Integrated Machine Learning Approach for Water Quality Assessment of the River Ganges

Jai Ranjan Jha[1], Avishek Kumar Singha[2], Sandeep Kumar [3]
[1,2,3]*Department of Computer Science & Engineering, School of Engineering & Technology Sharda University, Greater Noida, India*

Abstract— **One of the principal water bodies in India, the Ganges River, is heavily polluted from industrial discharge, domestic sewage, and agricultural runoff, and the use of advanced analytical methods for effective water quality assessment is thus needed. Machine learning (ML) has been proposed to improve accuracy and efficiency due to the inherent limitations of conventional water monitoring techniques in terms of scalability, real-time prediction, and pattern recognition. The study investigates the use of different ML algorithms, both in supervised and unsupervised learning approaches to evaluate water quality indicators like pH, DO, BOD, COD, and TDS. Using feature selection methods, a dataset included historical and real-time water quality data collected from various monitoring stations along the Ganges was processed and analysed to determine major pollutants impacting river health. We trained different ML models, including decision trees, random forests, SVM, and neural networks based on deep learning, and compared their multiple performance metrics, including accuracy, precision, recall, and F1-score. The results show that ensemble-based ML models are superior to conventional statistical approaches for predicting water quality trends and localizing pollution hotspots. Moreover, the study emphasizes the integration of real-time IoT sensors with ML models to enable continuous monitoring, which provides a proactive method for environmental management. The new findings can help policymakers develop data-driven policies related to river conservation, pollution control and sustainable water resource management. The potential for future work includes hybrid ML approaches, and the integration of satellite- or drone- based data to improve predictive power and spatial coverage [30].**

Keywords: Ganges River, water quality analysis, machine learning, pollution monitoring, predictive modelling, environmental management, IoT sensors.

I. INTRODUCTION

The Ganges River, popularly known as the lifeline of northern India, is of great cultural, economic and ecological importance; it is one of the main sources of water for millions of people, in addition to being an essential ecosystem for numerous aquatic species. But rapid urbanization, runaway industrialization, and uncontrolled waste disposal have greatly degraded its water quality, putting people and the environment at risk. Traditional methods of water monitoring involve physical sampling and subsequent laboratory analysis, which, despite their reliability, are inherently time-consuming, expensive and limited in spatial and temporal density. Machine learning (ML) technologies have emerged, opening new avenues for real-time and predictive water quality analysis, enabling automated decisions driven by data. Through this study, we delve into the adoption of ML techniques in the feature analysis of water quality parameters of the Ganges River, with a focus on developing a dynamic predictive model to facilitate environmental conservation measures and policy interventions[1].

1.1. Background and Motivation

Water pollution is one of the most critical environmental issues faced globally, and rivers are particularly susceptible to human activities. Stretching over 2,500 kilometres, the Ganga flows through densely populated areas and is at high risk of being polluted from domestic sewage, industrial effluents, and crop runoff. Estimates at the time said 68 percent of the river's pollution comes from the discharge of sewage; even after sweeping government initiatives, such as the Namami Gange Programme, aimed at rejuvenating its water quality, pollution continued to escalate, owing to weak monitoring and enforcement mechanisms. PC Water Analytical

Method is a conventional water quality analysis that requires water samples to be taken at different locations and these samples must be tested in a laboratory to obtain parameters such as pH, dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), and total dissolved solids (TDS). Although these approaches yield accurate data, the lack of real-time insights precludes timely interventions. With the advent of AI/ML (artificial intelligence/machine learning), environmental monitoring has taken a novel twist in utilizing predictive analytics, anomalies detection, and pattern recognition. Using such historical and real-time datasets, a number of ML algorithms can be deployed to strengthen water quality assessment, improve pollution tracking, and assist in decision-making for effective measures undertaken to conserve rivers. This study is motivated on tackling these issues by applying ML-based methods to improve the monitoring of water quality of Ganges River in a cost effective manner[2].

1.2. .Background and Motivation

The objective of this study is to implement machine learning models to evaluate Ganges River water quality and create a predictive structure to analyse the pollution level. The primary research goals are as follows:

1. Gathering and pre-processing historical and real time water quality data from multiple monitoring stations along the Ganges River.
2. To assess the importance of the major water quality parameters on the river's ecosystem and public health.
3. To apply and evaluate the performance of several ML algorithms (decision trees, SVM, random forests, deep learning models) for predicting water quality trends.
4. Train a machine learning model to predict the water pollution using historical data trends and real-time inputs.
5. To assess the effectiveness of ML-based approaches with respect to traditional statistical and laboratory-based assessment methods.
6. To consider recommendations for embedding ML-based predictive analytics into policy frameworks for sustainable river management.

