Journal of Integrated Earth Sciences, 2025, Vol.1(1), December 2025, pp. 61-71
https://doi.org/10.5281/zenodo.18587113

OPEN ACCESS

JOURNAL OF INTEGRATED EARTH SCIENCES

Comprehensive Water Quality Assessment using Ensemble Machine Learning in a Tropical Lake, Kerala, SW India

Sabu Joseph [a], S. Sukanya [a], and M.R. Vishnuprasad [a]

[1]Department of Environmental Sciences, University of Kerala, Karyavattom Campus, Trivandrum, Kerala-695581, India

ARTICLE INFO

Received. 15 January 2025
Revised. 06 February 2025
Accepted. 21 March 2025

KEYWORDS

Machine Learning; Surface Water; Vellayani Lake; Water Quality

How to cite

Joseph, S., S. Sukanya, & M.R. Vishnuprasad. (2026). Comprehensive Water Quality Assessment using Ensemble Machine Learning in a Tropical Lake, Kerala, SW India. Journal of Integrated Earth Sciences, 1(1), 61–71. https://doi.org/10.5281/zenodo.18587113

ABSTRACT

This study intends to evaluate and characterize the water quality of Vellayani Lake (VL), a tropical freshwater body in Kerala, Southwest India. Comprehensive analyses of physico-chemical parameters were conducted in surface water samples (n=13) during post-monsoon (January), pre-monsoon (May) and monsoon (July) periods. Key parameters viz., temperature, pH, Electrical Conductivity, Total Dissolved Solids, Dissolved Oxygen and nutrient concentrations were measured. The Water Quality Index (WQI) classification was employed to assess water quality. WQI values ranged from 'Excellent' to 'Poor', with a general decline of quality during post-monsoon. Additionally, a Random Forest based ensemble machine learning model was applied to validate the WQI results, achieving high accuracy (R^2 = 0.96) and identified phosphate, Dissolved Oxygen, Electrical Conductivity and Total Dissolved Solids as key predictors. This integrative approach provides a comprehensive understanding of the lake's water quality dynamics, emphasizing the need for spatially targeted interventions to reduce pollutant loads and stabilize the lake's ecological function.

CONTACT Author, Sabu Joseph - jsabu@keralauniversity.ac.in

Supplemental data for this article can be accessed online at our website

1. Introduction

Water quality, due to its dynamic nature, demands a multidimensional assessment strategy. As highlighted by Vasistha and Ganguly, 2020 and Naderian et al., 2024, a holistic evaluation framework is essential for capturing the inherent variability and long-term integrity of aquatic The complexity of water quality assessment arises from the necessity to measure a broad spectrum of parameters, as no single metric can comprehensively represent the overall condition of water bodies (Kwon and Jo., 2023). Consequently, synthesizing these diverse parameters into actionable insights presents a considerable challenge for stakeholders (Zhi et al., 2024). To address this challenge, the Water Quality Index (WQI) has been developed as a sophisticated tool to provide an integrated assessment of water quality (Mechal et al., 2024). WQI models employ aggregation functions that condense extensive spatiotemporal data into a singular, interpretable value , facilitating informed decision-making for policymakers and stakeholders (Akhtar et al., 2021).

However, the efficacy of WQI tools is contingent upon the availability of comprehensive water quality data. This data typically includes measurements of parameters such as pH, Dissolved Oxygen, Chemical Oxygen Demand, Electrical Conductivity, phosphorus etc. that require laboratory analysis, creating a significant barrier to timely and effective ecosystem management (Aldrees et al., 2022). Traditional WQI approaches, which assign weighted scores to individual water quality parameters, often encounter issues of complexity and uncertainty (Jha et al., 2020). Multivariate statistical methodologies, including principal component analysis, factor analysis, discriminant analysis, and cluster analysis have been employed to elucidate patterns in water quality data (Li et al., 2018; Sukanya and Sabu, 2020; Horvat et al., 2021).

Nonetheless, interpreting the complex interrelationships among multiple water quality parameters remains arduous without robust techniques. Advanced statistical techniques, while capable of handling certain uncertainties and improving accuracy, often struggle with high-dimensional data, non-linear relationships, and the significant computational resources required, limiting their effectiveness in providing practical conclusions (Ahmed et al., 2019).

Recent advances in machine learning (ML) offer promising tools to enhance the accuracy and depth of water quality assessments (Sukanya and Sabu, 2023). To enhance the realistic assessment of water bodies, it is imperative to integrate WQI models with advanced machine learning algorithms (Lap et al., 2023). This integration offers a practical and economical solution, reducing the challenges associated with water quality sampling and costs of labour and equipment, while also advancing traditional WQI models towards improved applicability.

The Random Forest model, an ensemble learning method that aggregates decision trees, is particularly well-suited for handling high-dimensional datasets and optimizing WQI models (Sukanya and Sabu, 2023; Talukdar et al., 2024). Focusing on Vellayani Lake (VL) in Southwest India, this research utilizes a Water Quality Index (WQI) model optimized with a Random Forest algorithm, aiming for a precise water quality evaluation. This innovative approach seeks to develop a cost-effective and efficient strategy for improving WQI, thereby providing critical support for water environment management decisions and demonstrating the profound potential of machine learning technologies in addressing complex environmental challenges.

2. Study Area

Vellayani Lake (VL), the second largest freshwater body in Kerala, Southwest India, spans approximately 3.15 km in length and covers an area of about 2.25 km². Situated around 15 km from Thiruvananthapuram city (N. Lat. 08.45°; E. Long. 76.98°), the lake lies roughly 4 km inland and runs parallel to the coastline, exhibiting an elongated, hammer-like shape with a general north–south orientation. Its catchment area extends over 38 km² (Fig. 1).Geologically, the lake basin consists predominantly (about 90%) of Precambrian crystalline basement rocks and Tertiary sediments. The eastern and southwestern zones are primarily composed of Precambrian formations, including khondalites.

