

# A decision support machine learning tool for environmental bioremediation as water safety in India

*Abhijit Debnath*<sup>1\*</sup>, *Satadru Jati*<sup>1</sup>, *Anuran Bhaumik*<sup>1</sup>, and *Mrinmay Das*<sup>1</sup>

<sup>1</sup>Department of Computer Science & Engineering, Institute of Engineering and Management, Newtown, UEM, Kolkata, B3, Newtown Road, Action Area- III, West Bengal, India, 700160.

**Abstract.** This study presents a novel, interpretable Decision Support System that integrates Stacking Ensemble learning with Shapley Additive exPlanations to predict water potability and recommend bioremediation strategies. Preparation and validation based on Indian datasets with over 4000 samples was carried out and the proposed Stacking Ensemble (RF, XGBoost, SVM, KNN, LR) achieved the highest accuracy of 74.0%, with Area under the Curve of 0.806, being at about ten percent above other classifiers, and maintaining state-of-the-art level standards. Further beyond prediction, the framework locates significant contamination drivers (such as pH and Sulfates) and maps them to viable ecological treatments such as phytoremediation and microbial degradation, which recent studies lack. This study demonstrates a scalable, transparent AI-driven pathway for real-time water quality management in resource-constrained environments.

**Keywords:** Water potability, Machine Learning, Physiochemical indicator, Environmental bioremediation, Decision support system.

---

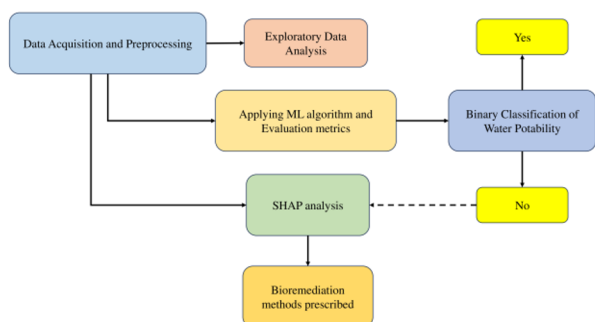
\* Corresponding author: [abhjit.dbnath@gmail.com](mailto:abhjit.dbnath@gmail.com)

## 1. Introduction

Access to safe water remains a challenge in India, with huge amounts of agricultural and industrial wastes released undetected and untreated. Most works examine the sources, toxicity, and remediation of contaminated water, highlighting physico-chemical treatment as a promising solution [1]. Some existing works [2] showed progress with stacking ensemble and CATBOOST. Many works in water quality prediction show increased risk of overfitting and also doesn't perform well on unseen data as their datasets are old or have limited data [2]. Many existing lab-tested tools are limited in scope, creating a need for a framework that can evaluate the potability level as well as recommend procedures for treating non-potable water. Recent works show that SHAP combined with XGBoost can handle complex spatial and non-spatial relationships [3, 4]. This study tries to fill the gap between detecting water quality and recommendations for treatment.

## 2. Methodology

We propose an interpretable framework for our work that integrates preprocessing, ensemble machine learning, and SHAP-driven bioremediation decision support system. The entire operational pipeline is visualized in **Fig. 1**.



**Fig. 1.** Integrated workflow of the AI-Driven Decision-Support framework.

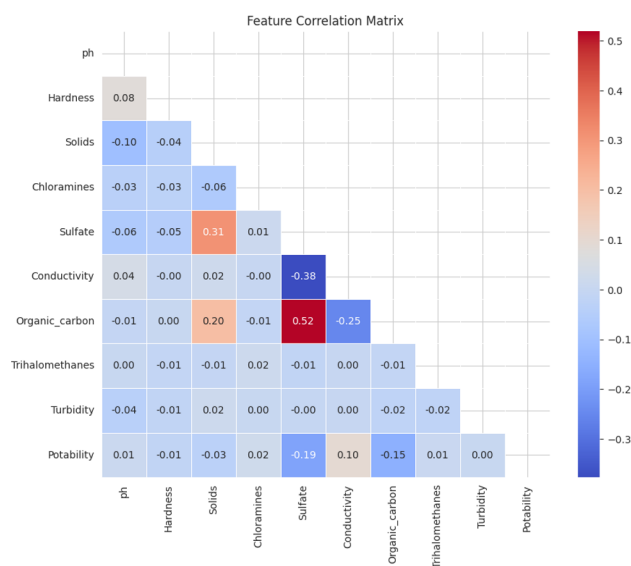
### 2.1 Data acquisition and preparation

