Enhancing Water Quality Classification in Agriculture Using a Voting Classifier Ensemble Approach

Sherilyn Kevin[1] Santosh Kumar Singh[2] Hrushi Bhola[3] Kunal Singh[4]
[1,2,3,4]Department of Information Technology
[1,2,3,4]Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India

Abstract — Water quality plays a critical role in agriculture, affecting crop yields, livestock health, and overall farm productivity. Traditional assessment methods are slow and resource-intensive, making them impractical for large-scale monitoring. This study explores the application of machine learning in water quality classification, leveraging a voting classifier that integrates Gradient Boosting, CatBoost, and AdaBoost to enhance accuracy and scalability. The dataset, pre-processed and augmented for robustness, was used to train and evaluate the model, achieving an accuracy of 92.03% on the test set. Results indicate that ensemble learning significantly improves prediction reliability over individual models. By providing a scalable and efficient approach to water quality monitoring, this research contributes to sustainable agriculture and better resource management. Future work will focus on expanding datasets, integrating environmental variables, and developing real-time IoT-based monitoring systems for enhanced decision-making.

Keywords: Water Quality, Machine Learning, Ensemble Learning, Voting Classifier

I. Introduction

Water quality is a crucial factor in agriculture, directly influencing crop yields, livestock health, and overall farm productivity. Clean water is essential for irrigation, fertilization, and animal hydration, while poor water quality can degrade soil, stunt crop growth, and increase the risk of diseases. Contaminated water—containing harmful chemicals, pathogens, and excess nutrients—can disrupt agricultural ecosystems, leading to long-term environmental and economic consequences. With increasing pressures from pollution, climate change, and a growing global population, the need for advanced water quality monitoring has become more urgent than ever.

Traditional water quality assessment methods rely on time-consuming field sampling and complex laboratory analyses, making them inefficient for large-scale, real-time monitoring. In response, machine learning has emerged as a powerful tool for processing vast datasets, identifying trends, predicting outcomes, and supporting data-driven decision-making. Numerous studies have demonstrated the potential of machine learning models in improving water quality prediction and classification.

Nasir et al. developed machine learning classifiers to predict the Water Quality Index (WQI) from complex datasets, finding that stack modeling significantly improved accuracy. Among various algorithms, CATBOOST demonstrated the best predictive performance [1]. Similarly, Khullar and Singh introduced a Bi-LSTM-based deep learning model (DLBL-WQA) for forecasting water quality in the Yamuna River, India. Their approach incorporated missing value imputation, feature extraction, and an optimized loss function, outperforming traditional models like SVR, random forest, ANN, LSTM, and CNN-LSTM [2].

Several studies have focused on identifying the most effective machine learning models for water quality assessment. Uddin et al. demonstrated that Decision Tree (DT), Extra Tree (ExT), and Extreme Gradient Boosting (XGB) were highly effective in predicting coastal WQI [3]. Asadollah et al. introduced an Extra Tree Regression (ETR) model for river WQI prediction, showing its superiority over Support Vector Regression (SVR) and Decision Tree Regression (DTR) [4]. Additionally, Chen et al. leveraged big data techniques to enhance surface water quality prediction by identifying key parameters for rapid monitoring [5].

Deep learning models have also shown promise in this field. Singha et al. proposed a groundwater quality prediction model that outperformed Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN) in accuracy [6]. Uddin et al. highlighted the issue of uncertainty in WQI models, advocating for robust methodologies like Monte Carlo simulations and Gaussian Process Regression to improve prediction reliability [7]. Azrour et al. demonstrated that multiple regression and artificial neural networks effectively classify water quality [8]. Sagan et al. explored remote sensing and deep learning for inland water quality monitoring, emphasizing the potential of multi-sensor data fusion [9].

In the context of smart water monitoring, CNN-based models have gained traction. Chen et al. developed a CNN model optimized with decision trees for near-infrared water pollution analysis in agricultural irrigation, improving prediction accuracy [10]. Haq et al. introduced hybrid deep learning models combining CNN with LSTM and GRU, which outperformed traditional models in aquaculture water quality prediction [11].

Despite these advancements, a major challenge remains scaling machine learning models for real-time, large-scale water quality monitoring across diverse agricultural regions. To address this, our research introduces a voting classifier that aggregates predictions from multiple base models to enhance classification accuracy and scalability. By leveraging the collective decision-making of diverse algorithms, this approach aims to improve the precision and reliability of water quality monitoring. Ultimately, it supports sustainable agriculture, safeguards water resources, and ensures food security amidst growing environmental challenges.

II. Methodology

The methodology is summarized in Fig. 1 and detailed in the following sections, with all analyses and coding performed in Google Colab.