The northeastern head region is characterized by rocks such as charnockite, charnockitic gneiss, and hypersthene-diopside gneiss (Banerji et al., 2021). Quaternary deposits dominate the northwestern area, while Tertiary sediments occur in the southern part of the basin. From a geomorphological perspective, Vellayani Lake is classified as a lowland lake (elevation between 10 and 100 meters), featuring ridge tops and moderate slopes. The eastern region contains distinct valleys, while the western part displays a more undulating terrain. Notably, the northern edge of the lake is predominantly used for agricultural irrigation.

3. Materials and Methods

Water quality data were collected from 13 monitoring stations around a tropical lake in Kerala, during three distinct seasons: post-monsoon (January), pre-monsoon (May) and monsoon (July). Water quality in India is primarily regulated by the Bureau of Indian Standards (BIS 10500, 2012) and Central Pollution Control Board (CPCB) standards. In this study, we evaluated water quality using 14 key parameters, chosen based on the regulatory benchmarks provided by these Indian standards as well as relevant international criteria (WHO, 2022). Water quality parameters including pH, electrical conductivity (EC), total dissolved solids (TDS), alkalinity, dissolved oxygen (DO), chloride, hardness, calcium, magnesium, sodium, potassium, ammonia, nitrate, and phosphate were measured (Table 1) and used to compute the Water Quality Index (WQI). The relative weight (RW) assigned to each WQ parameter was determined using equation:

$$RW = \frac{AW_i}{\sum_{i=1}^{n} AW_i} \quad \text{(Eq.1)}$$

where RW denotes relative weight, AW is assigned weight of i^{th} parameter, n represents the total number of parameters involved in assessment.

To compute the quality rating scale (Qi) for each parameter - excluding pH and DO, the following equation was used:

$$Q_i = \left(\frac{C_i}{S_i}\right) \times 100 \quad \text{(Eq. 2)}$$

In this context, Ci is the Observed concentration of i^{th} parameter, Si corresponds to its Standard permissible value.

For parameters viz., pH and DO, which have defined ideal values, a modified version of the quality rating formula was applied:

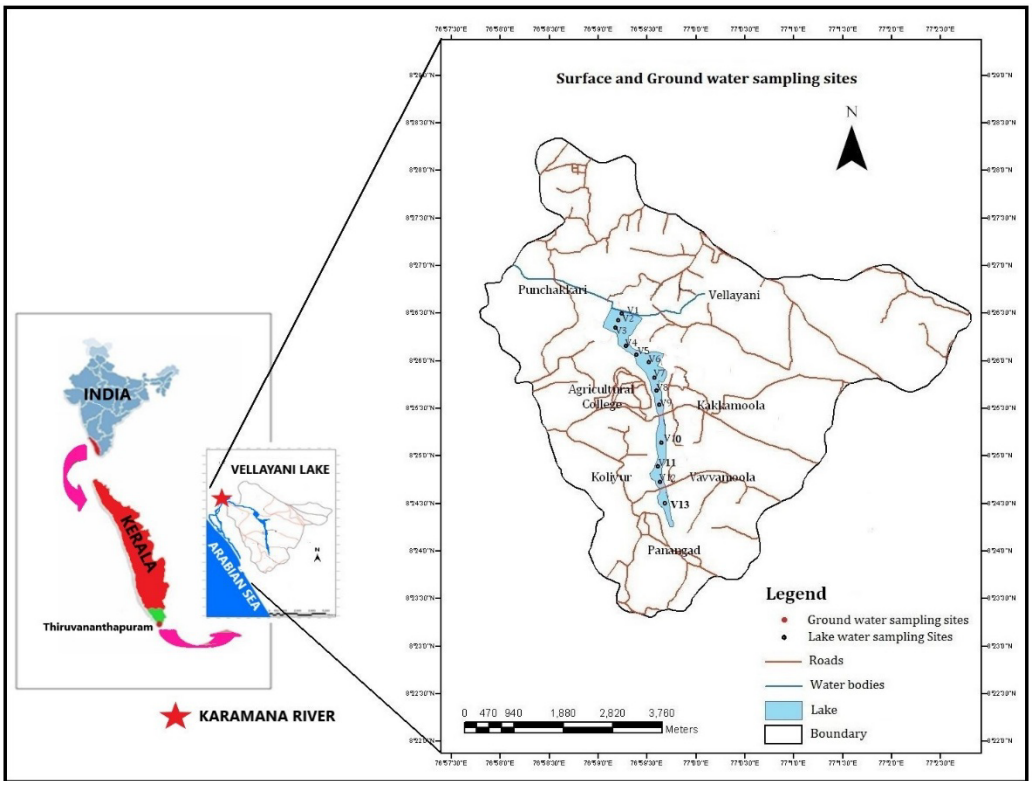

Fig. 1. Spatial map illustrating the location of Vellayani Lake and corresponding sampling points.

$$Q_i = \left(\frac{C_i V_i}{S_i V_i}\right) \times 100 \quad \text{(Eq. 3)}$$

where V_i refers to Ideal reference value which is taken as 7.0 for pH and 14.6 mg/L for DO.
Next, the sub-index (SI_i) for each parameter was evaluated by multiplying its relative weight with respective quality rating:

$$SI_i = RW \times Q_i \quad \text{(Eq. 4)}$$

The overall Water Quality Index (WQI) was then obtained by summing all individual sub-indices:

$$WQI = \sum_{i=1}^{n} SI_i \quad \text{(Eq. 5)}$$

The assigned and relative weights for each parameter are provided in Table 2.
The WQI values were categorized into the following classes: “Excellent” (0-25), “Good” (26-50), “Poor” (51-75) and “Very Poor” (76-100), “Unsuitable” (> 100) (Monira et al., 2024).