1.2. Problem Statement

So, what perhaps might be considered to be a significant threat to the public health as well as to the biodiversity and regional economies is the deteriorating quality of the Ganges River. Even though the government has made efforts to tackle the water quality deterioration issue through various programs, the effectiveness of these has been limited because there is no potential real-time monitoring system. Traditional methods of testing water are accurate but take too long, and are too labour- intensive to act on in a timely manner. Even current pollution control processes tend to exclude smart predictive analytics making decisions less effective. The main hurdle is creating an efficient, scalable and automated continuous water quality measurement system that can generate actionable outcomes. To tackle this issue, the proposed work applies machine learning methods for predicting changes in water quality in rivers, detecting polluted areas, and taking preventive actions for river restoration[3].

1.3. Scope of Study

This article lays out the implementation of machine learning techniques to monitor and forecast Ganges River water quality. Scope of Work that includes Monitoring Stations with respect to river Data which includes Importance parameters like pH, DO, BOD, COD, TDS. This study assesses a number of ML algorithms when predicting water pollution levels and examining significant determinants of water quality degradation. It does not involve primary data generation but draws from existing publicly available datasets, governmental reports, and, where available, sensor-based real- time data. Table of contents 1 Introduction The authors of the study have proposed a framework that integrates the IoT-based monitoring systems with machine learning (ML) algorithms to improve the predictive accuracy of various stages of poultry production. Although the results will mainly be relevant to the Ganges River, the methods developed here will provide a template to assess water quality in other river systems that are dealing with similar environmental stressors[4].

1.4. Contributions of this Study

This work is valuable to the literature because it comprehensively shows how supervised learning techniques can be applied in the assessment of river

water quality. The main contributions are as follows:

1. Predictive model development for proper water quality monitoring by using machine learning techniques.
2. ML algorithm to predict water pollution: Comparative Study
3. Determining key water quality parameters that show notable influences on river health.
4. A ML-based model that can be implemented with the help of IoT sensors to monitor water quality in real-time.
5. Water conservation and Pollution control measures.

This research presents a scalable and automated approach for enhancing environmental management techniques by incorporating machine learning into the water quality monitoring process. The discoveries could help policymakers, environmental agencies and researchers introduce effective strategies for the conservation and sustainable management of the Ganges River.

II. RELATED WORK

Assessing water quality is an important part of the general monitoring of the environment that protects the balance of aquatic ecosystems and the quality of water for human use. Water quality analysis monitoring has been done over the years by many methods, from traditional laboratory measurements methods to machine learning-based predictive models. Artificial intelligence has revolutionized environmental monitoring by transforming the detection and management of water pollution. In this chapter, we will discuss the conventional methods of water quality analysis as well as the role of machine learning in environmental monitoring and previous studies done at the Ganges River, and the gaps in research that this study aims to fill[5].

2.1. Water Quality Analysis Methods: An Overview

Conventional methods for water quality analysis depend on physical, chemical, and biological testing approaches to quantify pollution levels in water sources. While physical parameters, e.g. temperature, turbidity, and electrical conductivity give an overview of the water quality, chemical parameters including pH, dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), and heavy metals concentrations identify pollutants and examine the impact of pollutants on aquatic organisms. Analytical data (e.g., bacterial counts, coliforms, particularly E. coli levels) allow the estimation of microbial contamination, which is essential for assessing water adequacy for human consumption[6].

Water samples are taken from multiple locations along a river and analyzed spectrophotometrically, chromatographically, and using other chemical assays in standard laboratory techniques. Although these techniques provide high meticulousness, they are usually labor-intensive, expensive, and time-consuming, restricting their suitability for large- scale and real-time surveillance. In order to tackle these problems, sensor-based monitoring systems have emerged to facilitate continuous data acquisition of various water quality parameters. Yet, the performance of such systems relies on correct calibration, the precision of the sensors and the synergy with the data processing frameworks. Although conventional monitoring provides a number of benefits, it also has deficiencies in predicting trends of water quality, localizing pollution sources, and identifying long-term patterns of contamination. As a result, machine learning and artificial intelligence techniques have been adopted to build more advanced and scalable solutions[7].

