Predictive Modeling of Water Quality in Indian Rivers: A Machine Learning Approach for Sustainable Resource Management

Bela Shrimali[1][a], Shivangi Surati[2][b], Aditya Patel[1][c] and Rohit Kansagara[1][d]
[1]*Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India*
[2]*Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India*

Keywords: Machine Learning, Water Pollution, Water Quality Predictions, Indian Rivers

Abstract: Despite water being an essential constituent of life, water pollution is increased because of sewage, pesticides, and industrial waste. Polluted water creates a negative influence on the ecosystem, affecting not only human life but also aquatic life. River water pollution is one of the major concerns of recent days in emerging countries like India, Bhutan, Bangladesh, and many more. Hence, river water quality prediction becomes essential for sustainable resource management. In this paper, after describing various parameters to monitor water quality, an innovative Machine Learning (ML)-driven approach for prediction of the water quality of Indian rivers, is presented. The research involves the implementation of various machine learning models to predict diverse water quality parameters of the Indian rivers. These models are trained to address the intricate challenges associated with comprehending the complex dynamics of water quality. The efficacy of the trained models is experimented through evaluations of a huge dataset, comprising water samples from various Indian rivers. The outcomes of this research not only predict and monitor the accuracy of water quality through a robust framework but also contribute valuable insights and tools for sustainable resource management for Indian rivers.

1 INTRODUCTION

Water is an essential need in our lives, serving as a vital resource for drinking, industrial processes, and agriculture. Water of superior quality not only decreases the costs related to treatment but also enhances agricultural productivity. Nevertheless, the increasing need for water, influenced by factors such as population growth, changing agricultural methods, urban sprawl, and industrial advancement, presents a significant challenge. Water can become unfit for consumption, irrigation, and other uses due to both human activities and natural pollution. Hence, it is crucial to consistently evaluate and predict the quality of water to guarantee its appropriateness for particular purposes and implement necessary measures if standards are not met. Conventional practice involves examining numerous water quality parameters to measure the amount of dissolved substances. However, in developing nations like India, monitoring all such parameters together in a groundwater set-up or rivers is difficult and costly. Minimizing subjectivity and costs related to water quality evaluation is a major challenge. In recent times, numerous national and international organizations have suggested and created Water Quality Indexes (WQIs) in response to this realization. Prominent instances comprise the US National Sanitation Foundation WQI, Florida Stream WQI, British Columbia WQI, Canadian WQI, and Oregon WQI. These indices effectively evaluate the appropriateness of water for drinking.

In developing countries like India, agriculture is a major source of jobs and economic growth. It is also the biggest user of water, accounting for up to 80% of water consumption, and a significant source of pollution in water. As a consequence, effective and affordable planning and managing water is essential for sustainable agriculture. The important parameters for assessing water quality in Indian rivers are temperature, potential for Hydrogen (pH), B.O.D. (Biochemical Oxygen Demand) in mg/l, D.O. (Dissolved Oxygen) in mg/l, conductivity, NITRATE-

[a] https://orcid.org/0000-0002-7543-5389
[b] https://orcid.org/0000-0003-4381-5130
[c] https://orcid.org/0009-0005-9026-2083
[d] https://orcid.org/0009-0005-8843-283X

Shrimali, B., Surati, S., Patel, A. and Kansagara, R.
Predictive Modeling of Water Quality in Indian Rivers: A Machine Learning Approach for Sustainable Resource Management.
DOI: 10.5220/0013303800004646

In *Proceedings of the 1st International Conference on Cognitive & Cloud Computing (IC3Com 2024)*, pages 165-174
ISBN: 978-989-758-739-9

NAN N+ NITRITENANN in mg/l, FECAL coliform (MPN/100ml), and TOTAL coliform (MPN/100ml). Considering these characteristics collectively paints a comprehensive picture of the river's water quality, covering aspects of its physical, chemical, and microbiological composition. A thorough understanding of these elements is crucial not only for keeping tabs on the environment but also for assessing public health and making well-informed decisions regarding the management of water resources. It's essentially about having a holistic grasp of the river's health to ensure the decisions made are of benefit to both the eco-system and the communities relying on it.

