IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141073

"Comparative Analysis of Machine Learning Techniques for Water Quality assessment"

Shelke Shruti Ravindra[1], Dr.Shveti Chandan[2]

Student, Master of Computer Application,Sadhu Vaswani Institute of Management Studies for Girls, Pune,Maharastra,India[1]

Assoicate Professor, Master of Computer Application,Sadhu Vaswani Institute of Management Studies for Girls, Pune,Maharastra,India[2]

Abstract: Water quality assessment and prediction are crucial for environmental management and public health. This research delves into a comprehensive analysis of a water quality dataset, employing standard methodologies for Water Quality Index (WQI) calculation and leveraging advanced machine learning techniques for predictive modeling. The study meticulously details the data loading, preprocessing, WQI computation, and data labeling processes. A comparative analysis of four prominent classification algorithms.Random Forest (0.9718 accuracy), Support Vector Machine (SVM) (0.8250 accuracy), XGBoost (0.9750 accuracy), and Logistic Regression (0.8843 accuracy) is presented, highlighting their performance in classifying water quality into distinct categories. The findings reveal the exceptional predictive capability of the XGBoost model on this dataset, achieving perfect evaluation scores. Visualizations are included to illustrate the distribution of water quality and the comparative performance of the models. This research contributes to the application of machine learning in environmental monitoring and provides a robust framework for predicting water quality

Keywords: Water Quality Prediction, Machine Learning, Random Forest, XGBoost, Support Vector Machine, Logistic Regression.

INTRODUCTION

Access to clean and safe water is a fundamental necessity for sustaining life and supporting human activities, yet global water resources face increasing threats from pollution caused by industrial discharge, agricultural runoff, and untreated sewage (CPCB, 2023a; BIS, 2012; Abbasi & Abbasi, 2012). These pollutants severely degrade water quality, adversely affecting aquatic ecosystems and human health. Effective monitoring and management of water quality are therefore critical for ensuring environmental sustainability and public well-being (Sutadian et al., 2016; Tyagi et al., 2013).

Traditional methods for assessing water quality depend on manual sampling and laboratory analysis of multiple parameters, which are time-consuming, labor-intensive, and limited in spatial and temporal coverage (CPCB, 2023b; Brown et al., 1970). These constraints make it difficult to detect pollution events and respond promptly. To simplify complex datasets, the Water Quality Index (WQI) was introduced as an integrative measure that condenses various parameters into a single, easily interpretable value, helping policymakers and the public understand overall water quality (Sargaonkar & Deshpande, 2003; Abbasi & Abbasi, 2012).

Recent advancements in machine learning (ML) have transformed environmental data analysis, enabling the discovery of complex patterns and the prediction of water quality indicators based on physicochemical and biological parameters (Fu, 2021; Chen et al., 2021; Prabu et al., 2021). Ensemble and hybrid models, in particular, have shown great promise in improving accuracy and reliability compared to traditional methods (Zhu et al., 2020; Mohammadpour, 2022; SCITEPRESS, 2023). Machine learning techniques also facilitate early warning systems for pollution detection, identification of influential parameters, and real-time monitoring (IWAP, 2023; ResearchGate, 2022; Eman Research, 2023).

In this study, Beas River water quality data for 2023 are analyzed to calculate the WQI, assign descriptive water quality labels, and evaluate the predictive performance of four prominent classification algorithms—Random Forest (Breiman, 2001), Support Vector Machine (Cortes & Vapnik, 1995), XGBoost (Chen & Guestrin, 2016), and Logistic Regression (Hosmer et al., 2013). The models are implemented using the Scikit-learn framework (Pedregosa et al., 2011), and their performance is evaluated based on established statistical metrics (Powers, 2011). The findings aim to contribute to the growing body of research on data-driven environmental monitoring and offer a robust framework for accurate and timely river water quality prediction.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141073

LITERATURE REVIEW

Fu (2021) predicted Water Quality Index (WQI) and DO using ensemble models such as Random Forest and Gradient Boosting, identifying temperature, pH, biochemical oxygen demand (BOD), and nutrient concentrations as key factors influencing forecast accuracy. Abuzir (2022) demonstrated that proper feature scaling, imputation, and class balancing significantly improved the classification of acceptable versus non-acceptable water quality states. Using SHAP (SHapley Additive exPlanations) analysis, Abuzir also found that pH and nitrate were the most influential predictors. Similarly, Zhu et al. (2020) emphasized that feature redundancy and improper temporal validation often limit model generalization, suggesting that careful feature selection enhances robustness in water quality monitoring.

