.739479 * B2 (blue)) + (1987.915205 * B3 (green)) + (1269.052474 * B4 (red)) + (3971.197131 * B8 (NIR)) - (5537.775683 * B11 (SWIR 1)) + (1838.262847 * B12 (SWIR 2))	-0.002	472.106

Statistical analysis of water quality parameters indicates significant variability within the dataset.

Table 4: Statistical summary of water quality parameters

S.No	Parameter	Minimum	Mean	Maximum	Standard Deviation
1	Turbidity (NTU)	12.133	37.237	51.2	9.483
2	DO (mg/L)	1.675	6.208	14.22	2.954
3	BOD (mg/L)	0	104.319	1516.8	192.246
4	COD (mg/L)	0	173.865	2528	320.409
5	TSS (mg/L)	20	583.413	2740	653.678
6	Ph	6.657	8.185	8.567	0.421

The variability observed in parameters such as TSS, BOD, and COD highlights the complexity of the dataset and explains the difficulty in modelling non-optical parameters accurately.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Similarly, spectral band statistics provide insight into the variability of input features used in the models.

Table 5: Statistical summary of Sentinel-2 spectral bands

S.No.	Band	Minimum	Mean	Maximum	Std. Dev.
1	B2 (blue)	0.065714	0.226964	0.758354	0.244622
2	B3 (green)	0.089457	0.231177	0.681493	0.21303
3	B4 (red)	0.069947	0.203607	0.643184	0.20621
4	B8 (NIR)	0.032512	0.17906	0.659792	0.231913
5	B8A (narrow NIR)	0.024855	0.168002	0.660739	0.225755
6	B11 (SWIR)	0.011683	0.1385	0.561392	0.206505
7	B12 (SWIR)	0.008018	0.120473	0.529828	0.186576

The statistical characteristics of the spectral band values provide insight into the variability of the predictor variables used in the modelling process. The reflectance values of the selected Sentinel-2 bands vary across sampling locations, reflecting differences in water surface conditions.

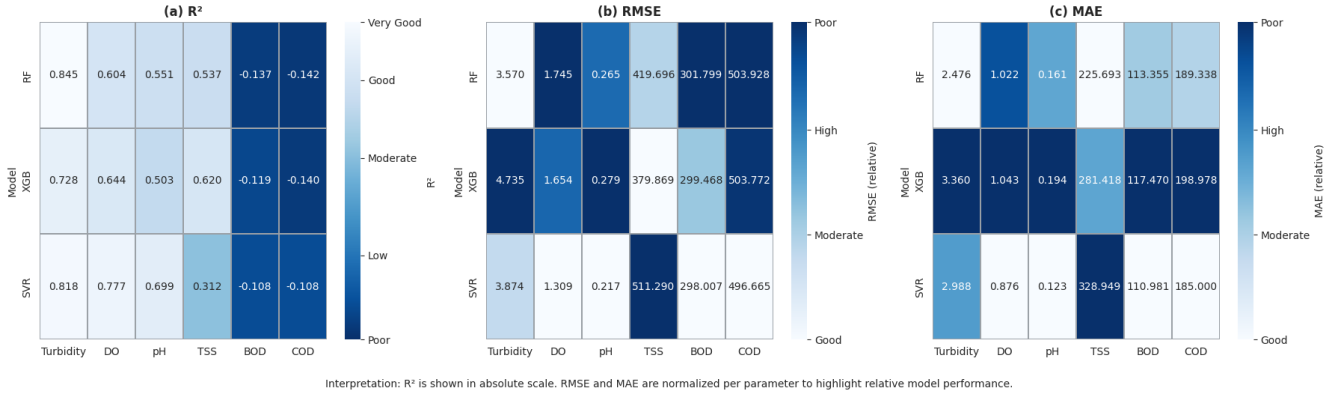
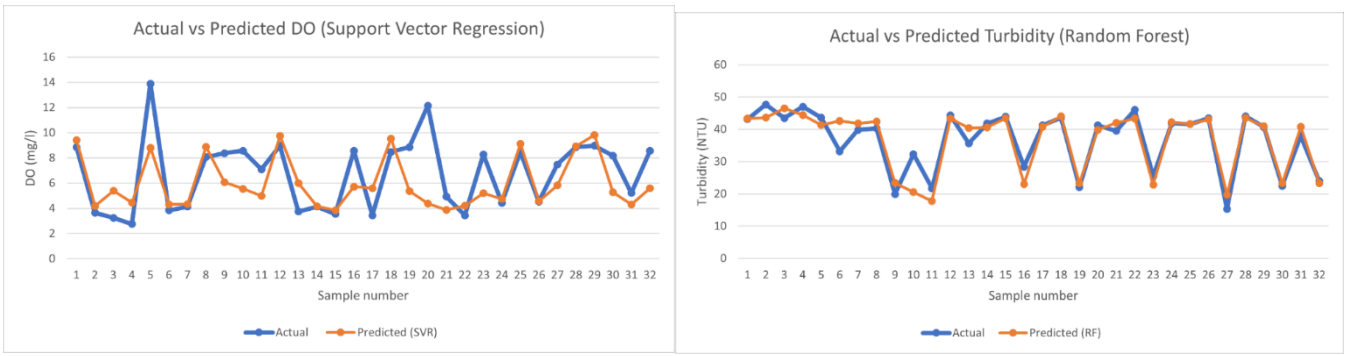


