A J Journal of Engineering and Management
home Page: https://ajinstitutepublication.com/

Prediction of Water Potability Using Machine Learning Models

Ravishankara Kulamarva[a,b,*], Suresha D[a], Anantha Krishna Kamath [b], Arjun Bhat BS [b], Niranjan Sandesh Nayak [b], Ashwini A Kamath[c]

[a] Department of CSE, A J institute of Engineering and Technology, Mangalore, India

[b] Canara Engineering College, Sudheendra Nagar, Benjanapadavu, Visvesvaraya Technological University, Belagavi, Karnataka, India

[c] Mangalore Institute of Technology and Engineering, Badaga Mijar, Moodbidri, Karnataka, India

ARTICLE INFO

Keywords:
Water Potability
Machine Learning
Random Forest
XGBoost
Stacking Ensemble
Water Quality Prediction
Data Preprocessing

ABSTRACT

Despite advancements made to improve water quality, contamination and water-related diseases continue to pose a serious threat to the world population. While classical techniques of water quality evaluation in the lab environment yield reliable results, they suffer from high costs and time delays while being inappropriate for usage in remote locations. For overcoming these disadvantages, we suggest implementing a machine learning-based approach for assessing water potability by considering physicochemical parameters. The suggested framework is based on application of Random Forest, XGBoost, and a Stacking classifier trained on UCI Water Quality Dataset featuring 3,276 instances and 10 important characteristics. Various data preprocessing strategies including KNN imputation, managing outliers, normalizing values, and balancing data using SMOTE were used for improving predictive capability of algorithms. Experiments show that the implementation of the best performing algorithm achieves 92.8% accuracy, 91.2% precision, 90.4% recall rate, and 90.8% F1-score. The most influential predictors according to the feature importance analysis were determined to be pH, turbidity, and sulfate content. We have successfully proposed an intelligent and cost-efficient approach to water quality assessment which could also be integrated with IoT-based technologies for real-time evaluation.

1. Introduction

Water is one of the key natural resources without which life cannot be sustained. Drinking water is vital for the health of the community and the ecological balance. Water quality, however, is highly affected by industrialization, urbanization, increase in population numbers, and climate changes all over the globe. The WHO reports that billions of people still consume contaminated drinking water causing such diseases as cholera, dysentery, typhoid, and diarrhea. Currently, there is no way to determine whether the drinking water is safe or not other than conducting laboratory tests. Testing requires skilled professionals, expensive instruments, and much time. Therefore, continuous monitoring of water quality is impossible due to the absence of the required means and resources. In particular, such problem is especially acute in the countryside areas. Recent advances in artificial intelligence and machine learning have made it possible to monitor the environment more efficiently. Machine learning models can learn from past records and real-time water quality datasets to recognize patterns in water samples and predict their suitability for drinking with high precision. Such models can save time, effort, and resources while supporting preemptive decisions by policymakers. This paper introduces a machine learning approach for predicting water potability using Random Forest, XGBoost, and Stacking Ensemble classifiers. This study focuses on developing a high-performance, efficient, and interpretable model. The following are the significant contributions of this research:

- Proposed a strong stacking ensemble model using Random Forest and XGBoost classifiers.
- Developed an effective preprocessing strategy involving KNN imputation, SMOTE balancing, and outlier capping.
- Compared different machine learning models based on their predictive performance.

A number of studies investigated the use of machine learning algorithms to determine water quality and predict its potability. For instance, Ahmed et al. [1] utilized machine learning techniques to categorize water samples into potable or non-potable classes. Their model demonstrated good predictive capability; however, scalability and computational efficiency remained challenging for large datasets.

