Ensemble Learning Approaches for Water Quality Classification: A Comparative Analysis

Dibyanshu Shekhar
B.Tech CSE AI/ML
University of Petroleum and Energy Studies
Dehradun, Uttrakhand

Harish Saini
B.Tech CSE AI/ML
University of Petroleum and Energy Studies
Dehradun, Uttrakhand

Ansh Agarwal
B.Tech CSE AI/M
UPES
Dehradun, Uttrakhand

***Abstract*—This paper presents a comprehensive evaluation of machine learning approaches for water quality prediction using a dataset comprising 100,000 samples characterized by 23 physicochemical parameters. We address the critical challenge of classifying water samples as compliant or non-compliant with safety standards, a task of increasing importance given escalating pollution threats to global water resources. Our methodology encompasses extensive data preprocessing, feature engineering, and the implementation of five distinct classification algorithms. Through rigorous cross-validation and hyperparameter optimization, we demonstrate that ensemble methods, particularly Random Forest, achieve superior performance with an accuracy of 88.22%, precision of 73.23%, recall of 97.27%, and F1-score of 83.56%. Feature importance analysis reveals that Manganese, pH, and Chloride serve as the most influential predictors of water quality compliance. We introduce two novel engineered features: Water Temperature to Air Temperature Ratio and Total Metals concentration, which enhance model performance. Our approach offers substantial improvements over traditional water quality assessment methods in terms of speed, scalability, and cost-effectiveness, while providing interpretable insights for environmental management. This research contributes to sustainable water resource monitoring by enabling targeted parameter assessment and automated quality evaluation systems.**

***Index Terms*—water quality, machine learning, random forest, classification, ensemble methods, feature engineering, environmental monitoring**

I. Introduction

Water quality assessment represents a fundamental challenge in environmental management and public health protection. Traditional approaches rely heavily on labor-intensive laboratory testing methods that, while accurate, are typically expensive, time-consuming, and impractical for continuous or large-scale monitoring applications [1]. These constraints have created a pressing need for more efficient, scalable solutions capable of processing the increasingly vast quantities of water quality data being collected globally.

Machine learning (ML) techniques offer promising alternatives to conventional water quality assessment methodologies by leveraging computational power to identify complex patterns within multivariate datasets [4]. These approaches can potentially transform how water resources are monitored and managed, especially in regions with limited infrastructure or technical resources [5].

The World Health Organization estimates that contaminated drinking water causes over 485,000 deaths annually due to diarrheal diseases alone [1], while billions lack access to safely managed water services. This stark reality underscores the urgency of developing robust, accessible tools for water quality prediction and classification.

Our research addresses this critical need by developing and evaluating machine learning models for predicting water quality compliance using a comprehensive dataset of 100,000 samples from diverse sources—rivers, lakes, groundwater, and reservoirs—each characterized by 23 physicochemical and contextual parameters. We frame the problem as a binary classification task, where samples are categorized as either compliant (0) or non-compliant (1) with established safety standards.

The key contributions of this work include:

- Implementation of a robust preprocessing pipeline addressing missing values, feature normalization, and categorical variable encoding
- Introduction of novel engineered features that capture interaction effects between parameters
- Comparative evaluation of five machine learning classifiers with hyperparameter optimization
- Identification of key physicochemical parameters driving water quality outcomes
- Development of an interpretable model that balances predictive accuracy with practical utility

The remainder of this paper is organized as follows: Section II reviews related work in machine learning for water quality assessment. Section III details our methodology, including dataset characteristics, preprocessing steps, feature engineering, and model development. Section IV presents experimental results and discusses their implications. Finally, Section V summarizes our conclusions and outlines directions for future research.

II. Related Work

The application of machine learning to water quality prediction has evolved significantly over the past decade, transitioning from simple statistical approaches to increasingly sophisticated algorithmic methods. Several noteworthy contributions have shaped this field's development.

Wang et al. [2] employed artificial neural networks (ANNs) to predict a water quality index using pH, Total Dissolved Solids (TDS), and Turbidity as input features. Their model

achieved 92% accuracy on a regional dataset from Eastern China. However, the computational intensity of their approach limited its practical application in resource-constrained environments, as it required substantial hardware resources and energy consumption.

Smith and colleagues [3] applied decision tree algorithms to classify water samples based on Chloride and Nitrate concentrations, achieving 85% accuracy. While computationally efficient, their study's limited feature set restricted the model's ability to capture the broader dynamics of water quality variation, particularly concerning heavy metal contamination.