3.1 Machine Learning Approach

Random Forest model training and evaluation

To comprehensively evaluate water quality trends and corroborate the outcomes of the Water Quality Index (WQI), a Random Forest (RF) model was implemented. Recognized for its ability to manage intricate relationships among numerous variables, this ensemble-based machine learning approach consolidates outputs from a collection of decision trees. This aggregation minimizes the risk of overfitting and improves the overall predictive performance of the model (Bakır et al., 2024).

Data Preparation: A data frame was created with the water quality data for the three seasons as predictors and the WQI categories as the response variable (Table 3). This method allows for capturing seasonal variations and their impact on water quality. The preprocessing of actual Water Quality Index (WQI) data aimed to refine the dataset for improved accuracy in the RF model predictions. Initially, raw WQI values were used as input. The first step in preprocessing involved data cleaning to rectify inconsistencies and address missing values, ensuring the integrity and completeness of the dataset.

Subsequently, normalization techniques were applied to standardize the variables, mitigating potential biases introduced by varying measurement scales. This normalization step was crucial in harmonizing the input features for the RF model, facilitating more accurate predictions by minimizing the influence of outliers and scaling differences among parameters. Moreover, exploratory data analysis techniques were employed to identify correlations and interactions among the water quality variables, guiding feature selection and ensuring that the most informative parameters were retained for model training. By systematically preparing the actual WQI data through these preprocessing steps, the RF model was primed to capture nuanced relationships and variations in water quality, thereby enhancing its predictive performance during different seasons and environmental conditions.

Model Training: The Random Forest model was developed using the Random Forest package in R, incorporating cross-validation (CV) to ensure reliable performance evaluation. This validation approach reduces the likelihood of overfitting and yields a more generalized estimate of the model's predictive capability. As an ensemble learning method, Random Forest strengthens both the stability and accuracy of predictions by combining the outputs of numerous decision trees, thereby improving overall model robustness.

Cross-Validation: A 5-fold cross-validation (CV) strategy was employed to resample the dataset, enabling a reliable and unbiased assessment of model performance. The data were partitioned into five subsets with sample sizes of 11, 9, 10, 10, and 12, respectively. Each subset was used once as a validation set while the remaining folds served as the training set in an iterative manner. This approach, combined with hyperparameter tuning, contributed to improved generalization and optimized the model's performance across varying data partitions.

Hyperparameter Tuning: Model tuning involved testing multiple *mtry* values, controlling the number of features sampled per split, to identify the configuration yielding the best predictive accuracy. Hyperparameter optimization was conducted to fine-tune the Random Forest model's parameters. Grid search combined with cross-validation was utilized to identify the optimal set of hyperparameters, ensuring the model's robustness and performance across various scenarios.

IncNodePurity Analysis*:* The Random Forest model's feature importance was quantified using the IncNodePurity metric, which assesses the purity gain achieved by each variable at each split in the decision trees. Variables exhibiting higher IncNodePurity scores were identified as having greater importance in predicting the target outcome, specifically the Water Quality Index (WQI) categories. This analysis provided insights into the relative importance of water quality parameters, highlighting critical factors driving water quality variations in the study area.

3.2 Model evaluation:

Performance Metrics: The Random Forest model's performance was comprehensively assessed using accuracy, Kappa statistics, MAE, RMSE, and R^2 metrics across different mtry values. Evaluating model performance with multiple metrics provides a comprehensive view of the model's predictive power and reliability.

Prediction: The trained Random Forest model was applied to classify the Water Quality Index (WQI) categories based on the input data, enabling a comprehensive evaluation of water quality conditions. Predictive modeling allows for proactive management strategies by forecasting water quality trends.

4. Results and Discussion

4.1 Water Quality Index (WQI) classification

The statistical summary of various water quality parameters analyzed are provided in Table 1. Table 3 displays the calculated WQI values and Supplementary Tables S1, S2, S3 shows the various physico-chemical data from which WQI values were estimated for surface water samples from the study area. During the post-monsoon period (January), WQI values ranged from 20.79 to 70.23 (mean = 38.95). Around 61.54% of the samples fell into the 'Good' category, and two stations, around 15.38%, were classified as 'Poor' while 23.08% were 'Excellent'. The highest WQI value of 70.23 was recorded at station V7, indicating the most degraded water quality in this season. During the pre-monsoon season (May), Water Quality Index values ranged from 22.35 to 63.79, with a mean of 39.87, indicating generally improved water conditions compared to measurements from the post-monsoon phase. This variation in water status is likely influenced by multiple environmental factors, particularly the decrease in surface runoff and the diminished effect of dilution. A significant portion of the samples (69.23%) fell within the 'Good' classification, while 23.08% were categorized as 'Excellent'. This distribution indicates that pre-monsoon hydrological conditions tend to favor higher water quality, possibly due to the reduced influx of nutrients and pollutants typically transported by runoff.

During the monsoon season (July), WQI values ranged from 23.46 to 62.79 (mean = 40.29). This period showed a mixed trend in water quality, with some locations experiencing a slight improvement and others a decline. However, most of the samples, i.e., 76.92%, fell under the 'Good' category, and 15.38% were found to be in the 'Excellent' category.