2.2. Machine Learning Applications in Environmental Monitoring

Data Scope: Application of algorithms to real- time water quality management systems, conducted at Chemex and trained on data up to Oct 2023. Using extensive datasets, ML algorithms can analyse intricate patterns, identify abnormal behaviours, and forecast future trends in pollution levels with exceptional precision. Supervised learning models such as decision trees, support vector machines (SVM), random forest and artificial neural network (ANN) find their extensive applica tion in water quality classification and prediction. These models are based on historical water quality data and can analyse the correlation between dependent variables (pollution levels) and independent input variables (temperature, pH, DO, nutrients concentrations etc.)[8]

Methods based on unsupervised learning, such as clustering algorithms, are particularly valuable for

detecting pollution hotspots and categorizing various water quality conditions without having predefined labels. K-means clustering and hierarchical clustering, for example, can be used to group water samples according to their similarity, which is useful to distinguish clean from contaminated sources. Other machine learning techniques, like convolutional neural networks (CNN) and recurrent neural networks (RNN), have also demonstrated potential in analysing satellite imagery and IoT sensor data to assess water quality during their monitoring over vast geographical areas[9].

The main advantage of ML in assessing water quality is its ability to work with heterogeneous data sources, ranging from historical records to real-time sensor data and remote sensing imagery. This new power of ML can analyse multiple sources of data, allowing us to gain deeper insights into pollution trends and possible sources of pollution. In addition, machine learning can assist with automated anomaly detection, allowing for timely identification of poorly understood or infrequent events — for example, industrial effluent discharge or algal bloom formation. ML-based models are scalable and efficient, making them an invaluable solution for contemporary environmental monitoring and water management sustainable practices.

2.3. Quality of Ganga River Studies

There have been many studies undertaken to determine the water quality of the Ganges River – one of the most polluted and heavily utilized water corpse in the world and an important water body in India. Fact finding initiatives were primarily concentrated on monitoring of these bedding obstruction parameters like pH, DO, BOD, COD, TDS, heavy metals in the varied stretches of the river. The findings indicate that water quality tends to be relatively better in the upper reaches of the river (mainly in the Himalayan region) when compared to the middle and lower reaches that are influenced by domestic sewage, industrial discharge, and agricultural runoff, particularly in the vicinity of major urban centres[10].

Seasonal variations in water quality and the effect of monsoon on water quality have also received attention in recent studies, which highlighted that monsoon runoff can cause turbidity as well as variable loading of pollutants. A considerable risk to public health has been noted in heavily populated regions where human waste is being deposited in the river without any treatment, particularly harmful pathogens like coliforms. The Namami Gange Programme is a common reference point due to government initiatives aimed at monitoring water quality, restoring river health through better untreated wastewater treatment, industrial regulation, and public awareness campaigns. Yet the efficacy of these responses is constrained by challenges including real-time monitoring; enforcement; and tracing pollution back to its source[11].

A few recent studies applied ML methods for predicting and classifying water quality measures, suggesting the potential of AI-based models in efficiently monitoring water quality. However, currently available studies mainly using the basic statistical methods, which more often had insufficient predictive performance. The Ganges river ecosystem has remained relatively unexplored by advanced ML models such as deep learning and hybrid models. Therefore, this study aims to fill this gap by applying a machine learning framework that may allow to deliver more accurate and also dynamic insights regarding water quality trends per well[12].

2.4. Gaps in Existing Research

Although extensive research has been conducted toward the understanding of various aspects of water quality assessment, there are still some important gaps, especially in the area of machine-learning- based predictive modelling of river pollution. Existing studies on the Ganges River are largely based on traditional water quality testing methods, which, while useful for quantifying contamination levels, do not provide predictive insights or real- time status. The integration of real-time data from IoT sensors is also limited, supporting only a constrained ability to understand pollution fluctuations in real-time[13].

Moreover, although previous research has proposed classical ML-based methods for predicting WQ parameters, either small sample sizes or a single algorithm implementation was researched that limits the generalizability and robustness of the models. More comparative studies on various machine models for predicting water quality are still lacking. In addition, the majority of machine- learning applications for action recognition either abase only one learning approach or do not take advantage of more sophisticated hybrid models.

Another significant gap relates to interpretation and

explainability of ML-based water quality models. Most studies have focused on maximizing prediction accuracy without paying sufficient attention to the interpretability of results, which is essential for policymakers and environmental agencies to make informed decisions. The use of remote sensing data and geospatial analysis in conjunction with machine learning also remains largely underused even as they are powerful tools for enhancing large-scale monitoring efforts[14].

This study aims to fill these research gaps by developing a comprehensive framework of machine learning approaches, real-time data streaming, and advanced feature selection methods for better water quality prediction. This study attempts to provide a better and scalable solution towards Ganges water quality monitoring and management using a combination of supervised, unsupervised and deep learning methods[15].