Fig 11: Heatmap of model performance

The heatmap provides a comparative visualization of model performance across parameters using R^2, RMSE, and MAE. It clearly highlights that turbidity is predicted with high accuracy, while DO and pH show moderate performance. TSS demonstrates acceptable prediction capability, whereas BOD and COD exhibit poor performance across all models, indicating the limitations of spectral-based prediction for non-optical parameters (Shah et al., 2024).



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

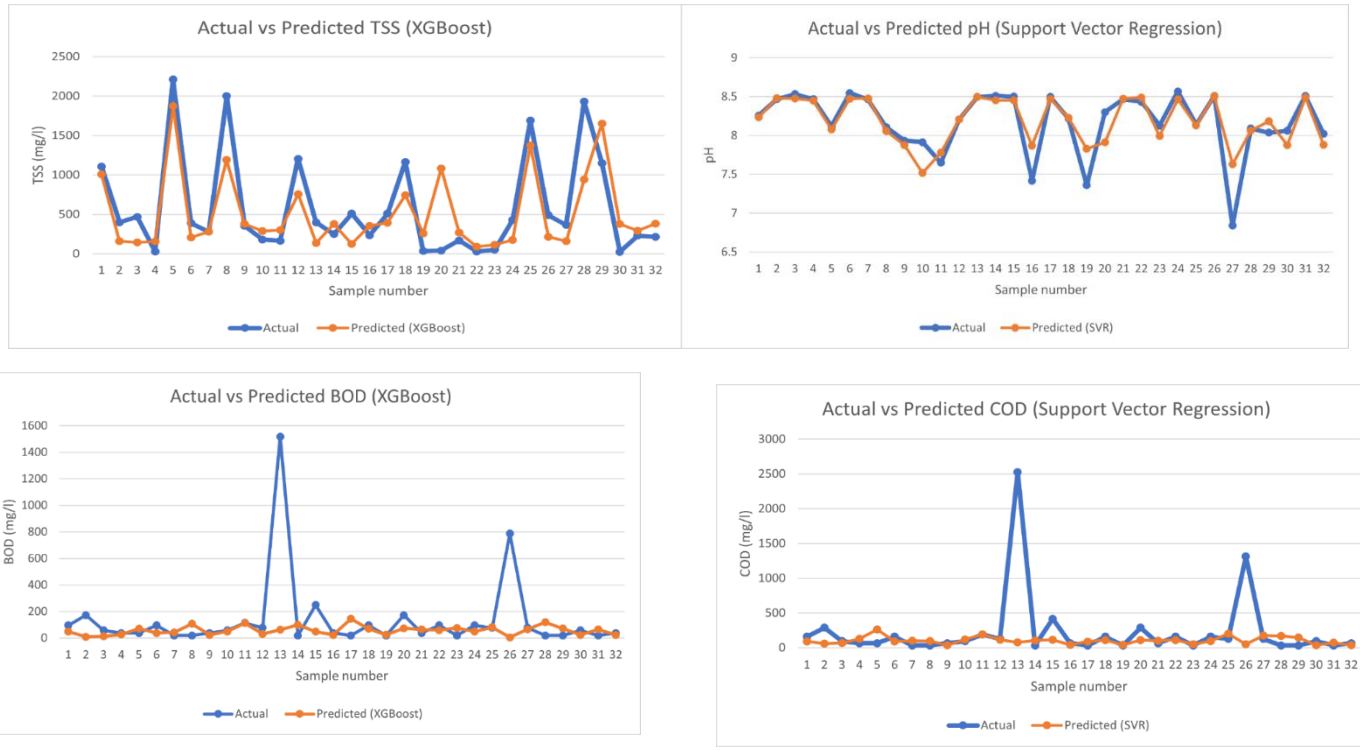
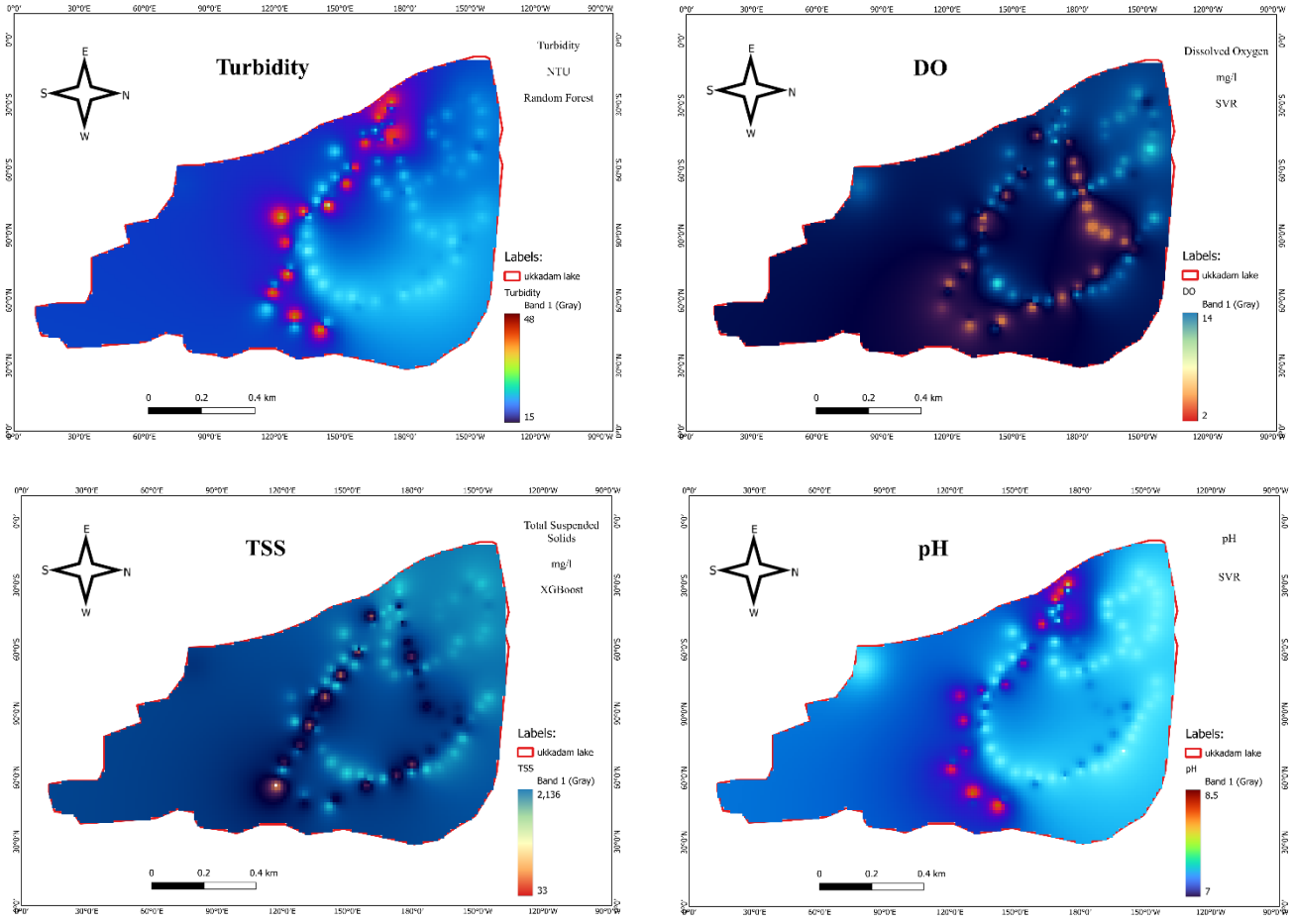


