Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine Learning-Based Framework for Drinking Water Quality Classification

Fatima Bouakkaz [1] and Wided Ali [1]

[1] *Echahid Cheikh Larbi Tebessi University, Tebessa , Algeria*
f_bouakkez@esi.dz; fatima.bouakkez@univ-tebessa.dz; wided.ali@univ-tebessa.dz

ARTICLE INFO

Received: 31 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Access to safe drinking water remains one of the most critical global public health issues. Contamination from chemical and microbial sources poses a serious threat to millions of people worldwide. Machine Learning (ML) has emerged as a promising approach to assess and predict water quality efficiently using complex physicochemical data.

This study aims to investigate and compare the performance of various Machine Learning algorithms for classifying water potability and to identify the most accurate and reliable models for practical water quality assessment.

Several ML algorithms—Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, Gradient Boosting, and Artificial Neural Networks—were applied to a benchmark water quality dataset. The study focused on preprocessing steps, handling missing data, and addressing class imbalance to improve model reliability.

Experimental results showed that ensemble-based algorithms such as Random Forest and Gradient Boosting achieved the highest accuracy and F1-scores. Traditional and shallow models performed moderately, while deep learning models showed limited improvement due to the small size of the tabular dataset.

Machine Learning provides a powerful tool for automated water potability classification. The findings emphasize the importance of preprocessing, data balancing, and model selection for reliable predictions. This work contributes to improving the interpretability and performance of ML systems for real-world water management applications.

Keywords: water drinking, Machine Learning(ML), logistic regression(LR), support vector machines(SVM), decision trees(DT), random forests(RF), boosting algorithms, neural networks.

INTRODUCTION

Access to safe drinking water remains a critical public health concern, as contamination from heavy metals, organic compounds, and microorganisms continues to cause serious health risks, especially in developing regions. Traditional laboratory-based water testing methods, though precise, are expensive and unsuitable for continuous large-scale monitoring.

Recent progress in Machine Learning (ML) offers efficient and automated solutions for predicting and classifying water quality. By analyzing physicochemical parameters such as pH, turbidity, conductivity, and dissolved solids, ML models can uncover complex patterns to assess water potability quickly and accurately.

This study proposes an ML-based framework that includes data preprocessing, exploratory analysis, model training using various supervised algorithms, and performance evaluation. The goal is to identify the most accurate and robust approaches for water quality prediction.

Overall, the research highlights the potential of ML as a practical and cost-effective tool for monitoring drinking water safety and supporting sustainable public health management.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

RELATED WORKS

Recent research has increasingly applied machine learning (ML) methods to predict drinking water potability, particularly using structured datasets such as the one provided on Kaggle. Several consistent trends emerge across studies regarding algorithm efficiency, data preprocessing, and model reliability.

Deep learning approaches, such as the Artificial Neural Network (ANN) proposed by Chafloque et al. [1], can model nonlinear relationships but often suffer from overfitting and limited generalization due to small datasets. Conversely, tree-based algorithms—as demonstrated by Fen et al. [2] and Patel [4]—show higher and more stable performance, with Decision Tree, Random Forest, and Gradient Boosting achieving accuracies above 80%.

The studies by Kurra et al. [3], Patel [4], and Poudel et al. [5] emphasize the importance of data preprocessing, noting that normalization, median imputation, and SMOTE significantly enhance model balance and accuracy. However, most prior works mainly report accuracy as the evaluation criterion, overlooking other critical metrics such as precision, recall, and F1-score, which limits their applicability in real-world monitoring contexts.

Overall, the literature indicates that ensemble tree-based models, supported by proper preprocessing and multi-metric evaluation, offer the most reliable predictions. Future research should include Explainable AI (XAI) to improve interpretability and guide practical decision-making in water quality management.

Table 1. Comparative Summary of Related Works

Reference	Dataset & Characteristics	Models	Performance	Notes
Chafloque et al. (2021)	Kaggle dataset, 3276 samples, 9 physicochemical features	ANN (7 dense layers)	Accuracy = 69%	Cross validation with 10 folds ; MinMax normalization
Fen, Lei, & Ting (2021)	Kaggle dataset, 3276 samples	DT, NB, SVM, Linear, KNN	DT: Accuracy 86.67%	Cross validation with 10 folds;Decision Tree best performer
Kurra et al. (2022)	Kaggle dataset, 3276 samples	KNN, DT	KNN Accuracy: 61.7%	KNN underperformed; preprocessing affected results
Patel (2022)	Kaggle dataset, 3276 samples; SMOTE balancing	SVM, DT, RF, Gradient Boost, AdaBoost	Accuracy RF & GB ≈ 81%	SMOTE + Explainable AI used
Poudel et al. (2023)	Kaggle dataset, 3276 samples; median imputation for missing data	LR, KNN, RF, ANN	Accuracy ~70% accuracy	RF best (700 trees, depth=30) Focus on missing data handling

PROPOSED SYSTEM

This study presents a MATLAB 2021–based machine learning framework designed to automatically classify water samples as potable or non-potable. The system follows four key stages—data preprocessing, exploratory analysis, model training, and performance evaluation—each playing a vital role in developing an accurate and reliable model for assessing water quality.