To ensure statistical robustness and regional relevance, we used a primary dataset of 3,276 samples [6] characterized by nine physicochemical indicators, further expanded on real-world Indian surface water data (CPCB, 2021–2023) [5]. Our preprocessing pipeline first addressed missing values (e.g., 14.9% in pH) through class-specific mean imputation to retain central tendencies, followed by Standard Scaling to normalize feature ranges. Next, we corrected the dataset's native imbalance, which originally contained 1,278 potable (39%) versus 1,998 non-potable (61%) samples. By applying the Synthetic Minority Over-sampling Technique (SMOTE), we expanded the dataset to a balanced distribution of records per class, enabling the model to learn safety patterns with equal priority and

preventing the majority-class bias common in these works.

### 2.2 Model strategy and evaluation

EDA found minor linear correlations between parameters (**Fig. 2**). This absence of multicollinearity explains why linear models such as Logistic Regression failed; the linear model cannot comprehend the non-linear complexity of water contamination. So, we utilized non-linear learners such as Random Forest and XGBoost [3] to analyze these complex environmental signals.



**Fig. 2.** Correlation matrix of ‘Physicochemical’ features.

To ensure the robustness essential for public health, we used Stacking Ensemble that actively corrected the errors of individual models, achieving a 74% accuracy. But we went beyond raw prediction by integrating SHAP analysis to dismantle the “black box” nature of AI algorithms. This transparency allows us to isolate the specific chemical culprits.

## 3. Results

We examined the physicochemical profile of the samples to investigate the environmental complexity and to train our predictive model. In addition, we established statistical baselines by computing the central tendency and variability for both classes:

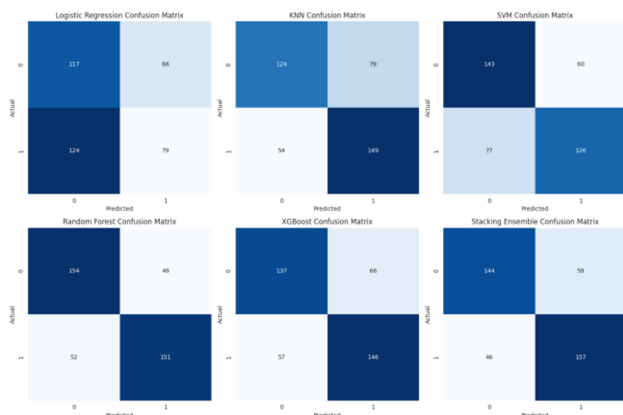
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \sigma = \sqrt{\sigma^2} \quad (1)$$

The data analysis of distribution showed significant environmental shifts. Differences in pH, Solids, Chloramines, Sulphate, Trihalomethanes, etc. were found between safe and unsafe samples. In a correlation analysis, the multicollinearity was found to be low, so the single feature only has some independence, however, the non-linear relationship between the variables requires strong modeling with e.g. tree-based ensembles rather than simple linear

regressors.

### 3.1 Model performance and selection

We verified: Logistic Regression, KNN, SVM, Random Forest, XGBoost using 5-fold cross-validation and standard metrics. Of these, Random Forest emerged as the most effective with 73% accuracy (AUC = 0.805) capturing the non-linear biological signatures of water data. While, Logistic Regression was close to random, achieving 50.7% accuracy, but could not generalize to complex environmental variables. Then, to alleviate these restrictions, we developed a Stacking Ensemble that integrates linear, distance-based, and tree-based learners as to be able to generate more accurate predictions, as the combination method attained the highest stability and prediction accuracy. Fig. 3 presents the confusion matrices contrasting the error distribution of the six models to clearly demonstrate the superior performance of the Stacking Ensemble in accuracy reduction in false classifications.