To accomplish this decision-making, numerous machine learning algorithms are trained in the existing literature to predict the quality of water based on its parameters (Ewuzie et al., 2022; Ibrahim et al., 2023; Khoi et al., 2022; Sakaa et al., 2022; Sakizadeh and Mirzaei, 2016; Singh et al., 2018; Tyagi et al., 2013). However, majority of the available methods for water quality predictions do not aim precise prediction for Indian rivers. In addition to that, available implementations cover limited water parameters to predict the quality of the water that is not robust enough to provide more accurate results in prediction. Hence, our contributions in this paper are as follows:

- Various parameters of rivers viz. temperature, pH level, electrical conductivity, Biochemical Oxygen Demand and Dissolved Oxygen that differ from region to region are described in the study literature.
- An innovative ML-driven approach for predictive modeling of the water quality of Indian rivers, with a primary focus on sustainable resource management, is presented. The research involves the implementation of machine learning models such as the Decision Tree , Logistic Regression, Random Forest (RF) classifier, K-Nearest Neighbor (KNN), Support Vector Classifier (SVC), Ada-boost classifier, Long Short-Term Memory (LSTM), and XGBoost (XGB) classifier to predict diverse water quality parameters of Indian rivers.
- These models are trained to address the intricate challenges associated with comprehending the complex dynamics of water quality.
- Through meticulous evaluations of a huge dataset, comprising water samples from various Indian rivers, the effectiveness of the proposed models is demonstrated by providing detailed and insightful assessments of water quality.
- Thus, the outcomes of this research contribute valuable insights and tools for sustainable resource management, presenting a robust framework for predicting and monitoring water quality in the context of Indian rivers.

The remaining paper is organized as follows: The existing literature is covered in section 2 as related work. The study on water quality parameters is explored in section 3. Machine learning models, datasets and data pre-processing methodology are presented in Section 4 and Section 5 respectively. Subsequently, the associated experimental results, discussions based on results, and conclusion are discussed in Sections 6 and 7, respectively.

2 RELATED WORK

Water pollution is a serious threat for the future, hence, there is a need to find different ways to manage water more effectively and affordably. Scientists and researchers have developed several tools for assessing water quality for irrigation purposes, including statistical-based approaches. These tools are useful, but they can be expensive and time-consuming to use. Therefore, model-based prediction tools are also being developed that can be used to assess water quality more quickly and cheaply. This could be a valuable tool for farmers, helping them to optimize their use of water resources.

The Irrigation Water Quality Guide (IWQG) software is developed in (Ewaid et al., 2019) that is based on the Food and Agriculture Organization (FAO) guidelines and work proposed in (Meireles et al., 2010). Al-Gharraf Canal (the southern region of Iraq) dataset of each month, for years 2013-15 (three years) is utilized to estimate the guide. While these models are very efficient tools for assessing the water quality index, they require a larger number of parameters and studies to be developed, that can be costly and time-consuming. Therefore, machine learning model can be utilized for water quality prediction that is important for agriculture, especially in developing countries like India. Two ML based models (Adaptive Neuro-Fuzzy Inference System (ANFIS) and SVM) are explored in (Ibrahim et al., 2023) for prediction of eight different irrigation water quality indices. The dry regions of El Kharga Oasis were selected for a case study and to apply machine learning algorithms. Similarly, twelve ML models (four based on ANN, three based on decision tree and five based on boosting) were experimented in (Khoi et al., 2022) to estimate the quality of surface water of La Buong River, Vietnam. Extreme gradient boosting performed outstanding by achieving the maximum accuracy that was useful in improved management of water quality.

A hybrid AI (Artificial Intelligence) model Sequential Minimal Optimization- Support Vector Machine (SMO-SVM) in addition to RF is constructed in (Sakaa et al., 2022). The aim was to predict water quality merits at the Wadi Saf-Saf river basin in Algeria. They utilized fifteen input parameters/datasets of water quality for developing and evaluating the predictive models. The RF model outperformed in predicting the quality index of water as compared to SMO-SVM model in this study.

Thus, ML models are proven to be successful in prediction of water quality across different regions of multiple countries.

For Indian rivers, Singh et al. (Singh et al., 2018) used Saaty's Analytic Hierarchy Process (SAHP) to develop a model for assessing water quality suitability, and showed that WQI is useful for irrigation water managers. In addition, a Logistic Regression model was developed to analyze the water quality of rivers in India in (Sharma et al., 2020), specifically for drinking reasons. This model used the water quality index and Multiple Linear Regression (MLR), focusing on five essential parameters that are provided through the dataset: temperature, pH, B.O.D. (mg/l), D.O. (mg/l), and conductivity. A parallel study intended to assess and map groundwater suitability using a variety of models such as Decision Tree, Random Forest Classifier, KNN, SVC, Adaboost Classifier, LSTM, and XGB Classifier.