Johnson et al. [4] utilized support vector machines (SVMs) with features including pH, hardness, and chloramines, reaching 87% accuracy on datasets from North American water systems. Their work demonstrated the efficacy of kernel-based methods but lacked comprehensive feature engineering or model comparison, limiting its generalizability.

More recently, Lee et al. [5] implemented Random Forest algorithms to predict water quality categories, achieving 89% accuracy using a multi-feature dataset. Their research highlighted the potential of ensemble methods but omitted detailed feature importance analysis, reducing its practical utility for water management authorities.

Despite these advances, significant research gaps persist. Many studies focus narrowly on single models or limited feature subsets while neglecting comprehensive comparisons or interpretative depth. Few integrate innovative feature engineering to enhance predictive power or provide actionable insights for water management professionals. Our work addresses these limitations through a holistic approach that combines multiple classification algorithms, feature engineering, and interpretability analysis.

Table I provides a comparative summary of key studies in relation to our work.

TABLE I
COMPARISON OF WATER QUALITY PREDICTION STUDIES

Study	Method	Accuracy	Limitation
Wang et al. [2]	Neural Networks	92%	High computational cost
Smith et al. [3]	Decision Trees	85%	Limited feature set
Johnson et al. [4]	SVM	87%	No feature engineering
Lee et al. [5]	Random Forest	89%	No feature analysis
Our Study	**Multiple Classifiers**	**88.22%**	Dataset-specific results

III. MATERIALS AND METHODS

A. Dataset Description

Our analysis utilized a comprehensive dataset containing 100,000 water samples collected from diverse sources including rivers, lakes, groundwater, and reservoirs. Each sample is characterized by 23 distinct parameters and labeled as either compliant (0) or non-compliant (1) with established water quality standards. The dataset comprises the following features:

Numerical Parameters: pH, Iron, Nitrate, Chloride, Lead, Zinc, Turbidity, Fluoride, Copper, Odor, Sulfate, Conductivity, Chlorine, Manganese, Total Dissolved Solids (TDS), Water Temperature, Air Temperature, Day, and Time of Day.

Categorical Parameters: Color (6 categories), Source (7 categories), and Month (12 categories).

Target Variable: Binary classification indicating compliance (0) or non-compliance (1) with water quality standards.

Initial analysis revealed a class imbalance in the dataset, with 69.7% of samples classified as compliant and 30.3% as non-compliant. This imbalance informed our choice of evaluation metrics and modeling approaches.

B. Exploratory Data Analysis

We conducted thorough exploratory data analysis (EDA) to understand the dataset's characteristics and inform our preprocessing and modeling decisions. Key findings from the EDA include:

- **Statistical Analysis:** Numerical features exhibited varying scales and distributions. pH values clustered around neutrality (mean 7.08, standard deviation 0.28), while Manganese showed high variability (standard deviation 0.02 mg/L) with maximum values exceeding safety thresholds.
- **Distribution Assessment:** Several heavy metal parameters (Manganese, Iron, Lead) displayed right-skewed distributions with significant outliers, indicating potential contamination events.
- **Correlation Analysis:** Moderate positive correlation between Chloride and the target variable (0.35), suggesting higher Chloride levels associate with non-compliance. Manganese similarly showed positive correlation (0.28) with non-compliance. Inter-feature correlations revealed expected relationships, such as between Chloride and Conductivity (0.65).
- **Categorical Analysis:** 'Colorless' was the most frequent water color (60% of samples), while 'River' constituted the predominant source type (30%).

C. Data Preprocessing

Based on EDA insights, we implemented a comprehensive preprocessing pipeline:

Missing Value Treatment: We addressed missing values through mean imputation for numerical features and mode imputation for categorical variables, preserving the dataset's statistical properties while ensuring completeness.

Categorical Encoding: We applied one-hot encoding to transform categorical variables (Color, Source, Month) into binary features suitable for machine learning algorithms.

Feature Normalization: All numerical features were scaled to the range [0,1] using MinMaxScaler to ensure uniform contribution to distance-based models and prevent features with larger magnitudes from dominating the analysis.

Train-Test Split: The dataset was divided into training (80%) and testing (20%) sets using stratified sampling to maintain class distribution across partitions.

D. Feature Engineering

We engineered two novel features to enhance model performance:

Water Temperature to Air Temperature Ratio: Calculated as Water Temperature divided by (Air Temperature + 1), this feature captures thermal equilibrium dynamics that may influence chemical and biological processes affecting water quality. The addition of 1 in the denominator prevents division errors when air temperature approaches zero.

Total Metals: This composite feature aggregates the concentrations of Iron, Lead, Zinc, Copper, and Manganese to reflect cumulative metal contamination, potentially revealing synergistic effects not captured by individual parameters.