Fig 12: Line graphs for actual and predicted values of parameters



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

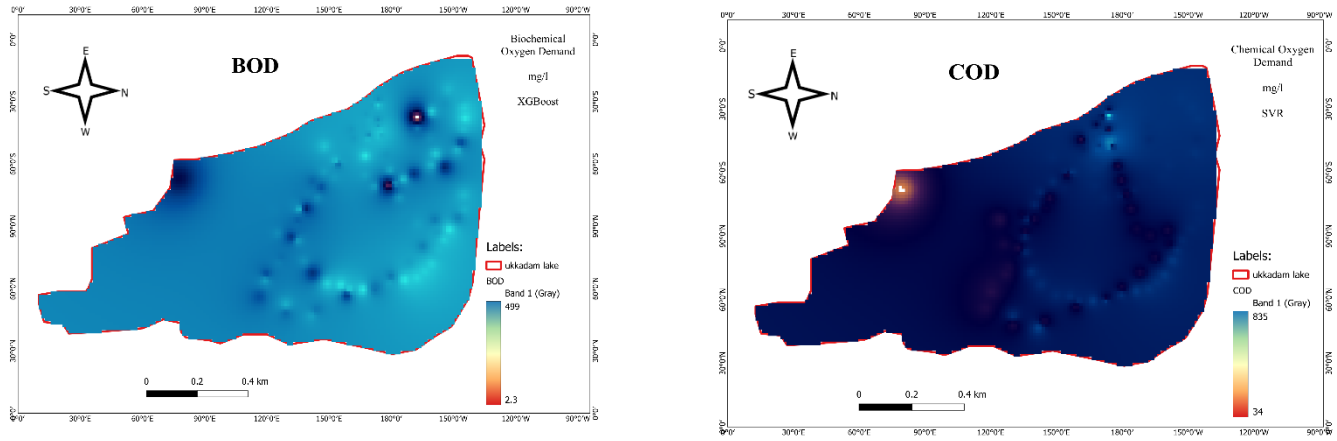


Fig 13: Spatial mapping predicted values of water quality parameters

Line graphs and spatial maps are presented only for the best-performing models for each parameter based on testing metrics (R^2, RMSE, and MAE). This avoids redundancy and improves clarity, as model performance has already been established through quantitative evaluation and scatter plot analysis. Parameters with poor predictive performance, such as BOD and COD, are excluded due to the absence of meaningful trends.

The Water Quality Index (WQI) was computed using the predicted values of selected water quality parameters to provide an integrated assessment of overall water quality conditions. The calculated WQI values ranged from 80.18 to 269.54, with an average value of 186.59 and a standard deviation of 46.13, indicating significant variability within the study area. The distribution of WQI values, as shown in Figure 14, reveals that a majority of the samples fall within higher WQI ranges, suggesting degraded water quality conditions. The classification of WQI further indicates that most samples fall under the poor and very poor categories, while only a limited number of samples are classified as good. This trend highlights the overall deterioration of water quality in the study region. The histogram representation illustrates a concentration of values in the higher WQI range, whereas the class distribution confirms the dominance of lower-quality categories.