In (Modaresi and Araghinejad, 2014), many machine learning algorithms viz. logistic regression, decision tree, random forest Classifier, KNN, SVC, Adaboost classifier, LSTM, and XGB classifier are trained on a combination of water quality parameters to predict the water quality. An Artificial Neural Network (ANN) has been successfully used to predict the appropriateness of groundwater for irrigation in India, incorporating factors such as temperature, pH, B.O.D. (mg/l), D.O. (mg/l), conductivity, NITRATENAN N+ NITRITENANN (mg/l), FECAL coliform (MPN/100ml), and TOTAL coliform (MPN/100ml).

Based on the study of the existing literature, there is a scope of improvement in predicting the water quality of Indian rivers by exploring ML algorithms on the maximum number of water quality parameters. These parameters are discussed in the next section.

3 WATER QUALITY PARAMETERS

Key thresholds for essential water quality parameters, pivotal in gauging the appropriateness of water for various uses are presented in this section.

3.1 Temperature

Temperature, which shows the degree of warmth within water, has a huge impact on many aspects of aquatic ecosystems. Aquatic organisms are significantly impacted by varying temperatures in the aquatic ecosystem that has different preferences for definite temperature ranges. Water temperature is a significant metric for regulating water quality and it is measured in degrees Celsius (°C) or Fahrenheit (°F).

3.2 pH

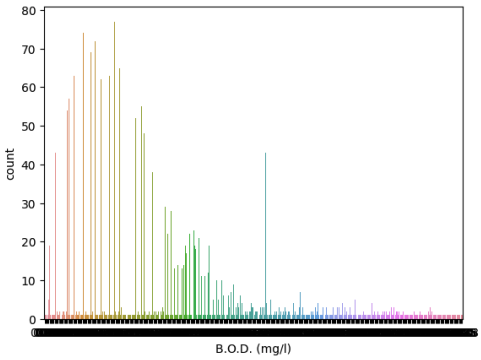
The concentration of hydrogen ions in water (represented by pH) is a vital parameter for determining its acidity or alkalinity. The dataset considered for Indian rivers and Aquatic organisms is adjustable to precise pH ranges, escalating their importance in water quality assessment. The pH scale of water, which ranges from 0 to 14, classifies values less than 7 as acidic in water, 7 as neutral, and 7 or higher as alkaline in water. This logarithmic scale has a significant impact on the solubility of various substances as well as microbial activity. pH is an important parameter because it provides insight into the chemical dynamics of water, guiding assessments and interventions to maintain optimal aquatic conditions.

3.3 Biochemical Oxygen Demand (B. O. D.)

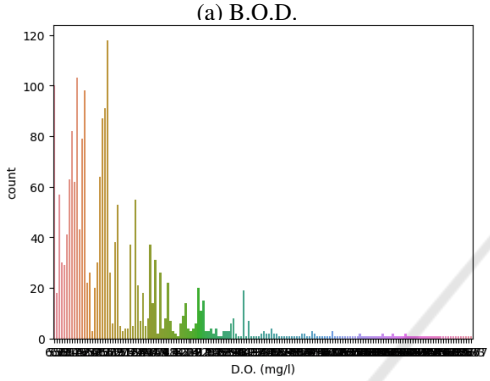
B.O.D. refers to the quantity of dissolved oxygen used by microorganisms while the organic substances are decomposed in water and in the aquatic ecosystem. This metric, which is typically measured over five days, quantifies the milligrams of oxygen consumed per liter of water (mg/l) as shown in Figure 1a. Elevated B.O.D. levels indicate increased organic pollution, indicating a greater risk of oxygen depletion and the potential can have a high impact on the Aquatic ecosystem.

3.4 Dissolved Oxygen (D. O.)

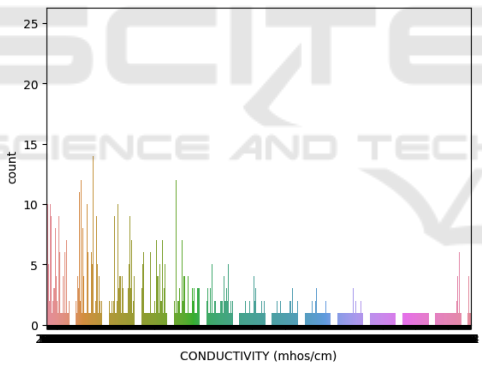
The term Dissolved Oxygen is a critical amount of the oxygen content of water, and it plays an important role in helping the aerobic organisms within aquatic ecosystems. Aquatic organisms, including fish and invertebrates, require adequate levels of dissolved oxygen to survive. Dissolved Oxygen is stated as milligrams of oxygen per liter of water (mg/l). Inadequate D.O. levels can cause hypoxia, which can harm fish and other aquatic organisms. Monitoring D.O. levels provides valuable insights into water