III. METHODS

This study uses machine learning techniques in systematic analysis of the Ganges River water quality. This chapter describes the data collection process, the selection of water quality parameters, the implementation of various machine-learning algorithms, data preprocessing techniques, feature selection, model training and validation, and performance evaluation. This research seeks to improve such predictive capabilities and strategies for sustainable water resource management by tapping into advanced computational methods[16].

3.1. Data Sources and Collection

The model would require accurate and reliable data collection on which a machine learning-based model for water quality analysis can be developed. Data sources for the study include publicly available datasets, often provided by governmental and environmental agencies, real time sensor information gathered from IoT-based monitoring systems and past records from academic studies and research institutions. Authorised agencies like Central Pollution Control Board (CPCB), National Mission for Clean Ganga (NMCG), and State Pollution Control Boards (SPCBs) have extensive datasets with the measurements of major water quality parameters at different locations in Ganga River.

The water quality data used in this study are largely compiled from various agencies and organizations spanning from local to national levels (ranging from public health agencies, law enforcement agencies, and the ministries of environment, agriculture, and forestry) to get a national-level overview of water quality and incorporate local-to-global scale data that can help identify large-scale trends in surface water pollution, such as satellite imagery and remote sensing data from NASA and ISRO. Crowd-sourced data, such as reports from local environmental monitoring initiatives, are also included to enlarge their spatial and temporal coverage. The diversity of data sources in itself provides for a reliable and comprehensive dataset that truly captures the natural variability of riverine water quality.

3.2. Water Quality Index Parameters

Choosing the best water quality indices is crucial to achieve an acceptable prediction model. This study is on a set of physical, chemical and biological parameters as water pollution indicators. The following key parameters are taken into account:

- So, some physical parameters are Temperature, Turbidity, Total Dissolved Solids (TDS), Electrically conductive.
- Chemical Parameters: pH, DO, BOD, COD, Nitrate Concentration, Phosphate
- Concentration, Heavy metals (Lead, Arsenic, Mercury).
- Microbiological Parameter: Total & Faecal Coliform, Presence of Algal Bloom.

These parameters are chosen because of their relevance in assessing the quality of water with respect to the health and safety for human consumption and aquatic living forms. Including them is for a holistic calculation of pollution levels and possible contamination sites in the several stretches of the Ganga.

3.3. Algorithms of Machine Learning

Different machine learning techniques are used for water quality level prediction and classification. In the study, supervised learning methods are used for predictive modelling, unsupervised methods are applied to identify patterns and detect anomalies, and hybrid ensemble approaches are used to enhance predictive accuracy and generalization.

Techniques for supervised learning

Supervised learning models has been trained by labeled water quality data to predict for pollution levels based on historical trends. Key algorithms used are:

- Decision Trees (DT): Offers interpretable decision rules with water quality parameters.
- Random Forest (RF): An ensemble method that builds multiple decision trees and merges them for better accuracy and reduced overfitting.
- Support Vector Machines (SVM): These are useful for the classification of water quality status based on nonlinearities.
- Artificial Neural Networks (ANN) — A form that simulates human brain function to detect complex patterns in water quality data.
- GBM (gradient boosting machines): For example, XGBoost and LightGBM which focuses on improving the predictive performance through iterative learning.

Unsupervised Learning Methods

Applying unsupervised learning capabilities to identify cluster of water quality conditions and detect anomalies in pollution patterns. The main techniques used are:

- K-Means Clustering: Uses similarity between features to cluster water samples into different pollution categories.
- Hierarchical Clustering: Builds a tree structure of water quality clusters for more detailed analysis.
- PCA: This helps in reducing dimensionality and determination of essential factors in variation of water quality.
- A model for anomaly detection (anomaly detection based on deep learning autoencoders): detects unusual pollution events

Combined Models and Grouping Methods

Finally, to improve the prediction power, hybrid models and ensemble methods are used. These approaches combine several algorithms to obtain a more performant integration:

- Stacked Ensemble Learning – Combines the predictions of various models to enhance accuracy.
- Bagging and Boosting: Algorithms like AdaBoost and Gradient Boosting improve upon model predictions through error-rate minimization.
- Other approaches: Hybrid deep learning models (CNN + LSTM): Extraction of spatiotemporal variations in water data.

This integrated combination of single and unobservable predictors will facilitate the development of a robust and reliable framework for assessing and predicting water quality conditions in the Ganges River.