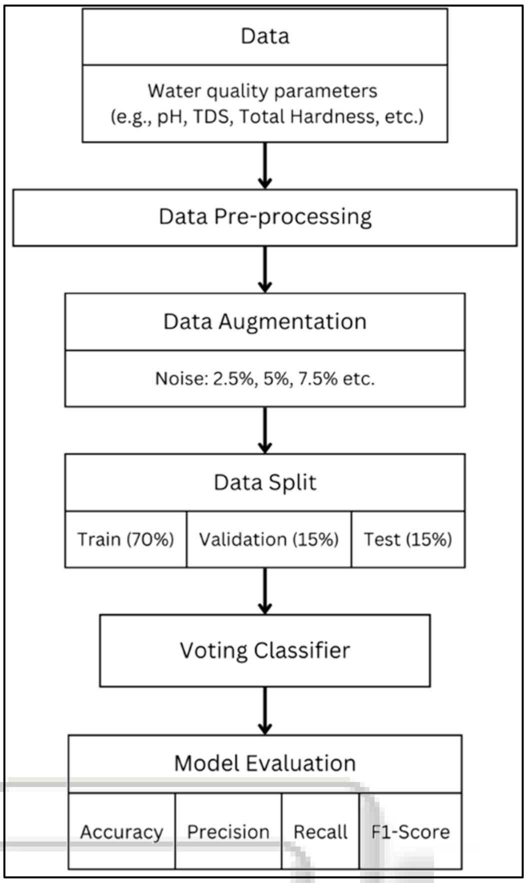

Fig 1. Overview of methodology

A. Data Collection

The dataset used in this research was provided by the Department of IT, Thakur College of Science and Commerce, in CSV format. Due to confidentiality agreements, the original source of the dataset cannot be disclosed. It includes key information relevant to our study, such as variables like Year, Calcium (Ca), Chloride (Cl), Carbonate (CO3), Electrical Conductivity (EC), Bicarbonate (HCO3), Potassium (K), Magnesium (Mg), Sodium (Na), Nitrate (NO3), pH Level, Sulfate (SO4), Total Dissolved Solids (TDS), Total Hardness (TH), Fluoride (F), Potability, and Infrastructure Suitability.

B. Data Pre-Processing

During the preprocessing phase, the dataset was carefully reviewed to ensure its suitability for analysis.

Key steps:

1) Removal of Irrelevant Columns: The Year and Infrastructure Suitability columns were removed as they did not align with the focus of our research.
2) Handling of CO3 (Carbonate) Parameter: Upon examination, it was found that all elds in the CO3 column contained values of zero. As a result, this parameter was deemed uninformative for the analysis and was removed from the dataset.
3) Missing Values and Data Consistency: Any rows containing missing values were dropped from the dataset to ensure data integrity. The features were standardized using to ensure that all variables contributed equally to the model. This step was crucial as it normalized the feature ranges, particularly for algorithms sensitive to the scale of input data.
4) Renaming of Target Variable: The target variable, originally labelled as "Potability," was renamed to "Agricultural Suitability" to better reflect the focus of the research on water quality classification for agricultural applications.

C. Data Augmentation

In this research, data augmentation was used to enhance the dataset and enhance model robustness. We introduced synthetic variations by applying random noise with variation probabilities of 2.5, 5, 7.5, 10, and 12.5 to these numeric features. Each variation created a slightly different dataset, and the Total Hardness (TH) was recalculated using the equation given below to maintain consistency. These synthetic datasets were amalgamated with the original data, resulting in a comprehensive augmented dataset. Each entry's agricultural suitability was reassessed and streamlined according to standard water quality criteria. This fortified dataset provides a broader range of conditions for further robust model training and evaluation.

$$TH = 2.5 \times Ca + 4.1 \times Mg$$ [12]

Noise levels were selected to simulate natural variations in water quality measurements. By conforming these parameters, the model was tested for adaptability under conditions of dimension noise, generally present in real-world data collection.

D. Exploratory Data Analysis (EDA):

During the exploratory data analysis (EDA) of the augmented CSV dataset, which consists of 2,676 samples, the primary focus was on agricultural suitability. The distribution plot revealed that the 'not suitable' class was the most prevalent, while the 'suitable' class was underrepresented, indicating a potential class imbalance. The statistical summary of the features provided insights into their central tendencies and variability. Visualization techniques, such as histograms and correlation heatmaps, were used to uncover significant relationships between features. The correlation matrix suggested potential multicollinearity, which will be addressed during the modelling phase. Outliers and missing values were also identified and handled to ensure the data's quality and robustness.

E. Models

In this study, a Voting Classifier was implemented to enhance the accuracy and robustness of water quality classification. The model combined three ensemble learning algorithms—Gradient Boosting Classifier (GBC), CatBoost Classifier, and AdaBoost Classifier—leveraging their collective strengths for improved predictive performance.

Before training, the dataset underwent data cleaning, where missing values were removed to ensure data integrity. Features were then standardized using StandardScaler, which normalized the data distribution, improving the performance of the classifiers. The dataset was split into training (70%), validation (15%), and test (15%) sets to facilitate model evaluation and generalization assessment.