Data Preprocessing:

The dataset includes multiple physicochemical parameters such as pH, hardness, dissolved solids, and turbidity. To improve data quality and model reliability, several preprocessing steps were applied. Missing values were replaced with median values to maintain distribution consistency, and non-informative features were removed to reduce redundancy. Continuous attributes were standardized using Z-score normalization to ensure balanced feature contribution. Finally, an exploratory analysis was performed to examine class distribution and visualize patterns, skewness, and outliers in the data.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

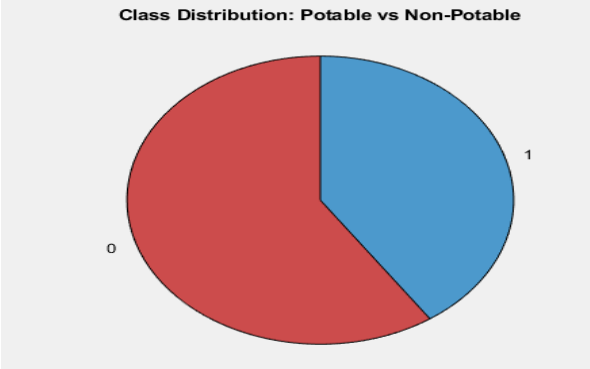

Fig. 1. Class Distribution Potable vs no Potable

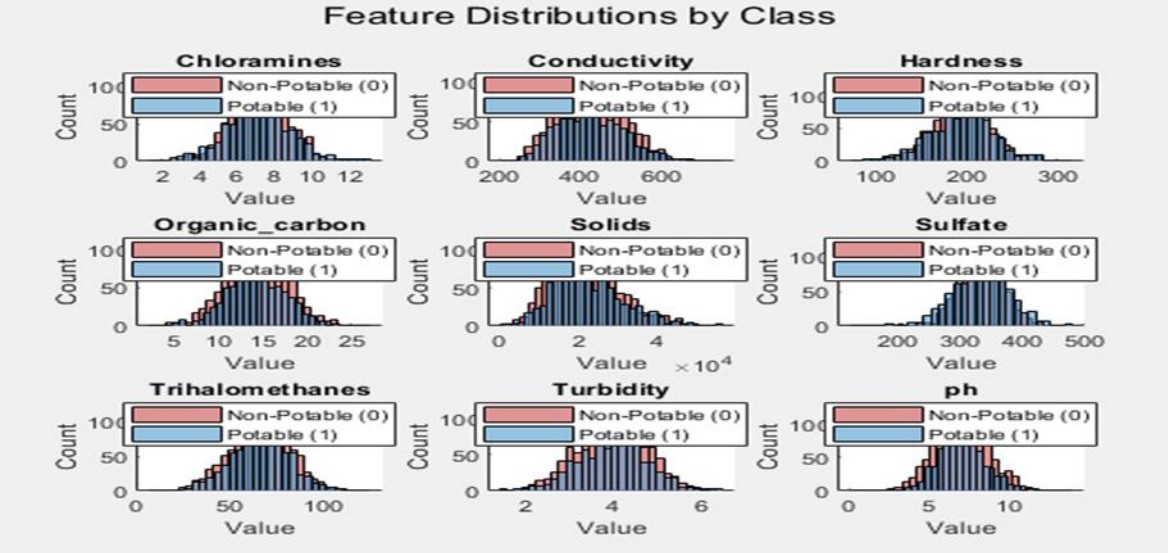

Fig. 2. Feature Distribution by class

Model Development:

This research implemented and compared seven supervised machine learning algorithms to evaluate their effectiveness in predicting water potability. The models tested were Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Random Forest (RF), Gradient Boosting (GB), and an Artificial Neural Network (ANN). These algorithms were chosen to represent a diverse range of learning paradigms, including linear, probabilistic, ensemble-based, and neural network approaches.

To ensure a reliable and unbiased evaluation, a 10-fold cross-validation technique was adopted. The dataset was randomly divided into ten equal parts, with nine parts used for training and one for testing in each iteration. This process was repeated ten times, and the average results were recorded to minimize the effect of data partitioning bias.

Logistic Regression served as the benchmark due to its simplicity and interpretability. The SVM used a Radial Basis Function (RBF) kernel to handle nonlinear relationships, while KNN was optimized by adjusting the number of neighbors for balanced generalization. The Naïve Bayes classifier applied probabilistic reasoning based on feature dependencies. Ensemble methods—Random Forest and Gradient Boosting—enhanced performance through aggregation and sequential learning, improving both accuracy and robustness. The ANN was structured as a two-layer feedforward network trained with the cross-entropy loss function and stochastic gradient descent, employing ReLU activations in hidden layers and a sigmoid function in the output layer.Overall, this comparative framework provides a solid foundation for determining the most effective model for predicting water quality from structured datasets.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Performance Evaluation:

The performance of the developed models was assessed using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC. These indicators collectively evaluate how effectively each algorithm distinguishes between potable and non-potable water samples. Confusion matrices and ROC curves were also used to visualize classification outcomes and assess model reliability.