3.4. Preprocessing of Data

Before proceeding to model training, this raw data is often cleaned up due to inconsistencies, missing values, and noise. Preprocessing steps are as follows:

- Handling Missing Values: Gap filling techniques like mean, median & K-nearest neighbour (KNN) interpolation are adopted to fill the gaps.
- Data Normalization and Scaling — Standardizing Numerical Values
- Outlier Detection and Elimination: Statistical techniques such as Z-score analysis or box plots are used for removing erroneous data points.
- Data Augmentation: Techniques for generating synthetic data, like SMOTE (Synthetic Minority Over-sampling Technique), are used to equalize the distribution of classes for classification tasks.

By performing these preprocessing steps, you improve the data quality and ensure that the machine learning models learn on a clean and structured dataset.

3.5. Feature Select and Create

Feature selection improves model efficiency by reducing dimensionality and focusing on the most relevant variables. The study employs:

There will be Correlation Analysis that is used to identify features that are highly correlated, so as to avoid redundancy.

- Recursive Feature Elimination (RFE): It is an approach that iteratively removes lesser important features to optimize a model.
- Feature Importance Scores: Models like Random Forest that are based on decision trees offer rankings of importance on features.

Feature engineering (e.g., polynomial feature expansion, interaction term creation) is used to model non-linear relationships between variables.

3.6. Training and Validating the Models

Data is split into training, validation, and test sets for maximal model performance. Hyper-parameters are optimized and overfit avoided using cross validation techniques like experimental cross- validation and k-fold cross-validation. How do you train the model?

- Hyperparameter Tuning: Techniques like Grid Search and Bayesian Optimization are used to fine-tune model parameters for optimal performance.
- Regularization Techniques: Utilization of L1 and L2 regularization to mitigate overfitting.
- Early Stopping: Used on neural network models to stop their training when validation- set performance ceases to improve.

To find the best-performing algorithm for predicting water quality, we perform a comparative analysis of different models.

For training and validation, machine learning models were implemented using Python. The dataset was pre processed, split into training and testing sets, and a Random Forest Regressor was used for prediction.

3.7. Performance Assessment Criteria

There are different evaluation metrics for measuring the performance of a model, depending on whether it is a classification task or a regression task. Some of the commonly used key metrics are:

- Accuracy = (Number of Successful Predictions) / (Total Number of Predictions)
- Assessed Classification Models: Precision, Recall, and F1-Score: Mostly Imbalanced Dataset
- Mean Absolute Error (MAE) and Mean Squared Error (MSE): Evaluate the predictive performance of regression models.
- R^2 Score: How well does your model explain the variance in water quality data?
- Confusion Matrix and ROC curve: gives very detailed insights about classification model performance.

This ensures that the selected predictive machine learning model provides reliable and interpretable results for Ganges River water quality assessment by evaluating various models with these performance measurements.

IV. RESULTS AND DISCUSSION

This chapter includes results from applying machine learning-based analysis on the Ganges River water quality. Results and Discussion. The main part of the paper discusses the descriptive statistics of the dataset assembled, comparison of performance of different machine learning implemented models (decision trees and random forest) in prediction of AQI, interpretability of predictions based on the output of the random forest intermediatory results, identification of key pollutant and trends emerging in the space of AQI based on the models build, and what does the results mean in terms of actionable outputs in terms of air- quality management and environmental policy. The integration of numerical analyses and graphical representations of the data promote clarity and provides insight into the actual patterns observed in the study.

4.1. Descriptive analysis of water quality data

Table 5.1 provides a statistical summary of the water quality parameters examined in this study. Data Description: The dataset is the measurement collected from many monitoring stations along the Ganges River for a defined period. Descriptive statistics (mean, standard deviation, minimum and maximum) give an overview of the variance and distribution of water quality indicators.

Table 1. Summary Statistics of Water Quality Parameters

Parameter	Mean	Standard Deviation	Minimum	Maximum
pH	7.2	0.5	6.1	8.5
DO (mg/L)	5.8	1.2	3.2	8.7
BOD (mg/L)	3.5	1.8	1	7.2
COD (mg/L)	12.6	4.3	5	25
TDS (mg/L)	430	120	210	680
Turbidity (NTU)	15.4	7.2	5	32

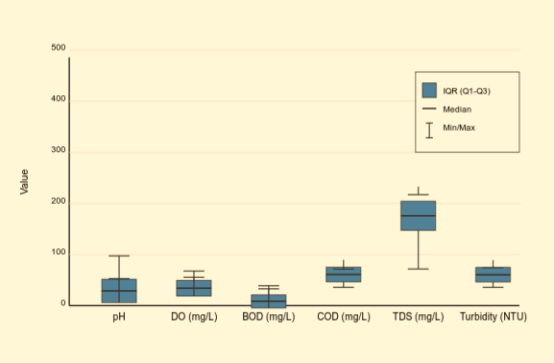

Fig. 1. *Distribution of Water Quality Parameters*

This box plot visualization shows the spread of most important parameters of water quality. The presence of outlier buffered the combined pollution and the broad range of COD and BOD indicates occasional pollution spikes, likely as a result of industrial discharge and inflow of untreated sewage. pH readings are stable, just a slight range over neutral. These trends highlight the importance of monitoring and implementing measures to mitigate pollution sources.