**Fig. 3.** Confusion matrices demonstrating the comparison of all the models.

### 3.2 Bioremediation diagnostic and real-world application

The “Black Box” predictions were turned into usable environmental insights with SHAP (Shapley Additive exPlanations). This approach quantitatively measures the contribution of each physicochemical parameter in the diagnosis, mathematically estimated

as:

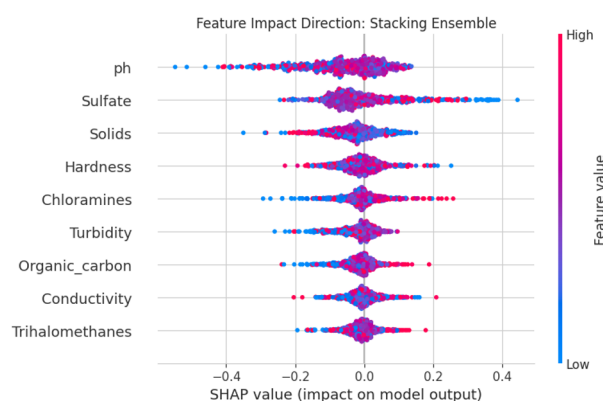
$$\Phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (2)$$

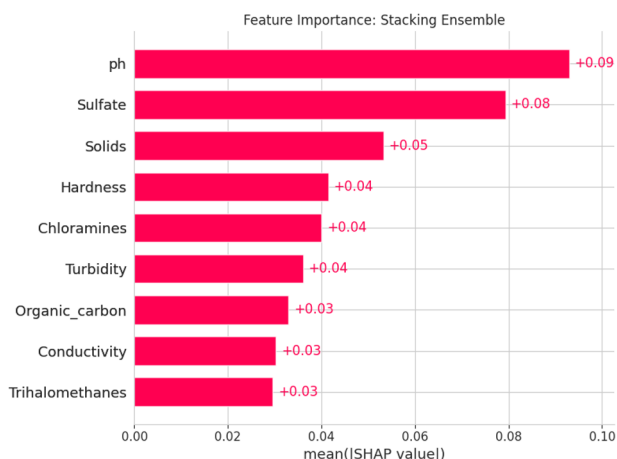
As shown in Fig. 4, pH and sulphate were water potability drivers, with Solids and Hardness ranking last. This shows that the model follows environmental chemistry principles, as acidity and dissolved salts play an important role in water safety.

**Table 1** provides us information on comparative performance metrics.

**Table 1.** Final performance metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.5075	0.5075	0.5075	0.5075	0.5341
K-Nearest Neighbors	0.6725	0.6612	0.7075	0.6836	0.7270
Support Vector Machine	0.6500	0.6508	0.6475	0.6491	0.7068
Random Forest	0.7300	0.7300	0.7300	0.7300	0.8050
XGBoost	0.6888	0.6855	0.6975	0.6914	0.7569
Stacking Ensemble	0.7400	0.7360	0.7220	0.7288	0.8060





**Fig. 4.** Stacking Ensemble Explainability. (a) SHAP Summary Plot showing feature impact distribution; (b) Feature Importance Bar Plot highlighting pH and Sulphate as key drivers.

To assess the usability of our system with field biotechnologists, we prepared a custom water sample with elevated turbidity (5.8 NTU), extreme solids (45,000 mg/L), and elevated organic carbon (15 mg/L). This sample was correctly rated as “Not Potable” by our model. In addition, the SHAP diagnostic module has helped in separating certain pollutants and correlating them to the correct bioremediation methods (see Table 2).