The observed WQI pattern is primarily influenced by parameters such as turbidity, TSS, and dissolved oxygen, which showed relatively better predictive performance. In contrast, the inclusion of parameters such as BOD and COD introduces uncertainty due to their lower prediction accuracy. Despite this limitation, the WQI provides a simplified yet effective representation of overall water quality status and demonstrates the practical applicability of the developed machine learning framework for environmental monitoring (Niazkar & Piraei, 2025).

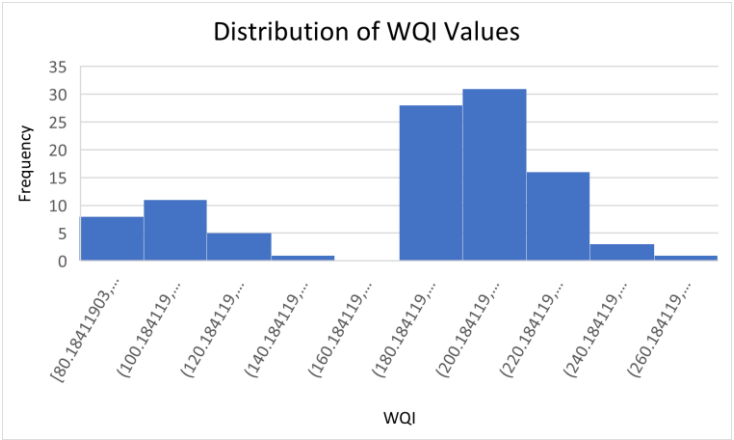


Fig 14: Distribution of Water quality index

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

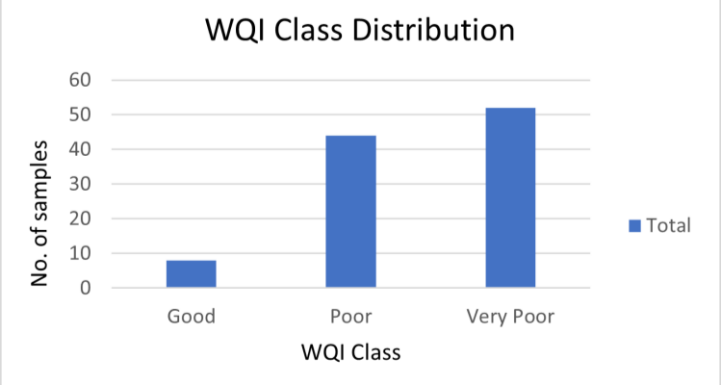


Fig 15: Classification of water quality index

4. CONCLUSION

This study evaluated the effectiveness of machine learning models integrated with Sentinel-2 satellite data for predicting key water quality parameters. The results demonstrate that machine learning approaches significantly outperform traditional regression models in capturing the relationship between spectral features and water quality variables.

Among the parameters analysed, turbidity exhibited the highest prediction accuracy due to its strong optical characteristics, with Random Forest identified as the most reliable model. Dissolved oxygen and pH showed moderate prediction performance, where Support Vector Regression provided better generalization compared to other models. Total suspended solids (TSS) achieved acceptable prediction accuracy, with XGBoost demonstrating relatively better performance. In contrast, biochemical oxygen demand (BOD) and chemical oxygen demand (COD) showed poor predictive capability across all models, indicating the limitations of using spectral data alone for parameters governed by complex biochemical processes.

The results highlight that model performance is strongly influenced by the optical nature of the parameter, with optically active variables yielding higher accuracy. The integration of derived features and auxiliary variables was found to improve model performance for moderately complex parameters. Spatial mapping and trend analysis further demonstrate the applicability of the developed approach for large-scale monitoring and visualization of water quality variations.

Overall, this study confirms that remote sensing combined with machine learning provides an efficient and scalable framework for water quality assessment. However, improving the prediction of non-optical parameters requires the incorporation of additional environmental variables and advanced modelling strategies. Future work may focus on multi-source data integration and hybrid modelling approaches to enhance prediction accuracy and expand the applicability of the proposed methodology.