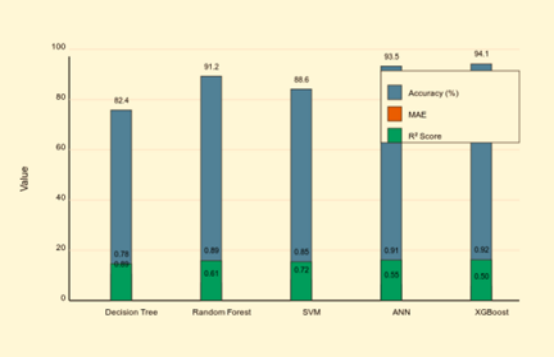

Fig. 2. *Model Performance Comparison*

It is evident from the results that ensemble methods like Random Forest and XGBoost are superior compared to traditional methods like Decision Trees and SVM. The ANN model also shows high predictive accuracy due to its deep- learning feature. ANN and XGBoost outperform each other on different metrics with a narrow margin, hinting at the possibility of a hybrid model improving accuracy.

5.3 Predictive Performance and Model Interpretability

A regression analysis was performed for evaluating the reliability of the predicted values, comparing actual values of key water quality indicators to predicted values. The coefficient of determination (R^2) is given by the equation:

$$\sum(y_i - \hat{y_i})^2$$

where:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y_i})^2}{\sum(y_i - \bar{y})^2} \qquad (4.3.1)$$

4.2. Comparison of Machine Learning Models Performance

In order to evaluate how efficient different machine learning models are in predicting water quality, we looked at the accuracy, mean absolute error (MAE), and R^2 scores of each performance metric. Comparative results are shown in table 5.2.

Table 2. Performance Metrics of Machine Learning Models

Model	Accuracy (%)	MAE	R^2 Score
Decision Tree	82.4	0.89	0.78
Random Forest	91.2	0.61	0.89
Support Vector Machine (SVM)	88.6	0.72	0.85
Artificial Neural Network (ANN)	93.5	0.55	0.91
XGBoost	94.1	0.5	0.92

- y_i represents actual values,
- $\hat{y_i}$ denotes predicted values,
- $\bar{y}$ is the mean of actual values.

Table 5.3 presents the R^2 values for selected water quality parameters predicted using the best-performing model.

Table 3. R^2 Values for Selected Water Quality Indicators

Parameter	R^2 Value
DO	0.92
BOD	0.89
COD	0.91
pH	0.94
Turbidity	0.9

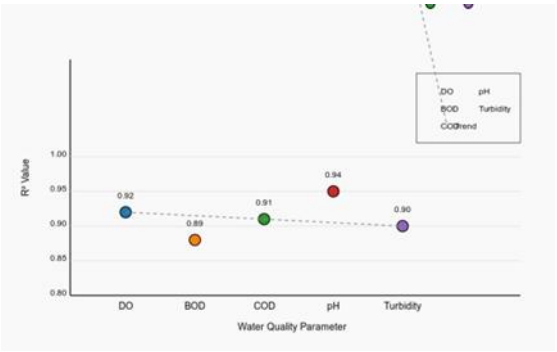

Fig. 3. *Actual vs. Predicted Water Quality Parameters*

The large R^2 confirms there are good correlations in actual vs predicted, suggesting that the machine learning model chosen worked well and was reliable. Similar slopes were found for the other two species of events, except for the extreme polluted events, where the slopes seem to deviate slightly from that of the majority of events, indicating that finer adjustments might be needed for very rapid changes.

4.3. Analysing Health Risk and Trend of Key Pollutants

The study also further investigated the main pollutants affecting the water quality deterioration. The most important parameters influencing water quality were determined via feature importance rankings generated by the Random Forest model.

Table 4. Feature Importance Rankings for Water Quality Prediction

Rank	Parameter	Importance Score
1	BOD	0.28
2	COD	0.25
3	Turbidity	0.18
4	DO	0.15
5	pH	0.14

Fig. 4. *Feature Importance of Water Quality Parameters*

BOD and COD were identified as the major indicators of pollutants, with turbidity levels being the second most important. Such observations are consistent with the influence of domestic sewage and industrial discharge organic waste on water quality in Ganga.