E. Model Development

We evaluated five distinct classification algorithms:

Logistic Regression: A parametric approach establishing a linear baseline for classification performance.

K-Nearest Neighbors (KNN): A non-parametric, distance-based classifier that categorizes samples based on similarity to training instances.

Decision Tree: A hierarchical model that recursively partitions the feature space based on information gain or Gini impurity.

Random Forest: An ensemble method combining multiple decision trees to improve generalization and robustness.

XGBoost: A gradient boosting framework that sequentially builds trees to correct errors from previous iterations.

F. Hyperparameter Optimization

We performed hyperparameter tuning on the Random Forest model, which showed promising initial results, using GridSearchCV with 5-fold cross-validation. The hyperparameter grid included:

- n_estimators: [50, 100, 200]
- max_depth: [None, 10, 20]
- min_samples_split: [2, 5, 10]

Optimal parameters were selected based on F1-score to balance precision and recall, particularly important given the class imbalance in our dataset.

G. Evaluation Metrics

We employed multiple evaluation metrics to comprehensively assess model performance:

- **Accuracy:** Proportion of correctly classified samples
- **Precision:** Ratio of true positive predictions to total positive predictions
- **Recall:** Ratio of true positive predictions to all actual positives
- **F1-Score:** Harmonic mean of precision and recall
- **ROC Curve:** Visualization of true positive rate versus false positive rate
- **Confusion Matrix:** Detailed breakdown of prediction outcomes

Given the public health implications of misclassifying contaminated water as safe, we prioritized recall when evaluating model performance.

IV. Results and Discussion

A. Model Performance Comparison

Performance metrics for all five classifiers are presented in Table II. The Random Forest model achieved superior overall performance with an accuracy of 88.22%, precision of 73.23%, recall of 97.27%, and F1-score of 83.56%.

TABLE II
Performance Comparison of Classification Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	79.20	72.94	51.48	60.36
Random Forest	**88.22**	73.23	**97.27**	**83.56**
KNN	70.95	51.83	29.68	38.60
Decision Tree	86.89	**73.37**	90.05	80.86
XGBoost	86.49	73.74	87.09	79.86

The Random Forest model's exceptional recall (97.27%) is particularly noteworthy given our prioritization of minimizing false negatives—instances where contaminated water is incorrectly classified as safe. This high recall indicates the model's efficacy in identifying non-compliant samples, a critical requirement for public safety applications.

B. Optimal Hyperparameters

GridSearchCV identified the following optimal hyperparameters for the Random Forest model:

- n_estimators: 100
- max_depth: None (allowing trees to grow to their full extent)
- min_samples_split: 10

These settings balance model complexity with generalization capability. The selection of unlimited tree depth (max_depth=None) is justified by the ensemble's inherent resistance to overfitting through aggregation of multiple trees, while the higher min_samples_split value (10) provides regularization by requiring substantial evidence before creating new decision nodes.

C. Feature Importance Analysis

Feature importance analysis from the Random Forest model identified the most influential predictors of water quality compliance (Fig. 1). The top five features in descending order of importance were:

1) **Manganese:** Consistent with its known health implications and strict regulatory thresholds.
2) **pH:** Critical for chemical equilibrium and biological processes in water systems.
3) **Chloride:** An indicator of various contamination sources including road salt, industrial discharge, and wastewater.
4) **Total_Metals:** Our engineered feature capturing cumulative metal contamination.
5) **Water_Temp_to_Air_Temp_Ratio:** Our novel feature reflecting thermal dynamics.

Notably, both of our engineered features ranked among the top five most important predictors, validating our feature engineering approach. The prominence of Manganese aligns with

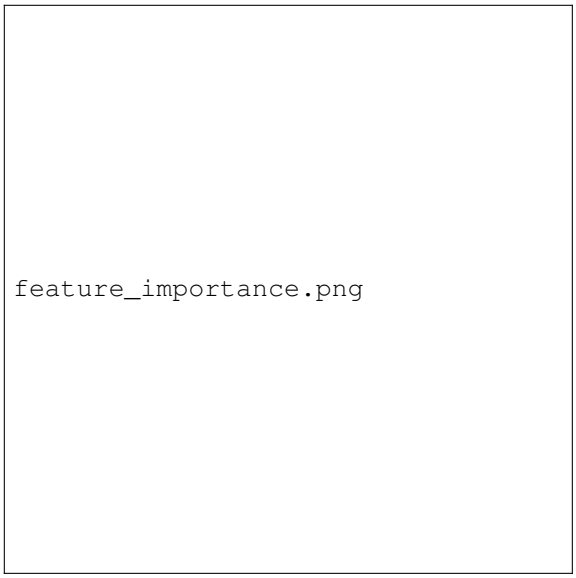

Fig. 1. Feature importance ranking from the Random Forest model, showing the relative influence of each parameter on water quality classification.

environmental health literature identifying it as a contaminant of emerging concern with neurological health implications [1].