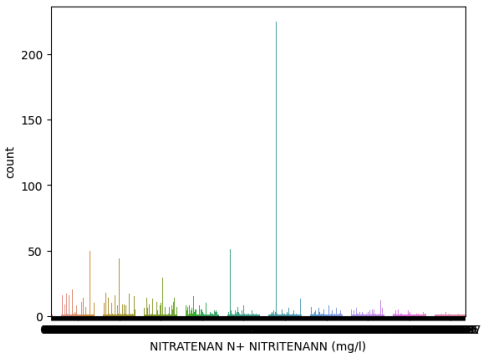
(a) B.O.D.



(b) D.O.



(c) CONDUCTIVITY.



(d) NITRATENAN N+ NITRITENANN.

Figure 1: The count of various water parameters.

quality, indicating potential pollution or a deficiency in oxygen-producing organisms and contributing to overall ecosystem health assessment. The count of D.O. is depicted in the Figure 1b.

3.5 Conductivity

Conductivity is a crucial player in assessing water quality because it gauges the water's capacity to carry electrical currents, influenced by the presence of dissolved ions in aquatic ecosystems. Its significance is amplified in freshwater environments, acting as a valuable gauge for salinity, nutrient concentrations, and overall water quality. Higher values of it shows excessive amount of minerals and dissolved salts. Examining conductivity levels using sensors helps to sense any variation in ion matter identifying presence of water pollution. It is measured in microsiemens per centimeter (S/cm). The conductivity amount is depicted in Figure 1c.

3.6 NITRATENAN N+ NITRITENANN

Substances like Nitrates and Nitrites are examined from the Nitrogen compounds present in the water samples from Indian rivers. These components are essential for the smooth growth of the plants. These nitrogen compounds are measured as a milligrams of nitrogen per liter of water (mg/l). It is responsible for effectively managing nutrients in aquatic environments. However, it's crucial to strike a balance because elevated concentrations of these compounds pose risks and can harm both plants and humans. The sources of these heightened levels vary and can be attributed to factors like fertilizer use, agricultural runoff, and contamination of water sources by sewage. Maintaining optimal nitrogen levels is paramount to ensuring a balanced and thriving aquatic ecosystem. The count of NITRATENAN N + NITRITENANN is depicted in Figure 1d.

3.7 Fecal Coliform

Fecal coliform is a microbiological parameter that indicates the existence of fecal contamination in river water for the waste that is added to the rivers. Fecal coli-form includes bacteria from warm-blooded animals' intestines, and elevated lev-els to indicate possible contamination, posing potential health risks in humans. The existence of fecal coliform bacteria in water specifies the presence of sewage or other fecal matter in the Indian River has a higher chance in the water sample, raising concerns about the safety of drinking water in most of the human ecosystem.

The most Probable Number per 100 milliliters of water (MPN/100ml) is mainly used to measure fecal coliform levels in samples of river water. The elevated fecal coliform levels presence makes it necessary to further investigate and emphasize the significance of water treatment and remediation measures in ensuring the safety of water resources. If the level of fecal coliform is higher from the sample, it needs to be filtered out before it is sent further for drinking or farming purposes.

3.8 TOTAL Coliform

Total coliform bacteria are present in human, and animal waste, soil and water. Their presence particularly in a water indicates poor or below-average water quality. It is measured as the MPN/100ml. These identified standards help as reference points, helping to identify and evaluate the health and environment health issues occurred due to water quality.

As an illustration, the temperature limit for acceptable water temperature is 25°C. For appropriate aquatic ecosystems, the dissolved oxygen levels in water should be above 5 mg/l, and the acceptable acidity or alkalinity range is maintained by pH threshold of 7.5. In addition to that, the conductivity threshold is set at 1500 mhos/cm, as allowable electrical conductivity in river water. These standards are helpful to maintain water quality within appropriate and sustainable limits. Moreover, range of B.O.D. is limited to 3 mg/l, presenting the maximum oxygen requirement for the decomposition of organic matters. The limit for Nitrate+Nitrite concentration is set at 5 mg/l to reduce/control water pollution. Furthermore, the thresholds for fecal coliform and total coliform at 1000 MPN/100ml and 5000 MPN/100ml are the most acceptable values for bacterial contamination for Indian rivers, respectively.