REFERENCE

[1] Awasthi, A., Rana, S., Thakur, A., Thakur, S. K., & Jaswal, D. (2025). Groundwater quality and health risk assessment in the Baddi-Barotiwala-Nalagarh industrial belt of the northwestern Himalayas. *Scientific Reports*, *16*(1), 3388. https://doi.org/10.1038/s41598-025-33393-w

[2] Baharudin, F., Kassim, J., Imran, S. N. M., & Wahab, M. A. (2021). Water Quality Index (WQI) classification of rivers in agriculture and aquaculture catchments. *IOP Conference Series: Earth and Environmental Science*, *646*(1), 012023. https://doi.org/10.1088/1755-1315/646/1/012023

[3] Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[4] Campos, D., Galvão, V., De Rezende, M. L., Braga, A., Bodini, M., Aires, U. R. V., Yonaba, R., & Goliatt, L. (2026). Automated machine learning achieves accurate water quality prediction with reduced parameter requirements. *Scientific Reports*, *16*(1), 4431. https://doi.org/10.1038/s41598-025-34448-8

[5] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[6] Chen, W., Shao, Y., Xu, Z., Zhou, B., Cui, S., Dai, Z., Yin, S., Gao, Y., & Liu, L. (2026a). Ensemble Machine Learning for Operational Water Quality Monitoring Using Weighted Model Fusion for pH Forecasting. *Sustainability*, *18*(3), 1200. https://doi.org/10.3390/su18031200

[7] Chen, W., Shao, Y., Xu, Z., Zhou, B., Cui, S., Dai, Z., Yin, S., Gao, Y., & Liu, L. (2026b). Ensemble Machine Learning for Operational Water Quality Monitoring Using Weighted Model Fusion for pH Forecasting. *Sustainability*, *18*(3), 1200. https://doi.org/10.3390/su18031200

[8] Choudhary, R., Kumar, A., C., P., Naik, M. M., Choudhury, M., & Khan, N. A. (2025). Predicting water quality index using stacked ensemble regression and SHAP based explainable artificial intelligence. *Scientific Reports*, *15*(1), 31139. https://doi.org/10.1038/s41598-025-09463-4

[9] Chowdhury, M., De La Calle, I., Laiz, I., & Ruescas, A. B. (2025). Near-Real-Time Turbidity Monitoring at Global Scale Using Sentinel-2 Data and Machine Learning Techniques. *Remote Sensing*, *17*(22), 3716. https://doi.org/10.3390/rs17223716

[10] Dawn, A., Hinge, G., Kumar, A., Nikoo, M. R., & Hamouda, M. A. (2025a). Assessment of Water Quality in Urban Lakes Using Multi-Source Data and Modeling Techniques. *Sustainability*, *17*(16), 7258. https://doi.org/10.3390/su17167258

[11] Dawn, A., Hinge, G., Kumar, A., Nikoo, M. R., & Hamouda, M. A. (2025b). Assessment of Water Quality in Urban Lakes Using Multi-Source Data and Modeling Techniques. *Sustainability*, *17*(16), 7258. https://doi.org/10.3390/su17167258

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[12] Deng, Y., Zhang, Y., Pan, D., Yang, S. X., & Gharabaghi, B. (2024). Review of Recent Advances in Remote Sensing and Machine Learning Methods for Lake Water Quality Management. *Remote Sensing*, *16*(22), 4196. https://doi.org/10.3390/rs16224196

[13] França, V. F. C. de, Silva, L. O. B. da, & De Andrade, H. A. (2026). Impact of variable selection and model complexity on the prediction of water quality parameters for Penaeus vannamei aquaculture in a short dataset context. *Aquacultural Engineering*, *112*, 102640. https://doi.org/10.1016/j.aquaeng.2025.102640

[14] Gao, L., Shangguan, Y., Sun, Z., Shen, Q., & Shi, Z. (2024). Estimation of Non-Optically Active Water Quality Parameters in Zhejiang Province Based on Machine Learning. *Remote Sensing*, *16*(3), 514. https://doi.org/10.3390/rs16030514