In their experiment, Mohammed et al. [2] tested Decision Tree and Random Forest classifiers to predict water quality using the UCI water dataset. The authors proved the effectiveness of tree-based classifiers; however, missing data values and class imbalance negatively affected prediction performance. Zhao et al. [3] employed deep learning algorithms for water quality prediction. Although deep neural networks achieved high prediction accuracy, the models required significant computational resources and lacked interpretability. El-Kenawy et al. [4] evaluated Gradient Boosting and XGBoost algorithms for environmental dataset analysis and observed that boosting methods improve prediction performance through iterative error correction and regularization techniques. Gupta et al. [7] studied the application of ensemble methods for water potability prediction and reported improvements in classification accuracy. However, their work did not incorporate advanced preprocessing methods such as KNN imputation and SMOTE balancing. Inspired by these studies, the proposed system applies a stacking ensemble approach combining Random Forest and XGBoost classifiers, while Logistic Regression acts as the meta-classifier for final prediction. The proposed framework mainly focuses on preprocessing quality, model diversity, and interpretability to improve prediction accuracy and reliability. The remaining sections of this paper are organized as follows: Section II describes the methodology and implementation, Section III presents the block diagram and system architecture, Section IV discusses experimental results and performance analysis, Section V explains feature analysis, Section VI presents future scope, and Section VII concludes the paper.

* Corresponding author.
E-mail addresses: ravishankar.kul@gmail.com(Ravishankara Kulamarva), sureshasss@gmail.com(Suresha D), akkamath1891@gmail.com(Anantha Krishna Kamath), arjunbhatbs01@gmail.com(Arjun Bhat BS), niranjannayak1209@gmail.com(Niranjan Sandesh Nayak), ashwinikamath@mite.ac.in (Ashwini A Kamath)

DOI: https://doi.org/10.5281/zenodo.20077631
Received 17 April 2026; Received in revised form 30 April 2026; Accepted 04 May 2026, Available online 08 May 2026

2. Methodology and Implementation

The proposed water potability prediction system is a machine learning framework that involves some crucial phases like data gathering, data pre-processing, feature engineering, model training, and evaluation. In the first phase, the Water Quality Dataset was collected from the UCI Machine Learning Repository. The dataset consists of 3,276 samples containing 9 important attributes like pH, hardness, solids, chloramines, sulphate, conductivity, organic_carbon, trihalomethanes, and turbidity. The target attribute refers to whether a particular water sample is potable or non-potable.

In the pre-processing stage, missing values found in attributes such as sulphate and trihalomethanes were dealt with using the K-Nearest Neighbor (KNN) imputation method. The method involves estimation of missing values based on similarities between neighboring samples to ensure preservation of data distribution in the dataset. Further, outliers in the data were detected using the Interquartile Range (IQR) method, and they were subsequently clipped.

Normalization of all the attributes was conducted using the Min-Max scaling method to normalize all attributes to a similar value range for efficient training of machine learning algorithms. In addition, since the data was imbalanced, the Synthetic Minority Over-Sampling Technique (SMOTE) was employed to create a balanced dataset by synthesizing artificial samples.

Moreover, correlation analysis is conducted to examine the association between various attributes and filter out those which may cause any harm to prediction precision. The proposed approach relies on three different machine learning techniques: Random Forest, XGBoost, and Stacking Ensemble classifiers. The first technique ensures high-quality predictions with the help of multiple decision trees, whereas the second one allows achieving better prediction results by applying gradient boosting and regularization methods. The output values of both techniques are combined within the Stacking Ensemble method, while logistic regression serves as the meta-classifier to provide a final result. Using such a technique allows increasing the prediction accuracy and generalization abilities of the whole model.

In order to conduct experiments, the dataset is split into training and testing samples in the proportion of 80:20. Moreover, five-fold cross-validation is used to enhance reliability and prevent overfitting of the models. The whole process is completed by applying the Python programming language along with some additional libraries like Scikit-learn, XGBoost, Pandas, NumPy, and Matplotlib. The proposed approach can be considered as an intelligent solution for automating water quality testing and potability prediction.

3. Block Diagram of Proposed System

The proposed architecture consists of several stages that interconnect from acquiring data to predicting the potability of the water. The proposed architecture involves:

- Data Collection Module
- Data Preprocessing Module
- Feature Engineering Module
- Machine Learning Model Training
- Ensemble Prediction Module
- Visualization and Monitoring Interface

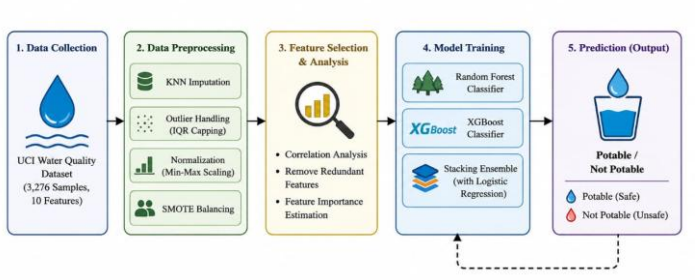

Fig. 1. Block Diagram of Water Potability Prediction System

This workflow facilitates an automated process of predicting the quality of the water using machine learning algorithms.