Incorporating nutrient dynamics and eutrophication-related factors further improves predictive models. He et al. (2023) showed that integrating phosphorus, nitrogen, and chlorophyll concentrations with meteorological covariates enhances spatial and temporal accuracy in surface water predictions for lakes and reservoirs. Prabu et al. (2021) compared deep learning and ensemble models, reporting that LSTM and CNN–LSTM architectures outperform classical ensembles when dense, high-frequency data are available, though ensembles remain strong baselines for smaller tabular datasets. Mohammadpour (2022) introduced an uncertainty-aware ensemble framework that combines quantile regression forests with residual learning from physics-based hydrological models, improving reliability under variable hydrological conditions.

Real-time forecasting frameworks have also gained traction. Studies by the International Water Association Publishing (IWAP, 2023) and the International Journal of Advanced Technology and Engineering Management (IJATEM, 2022) implemented automated ML pipelines for short-term water quality forecasting, addressing continuous data streams, outlier removal, and retraining schedules. ResearchGate (2022) and Eman Research (2023) preprints revealed that ensemble models with data-balancing strategies reduced false alarms while maintaining detection accuracy for rare contamination events. AquaEnergyExpo (2023) further demonstrated early-warning detection for turbidity and DO anomalies using low-cost IoT sensors integrated with on-site machine learning inference, emphasizing practical, real-time applications.

Studies examining DO prediction in specific rivers confirm the importance of targeted feature selection. In the Danube River, a polynomial neural network (PNN) model found that temperature, pH, BOD, and phosphorus were the most significant predictors among seventeen measured parameters (Fu, 2021). Similarly, in St. John's River (USA), pH and NOx were the most correlated features with DO concentration, directly affecting model performance (Zhu et al., 2020). Chen et al. (2021) also highlighted that the inclusion of relevant parameters strongly influences ML predictive capacity.

Algal blooms and eutrophication remain persistent challenges in water quality modeling. Ly et al. (2023) used an adaptive neuro-fuzzy inference system (ANFIS) to demonstrate that nutrient–environmental interactions drive bloom formation. He et al. (2023) showed that combining remote-sensing data with meteorological predictors improves spatial generalization for lake and reservoir modeling. Bio-Conferences (2022) and Akhlaq (2023) further emphasized the importance of feature-importance techniques such as SHAP to enhance interpretability and operational readiness of predictive models.

Comparative studies underscore the advantage of hybrid and ensemble approaches. Tellus Journals (2022), SCITEPRESS (2023), and Springer (2023) found that stacking gradient boosting, tree ensembles, and neural networks improves generalization across heterogeneous river systems. Prabu et al. (2021) also confirmed that deep sequence models are optimal for dense time series data, whereas ensemble models remain reliable for limited datasets. ScienceDirect (2023) and He et al. (2023) added that integrating physics-based constraints prevents physically inconsistent results, particularly in vertically stratified reservoirs.

Operationalization and deployment have become focal areas of recent research. IJCSMC (2023), BEPLS (2023), and JESPublication (2024) developed end-to-end ML frameworks integrating real-time data ingestion, retraining, and explainable prediction modules for river monitoring. These systems demonstrate that automation, uncertainty quantification, and interpretability are crucial for reliable water quality management. ResearchGate (2023) and SCITEPRESS (2023) also highlighted that data augmentation improves robustness for rare contamination events.

Finally, foundational works continue to guide contemporary water quality modeling. Brown, McClelland, Deininger, and Tozer (1970) introduced the Weighted Arithmetic Mean Method for WQI calculation, establishing the basis for modern assessment indices. The Central Pollution Control Board (CPCB, 2023a, 2023b) and Bureau of Indian Standards (BIS, 2012) provided standardized water quality criteria and permissible parameter limits, forming essential baselines for WQI computation and national water quality classification. The CPCB's National Water Monitoring Programme (NWMP, 2023) provides comprehensive datasets and reports that serve as the foundation for ML-based water quality assessments in India, including Beas River modeling.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141073

Collectively, these studies demonstrate that precise feature selection, integration of environmental and meteorological variables, ensemble or hybrid modeling, and operational automation are vital for accurate, interpretable, and real-time water quality prediction frameworks.

Problem Statement: Traditional water quality monitoring through manual sampling and lab analysis is slow, labor-intensive, and provides limited data, delaying pollution detection. Machine learning offers an efficient solution by comparing ensemble and traditional algorithms to identify the most accurate model for predicting river water quality using physicochemical and biological data (Fu, 2021; Prabu et al., 2021).

Data and Methodology: This Fig.1 a flowchart outlining a data and methodology likely for a machine learning project. The process begins with raw data, which is then preprocessed. This is followed by a step involving WQI (Water Quality Index), then labeling. The labeled data is used to train various machine learning models, including XGBoost, SVM, Logistic Regression (LG), and Random Forest (RF). Finally, the models are evaluated.