Notably, this period had a higher percentage of locations classified under excellent water quality compared to others. The highest WQI value recorded was 62.79 at station V10 from the northern part of the lake, indicating the influence of increased nutrient runoff during heavy rainfall. The improved mean water quality during the pre-monsoon period, despite potential nutrient concentration, can be attributed to reduced external nutrient loading from surface runoff. In contrast, the monsoon period exhibits mixed water quality dynamics due to variable rainfall intensities. High rainfall in some areas dilutes nutrients, whereas localized runoff and point-source pollution, particularly at station V10, contribute to nutrient enrichment. Seasonal trends in WQI indicate that both pre-monsoon and monsoon periods generally support better water quality than the post-monsoon season, as reflected in higher percentages of samples in the 'Good' and 'Excellent' categories. However, the post-monsoon period shows the most degraded conditions, likely due to residual nutrient buildup and reduced dilution. The higher percentage of 'Poor' water quality samples (15.38%), suggests localized issues driven by nutrient-laden groundwater discharge from monsoon-recharged aquifers, compounded by reduced flushing and stagnant conditions.

These findings highlight the interplay between groundwater discharge and surface runoff as critical regulators of lake water quality across seasons, calling for strategic monitoring and mitigation efforts tailored to post-monsoon conditions. The mean WQI values across all seasons further highlight the spatial variability in water quality. Station V3 from northern part had the lowest mean WQI of 22.9, indicating the best water quality among the sampled locations. Conversely, station V2 had the highest mean WQI of 62.1, reflecting consistent water quality issues. The average WQI values for post-monsoon, pre-monsoon and monsoon periods were 38.7, 39.2, and 39.7 respectively (Fig. 2). In a study by Singh et al. (2016), 'Poor' and 'Unfit' category samples were observed during the post-monsoon period in an urban lake in Bhopal, India. Similarly, high WQI values (> 100) indicating low water quality were observed during post-monsoon in Hebbal Lake, South India. Higher WQI values observed during monsoon are likely the result of pollutant influx carried by heavy rainfall, which transports contaminants into the lake system. As water levels decline, these substances tend to settle and become more concentrated. In relatively shallow lakes, the post-monsoon period is often marked by sediment disturbance and resuspension, which acts as a key internal source of contamination (Yin et al., 2024).

In contrast, the pre-monsoon phase is characterized by limited surface runoff, a greater influence of groundwater inputs, and enhanced natural cleansing mechanisms within the lake; all of which contribute to better overall water conditions. These findings suggest the need for targeted management strategies, especially during dry seasons to mitigate nutrient enrichment and preserve water quality.

4.2 Ensemble machine learning and WQI

To enhance the understanding of water quality patterns and validate the WQI results, Random Forest, a robust ensemble machine learning technique was chosen for its ability to handle complex interactions between multiple water quality parameters and provide reliable predictions. The study involved rigorous preprocessing of a comprehensive dataset encompassing critical water quality parameters such as pH, electrical conductivity (EC), total dissolved solids (TDS), dissolved oxygen (DO) and concentrations of nutrients like ammonia, nitrate, and phosphate. The dataset was split into training (70%) and testing (30%) sets to ensure unbiased evaluation. These parameters served as input features for training a Random Forest model, which was meticulously optimized through a grid search with cross-validation. This approach ensured that hyperparameters, including the number of trees (mtry), maximum depth, and minimum samples per leaf, were tuned to maximize predictive accuracy. The model's performance was evaluated using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). Utilizing 5-fold cross-validated resampling, the model was systematically evaluated across various mtry values, revealing a robust performance landscape. Notably, the model achieved its pinnacle with mtry = 2, attaining an impressive accuracy with an average Mean Absolute Error (MAE) of approximately 2.24 units, Root Mean Squared Error (RMSE) of 2.93 units, and a coefficient of determination (R-squared) of 0.96, indicating robust agreement between predicted and observed values across varied environmental conditions (Fig. 3).

Furthermore, Cohen's kappa predicted and observed classifications beyond chance, substantiated the model's reliability. Our findings are in line with those of Zhang et al., 2024, who also utilized an RF model for optimizing WQI predictions, achieving an R^2 of 0.98. This high accuracy reinforces the effectiveness of RF models in environmental monitoring and their capability to handle complex, nonlinear interactions in water quality data.The predictive capabilities of the model were reflected by its ability to forecast higher water quality index values-indicative of poorer water quality at specific stations (V2, V7, and V10). These predictions aligned closely with actual WQI calculations, validating the model's efficacy in capturing real-world variations and trends.

Temporal analysis further elucidated seasonal patterns, revealing a consistent deterioration in water quality post-monsoon, alongside modest improvements during pre-monsoon and monsoon periods. These findings correlated well with observed fluctuations in nutrient levels and runoff dynamics, emphasizing the model's robustness in capturing complex environmental interactions.

Performance evaluation of the model was further supported by metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), which provided a detailed assessment of predictive reliability. For both pre-monsoon and post-monsoon periods, the model yielded an MAE of 2.54 and an RMSE of 3.32, suggesting a high level of predictive accuracy within acceptable error thresholds for water quality evaluation. These results demonstrate the model's effectiveness in estimating the WQI while accounting for the natural variability associated with water quality conditions. During the monsoon season, the RF model excelled with high R^2 value of 0.96, similar to other seasons, but with notably lower MAE of 1.64 and RMSE of 2.16. These reduced errors suggested the model's enhanced precision in predicting monsoon WQI values, capturing seasonal variations more effectively. Moreover, the high R^2 value highlighted the model's capability to explain 96% of the variance in WQI values, reinforcing its reliability for predictive applications. Misclassifications were minimal and primarily occurred near category boundaries, where WQI values were close to threshold limits. The model's ability to highlight stations with persistently high WQI values, such as V2 (northern part of VL), V7 (central part), and V10 (southern part), aligns with the manual WQI calculations, suggesting these locations require targeted management interventions.