4 MACHINE LEARNING MODELS

Machine Learning models are classified into two main categories i.e., supervised learning and unsupervised learning. A rigorous evaluation using the eight supervised machine learning models is implemented for the numerical prediction of the given water quality parameters.

The models used in machine learning include both traditional and cutting-edge methodologies, offering a wide range of tools for predictive tasks that help to predict the water quality of the Indian River waters (Ewuzie et al., 2022). Logistic regression known as a fundamental statistical approach, gives outstanding performance in the case of binary classification. The decision tree is a type of algorithm that makes decisions by analyzing input parameters and is known for their easily described structures. The Random Forest Classifier is an ensemble technique that leverages the collective strength of multiple Decision Trees to achieve robust predictions of the Water Quality for the dataset that is provided to it to perform the task. The KNN algorithm is utilized to make predictions by calculating the average of 'k' neighboring instances of the parameters that are provided to perform. The SVC is a machine learning algorithm that constructs a hyperplane to optimize the classification process of the model and it describes the Indian river data. This algorithm is recognized for its capability to handle various types of data distributions, making it adaptable in ML.

The Ada-boost Classifier utilizes ensemble learning techniques to improve the predictive accuracy of the model by combining multiple weak learners. The LSTM model- a form of Recurrent Neural Network (RNN) design, demonstrates exceptional proficiency in capturing temporal dependencies. This characteristic renders in it is highly suitable for analyzing sequential data, particularly time series. The XGB Classifier, also known as Extreme Gradient Boosting, effectively integrates decision trees to enhance the accuracy of classification tasks. Each model is carefully chosen and customized, considering its inherent capabilities and appropriateness for the intricate task of predicting IWQ. Although the study offers a thorough examination of these models, it does not conduct a comprehensive algorithmic evaluation.

Here, the implementations investigate three pioneering machine learning models: the AdaBoost Classifier, Long Short-Term Memory, and XGB Classifier. By identifying the weaker learners through ensemble approaches, the Ada-boost Classifier increases the overall prediction of water quality accuracy. A modified version of a recurrent neural network that is particularly appropriate for identifying temporal correlations in sequential water quality data is the Long Short-Term Memory (LSTM). Due to this, it is used for identifying minute patterns that alter over time. The XGB Classifier effectively melds decision trees and increases predicting accuracy by utilizing ensemble methods and gradient boosting. Extensive analyses performed on the Indian River dataset demonstrate the effectiveness of these models in offering comprehensive insights into water quality. This work provides a better understanding of the complex dynamics of water quality and significantly advances predictive modeling used for environmental studies.

The models were evaluated and tested with a standardized dataset comprising water samples collected from all rivers in India.

The commonly used Anaconda distribution is used to implement these machine learning models. Anaconda is an open-source well known and popular platform for using Python modules for machine learning and data science. For this work, ML models are implemented on 1991 river water samples that were gathered from several rivers around India. The data that has been previously utilized is divided into two distinct categories: a dataset that is used for training and testing purposes, and another dataset that is used for validation. Out of the 1991 samples, 1393 (70%) are allocated for training purposes, while the remaining 598 samples are used for model validation. The sample data was collected from all the rivers in India, and the parameters were measured and recorded.

5 DATASET AND DATA PREPROCESSING

After discussion of existing literature, water quality parameters and ML models, dataset and preprocessing are discussed in detail in this section.

5.1 Dataset

A dataset 'waterdataX-1.csv' devoted to the evaluation of water quality includes several parameters related to indicators of water quality. These parameters are Temperature (Temp), Dissolved Oxygen, pH Value, Conductivity, Biochemical Oxygen Demand, Nitrate and Nitrite Levels (NITRATENAN N+ NITRITENANN), Fecal coliform and Total coliform. Each specific variable plays a significant role in assessing the quality of the water. Additionally, 'Water Quality' is a target variable in the dataset. Three different classes of water quality are distinguished by this target variable: Good, Moderate, and Poor.

5.2 Data Preprocessing Workflow

It discusses the various preprocessing steps performed to make the dataset appropriate. The operations are as follows:

Dataset Loading. The water quality dataset is loaded from the dataset file and it is read into a DataFrame.

Data Exploration and Cleaning. df.head(), df.info(), df.describe(), df.columns, df.shape, df.isnull().sum(), and visualisations such as heatmaps, histograms, and count plots (sns.countplot()) are used to examine the structure and content of the dataset. Missing values are determined (df.isnull().sum()) and they are dealt with by either dropping rows or columns or, if necessary, filling the gaps with means.