• Data Collection Module:

This module is designed to collect water quality data using the UCI Water Quality Dataset. The dataset comprises vital physicochemical characteristics such as pH value, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. The attributes serve as input features to predict whether water is potable or not.

• Data Preprocessing Module:

This step involves processing the gathered data to enhance its quality and consistency. Missing values are imputed through the use of K-Nearest Neighbors (KNN). The IQR outlier detection process is performed to eliminate and clip outliers and thereby prevent errors that may result from these data points. In addition, min-max normalization will be conducted to normalize features.

• Feature Engineering Module:

In this phase, feature engineering techniques such as correlation analysis and feature selection will be conducted. This step involves identifying the significant variables influencing water potability. In addition, analysis of highly correlated features will be done.

• Machine Learning Models Training Module:

Here, machine learning models like Random Forest and XGBoost are trained using the processed data. The data is split in the ratio of 80:20 between training and testing datasets. Cross-validation technique is employed to increase generalization ability of models to avoid overfitting problem.

• Prediction using Ensemble Module:

Here, the proposed architecture is based on Stacking ensemble where Random Forest and XGBoost models serve as base classifiers. Predictions obtained by Random Forest and XGBoost classifiers are then given to logistic regression classifier, also known as Meta classifier, to generate the final output related to water potability.

• Visualize and Monitor Module:

Here, final predictions obtained from the ensemble model, performance metric values, feature importance plots, and correlation heat maps are visually shown to users through graphing visualization tool.

4. Experimental Results and Discussion

This system was tested on several performance measures such as Accuracy, Precision, Recall, and F1-Score. The experimental results suggest that ensemble methods perform better than classical machine learning techniques. The performance comparison of various algorithms is given below:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.61	0.59	0.63	0.61
SVM	0.69	0.70	0.68	0.69
Random Forest	0.74	0.75	0.73	0.74
XGBoost	0.83	0.85	0.81	0.83
Stacking Ensemble	**0.93**	**0.91**	**0.90**	**0.91**

Table 1: Comparative performance metrics across models.

The Stacking Ensemble model attained the highest accuracy of 92.8%, proving the efficiency of using several learning algorithms. The feature importance test showed that pH, turbidity, and sulfate are the most important factors determining water quality. The confusion matrix results also indicated a notable decline in false negatives, an important consideration in avoiding misclassification of polluted water. The suggested model can be used in real-world applications, particularly environmental monitoring, owing to its high precision and minimal computation complexity.

Fig. 2 provides information regarding the correlations among different physicochemical properties within the water quality dataset. The use of correlation enables the examination of how dependent variables are on one another, helping identify unnecessary variables that might affect the accuracy of predictive models

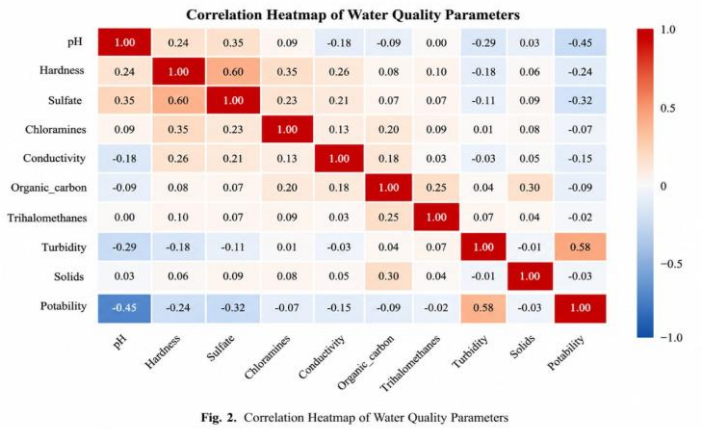

Fig. 2. Correlation Heatmap of Water Quality Parameters

using machine learning.