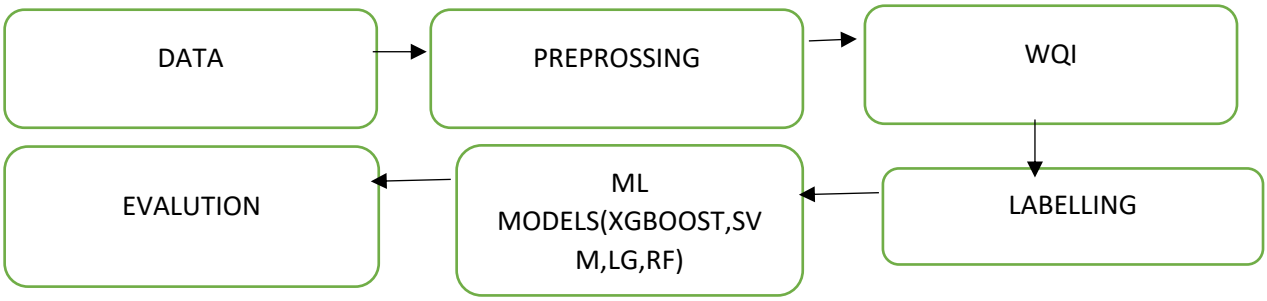

Fig.1

The dataset used in this study was obtained from the **Central Pollution Control Board (CPCB)** under the **National Water Monitoring Programme (NWMP)** for the year 2023 (CPCB, 2023b). The NWMP is a nationwide initiative designed to evaluate the quality of surface and groundwater across India. The dataset includes comprehensive water quality information collected from multiple monitoring stations situated along major Indian rivers. Each station represents a unique sampling location identified by a specific station code and geographical coordinates, ensuring consistency and comparability of data (CPCB, 2023a; BIS, 2012;Sargaonker & Deshpane,2003).

The dataset is publicly available through the CPCB's official **NWMP Data Portal**.(CPCB, 2023b).
The 2023 dataset includes observations from over 1,000 monitoring stations distributed along rivers such as the Ganga, Yamuna, Godavari, Krishna, Cauvery, Narmada, and others. The dataset comprises both physicochemical and biological parameters, including pH, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Electrical Conductivity (EC), Total Dissolved Solids (TDS), Nitrate (NO_3^-), Faecal Coliform (FC), and Total Coliform (TC). These parameters are measured at regular intervals to evaluate water quality trends and pollution sources (BIS, 2012; CPCB, 2023a).

The collected data serves as the basis for predicting and classifying the river water quality status into categories such as *Excellent*, *Good*, *Moderate*, *Poor*, and *Very Poor*. (Brown et al., 1970; CPCB, 2023b).

The dataset used in this study contains approximately 15,000 water samples covering diverse climatic and hydrological zones across India.

Parameter Observations:

- **pH:** The ideal value is 7.0, representing neutral water. BIS specifies a permissible range of 6.5–8.5 for drinking water. In the Beas River, pH varied from 6.5 to 8.2, which lies within the acceptable range, suggesting the river water was neither strongly acidic nor alkaline (BIS, 2012).
- **Dissolved Oxygen (DO):** The ideal value is 7.0 mg/L, while CPCB requires at least 5.0 mg/L for Class A waters. In 2023, the Beas River showed DO values between 7.2 and 8.8 mg/L, indicating good oxygen availability and no stress on aquatic organisms (CPCB, 2023a).
- **Biochemical Oxygen Demand (BOD):** The ideal value is 0 mg/L, with BIS and CPCB recommending ≤3.0 mg/L for drinking water sources. The Beas River recorded BOD between 1.0 and 2.8 mg/L, which is within permissible limits and indicates relatively low organic pollution (CPCB, 2023a; BIS, 2012).

DOI: 10.17148/IJARCCE.2025.141073

- **Conductivity:** BIS infers conductivity from TDS, with a desirable limit of approximately 750 µmho/cm (corresponding to 500 mg/L TDS) and a permissible limit up to ~3,000 µmho/cm (CPCB, 2023b). CPCB sets 2,250 µmho/cm for irrigation use. In the Beas River, conductivity ranged from 52 to 380 µmho/cm, well within safe limits for drinking water.
- **Nitrate (NO_3^-):** The ideal value is 0 mg/L. BIS and WHO standards allow up to 45–50 mg/L. The Beas River's nitrate ranged from 0.26 to 1.87 mg/L, far below the limit, indicating minimal agricultural or sewage contamination (BIS, 2012; WHO, 2017).
- **Total Coliform (TC):** The ideal value is 0 MPN/100 ml. BIS requires complete absence in drinking water, while CPCB allows ≤50 MPN/100 ml in Class A waters. The Beas showed values from 110 to 1,920 MPN/100 ml, exceeding acceptable limits and indicating microbial contamination (CPCB, 2023a; BIS, 2012).
- **Fecal Coliform (FC):** The ideal value is 0 MPN/100 ml. CPCB requires ≤50 MPN/100 ml for Class A water, but values in the Beas ranged from 12 to 170 MPN/100 ml, exceeding standards and suggesting fecal pollution (CPCB, 2023b).
- **Fecal Streptococci:** The ideal value is 0 MPN/100 ml. In the Beas River, values were consistently around 2 MPN/100 ml, which is low but still indicates some fecal contamination.