1) Voting Classifier Implementation

A hard voting strategy was employed, where each base classifier independently made predictions, and the final output was determined by majority voting. The models were configured with the following hyperparameters:

- Gradient Boosting Classifier (GBC): 100 estimators, max depth = 3, learning rate = 0.01, subsample = 0.8
- CatBoost Classifier: 100 iterations, depth = 3, learning rate = 0.01
- AdaBoost Classifier: 100 estimators, learning rate = 0.01

F. Model Evaluation

The voting classifier is evaluated using accuracy, precision, recall, and F1-score on a test set to assess its ability to predict water suitability for agriculture. Accuracy measures overall correctness, while precision and recall focus on correctly identifying suitable and unsuitable water. The F1-score balances precision and recall for a comprehensive performance measure. Comparing these metrics helps evaluate model robustness, generalization, and effectiveness in distinguishing water quality.

III. RESULTS AND OBSERVATION

Train	Val	Test
92.47%	92.01%	92.03%

Table I: Training, Validation, Testing accuracy of Voting Classifier model

As seen in Table II, the Voting Classifier exhibited strong performance in classifying water quality, achieving accuracies of 92.47% on the training set, 92.01% on the validation set, and 92.03% on the test set. The minimal variation across these datasets indicates that the model generalizes well without overfitting. By leveraging Gradient Boosting, CatBoost, and AdaBoost classifiers in a hard voting strategy, the model effectively combined the strengths of multiple algorithms, enhancing prediction reliability. The consistency in accuracy suggests that standardized preprocessing, including feature scaling, contributed to stable performance across different data splits. Additionally, the model effectively distinguished between suitable and unsuitable water for agriculture, demonstrating its reliability in real-world applications. The slight drop in validation and test accuracy compared to the training set indicates minor dataset variations but does not significantly impact overall performance. These results underscore the effectiveness of ensemble learning in improving classification accuracy and robustness for water quality assessment.

IV. CONCLUSION:

This research demonstrated the effectiveness of a voting classifier in accurately assessing water quality for agricultural use. By combining multiple base models, the proposed approach addressed challenges posed by heterogeneous datasets, achieving higher accuracy and robustness than traditional methods. Standardized preprocessing, including feature scaling, further improved model consistency and reliability. The findings underscore the potential of ensemble learning in transforming real-time water quality monitoring, enabling better resource management and more informed decisions for irrigation and livestock health. Future research can enhance predictive capabilities by incorporating additional environmental factors such as climate and soil conditions, expanding datasets across diverse geographic regions, and integrating the framework with IoT-enabled smart monitoring systems and edge computing for real-time, automated decision-making. These advancements will further strengthen agricultural sustainability and promote efficient water resource management.

REFERENCES

[1] Nasir N, Kansal A, Alshaltone O, Barneih F, Sameer M, Shanableh A, Al-Shamma'a A. Water quality classification using machine learning algorithms. Journal of Water Process Engineering. 2022 Aug 1;48:102920.
[2] Khullar S, Singh N. Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation. Environmental Science and Pollution Research. 2022 Feb;29(9):12875-89.
[3] Uddin MG, Nash S, Diganta MT, Rahman A, Olbert AI. Robust machine learning algorithms for predicting coastal water quality index. Journal of Environmental Management. 2022 Nov 1;321:115923.
[4] Asadollah SB, Sharafati A, Motta D, Yaseen ZM. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. Journal of environmental chemical engineering. 2021 Feb 1;9(1):104599.
[5] Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, Liu F, Zuo M, Zou X, Wang J, Zhang Y. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water research. 2020 Mar 15;171:115454.
[6] Singha S, Pasupuleti S, Singha SS, Singh R, Kumar S. Prediction of groundwater quality using efficient machine learning technique. Chemosphere. 2021 Aug 1;276:130265.
[7] Uddin MG, Nash S, Rahman A, Olbert AI. A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches. Water Research. 2023 Feb 1;229:119422.
[8] Azrour M, Mabrouki J, Fattah G, Guezzaz A, Aziz F. Machine learning algorithms for efficient water quality prediction. Modeling Earth Systems and Environment. 2022 Jun;8(2):2793-801.
[9] Sagan V, Peterson KT, Maimaitijiang M, Sidike P, Sloan J, Greeling BA, Maalouf S, Adams C. Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. Earth-Science Reviews. 2020 Jun 1;205:103187.
[10] Chen H, Chen A, Xu L, Xie H, Qiao H, Lin Q, Cai K. A deep learning CNN architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources. Agricultural Water Management. 2020 Oct 1;240:106303.
[11] Haq KR, Harigovindan VP. Water quality prediction for smart aquaculture using hybrid deep learning models. Ieee Access. 2022 Jun 6;10:60078-98.

[12] Pal A, Pal M, Mukherjee P, Bagchi A, Raha A. Determination of the hardness of drinking packaged water of Kalyani area, West Bengal. Asian Journal of Pharmacy and Pharmacology. 2018;4(2):203-6.