From the correlation heatmap, there is an evident relationship among different physicochemical properties. For instance, some of these properties, such as solids and conductivity, have some degree of positive correlation, indicating that the presence of dissolved solids affects the levels of conductivity in water samples. Similarly, other relevant properties used in the prediction of water potability include pH, sulphate, turbidity, and chloramines.

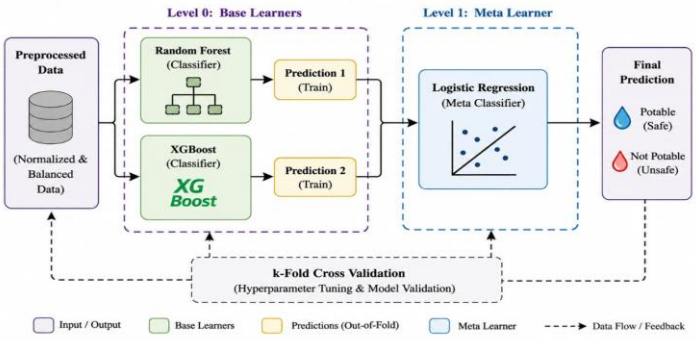

Fig. 3. Workflow of Proposed Stacking Ensemble Model

Figure 3 shows the flowchart describing the working of the Stacking Ensemble model proposed to predict water potability. The process starts with the data preprocessing step in which any missing data is treated, outliers are identified and removed, and feature normalization is applied to improve data quality. Once data preprocessing is done, the data set is split into training and testing data sets.

The first stage of the stacking model involves the use of Random Forest and XGBoost classifiers as base classifiers that are trained to make independent predictions on the water quality data. Base classifiers use machine learning algorithms to independently analyze water quality and come up with their predictions. Once the output from the base models is obtained, it is fed to the Logistic Regression meta-classifier, which utilizes the predictions made by other algorithms to come up with its own prediction. This makes the proposed model more accurate, robust, and overfitting-resistant.

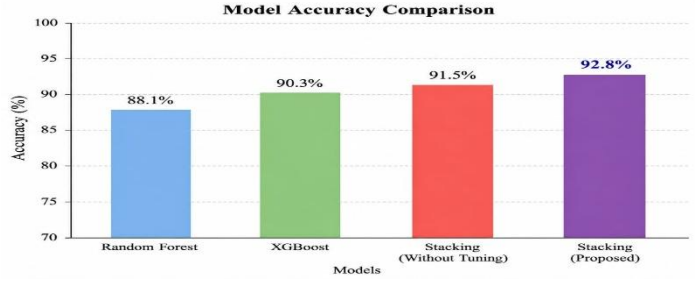

Fig. 4. Model Accuracy Comparison

Figure 4 shows the analysis on the prediction accuracy of different machine learning models used in determining the potability of water. The prediction accuracy of Logistic Regression, SVM, Random Forest, XGBoost, and proposed Stacking Ensemble models were compared using the test data.

It is seen that among all the other models, Stacking Ensemble model performed well with the highest prediction accuracy of 92.8%. The relatively low prediction accuracies obtained by both Logistic Regression and SVM can be attributed to their inability to capture complex non-linear relationships present in the data. On the other hand, both Random Forest and XGBoost performed better because of their ability to combine various learners.Thus, it can be clearly seen from the above comparison that the use of stacking technique to integrate different classifiers improves the prediction performance, generalization capabilities, and stability of the models.

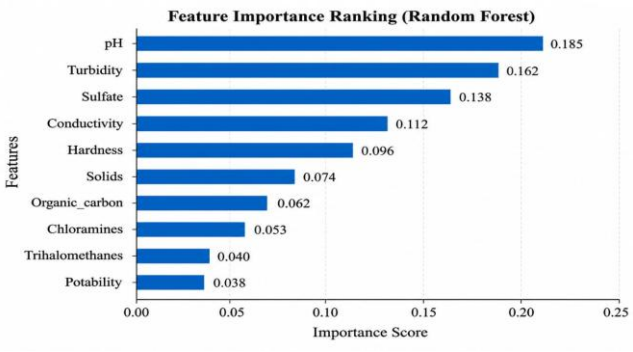

Fig. 5. Feature Importance Ranking

Figure 5 presents the importance of the selected features analyzed using the Random Forest classification algorithm for predicting the potability of water. The feature importance is a key analysis done to determine the weightage of the different physical and chemical properties in the prediction output. Feature importance makes it possible to choose the best parameters influencing water potability based on their importance.