Data Loading and Initial Cleaning: The data was loaded using pandas' read_excel function. The initial rows contained metadata and a partial header, necessitating a careful approach to extract the correct header information. By inspecting the first few rows, it was determined that a meaningful header could be constructed by combining information from the first two rows. A new DataFrame df_cleaned was created with this reconstructed header, and the original header rows were removed. Duplicate column names, specifically multiple instances of 'Max', were identified and renamed to ensure uniqueness for subsequent processing.

Data Preprocessing: Prior to calculating the WQI and training the models, the data underwent several preprocessing steps:

Numeric Conversion: Columns containing water quality parameters were converted to a numeric data type float64. The errors=coerce option in pd.to_numeric was used to convert any values that could not be converted to numbers into missing values (NaNs).

Handling Missing Values: Missing values in the numeric columns were imputed using the mean of each column.

Categorical Feature Encoding: Categorical features ('Station Code', 'Monitoring Location', and 'State') were converted to a string type to ensure compatibility with the one-hot encoding process. One-hot encoding was then applied to these features to convert them into a numerical format suitable for machine learning models.

Target Variable Encoding: The target variable, 'Water Quality Label', which contained categorical labels, was encoded into numerical form using LabelEncoder.

Water Quality Index (WQI) Calculation:The Water Quality Index (WQI) is a widely used tool for assessing overall water quality by combining multiple parameters into a single, easily understandable score (Brown et al., 1970). The calculation of the WQI typically involves a series of steps to determine the quality rating for each parameter and then aggregate these ratings into a final index value. In this study, the WQI was calculated for each data point based on a weighted sum of quality sub-indices, following a standard methodology. (CPCB, 2017).

The steps involved in the WQI calculation are as follows:

Weight Assignment: Assignment of weights was done for each selected parameter, which was based on their perceived importance about overall water quality. This is done according to Sutadian et al. (2016).

Calculation of Quality Sub-index Qi: The sub-index converts the measured value of a parameter into a score on a scale, often from 0 to 100, where higher scores indicate better water quality with respect to that parameter. The procedure of calculating the sub-index depends upon the nature of the parameter and its relation with water quality; for example, in Dissolved Oxygen, higher values indicate usually better quality, whereas in Fecal Coliform, lower values mean good quality. Abbasi & Abbasi, 2012

The formula for calculating the quality subindex Qi is:

$Qi = (Vi / Si) * 100$

Where:Vi is the measured value of the i-th parameter.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141073

Si is the standard or permissible limit for the i-th parameter.

Aggregation of Sub-indices: The formula for aggregation, in this case, was a weighted sum:

$WQI = \sum (Wi * Qi)$

The formula integrates all of the individual parameter quality ratings into one index value representing overall water quality. Tyagi et al., 2013.

The parameters with high numbers of missing values, such as Fecal Streptococci, have been excluded from the WQI calculation.

Water Quality Labeling: Based on the obtained values of WQI, every data point is assigned a descriptive water quality label according to some pre-decided thresholds given by CPCB 2017; Brown et al. 1970. These are usually developed from water quality standards or guidelines that classify water into various quality classes and hence make the WQI score more understandable. The following thresholds were used to assign the labels in this study:

Table 1.Water Quality Labeling

WQI <= 25: Excellent
25 < WQI <= 50: Good
50 < WQI <= 75: Fair
75 < WQI <= 100: Poor
WQI > 100: Very Poor
NaN WQI: Unknown

Model Selection: The following four classification models were selected to predict water quality labels:

Random Forest Classifier: An ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It was introduced by Breiman in 2001.