[15] Gikas, G. D., Lergios, D., & Tsihrintzis, V. A. (2023). Comparative Assessment of the Application of Four Water Quality Indices (WQIs) in Three Ephemeral Rivers in Greece. *Water*, *15*(8), 1443. https://doi.org/10.3390/w15081443

[16] Hridoy, Md. A. A. M., Shawkat, A. I., Bordin, C., Acharjee, M. R., Masood, A., Baki, A. O., & Al Mamun, Md. A. (2025). Advanced machine learning models for accurate water quality classification and WQI prediction: Implications for aquatic disease risk management. *Science of The Total Environment*, *1008*, 180965. https://doi.org/10.1016/j.scitotenv.2025.180965

[17] Jena, P. K., Rahaman, S. M., Das Mohapatra, P. K., Barik, D. P., & Patra, D. S. (2023). Surface water quality assessment by Random Forest. *Water Practice and Technology*, *18*(1), 201–214. https://doi.org/10.2166/wpt.2022.156

[18] Kim, Y. H., Im, J., Ha, H. K., Choi, J.-K., & Ha, S. (2014). Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GIScience & Remote Sensing*, *51*(2), 158–174. https://doi.org/10.1080/15481603.2014.900983

[19] Lencha, S. M., Tränckner, J., & Dananto, M. (2021). Assessing the Water Quality of Lake Hawassa Ethiopia—Trophic State and Suitability for Anthropogenic Uses—Applying Common Water Quality Indices. *International Journal of Environmental Research and Public Health*, *18*(17), 8904. https://doi.org/10.3390/ijerph18178904

[20] Li, N., Ning, Z., Chen, M., Wu, D., Hao, C., Zhang, D., Bai, R., Liu, H., Chen, X., Li, W., Zhang, W., Chen, Y., Li, Q., & Zhang, L. (2022). Satellite and Machine Learning Monitoring of Optically Inactive Water Quality Variability in a Tropical River. *Remote Sensing*, *14*(21), 5466. https://doi.org/10.3390/rs14215466

[21] Liu, X., Zhang, Z., Jiang, T., Li, X., & Li, Y. (2021). Evaluation of the Effectiveness of Multiple Machine Learning Methods in Remote Sensing Quantitative Retrieval of Suspended Matter Concentrations: A Case Study of Nansi Lake in North China. *Journal of Spectroscopy*, *2021*, 1–17. https://doi.org/10.1155/2021/5957376

[22] Niazkar, M., & Piraei, R. (2025). Enhancing estimation of water quality index using stacking machine learning techniques: The case of Southern Bug River. *Science of The Total Environment*, *1003*, 180744. https://doi.org/10.1016/j.scitotenv.2025.180744

[23] Padilla-Mendoza, C., Torres-Bejarano, F., Campo-Daza, G., & González-Márquez, L. C. (2023). Potential of Sentinel Images to Evaluate Physicochemical Parameters Concentrations in Water Bodies—Application in a Wetlands System in Northern Colombia. *Water*, *15*(4), 789. https://doi.org/10.3390/w15040789

[24] Prasad, D. V. V., Venkataramana, L. Y., Kumar, P. S., Prasannamedha, G., Soumya, K., & Poornema, A. J. (2021). Prediction on water quality of a lake in Chennai, India using machine learning algorithms. *Desalination and Water Treatment*, *218*, 44–51. https://doi.org/10.5004/dwt.2021.26970

[25] Salas, E. A. L., Kumaran, S. S., Bennett, R., Partee, E. B., Brownknight, J., Schrack, K., & Willis, B. (2025). Integration of Google Earth Engine, Sentinel-2 images, and machine learning for temporal mapping of total dissolved solids in river systems. *Scientific Reports*, *15*(1), 27555. https://doi.org/10.1038/s41598-025-12548-9