The integration of Random Forest based ensemble machine learning techniques provided a powerful tool for water quality assessment, offering detailed insights and reinforcing the empirical findings. A notable advantage of the Random Forest (RF) algorithm lies in its capacity to evaluate the relative contribution of individual water quality parameters. The feature importance analysis generated by the RF model offers a reliable means to rank and highlight parameters that play a significant role in determining overall water quality, thereby aiding in more targeted and effective assessments. The feature importance analysis identified phosphate and Dissolved oxygen (DO) as the most significant predictors, highlighting its crucial role in water quality assessment. Certain parameters were found to be exhibiting substantial importance, quantified by their IncNodePurity values. For instance, phosphate and DO emerged as highly influential with IncNodePurity values of 156.92 and 121.18 respectively (Fig. 4). This highlights the crucial role of these parameters in aquatic ecosystems where oxygen availability and phosphate loading directly impacts biotic communities and biochemical processes.

The prominence of pH with an IncNodePurity of 69.39 signifies its pivotal role in regulating chemical reactions, nutrient availability, and overall water chemistry. Similarly, high IncNodePurity values for sodium (66.84), chloride (59.48) and TDS (61.20) highlight their significant contributions to salinity, ion balance, and pollutant load in water bodies, crucial for assessing water quality and ecological health. Parameters such as calcium (54.80), and hardness (51.62) also demonstrated considerable importance, reflecting their roles in nutrient dynamics, mineral content, and water hardness, respectively. Their elevated IncNodePurity values underscore their influence on aquatic habitats and the potential implications for ecosystem integrity and water resource management. Alqahtani et al. (2022) reported that ensemble learning approaches such as RF deliver higher predictive accuracy and consistency in estimating water quality parameters, achieving an impressive coefficient of determination (R^2) of 0.98. This performance notably surpasses that of standalone machine learning models, including gene expression programming (GEP) and artificial neural networks (ANN). Supporting this, Sakaa et al. (2022) highlighted the advantages of RF over hybrid optimization techniques, emphasizing its efficiency, lower computational demand, and suitability for cost-effective water quality monitoring, making it a valuable tool for advancing sustainable water resource management. In another recent research, the efficacy of ensemble learning models, particularly Random Forest (RF), was highlighted in customizing the water quality index (WQI) to suit specific environmental contexts and management needs (Lee et al., 2023).

However, to reduce overfitting in Random Forest (RF) models, optimizing tree depth, increasing minimum samples per leaf, conducting careful feature selection to exclude irrelevant variables, employing cross-validation, and monitoring ensemble size are critical strategies to enhance model generalization and performance (Huang and Boutros, 2016; Ahmad et al., 2018; Huang et al., 2021; Bakır et al., 2024).

5. Conclusions

The assessment of water quality in Vellayani Lake revealed pronounced seasonal and spatial variability, strongly shaped by monsoonal influences. While the monsoon and pre-monsoon periods exhibited better water quality in general, the post-monsoon period stood out due to higher incidence of 'Poor' water quality based on water quality index (WQI), particularly at localized sites. Employing a Random Forest based ensemble machine learning model offered reliable support for the WQI outcomes. The model achieved a strong predictive performance ($R^2 = 0.96$), successfully modeling the nonlinear relationships among multiple water quality parameters.

Table 1. Statistical Summary of Physico-chemical Parameters during January, May and July

Sl. No.	Parameter	January (Post-monsoon)				May (Pre-Monsoon)				July (Monsoon)			
		Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
1	Temp. (^{0}C)	27	29	28.46	0.66	30	32	31.31	0.75	26	27	26.6	0.48
2	pH	5.7	6.9	6.48	0.33	6.1	6.9	6.6	0.24	6	6.5	6.2	0.18
3	EC (µS)	127.5	132.4	130.24	1.54	134.9	150.4	141.85	5.04	124.4	143.3	131.44	4.79
4	TDS (mg/L)	64.14	68.52	66.69	1.34	66.23	71.11	69.15	1.41	63	71.77	65.51	2.27
5	Alkalinity (mg/L)	25	35	31.46	3.07	25	40	33.85	5.45	30	38	33.38	2.81
6	DO (mg/L)	3	8	5.73	1.67	3	7	5.43	0.99	4	8.2	6.18	1.40
7	Chloride (mg/L)	38	55	42.54	4.41	32	46	37.38	3.75	27	39	32.23	3.34
8	Hardness (mg/L)	15	41	28.69	9.26	14	40	26.08	7.79	14	40	25.84	7.30
9	Calcium (mg/L)	4	7.5	5.55	1.11	4	7.5	5.18	1.05	3	6.2	4.56	1.01
10	Magnesium (mg/L)	2.4	8.6	5.62	2.17	2.4	8.3	5.09	1.76	3	7.3	4.50	1.33
11	Sodium (mg/L)	13.6	18.4	16.85	1.20	12.4	17.2	15.67	1.31	11.4	16.7	14.68	1.37
12	Potassium (mg/L)	2.5	4.5	3.16	0.53	2.5	4	3.26	0.39	2	4	3.00	0.61
13	Ammonia (mg/L)	0.04	0.31	0.11	0.09	0.04	0.31	0.11	0.08	0.03	0.94	0.23	0.32
14	Nitrate (mg/L)	1.03	1.12	1.06	0.03	1.03	1.12	1.06	0.02	0.93	1.03	0.98	0.03
15	Phosphates (mg/L)	0.04	0.4	0.151	0.09	0.04	0.4	0.15	0.09	0.03	0.99	0.34	0.36

Table 2. Assigned Weights and Relative Weights of Parameters for Water Quality Index

Parameter	Assigned Weights	Relative Weights
pH	2.7	0.069
TDS	3	0.077
EC	3.5	0.089
Alkalinity	3	0.077
DO	4	0.102
Chloride	3	0.077
Hardness	3	0.077
Calcium	3	0.077
Magnesium	3	0.077
Sodium	3	0.077
Potassium	1	0.026
Ammonia	1	0.026
Nitrate	1	0.026
Phosphate	5	0.128