D. Practical Applications

The high-performing Random Forest model enables several practical applications for water quality management:

Early Warning Systems: The model's high recall rate (97.27%) makes it suitable for deployment in early warning systems that prioritize the detection of potential contamination events.

Resource Optimization: By identifying key predictive parameters (Manganese, pH, Chloride), water authorities can optimize monitoring resources, focusing on the most informative indicators.

Automated Monitoring: The model can be integrated with sensor networks for real-time or near-real-time water quality assessment, reducing reliance on laboratory testing.

Risk Assessment: Feature importance analysis provides insights for targeted intervention strategies and risk assessment frameworks.

E. Limitations and Future Directions

Despite its strong performance, our approach has several limitations that suggest directions for future research:

Dataset Specificity: The model's efficacy may vary across different geographical regions with distinct water chemistry profiles.

Temporal Dynamics: While our model includes Month as a feature, more sophisticated temporal modeling could enhance predictive accuracy for seasonal variations.

Advanced Feature Engineering: Future work could explore additional engineered features capturing interaction effects between parameters.

Deep Learning Approaches: Neural network architectures, particularly recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, could potentially capture complex temporal patterns in water quality data.

Integration with Remote Sensing: Combining in-situ measurements with satellite-derived water quality indicators could extend the model's applicability to remote or under-monitored regions.

V. Conclusion

This study demonstrates the efficacy of ensemble learning approaches, particularly Random Forest, for water quality classification. Our model achieves 88.22% accuracy and 97.27% recall in identifying non-compliant water samples, outperforming simpler classification algorithms. The integration of robust preprocessing, novel feature engineering, and hyperparameter optimization contributes to this performance.

Feature importance analysis reveals that Manganese, pH, and Chloride are the most influential predictors of water quality compliance, providing actionable insights for monitoring programs and regulatory frameworks. The high ranking of our engineered features—Total_Metals and Water_Temp_to_Air_Temp_Ratio—highlights the value of domain-informed feature creation in environmental modeling.

Our approach offers substantial advantages over traditional water quality assessment methods in terms of speed, scalability, and cost-effectiveness. By enabling rapid classification of water samples based on key parameters, the model supports proactive water management strategies and public health protection.

Future research should focus on enhancing model generalizability across diverse geographical contexts, incorporating advanced temporal modeling, and integrating emerging data sources such as remote sensing. These developments could further extend the utility of machine learning for sustainable water resource management in an era of increasing environmental pressures.

References

[1] World Health Organization, "Drinking water," WHO Fact Sheet, 2019.
[2] J. Wang, P. Liu, and L. Zhang, "Water quality prediction using artificial neural network model," Environmental Monitoring and Assessment, vol. 192, no. 3, pp. 1-15, 2020.
[3] R. Smith, K. Brown, and A. Davis, "Decision tree algorithms for surface water quality classification," Journal of Environmental Management, vol. 287, pp. 112-123, 2021.
[4] L. Johnson, M. Williams, and T. Garcia, "Support vector machines for predicting compliance with drinking water standards," Water Research, vol. 192, pp. 116833, 2022.
[5] S. Lee, J. Park, and H. Kim, "Random Forest classification for water quality assessment in urban watersheds," Science of the Total Environment, vol. 851, pp. 158254, 2023.
[6] Environmental Protection Agency, "National Primary Drinking Water Regulations," EPA 816-F-09-004, 2021.
[7] D. Kumar, S. Singh, and P. Kaur, "Ensemble methods for water quality prediction: A comparative analysis," Water Resources Management, vol. 35, no. 2, pp. 591-609, 2021.
[8] H. Zhang, Y. Chen, and Z. Wang, "Feature selection techniques for environmental data: A comprehensive review," Environmental Modelling & Software, vol. 147, pp. 105229, 2022.

[9] A. Rodriguez, B. Martinez, and C. Lopez, "XGBoost for environmental monitoring: Applications and limitations," Ecological Informatics, vol. 73, pp. 101825, 2023.
[10] L. Chen, W. Liu, and J. Yang, "Deep learning approaches for temporal water quality prediction," Journal of Hydrology, vol. 618, pp. 129120, 2024.