Feature Engineering. A function called classify-water-quality is implemented to categorize water quality according to various parameters' threshold values. A new column called "Water Quality" is created and classified as the labels "Good", "Moderate", and "Poor".

Data Preprocessing. Feature scaling is performed on numerical features to standardize them. Categorical labels are encoded into numerical values using LabelEncoder. The dataset is split into features (X) and target variables (y) for model training and testing.

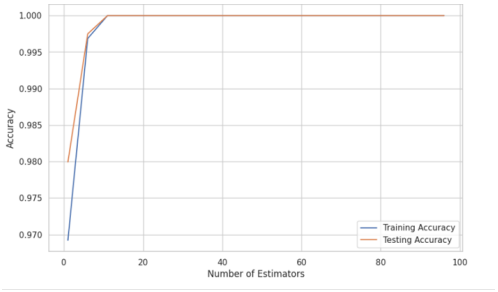
Model Training and Evaluation. Multiple ML models (Logistic Regression, Decision Tree, Random Forest, KNN, SVC, Adaboost, LSTM, XGBoost) are selected for classification based on the nature of the problem. Each model is trained using the training set (fit() method). The performance of model is calculated using metrics- accuracy, precision, recall, and F1-score (accuracy_score, classification0_report) on the test set.

Model Comparison and Analysis. Performance of different models are compared using visualizations (bar plots, tables) to construct metrics such as accuracy, precision, recall, and F1-score for each model. The best-performing model is identified based on the evaluation metrics for water quality problems.

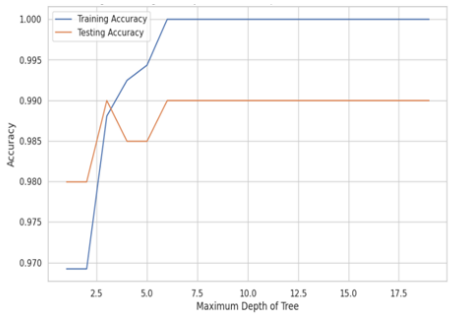
5.3 Comparative Analysis of Models

Various machine learning algorithms viz. decision tree, logistic regression, random forest Classifier, KNN, SVC, Adaboost classifier, LSTM, and XGB classifier are explored in the existing state-of-an-art to predict the water quality on the dataset of multiple rivers of various countries. Particularly, this research involves the data of Indian rivers with multiple/different parameters to predict the water quality of Indian rivers. Hence, the results are not compared with the results of existing implementations.

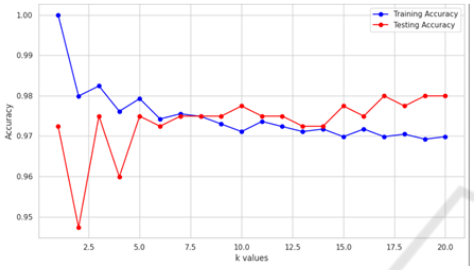
Accuracy Comparison. AdaBoost Classifier and XGBClassifier achieved the highest accuracy of 100%. Random Forest Classifier closely follows with an accuracy of 99.50%. Decision Tree and KNN also achieved high accuracy levels, scoring 99.00% and 97.49%, respectively. Moreover, training and testing accuracy comparison of various models is presented in Figure 2.



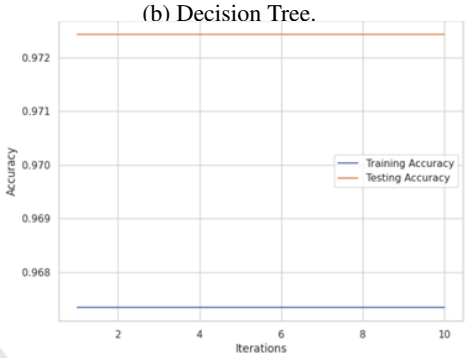
(a) Adaboost Classifier.



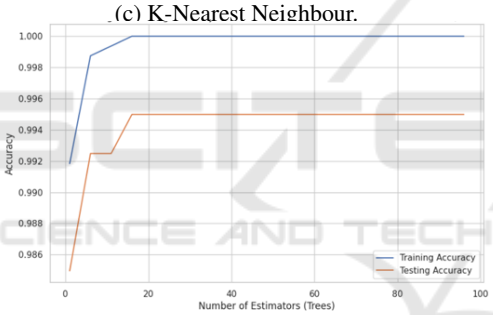
(b) Decision Tree.



(c) K-Nearest Neighbour.