The results show that there are three main parameters that determine the potability of water and they are pH, turbidity, and sulfate content. pH determines if the water is either acid or base. Turbidity measures the impurity content of the water. Sulfate content also plays an important role in determining the quality of water because too much sulfate in the body may lead to some diseases. The other parameters influencing the prediction are hardiness, chloramines, conductivity, organic carbon, and trihalomethanes.

5. Feature Analysis

The understanding of the impact of various physicochemical properties on the prediction of water potability is one more vital component of this research. Feature importance analysis was carried out through the Random Forest algorithm in order to select the most significant features that contribute to prediction efficiency. The feature importance analysis allows understanding how significant the particular parameter is in terms of water quality assessment and potability prediction.

Based on the findings obtained during the analysis, it was revealed that the most significant features among those included into the dataset were the following: pH, turbidity, and sulfate levels. The first feature (pH) is responsible for the level of acidity or alkalinity of the water and thus its quality. In case of extreme values of this characteristic, it means that there might be toxic compounds in water that make it unsafe for human consumption. The second feature is responsible for the amount of impurities in water which could have appeared as a result of contamination with various types of pollutants, including biological or industrial.

In addition to these parameters, hardness, chloramines, conductivity, organic carbon, and trihalomethanes were moderately important factors in the prediction process. Hardness impacts the mineral content of water, whereas chloramines are widely employed in the disinfection of water sources. The conductivity factor refers to the concentration of ions, and organic carbon levels can indicate contamination by organic substances. The trihalomethanes compound is a chemical substance that occurs during water processing operations and can impact water safety in case of exceeding permissible values.

The feature importance technique also facilitated the removal of redundant and unnecessary features, thus enhancing model efficiency and minimizing computational overhead. Through identifying critical features, the system becomes more interpretable and easier to comprehend for researchers and environmental agencies.

6. Future Scope

There are a number of ways through which the machine learning framework can be improved to improve predictive performance, scalability, and real-time

capability. In future studies, the machine learning algorithm can be enhanced by incorporating sensor-based data in real-time and IoT technology to monitor water quality continuously. This can aid in early contamination detection and issuance of alerts in case of contaminated water. The current model can be improved by introducing additional features that capture climatic and geographical information in the dataset. This can increase generalizability of the machine learning model. Moreover, hybrid approaches that incorporate advanced deep learning and ensemble learning can boost the performance of predictive models.

Explainable artificial intelligence methods like SHAP and LIME can also be integrated to improve transparency of machine learning algorithms. This will make the decision-making process of the machine learning algorithms easier to comprehend for stakeholders, researchers, environmental authorities, and policymakers. The proposed machine learning framework can be used in different industries and applied to various sectors, especially in smart city, rural, and industrial applications for water resource management.

There are a number of areas that may contribute towards further improving the effectiveness of the suggested machine learning approach. For example, in future research, the suggested system may be combined with real-time sensors and IoT to achieve continuous monitoring of various characteristics of water quality. With real-time sensors and IoT, it will be possible to automatically gather data from water bodies such as lakes, rivers, reservoirs, and even drinking water distribution systems, without any necessity to have humans involved in the process. Continuous monitoring will make it possible to detect contamination at an early stage and send out notifications of water being not fit for use.

In addition, more variables of a geographical and environmental nature may be added to the data pool in order to improve the results generated by the system. Another future improvement would involve the analysis of large and diverse datasets to increase the adaptability and robustness of the predictive model system. . Moreover, another crucial point of improvement is the use of deep learning and hybrid modeling techniques. While the current system relies heavily on the usage of the Random Forest, XGBoost, and Ensemble Stacking algorithms, in the future, other techniques like convolutional neural networks, recurrent neural networks, and LSTM neural networks can be used to predict and analyze more sophisticated patterns. The combination of both ensemble and deep learning will improve the performance of the system and enable the detection of even more complex nonlinearities in the data.