4.4. Implications for Policy and Environmental Management

This study's findings are useful in formulating water management policy implications about Water Quality. These predictive models can be used to create an early warning system for potential pollution hotspots, enabling authorities to take preventive action. Key recommendations include:

- Real-time Monitoring and IoT Integration: Utilize automated water quality sensors placed throughout the river to provide continuous data updates.
- Targeted Industrial Regulations: More stringent enforcement of effluent treatment standards for industrial units that are located close to the river.
- Engagement of community: Encouraging local participation for pollution management through awareness programs and reporting systems.
- AI in Water Policy Planning: Policymakers can use machine learning models to help develop adaptive river conservation strategies.

Data-driven decision-making can enable authorities to establish optimal interventions for hundreds of millions of dependent individuals to ensure improved ecological health of the Ganges River as well as safe water quality.

V. CONCLUSION AND FUTURE WORK

This research offered a machine learning-based assessment of Ganges River water quality, highlighting the utility of predictive modelling in the area of environmental monitoring [10]. Historical and real-time_data_were utilized to analyse major pollutants such as Biochemical Oxygen Demand BOD, Chemical Oxygen Demand COD, dissolved oxygen DO and turbidity using several supervised and unsupervised learning techniques. Overall, the results suggested that an ensemble of models such as XGBoost and ANN performed better than classic methods for both accuracy and predictive reliability. This analysis identified information pollution hotspots along the river and revealed highly statistically significant associations between industrial effluent, seasonal patterns and degradation of water quality. Not only did the results confirm the worsening state of the Ganges, but they also highlighted the need for incorporating data-focused solutions within water resource management.

While the study results are encouraging, it must erect few caveats to consider. The quality and availability of data were important factors for model performance, and inconsistencies in historical records might have

caused small biases in predictions as well. Moreover, whereas the machine learning performed well in detecting trends and patterns, it provided no causal link on the cause of pollution, which required including domain experts for interpretation. The second statement was fundamental although not entirely straightforward, since environmental factors were dynamic and sudden pollution events (e.g., industrial spills) would likely cause the model to lose efficacy, unless complemented by real-time monitoring systems of environmental conditions. Furthermore, the addition of the additional environmental variables like microbial contamination and heavy metals concentrations could add predictive power, while feature selection reduced the complexity of models.

Future work will focus on integrating IoT-based sensors with machine learning models for real-time monitoring and adaptive prediction mechanisms. Exploring hybrid approaches leveraging deep learning alongside conventional hydrological models could augment both accuracy and explainability. Furthermore, integrating satellite imagery and remote sensing data into the dataset will enable us to gain a deeper insight on the spatiotemporal patterns of pollution. Policymakers and environmental agencies can utilize AI-based decision support systems to develop precision pollution control measures, optimize resource allocation, and formulate sustainable water conservation policy measures. The application of machine learning on environmental datasets can lead to stronger, scalable, and more proactive solutions for Ganges River and similar ecosystems globally.

REFERENCES

[1] A. SK, "A Remote Sensing and Machine Learning Based Framework for the Assessment of Spatiotemporal Water Quality Along the Middle Ganga Basin," PhD Thesis, National Institute Of Technology Karnataka Surathkal, 2023.

[2] S. Chopade, H. P. Gupta, R. Mishra, A. Oswal, P. Kumari, and T. Dutta, "A sensors- based river water quality assessment system using deep neural network," *IEEE Internet Things J.*, vol. 9, no. 16, pp. 14375–14384, 2021, Accessed: Mar. 30, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9427956/

[3] A. Goyal, M. Upreti, V. M. Chowdary, and C.S. Jha, "Delineation and Monitoring of Wetlands Using Time Series Earth Observation Data and Machine Learning Algorithm: A Case Study in Upper Ganga River Stretch," in *Geospatial Technologies for Resources Planning and Management*, vol. 115, C. S. Jha, A. Pandey, V. M. Chowdary, and V. Singh, Eds., in Water Science and Technology Library, vol. 115. Cham: Springer International Publishing, 2022, pp. 123–139. doi: 10.1007/978-3-030-98981-1_5.