**Table 2.** Interpretability & Remediation Output for Real-Life Case

Parameter	Value	SHAP Impact	Diagnosis	Recommended Bioremediation
Solids	45,000.00	0.2213	Severe inorganic contamination	Phytoremediation / Ion-exchange
pH	6.80	0.1242	Sub-optimal acidity	Neutralization (Liming)
Organic Carbon	15.00	0.0878	High microbial growth risk	Microbial Degradation / Biofilm Treatment

For a broader aspect of our study, we compared our work and novelty against the existing State-of-the-Art works on similar grounds and ideas is presented in Table 3.

We analyzed the time complexity of our model and found the training complexity is dominated by the Random Forest and XGBoost, approximating  $O(N \cdot M \cdot \log(N))$  where  $N$  is sample size and  $M$  is feature count. But, the inference time (prediction speed) is  $O(M \cdot T)$ , where  $T$  is the number of decision trees. The average inference time per sample in our studies was in milliseconds, stating that the model is lightweight enough to be used on edge devices.

**Table 3.** Comparison with State-of-the-Art (SOTA) Methods

Study	Methodology	Focus	Limitations Addressed by Our Work
Nasir et al. (2022)	Random Forest, KNN	Classification Accuracy	“Black Box” predictions; No remediation suggestions.
Dodig et al. (2024)	LSTM, XGBoost	Time-series Prediction	High complexity; requires historical temporal data not available in rural surveys.
Proposed Method	Stacking Ensemble, SHAP	Decision Support & Bioremediation	Provides interpretability (Why is it unsafe?) and Actionability (How to fix it?).

## 4. Conclusions

Our proposed system provides both potability classification by predictive modelling and practical guidance regarding contamination treatment through SHAP-based explanation and bioremediation mapping. Even in the absence of microbial or geospatial data, the framework clearly offers scope to support scalable, transparent water quality decision support when applied to resource-limited regions. SHAP made the solution explainable and we were able to follow the decision to indicators of high organic carbon or other stress factors. So as well as just pointing out that the water is unsafe, the model gives us hints of the type of bioremediation needed. The approach also utilizes chemical proxies, so it still lacks the sensitivity of a direct microbial genomic test. A further drawback is that no dataset depicts all aspects of

the broad hydrological diversity in India. Thus, to facilitate real-time deployment, we propose a physical architecture with a sensing layer with IoT probes (pH, TDS and turbidity sensors) inside water bodies, an edge processing framework with an Arduino/Raspberry Pi microcontroller hosting our model, and an action layer where the model sends a bioremediation protocol directly to a mobile dashboard for field engineers.

## References

1. M. Rafique, S. Hajra, M.B. Tahir, S.S.A. Gillani and M. Irshad, A review on sources of heavy metals, their toxicity and removal technique using physico-chemical processes from wastewater. *Environ. Sci. Pollut. Res.* **29**, 16772–16781 (2022).  
<https://doi.org/10.1007/s11356-022-18638-9>
2. N. Nasir, A. Kansal, O. Alshaltone, F. Barneih, M. Sameer, A. Shanableh and A. Al-Shamma'a, Water quality classification using machine learning algorithms. *J. Water Process Eng.* **48**, 102920 (2022).  
<https://doi.org/10.1016/j.jwpe.2022.102920>
3. Z. Li, Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Comput. Environ. Urban Syst.* **96**, 101845 (2022).  
<https://doi.org/10.1016/j.compenvurbsys.2022.101845>
4. R. Dwivedi, D. Dave, H. Naik, S. Singhal, O. Rana, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan et al., Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Comput. Surv.* **55**, 1–33 (2023).  
<https://doi.org/10.1145/3561048>
5. R. Chitloor, Indian Water Quality Data (2021–2023). Kaggle Repository (2023).  
<https://www.kaggle.com/datasets/rishabchitloor/indian-water-quality-data-2021-2023>  
(Accessed: December 2025)
6. N. Pourmoradi, Water Potability Analysis. Kaggle Repository.  
<https://www.kaggle.com/code/nimapourmoradi/water-potability>  
(Accessed: December 2025)