Table 3. Calculated Water Quality Index (WQI) values for various seasons

Site	Post-Monsoon	Pre-Monsoon	Monsoon	Mean WQI
V1	24.12	30.46	29.79	28.12
V2	60.12	63.79	62.46	62.12
V3	20.79	22.35	25.68	22.94
V4	32.79	33.46	34.68	33.64
V5	25.57	22.68	23.46	23.9
V6	40.68	45.12	50.79	45.53
V7	70.23	60.46	50.12	60.27
V8	37.46	34.79	33.46	35.23
V9	28.35	30.68	32.12	30.38
V10	55.68	60.35	62.79	59.6
V11	35.79	36.12	40.57	37.49
V12	35.79	38.23	37.57	37.2
V13	35.46	31.79	32.35	33.2
Min	20.79	22.35	23.46	22.94
Max	70.23	63.79	62.79	62.12

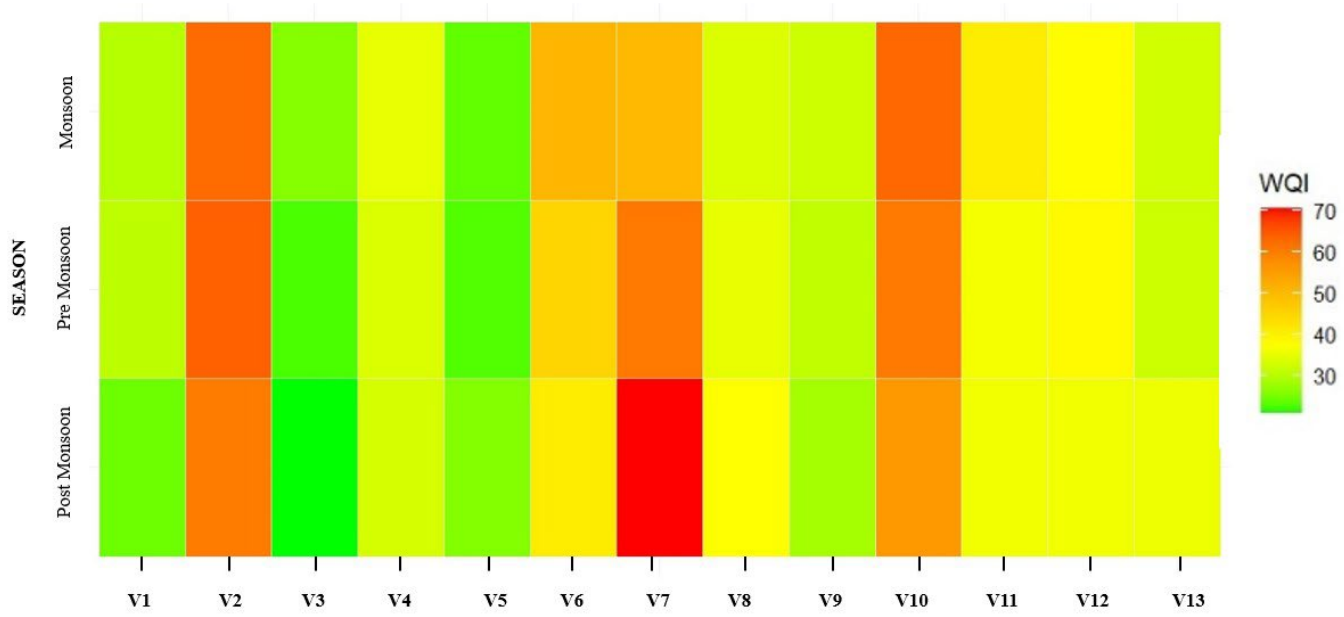

Fig. 2 Heat map of WQI — surface water samples, Vellayani Lake

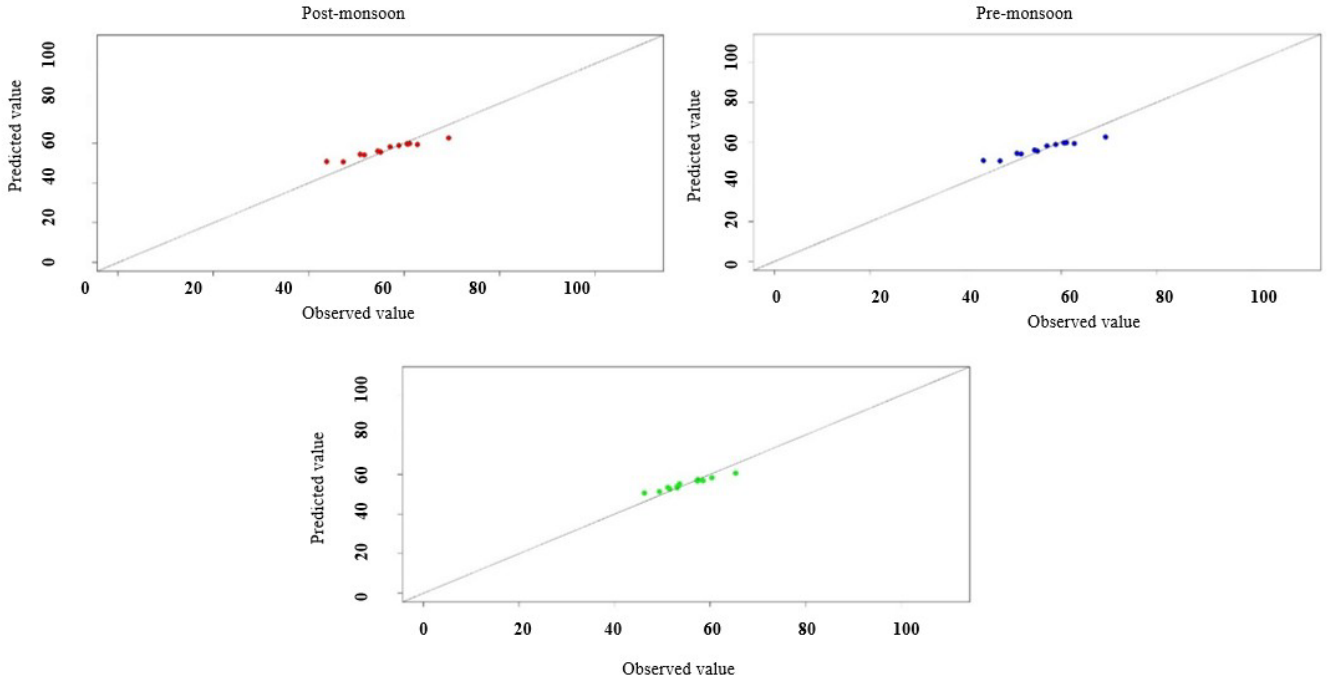

Fig. 3 Observed values vs. Predicted values for RF model

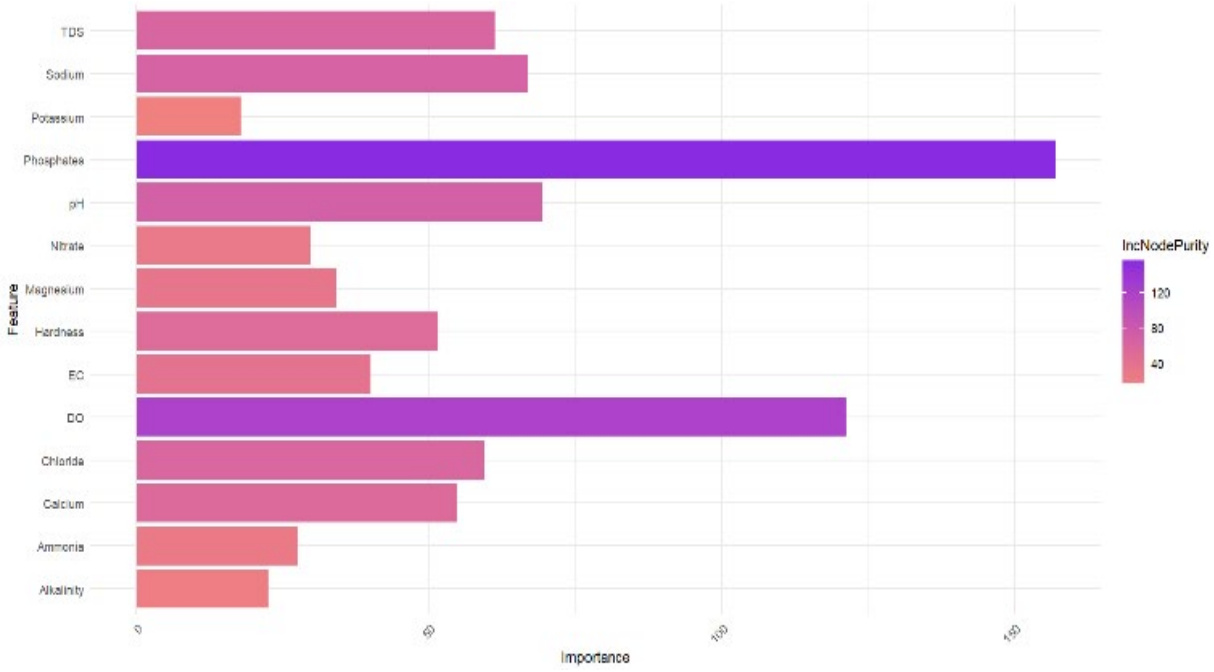

Fig. 4 Feature Importance Plot for Water Quality Parameters based on RF Model

Key influencing factors such as phosphate, dissolved oxygen, electrical conductivity, and total dissolved solids were identified, consistent with observed field data. The model's predictions of higher WQI values at stations V2 (northern part of VL), V7 (central), and V10 (southern part) corroborated the manual calculations, highlighting areas with persistent water quality issues. By combining conventional WQI assessment with machine learning validation, the study delivered an integrated perspective on water quality dynamics in Vellayani freshwater Lake. It highlighted specific periods and hotspots that warrant targeted management efforts to address nutrient loading and water quality deterioration. These findings reinforce the need for sustained monitoring and the use of advanced analytical approaches to support the long-term sustainability of freshwater ecosystems.