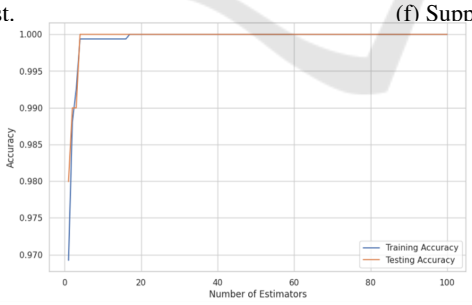


(d) Linear Regression.



(e) Random Forest.

(f) Support Vector Classifier.



(g) XGBoost.

Figure 2: Accuracy comparison of ML models.

5.3.1 Precision, Recall, and F1-score

AdaBoost Classifier and XGBClassifier performed exceptionally well, achieving perfect scores (1.0) for precision, recall, and F1-score. Random Forest Classifier also achieved perfect scores in precision, recall, and F1-score. Logistic Regression, Decision Tree, KNN and SVC also achieved high scores, indicating their effectiveness in classifying ‘Good’, ‘Moderate’, and ‘Poor’ water quality.

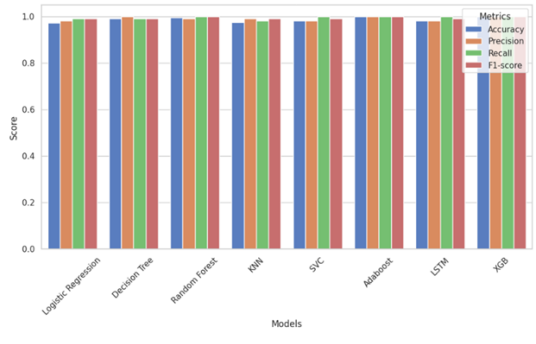


Figure 3: Prediction results and comparative analysis of models.

5.3.2 Support

The 'support' metric represents the number of instances for each class ('Good', 'Moderate', and 'Poor'). In this case, it seems to be consistent across all models at 391 instances for each class.

The above prediction results and comparative analysis of models are summarized in Table 1.

6 RESULTS AND DISCUSSION

6.1 Results

Accuracy, precision, recall, F1-score and support all these parameters' values are evaluated on the basis of the models in Figure 3.

6.1.1 Adaboost Classifier and XGBClassifier

- Achieved the highest accuracy of 100%.
- Demonstrated perfect precision, recall, and F1-score, indicating accurate classification across 'Good', 'Moderate', and 'Poor' water quality classes.
- These models exhibit exceptional performance and could be considered as prime choices for accurate water quality assessment.

6.1.2 Random Forest Classifier

- Achieved a high accuracy of 99.50% and demonstrated perfect precision, recall, and F1-score.
- Showed strong performance in accurately classifying water quality.

6.1.3 Decision Tree, Logistic Regression, KNN, SVC, and LSTM

- Each model achieved an accuracy ranging from 97.24% to 99.00%.
- Displayed consistent and reliable precision, recall, and F1-score metrics, indicating their effectiveness in water quality classification.

6.2 Discussion

6.2.1 Adaboost and XGBClassifier's Perfect Scores

These models were flawless in every metric, demonstrating their resilience in managing the evaluation of water quality. Their ensemble learning techniques, which combine several weak learners to produce a stronger predictive model, may be the cause of this result.

6.2.2 Random Forest Classifier

It showed good predictive power and accuracy, trailing closely behind the best-performing models.

6.2.3 Consistency in Performance

The decision tree, logistic regression, KNN, SVC, and LSTM models all performed consistently and dependably, demonstrating how well they could classify the quality of water.

6.2.4 Consideration for Application

When choosing a model for real-world applications, factors like computational complexity, interpretability, and scalability should also be taken into account, even though models like Adaboost, XGBClassifter, and Random Forest Classifier demonstrated remarkable performance.

Thus, according to the analysis, certain models such as the Random Forest Classifier, XGB Classifier, and Adaboost Classifier are very effective at accurately predicting the quality of water. They did remarkably well on the tests. However, when selecting the optimal model for actually applying it to evaluate water quality, there is need to consider factors like knowledge of the appearance of the data, the amount of processing power or time required by the model, and simplicity to comprehend how the model makes its decisions. Each of these factors is crucial in order to select the model that will perform the best in practical scenarios involving the evaluation of water quality.