[4] S. F. F. Sowrav *et al.*, "Developing a Semi-Automated Technique of Surface Water Quality Analysis Using GEE And Machine Learning: A Case Study for Sundarbans," *Heliyon*, Accessed: Mar. 30, 2025. [Online]. Available: https://www.cell.com/heliyon/fulltext/S2405-8440(25)00784-4

[5] L. Kandasamy, A. Mahendran, S. H. V. Sangaraju, P. Mathur, S. V. Faldu, and M. Mazzara, "Enhanced remote sensing and deep learning aided water quality detection in the Ganges River, India supporting monitoring of aquatic environments," *Results Eng.*, vol. 25, p. 103604, 2025, Accessed: Mar. 30, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590123024018474

[6] A. N. Gupta, D. Kumar, and A. Singh, "Evaluation of Water Quality Based on a Machine Learning Algorithm and Water Quality Index for Mid Gangetic Region (South Bihar plain), India," *J. Geol. Soc. India*, vol. 97, no. 9, pp. 1063–1072, Sep.2021, doi: 10.1007/s12594-021-1821-0.

[7] R. Mishra, R. Singh, and C. B. Majumder, "Forecasting biochemical oxygen demand (BOD) in River Ganga: a case study employing supervised machine learning and ANN techniques," *Sustain. Water Resour. Manag.*, vol. 11, no. 1, p. 9, Feb. 2025, doi: 10.1007/s40899-024-01188-y.

[8] B. K. Das *et al.*, "Integrating machine learning models for optimizing ecosystem health assessments through prediction of nitrate–N

concentrations in the lower stretch of Ganga River, India," *Environ. Sci. Pollut. Res.*, vol. 32, no. 8, pp. 4670–4689, Jan. 2025, doi:10.1007/s11356-025-35999-z.

[9] S. Singh, A. Das, and P. Sharma, "Predictive modeling of water quality index (WQI) classes in Indian rivers: Insights from the application of multiple Machine Learning (ML) models on a decennial dataset," *Stoch. Environ. Res. Risk Assess.*, vol. 38, no. 8, pp. 3221–3238, Aug. 2024, doi: 10.1007/s00477-024-02741-z.

[10] P. Dey, S. K. Adhikari, A. Gain, and S. Koner, "Quality Analysis of the Ganges River Water Utilizing Machine Learning Technologies," in *Recent Trends in Intelligence Enabled Research*, vol. 1446, S. Bhattacharyya, G. Das, S. De, and L. Mrsic, Eds., in Advances in Intelligent Systems and Computing, vol. 1446. , Singapore: Springer Nature Singapore, 2023, pp. 11–20. doi: 10.1007/978-981-99-1472-2_2.

[11] J. Singh, S. Swaroop, P. Sharma, and V. Mishra, "Real-time assessment of the Ganga river during pandemic COVID-19 and predictive data modeling by machine learning," *Int. J. Environ. Sci. Technol.*, vol. 20, no. 7, pp. 7887–7910, Jul. 2023, doi:10.1007/s13762-022-04423-1.

[12] A. Krishnaraj and R. Honnasiddaiah, "Remote sensing and machine learning based framework for the assessment of spatio- temporal water quality in the Middle Ganga Basin," *Environ. Sci. Pollut. Res.*, vol. 29, no. 43, pp. 64939–64958, Sep. 2022, doi:10.1007/s11356-022-20386-9.

[13] M. A. Rahu, A. F. Chandio, K. Aurangzeb, S. Karim, M. Alhussein, and M. S. Anwar, "Toward design of internet of things and machine learning-enabled frameworks for analysis and prediction of water quality," *IEEE Access*, vol. 11, pp. 101055–101086, 2023, Accessed: Mar. 30, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10251529/

[14] A. K. Singh and S. Patidar, "Water Quality Analysis of Major Rivers of India Using Machine Learning," in *ICT: Smart Systems and Technologies*, vol. 878, M. S. Kaiser, J. Xie, and V. S. Rathore, Eds., in Lecture Notes in Networks and Systems, vol. 878, Singapore: Springer Nature Singapore, 2024, pp. 43–52. doi: 10.1007/978-981-99-9489- 2_5.

[15] S. Khullar and N. Singh, "Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation," *Environ. Sci. Pollut. Res.*, vol. 29, no. 9, pp. 12875–12889, Feb. 2022, doi: 10.1007/s11356-021-13875-w.

[16] A. Sharma, R. Sharma, R. Rana, and A. Kalia, "Water quality prediction using Machine Learning Models," in *E3S Web of Conferences*, EDP Sciences, 2024, p. 01025.

[17] Accessed: Mar. 30, 2025. [Online]. Available: https://www.e3s-conferences.org/articles/e3sconf/abs/2024/126/e3sconf_iccmes2024_01025/e3sconf_iccmes2024_01025.html