[26] Shah, F. U., Khan, A. U., Khan, A. W., Ullah, B., Khan, M. R., & Javed, I. (2024). Comparative analysis of ensemble learning algorithms in water quality prediction. *Journal of Hydroinformatics*, *26*(12), 3041–3059. https://doi.org/10.2166/hydro.2024.071

[27] Silveira Kupssinskü, L., Thomassim Guimarães, T., Menezes De Souza, E., C. Zanotta, D., Roberto Veronez, M., Gonzaga, L., & Mauad, F. F. (2020). A Method for Chlorophyll-a and Suspended Solids Prediction through Remote Sensing and Machine Learning. *Sensors*, *20*(7), 2125. https://doi.org/10.3390/s20072125

[28] Singh, K. A., Ryu, D., Arora, M., Tiwari, M. K., & Sahoo, B. (2026). Machine learning and remote sensing-based hierarchical framework for assessing non-optically observable riverine water quality parameters. *Journal of Hazardous Materials Advances*, *22*, 101111. https://doi.org/10.1016/j.hazadv.2026.101111

[29] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

[30] Tian, S., Guo, H., Xu, W., Zhu, X., Wang, B., Zeng, Q., Mai, Y., & Huang, J. J. (2022). Remote sensing retrieval of inland water quality parameters using Sentinel-2 and multiple machine learning algorithms. *Environmental Science and Pollution Research*, *30*(7), 18617–18630. https://doi.org/10.1007/s11356-022-23431-9

[31] Tiyasha, T., Tung, T. M., Bhagat, S. K., Tan, M. L., Jawad, A. H., Mohtar, W. H. M. W., & Yaseen, Z. M. (2021). Functionalization of remote sensing and on-site data for simulating surface water dissolved oxygen: Development of hybrid tree-based artificial intelligence models. *Marine Pollution Bulletin*, *170*, 112639. https://doi.org/10.1016/j.marpolbul.2021.112639

[32] Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023). A sophisticated model for rating water quality. *Science of The Total Environment*, *868*, 161614. https://doi.org/10.1016/j.scitotenv.2023.161614

[33] Xiao, Y., Guo, Y., Yin, G., Zhang, X., Shi, Y., Hao, F., & Fu, Y. (2022). 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*, *14*(14), 3272. https://doi.org/10.3390/rs14143272

[34] Yee Wong, W., Hasikin, K., Salwa Mohd Khairuddin, A., Abdul Razak, S., Farzana Hizaddin, H., Istajib Mokhtar, M., & Mokhzaini Azizan, M. (2023). A Stacked Ensemble Deep Learning Approach for Imbalanced Multi-Class Water Quality Index Prediction. *Computers, Materials & Continua*, *76*(2), 1361–1384. https://doi.org/10.32604/cmc.2023.038045

[35] Zhang, K., Xia, R., Wang, Y., Chen, Y., Wang, X., & Dou, J. (2025). Stack Coupling Machine Learning Model Could Enhance the Accuracy in Short-Term Water Quality Prediction. *Water*, *17*(19), 2868. https://doi.org/10.3390/w17192868

[36] Zhao, Y., He, X., Pan, S., Bai, Y., Wang, D., Li, T., Gong, F., & Zhang, X. (2024). Satellite retrievals of water quality for diverse inland waters from Sentinel-2 images: An example from Zhejiang Province, China. *International Journal of Applied Earth Observation and Geoinformation*, *132*, 104048. https://doi.org/10.1016/j.jag.2024.104048

[37] Zheng, Y., Li, C., Zhang, X., Zhao, W., Yang, Z., & Cao, W. (2024). Estimation of water quality parameters based on time series hydrometeorological data in Miaowan Island. *Ecological Indicators*, *159*, 111693. https://doi.org/10.1016/j.ecolind.2024.111693

[38] Zhu, X., Wen, Y., Li, X., Yan, F., & Zhao, S. (2023). Remote Sensing Inversion of Typical Water Quality Parameters of a Complex River Network: A Case Study of Qidong's Rivers. *Sustainability*, *15*(8), 6948. https://doi.org/10.3390/su15086948