CRediT authorship contribution statement.

Sabu Joseph: Investigation, Formal analysis, Data curation, Writing - original draft, Writing – review & editing; Supervision, Funding acquisition; **S.Sukanya**: Conceptualization, Methodology, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization; **M.R. Vishnuprasad**: Investigation, Formal analysis, Data curation, Writing - original draft, Visualization.

Declaration of competing interest

The authors declare that they have no known financial or personal conflicts of interest that could have influenced the work reported in this paper.

Acknowledgements

The authors acknowledge the research funding from University of Kerala, India and Department of Science & Technology (DST), Govt. of India for conducting this work.

References

Ahmad, M. W., Reynolds, J., & Rezgui, Y. (2018). Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of cleaner production*, *203*, 810-821.

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., ... & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, *578*, 124084.

Akhtar, N., Ishak, M. I. S., Ahmad, M. I., Umar, K., Md Yusuff, M. S., Anees, M. T., Qadir, A. & Ali Almanasir, Y. K. (2021). Modification of the water quality index (WQI) process for simple calculation using the multi-criteria decision-making (MCDM) method: a review. *Water*, *13*(7), 905.

Aldrees, A., Awan, H. H., Javed, M. F., & Mohamed, A. M. (2022). Prediction of water quality indexes with ensemble learners: Bagging and Boosting. *Process Safety and Environmental Protection*, *168*, 344-361.