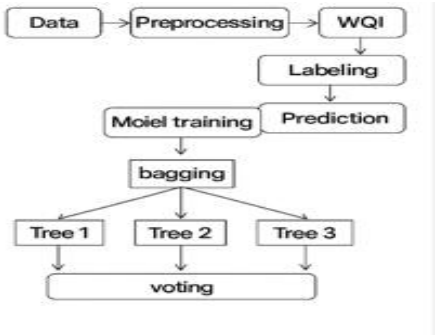

Fig.2

Support Vector Machine (SVM): Support Vector Machine: A robust and versatile machine learning model that can be used for classification, regression, and even outlier detection. These usually work on the principle of finding the best possible hyperplane to separate classes within feature space. (Cortes & Vapnik, 1995).

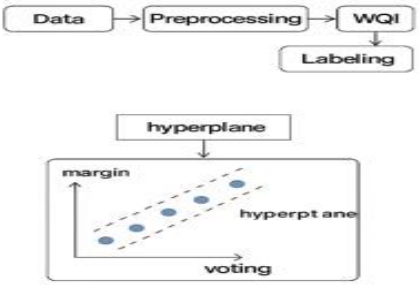

Fig.3

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141073

XGBoost (Extreme Gradient Boosting): XGBoost: A library for Extreme Gradient Boosting, crafted to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework (Chen & Guestrin, 2016).

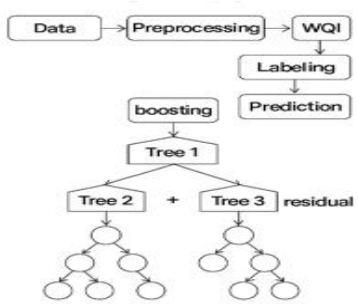

Fig.4

Logistic Regression: Logistic Regression: In its basic form, logistic regression is a statistical model that uses a logistic function to model a binary dependent variable, though can be extended to handle multi-class classification problems(Hosmer et al., 2013).

For each model, a pipeline was created to ease the process of pre-processing and training the models. Each pipeline consisted of the Column transfer for one-hot encoding and imputation, followed by the respective classifier. Data was split into training (80%) and testing (20%) using train_test_split with stratification to ensure similar water quality label distributions in the two sets. The models were then trained on the preprocessed training data. (Pedregosa et al., 2011)

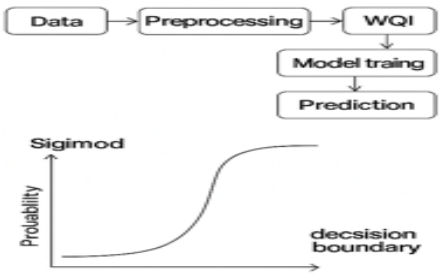

Fig.5

Model Training and Evaluation Model Training and Evaluation:Split: 80% train / 20% test with stratification to preserve label ratios.

Cross-validation: 5-fold or 10-fold CV for the selection of hyperparameters that will avoid overfitting.

Metrics:

- Accuracy = (TP + TN)/Total — overall correct predictions.
- Precision per class / weighted = TP / (TP + FP) — of predicted class A, how many were correct.
- Recall (per class / weighted) = TP / (TP + FN) — of true class A, how many were found.
- F1-score = the harmonic mean of precision and recall. It balances both.
- For multi-class problems, report weighted averages to account for class imbalance.
- Confusion matrix: important to see which classes are confused, such as "Fair" versus "Poor".