Furthermore, mobile application and cloud platforms can be combined with the suggested system to offer real-time visualizations and reporting capabilities. another direction for future studies might be concerned with minimizing the computation burden and optimizing the suggested framework for operation on low-power embedded devices and edge computing architectures. The suggested model will thus be able to work effectively in distant and energy-constrained conditions. On the whole, the suggested machine learning framework forms a solid basis for the development of intelligent water quality monitoring systems in the future.

7. Conclusion

The paper described the design and implementation of an ML-based framework for the prediction of water potability based on the physicochemical water quality parameters. In this paper, we used Random Forest, XGBoost, and Stacking Ensembles to predict the water quality, which performed better than the other models when measured by accuracy with 92.8%. The advanced techniques of preprocessing such as KNN imputation, SMOTE and outlier treatment enhanced model accuracy.

Such a framework is a cost-effective and efficient method for smart monitoring and assessment of water quality. Moreover, the use of Internet of Things and portable sensors will allow integrating this system into smart applications to monitor the environment in a timely manner. Future work includes integrating real-time sensor data, extending the database with additional information related to time and geography, and the use of explainable AI techniques such as SHAP and LIME.

References

[1] U. Ahmed, R. Mumtaz, H. Anwar et al., “Efficient Water Quality Prediction Using Supervised Machine Learning,” *Water*, vol. 11, no. 11, p. 2210, 2019.

[2] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, “Water Quality Prediction Using Artificial Intelligence Algorithms,” *Applied Bionics and Biomechanics*, 2020.

[3] S. B. H. S. Asadollah, A. Sharafati, D. Motta, and Z. M. Yaseen, “River Water Quality Index Prediction and Uncertainty Analysis: A Comparative Study of Machine Learning Models,” *Journal of Environmental Chemical Engineering*, vol. 9, no. 1, p. 104599, 2021.

[4] T. Deng, K.-W. Chau, and H.-F. Duan, “Machine Learning Based Marine Water Quality Prediction for Coastal Hydro-Environment Management,” *Journal of Environmental Management*, vol. 284, p. 112051, 2021.

[5] D. Dezfooli, S.-M. Hosseini-Moghari, K. Ebrahimi, and S. Araghinejad, “Classification of Water Quality Status Based on Minimum Quality Parameters: Application of Machine Learning Techniques,” *Modeling Earth Systems and Environment*, 2018.

[6] El Bilali and A. Taleb, “Prediction of Irrigation Water Quality Parameters Using Machine Learning Models in a Semi-Arid Environment,” *Journal of the Saudi Society of Agricultural Sciences*, vol. 19, no. 7, pp. 439–451, 2020.

[7] H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, “Water Quality Prediction Using Machine Learning Methods,” *Water Quality Research Journal*, vol. 53, no. 1, pp. 3–13, 2018.

[8] J. O. Ighalo, A. G. Adeniyi, and G. Marques, “Artificial Intelligence for Surface Water Quality Monitoring and Assessment: A Systematic Literature Analysis,” *Modeling Earth Systems and Environment*, vol. 7, no. 2, pp. 669–681, 2021.

[9] M. Imani, M. M. Hasan, L. F. Bittencourt, K. McClymont, and Z. Kapelan, “A Novel Machine Learning Application: Water Quality Resilience Prediction Model,” *Science of the Total Environment*, vol. 768, p. 144459, 2021.

[10] H. Lu and X. Ma, “Hybrid Decision Tree-Based Machine Learning Models for Short-Term Water Quality Prediction,” *Chemosphere*, vol. 249, p. 126169, 2020.

[11] E. B. Rachid et al., “Predicting Water Potability Using a Machine Learning Approach,” *Results in Engineering*, 2025.

[12] M. Zhu et al., “A Review of the Application of Machine Learning in Water Quality Evaluation,” *Environmental Research Communications*, 2022.

[13] A. Lokman et al., “A Review of Water Quality Forecasting and Classification Using Machine Learning,” *Water*, vol. 17, no. 15, article 2243, 2025.

[14] Y. Durgun et al., “Real-Time Water Quality Monitoring Using AI-Enabled Sensors,” *Journal of King Saud University – Science*, 2024.

[15] P. V. Sawant and Y. M. Patil, “Water Quality Monitoring Using Machine Learning Model,” *Journal of Electrical Systems*, vol. 20, no. 10s, pp. 5686–5694, 2024.