Alqahtani, A., Shah, M. I., Aldrees, A., & Javed, M. F. (2022). Comparative assessment of individual and ensemble machine learning models for efficient analysis of river water quality. *Sustainability*, *14*(3), 1183.

Bakır, R., Orak, C., & Yüksel, A. (2024). Optimizing hydrogen evolution prediction: A unified approach using random forests, lightGBM, and Bagging Regressor ensemble model. *International Journal of Hydrogen Energy*, *67*, 101-110.

Banerji, U. S., Shaji, J., Arulbalaji, P., Maya, K., Mohan, S. V., Dabhi, A. J., ... & Padmalal, D. (2021). Mid-late Holocene evolutionary history and climate reconstruction of Vellayani lake, south India. *Quaternary International*, *599*, 72-94.

BIS. (2012). *Indian Standard Drinking Water — Specification* (Second Revision). Bureau of Indian Standard (2012), 1–8. http://cgwb.gov.in/Documents/WQ-standards.pdf

Horvat, M., Horvat, Z., & Pastor, K. (2021). Multivariate analysis of water quality parameters in Lake Palic, Serbia. *Environmental Monitoring and Assessment*, *193*(7), 410.

Huang, B. F., & Boutros, P. C. (2016). The parameter sensitivity of random forests. *BMC bioinformatics*, *17*, 1-13.

Huang, H., Jia, R., Shi, X., Liang, J., & Dang, J. (2021). Feature selection and hyper parameters optimization for short-term wind power forecast. *Applied Intelligence*, 1-19.

Jha, M. K., Shekhar, A., & Jenifer, M. A. (2020). Assessing groundwater quality for drinking water supply using hybrid fuzzy-GIS-based water quality index. *Water Research*, *179*, 115867.

Kwon, H. G., & Jo, C. D. (2023). Water quality assessment of the Nam River, Korea, using multivariate statistical analysis and WQI. *International Journal of Environmental Science and Technology*, *20*(3), 2487-2502.

Lap, B. Q., Du Nguyen, H., Hang, P. T., Phi, N. Q., Hoang, V. T., Linh, P. G., & Hang, B. T. T. (2023). Predicting Water Quality Index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system. *Ecological Informatics*, *74*, 101991.

Lee, H., Park, S., Hang, V., Nguyen, M., & Shin, H. S. (2023). Proposal for a new customization process for a data-based water quality index using a random forest approach. *Environmental Pollution*, *323*, 121222.

Li, T., Li, S., Liang, C., Bush, R. T., Xiong, L., & Jiang, Y. (2018). A comparative assessment of Australia's Lower Lakes water quality under extreme drought and post-drought conditions using multivariate statistical techniques. *Journal of Cleaner Production*, *190*, 1-11.

Mechal, A., Fekadu, D., & Abadi, B. (2024). Multivariate and Water Quality Index Approaches for Spatial Water Quality Assessment in Lake Ziway, Ethiopian Rift. *Water, Air, & Soil Pollution*, *235*(1), 78.

Monira, U., Sattar, G. S., & Mostafa, M. G. (2024). Assessment of surface water quality using the Water Quality Index (WQI) and multivariate statistical analysis (MSA), around tannery industry effluent discharge areas. *H2Open Journal*, *7*(2), 130-148.

Naderian, D., Noori, R., Heggy, E., Bateni, S. M., Bhattarai, R., Nohegar, A., & Sharma, S. (2024). A water quality database for global lakes. *Resources, Conservation and Recycling*, *202*, 107401.

Sakaa, B., Elbeltagi, A., Boudibi, S., Chaffaï, H., Islam, A. R. M. T., Kulimushi, L. C., ... & Wong, Y. J. (2022). Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin. *Environmental Science and Pollution Research*, *29*(32), 48491-48508.

Singh, S. K., Singh, P., & Gautam, S. K. (2016). Appraisal of urban lake water quality through numerical index, multivariate statistics and earth observation data sets. *International journal of environmental science and technology*, *13*, 445-456.

Sukanya, S., & Sabu, J. (2020). Water quality assessment using environmetrics and pollution indices in a tropical river, Kerala, SW Coast of India. *Current World Environment*, *15*(1), 11.

Sukanya, S., & Sabu, J. (2023). Climate change impacts on water resources: An overview. *Visualization Techniques for Climate Change with Machine Learning and Artificial Intelligence*, 55-76.

Talukdar, S., Bera, S., Naikoo, M. W., Ramana, G. V., Mallik, S., Kumar, P. A., & Rahman, A. (2024). Optimisation and interpretation of machine and deep learning models for improved water quality management in Lake Loktak. *Journal of Environmental Management*, *351*, 119866.

Vasistha, P., & Ganguly, R. (2020). Water quality assessment of natural lakes and its importance: An overview. *Materials Today: Proceedings*, *32*, 544-552.

World Health Organization. (2022). *Guidelines for drinking water quality: incorporating the first and second addenda*. World Health Organization.

Yin, Y., Xia, R., Liu, X., Chen, Y., Song, J., & Dou, J. (2024). Spatial response of water level and quality shows more significant heterogeneity during dry seasons in large river-connected lakes. *Scientific Reports*, *14*(1), 8373.

Zhang, H., Ren, X., Chen, S., Xie, G., Hu, Y., Gao, D., ... & Wang, H. (2024). Deep optimization of water quality index and positive matrix factorization models for water quality evaluation and pollution source apportionment using a random forest model. *Environmental Pollution*, *347*, 123771.

Zhi, W., Appling, A. P., Golden, H. E., Podgorski, J., & Li, L. (2024). Deep learning for water quality. *Nature Water*, *2*(3), 228-241.