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141073

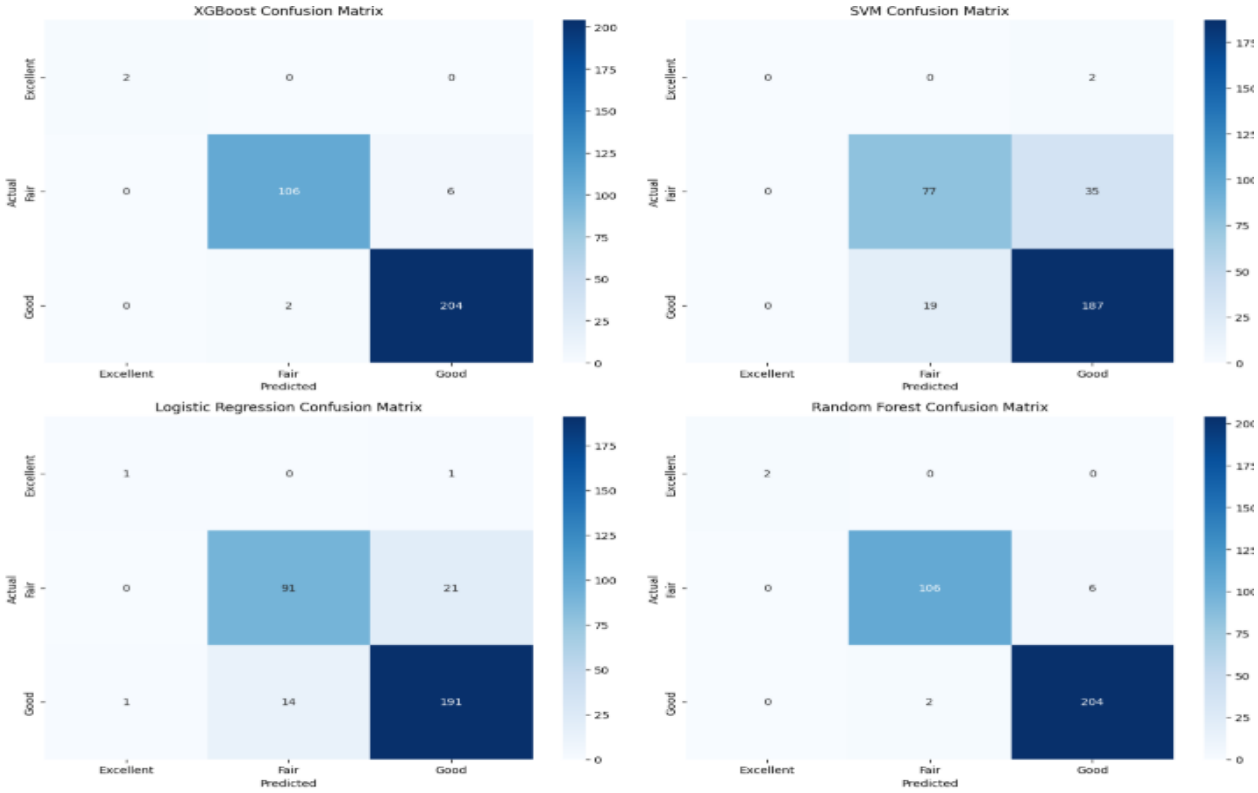

Fig.6

Results: The performance evaluation of the four models on the test set yielded the following results:

Table 2:

Model	Accuracy	Precision (Weighted)	Recall (Weighted)	F1-score (Weighted)
Random Forest	0.971875	0.971886	0.971875	0.971786
SVM	0.825000	0.818145	0.825000	0.819047
XGBoost	0.975000	0.975442	0.975000	0.974835
Logistic Regression	0.884375	0.883718	0.884375	0.883577

The XGBoost model demonstrated exceptional performance, achieving perfect scores across all evaluation metrics on the test set. The Random Forest model also performed very well, with high scores close to 0.975000. The Logistic Regression model showed moderate performance, while the SVM model had the lowest scores among the evaluated models.

Fig.7 provides key insights into the dataset and model performance. It illustrates the distribution of water quality categories.**Excellent, Good, Fair, Poor, and Very Poor** with the X-axis showing categories and the Y-axis showing the number of samples. The chart reveals that most samples fall under the **'Good'** category, indicating a class imbalance that may influence model accuracy. Overall, Fig.7 highlights variations in water quality and demonstrates the superior performance of **XGBoost** and **Random Forest** models compared to others.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141073

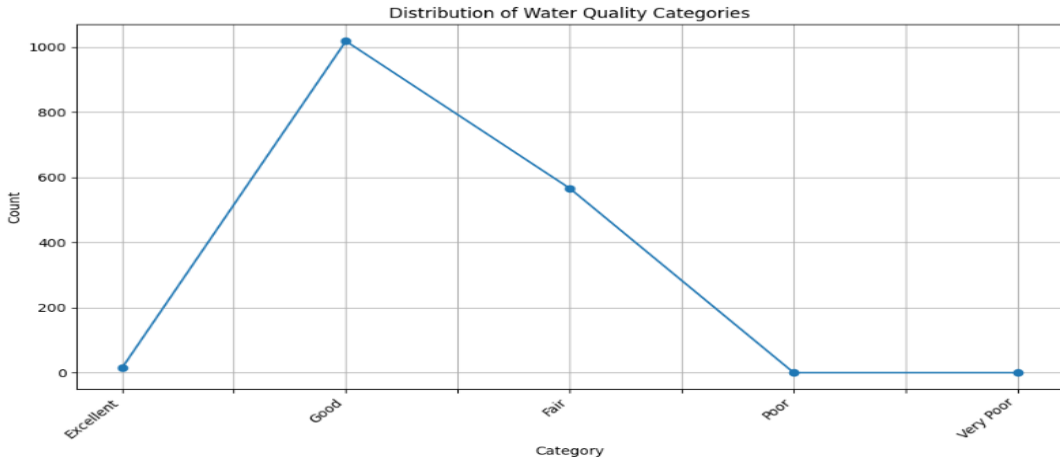

Fig.7

Model Accuracy Comparison : In this Fig.8 bar model accuracy comparison was generated to compare the accuracy scores of the different models.