7 CONCLUSION

This research work signifies a substantial advancement in the field of predictive modeling of water qual-

Table 1: Prediction results and comparative analysis of models

Name of the Model	Accuracy	Precision	Recall	F1-score	Support
Logistic Regression	97.24%	0.98	0.99	0.99	391
Decision Tree	99.00%	1	0.99	0.99	391
Random Forest Classifier	99.50%	0.99	1	1	391
KNN	97.49%	0.99	0.98	0.99	391
SVC	97.99%	0.98	1	0.99	391
Adaboost Classifier	**100.00**%	1	1	1	391
LSTM	97.99%	0.98	1	0.99	391
XGB Classifier	**100.00**%	1	1	1	391

ity in Indian rivers by means of machine learning, specifically aimed at promoting sustainable resource management and the equivalent water parameters that are needed to predict water quality. The application of sophisticated ML models such as Logistic Regression, Decision Tree, Random forest classifier, KNN, SVC, AdaBoost Classifier, LSTM, and XGB Classifier, shows the dedication to tackling the challenges involved in understanding the intricate dynamics of water quality. The aforementioned models have been thoroughly evaluated using a comprehensive dataset, which includes a wide variety of water samples across diverse rivers of India. The results of these evaluations demonstrate the effectiveness of the models. The implementation results depict that the AdaBoost and XGB Classifier outperform with 100% accuracy. Whereas the other models such as Logistic Regression, Decision Tree, Random Forest classifier, KNN, SVC, and LSTM predict the water quality with an accuracy of 97.24%, 99%, 99.50%, 97.49%, 97.99% and 97.99% respectively.

In addition to that, various regional disparities parameters in the dataset are highlighted in this paper, providing an insight into the intricate environmental elements that affect water quality in different regions of India. The water quality prediction results also give an insight view of the different types of water quality in the different regions of India. These predictions can help in the perspective of farming areas to grow different types of crops in suitable areas. Future work may involve exploring more datasets and the effect of the ensemble approach on ML models.

ACKNOWLEDGEMENTS

The authors wish to thank Mr. N. L. Chauhan for guiding about water parameters. Thanks are also to the management of Nirma University and Pandit Deendayal Energy University for providing resources to carry out research.

REFERENCES

Ewaid, S. H., Kadhum, S. A., Abed, S. A., and Salih, R. M. (2019). Development and evaluation of irrigation water quality guide using iwqg v. 1 software: A case study of al-gharraf canal, southern iraq. *Environmental technology & innovation*, 13:224–232.

Ewuzie, U., Bolade, O. P., and Egbedina, A. O. (2022). Application of deep learning and machine learning methods in water quality modeling and prediction: a review. *Current trends and advances in computer-aided intelligent environmental data engineering*, pages 185–218.

Ibrahim, H., Yaseen, Z. M., Scholz, M., Ali, M., Gad, M., Elsayed, S., Khadr, M., Hussein, H., Ibrahim, H. H., Eid, M. H., et al. (2023). Evaluation and prediction of groundwater quality for irrigation using an integrated water quality indices, machine learning models and gis approaches: A representative case study. *Water*, 15(4):694.

Khoi, D. N., Quan, N. T., Linh, D. Q., Nhi, P. T. T., and Thuy, N. T. D. (2022). Using machine learning models for predicting the water quality index in the la buong river, vietnam. *Water*, 14(10):1552.

Meireles, A. C. M., Andrade, E. M. d., Chaves, L. C. G., Frischkorn, H., and Crisostomo, L. A. (2010). A new proposal of the classification of irrigation water. *Revista Ciência Agronômica*, 41:349–357.

Modaresi, F. and Araghinejad, S. (2014). A comparative assessment of support vector machines, probabilistic neural networks, and k-nearest neighbor algorithms for water quality classification. *Water resources management*, 28:4095–4111.

Sakaa, B., Elbeltagi, A., Boudibi, S., Chaffaï, H., Islam, A. R. M. T., Kulimushi, L. C., Choudhari, P., Hani, A., Brouziyne, Y., and 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.

Sakizadeh, M. and Mirzaei, R. (2016). A comparative study of performance of k-nearest neighbors and support vector machines for classification of groundwater. *Journal of Mining and Environment*, 7(2):149–164.

Sharma, A., Jain, A., Gupta, P., and Chowdary, V. (2020). Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9:4843–4873.

Singh, S., Ghosh, N., Gurjar, S., Krishan, G., Kumar, S., and Berwal, P. (2018). Index-based assessment of suitability of water quality for irrigation purpose under indian conditions. *Environmental monitoring and assessment*, 190:1–14.

Tyagi, S., Sharma, B., Singh, P., and Dobhal, R. (2013). Water quality assessment in terms of water quality index. *American Journal of water resources*, 1(3):34–38.