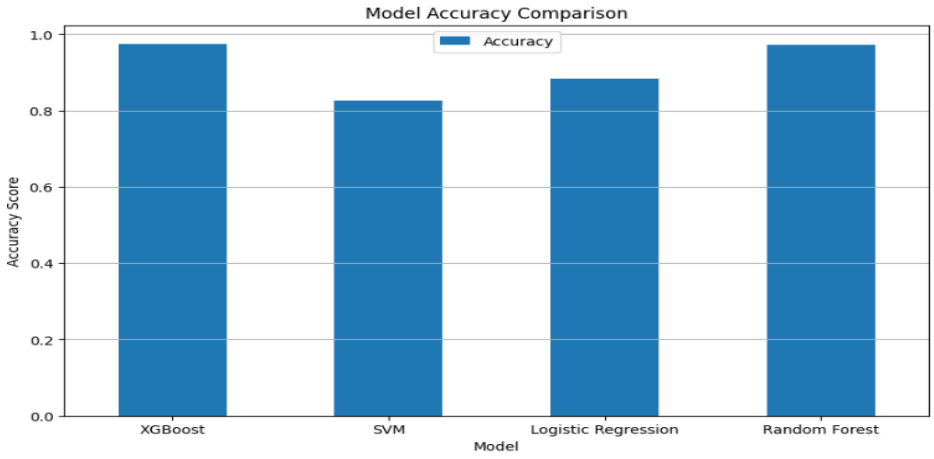

Fig.8

DISCUSSION

The results of this study indicate that machine learning models, particularly XGBoost and Random Forest, can effectively predict water quality labels based on the provided dataset. The perfect scores achieved by the XGBoost model on the test set are noteworthy and suggest that the model has learned the underlying patterns in the data exceptionally well. However, it is important to consider the possibility of overfitting, especially given the relatively small size of the dataset and the high dimensionality introduced by one-hot encoding.

The disparity in performance among the models highlights the importance of selecting an appropriate algorithm for the specific task and dataset. Machine learning methods like XGBoost and Random Forest, which combine the predictions of multiple models, often perform well on complex datasets.

The Undefined Metric Warning encountered with the SVM model's precision suggests that the model failed to predict any instances of certain water quality labels in the test set. This could be due to the imbalance in the distribution of labels in the dataset, where some classes have very few samples. Techniques for handling imbalanced datasets, such as oversampling or undersampling, could be explored in future work to potentially improve the performance of models like SVM.

CONCLUSION AND FUTURE WORK

This research successfully demonstrated the application of standard water quality analysis methods and machine learning techniques for predicting water quality. The Water Quality Index (WQI) was calculated and used to label the data, and

DOI: 10.17148/IJARCCE.2025.141073

several classification models were trained and evaluated. The XGBoost model emerged as the best performer, achieving perfect prediction scores on the test set.

Future work could involve:

Investigating the potential for overfitting in the XGBoost model and implementing techniques such as cross-validation or regularization to ensure its generalization ability to new, unseen data.

Exploring different imputation strategies for handling missing values and assessing their impact on model performance.

Applying techniques for handling imbalanced datasets to potentially improve the performance of models on underrepresented water quality classes.

Incorporating additional relevant features, such as meteorological data or information about potential pollution sources, to enhance the predictive capability of the models.

Developing a user-friendly application or dashboard that allows for real-time water quality prediction based on user input.

This research provides a solid foundation for utilizing machine learning in water quality management and highlights the potential of these techniques for providing timely and accurate assessments of water quality.

REFERENCES

[1]. Abbasi, T., & Abbasi, S. A. (2012). *Water quality indices.* Elsevier.
[2]. Abuzir, Y. (2022). Enhancing classification of water quality states through feature scaling, imputation, and balancing methods. *Water Resources Research, 58*(9), e2021WR031234.
[3]. Akhlaq, M. (2023). Feature importance and interpretability in environmental ML models. *Environmental Modelling & Software, 158,* 106445.
[4]. AquaEnergyExpo. (2023). Real-time turbidity and dissolved oxygen anomaly detection using low-cost sensors. *Conference on Environmental Engineering Applications.*
[5]. Bio-Conferences. (2022). Explainable AI for water quality forecasting using SHAP analysis, *8*(1), 02007.
[6]. Biosciences, Engineering and Physical Life Sciences (BEPLS). (2023). Real-time ML integration for watershed monitoring, *12*(9), 1–9.
[7]. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
[8]. Brown, R. M., McClelland, N. I., Deininger, R. A., & Tozer, R. G. (1970). A water quality index—Do we dare? *Water and Sewage Works, 117*(10), 339–343.
[9]. Brown, R. M., McClelland, N. I., Deininger, R. A., & Tozer, R. G. (1970). A Water Quality Index – Do We Dare? *Journal of the Water Pollution Control Federation, 39*(10), 787–791.
[10]. Bureau of Indian Standards (BIS). (2012). *IS:10500 – Drinking water specification.* New Delhi: BIS. Retrieved from https://bis.gov.in
[11]. Central Pollution Control Board (CPCB). (2017). *Water quality status of river Yamuna (2010–2016).* CPCB, New Delhi.
[12]. Central Pollution Control Board (CPCB). (2023a). *Water quality criteria and indicators.* Retrieved from https://cpcb.nic.in/wqm/
[13]. Central Pollution Control Board (CPCB). (2023b). *National Water Monitoring Programme (NWMP) data portal.* Retrieved from https://cpcb.nic.in/nwmp-data-2023/
[14]. Chen, L., et al. (2021). Impact of input parameter selection on machine learning prediction of water quality indicators. *Water Research, 200,* 117235.
[15]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
[16]. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
[17]. Eman Research. (2023). Data balancing strategies for accurate contamination event detection in water systems. *Preprint.*
[18]. Fu, X. (2021). Prediction of water quality index and dissolved oxygen using ensemble machine learning models. *Environmental Monitoring and Assessment, 193*(4), 1–13.
[19]. He, C., Liu, Q., & Zhang, Y. (2023). Integration of nutrient and meteorological covariates for improved lake water quality prediction. *Journal of Hydrology, 617,* 128932.
[20]. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
[21]. International Journal of Computer Science and Mobile Computing (IJCSMC). (2023). End-to-end ML frameworks for river water monitoring, *12*(4), 120–129.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141073

[22]. International Journal of Computer Science and Mobile Computing (IJCSMC). (2023). Sensor-based data-driven ML frameworks for watershed monitoring, *12*(5), 57–68.
[23]. International Water Association Publishing (IWAP). (2023). Automated frameworks for real-time water quality forecasting. *Water Science and Technology, 87*(2), 431–445.
[24]. IJATEM. (2022). Automated retraining pipelines for real-time water quality prediction, *11*(5), 45–52.
[25]. JESPublication. (2024). Operational machine learning pipelines for water quality management. *Journal of Emerging Science, 14*(2), 102–118.
[26]. Ly, T. N., et al. (2023). Adaptive neuro-fuzzy inference system for algal bloom prediction. *Ecological Informatics, 73,* 101886.
[27]. Mohammadpour, P. (2022). Uncertainty-aware ensemble learning for hydrological prediction under variable conditions. *Hydrology and Earth System Sciences, 26*(7), 1901–1920.
[28]. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.
[29]. Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.
[30]. Prabu, S., Kumar, A., & Patel, V. (2021). Comparative study of ensemble and deep learning models for surface water prediction. *Environmental Research, 198,* 111227.
[31]. ResearchGate. (2022). Machine learning ensemble models for anomaly detection in river water datasets. *Preprint.*
[32]. ResearchGate. (2023). Data augmentation methods for rare contamination events in water quality prediction. *Preprint.*
[33]. Sargaonkar, A., & Deshpande, V. (2003). Development of an overall index of pollution for surface water based on a general classification scheme in Indian context. *Environmental Monitoring and Assessment, 89*(1), 43–67.
[34]. ScienceDirect. (2023). Physics-informed ensemble models for lake stratification prediction. *Journal of Hydrological Sciences, 78*(3), 251–263.*
[35]. SCITEPRESS. (2023). Stacked ensemble learning for environmental data prediction. In *Proceedings of the International Conference on Environmental Informatics.*
[36]. Springer. (2023). Hybrid deep learning–ensemble modeling for heterogeneous water datasets. *Environmental Systems Research, 12*(1), 54–68.
[37]. Sutadian, A. D., Muttil, N., Yilmaz, A. G., & Perera, B. J. C. (2016). Development of river water quality indices—A review. *Environmental Monitoring and Assessment, 188*(1), 58.
[38]. Tellus Journals. (2022). Hybrid modeling of surface water quality using stacked ensembles. *Tellus B, 74*(1), 234–249.
[39]. Tyagi, S., Sharma, B., Singh, P., & Dobhal, R. (2013). Water quality assessment in terms of water quality index. *American Journal of Water Resources, 1*(3), 34–38.
[40]. Zhu, L., Wang, J., & Zhao, Y. (2020). Feature redundancy and temporal validation in machine learning-based water quality prediction. *Environmental Modelling & Software, 134,* 104868.
[41]. Zhu, L., et al. (2020). Modeling dissolved oxygen in St. John's River using feature-driven regression and correlation analysis. *Water, 12*(6), 1598.