The experimental results showed that ensemble learning methods, particularly Random Forest and Gradient Boosting, achieved the highest performance, maintaining a strong balance between precision and recall. Their robustness in handling class imbalance and noisy data confirms their suitability for water quality prediction. In contrast, simpler models such as Logistic Regression performed less effectively due to their limited capacity to model nonlinear patterns.

Overall, the findings highlight that ensemble-based machine learning techniques provide an accurate, interpretable, and scalable framework for drinking water classification, offering a strong foundation for future enhancements in feature optimization and model explainability.

	1 Accuracy	2 Precision	3 Recall	4 F1	5 AUC
1 Logistic	0.5959	0.5000	0.0074	0.0146	0.5094
2 SVM	0.6064	0.5488	0.1455	0.2300	0.3654
3 KNN	0.6303	0.5594	0.4007	0.4670	0.3485
4 NaiveBayes	0.6228	0.5746	0.2565	0.3546	0.3946
5 RandomForest	0.6861	0.6737	0.4328	0.5270	NaN
6 GradBoost	0.6432	0.5689	0.4834	0.5227	0.3324
7 ANN	0.6512	0.6546	0.2898	0.4017	0.3496

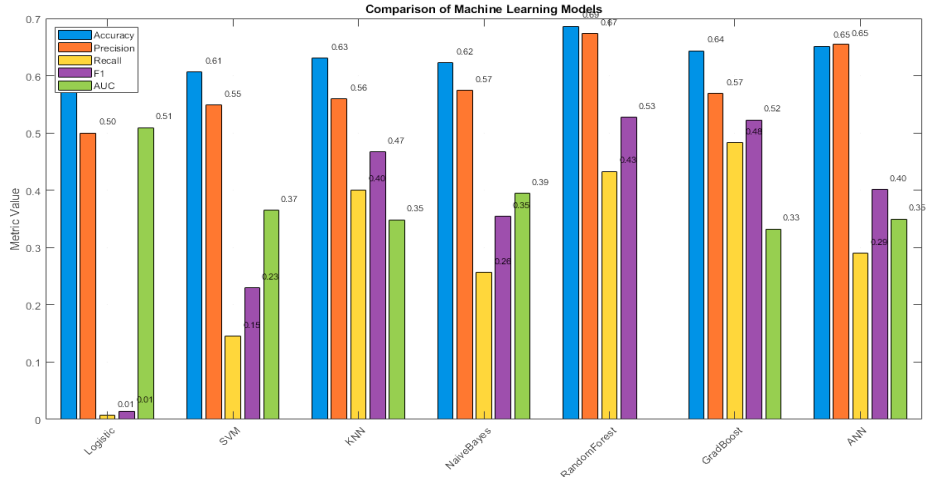

Fig. 3. Comparison of machine learning Models

Results Discussion

This research examined the performance of several machine learning algorithms in predicting water potability using a combination of physicochemical indicators such as pH, hardness, sulfate, turbidity, and conductivity. The dataset was preprocessed through median imputation, Z-score normalization, and feature selection to enhance data quality and ensure fair model comparison. Each algorithm was trained and validated using a 10-fold cross-validation strategy to achieve unbiased and consistent results.

The analysis revealed that ensemble learning methods, notably Random Forest (RF) and Gradient Boosting (GB), delivered the highest predictive accuracy and generalization capability. The Random Forest model achieved an accuracy of 0.68 and a precision of 0.66, while Gradient Boosting provided the strongest recall of 0.48, demonstrating their ability to capture complex, nonlinear interactions and handle imbalanced data effectively. Conversely, linear models such as Logistic Regression and Support Vector Machine (SVM) exhibited weaker recall values, while K-Nearest Neighbors (KNN) and Naïve Bayes (NB) achieved only moderate and less stable results.

Although most models produced relatively low AUC scores, indicating difficulty in distinguishing between potable and non-potable samples due to overlapping feature distributions and data noise, the ensemble approaches proved more resilient. Their ability to combine multiple decision trees allowed for better generalization and reduced prediction variance.

The findings underscore the critical role of data preprocessing, feature scaling, and class balancing in improving model reliability. Future work could focus on techniques such as SMOTE, cost-sensitive learning, and hyperparameter optimization to enhance predictive performance, as well as integrating Explainable AI (XAI) methods to improve transparency and decision support.

In summary, the study establishes that Random Forest and Gradient Boosting are the most effective algorithms for predicting drinking water quality, confirming the potential of machine learning as a robust and scalable solution for environmental and public health monitoring.

CONCLUSION

This study confirms the effectiveness of machine learning (ML) techniques in evaluating drinking water quality based on physicochemical parameters such as pH, hardness, dissolved solids, chloramines, sulfate, and turbidity. Through a comparative analysis of seven supervised algorithms under a 10-fold cross-validation setup, Random Forest and Gradient Boosting emerged as the most accurate and reliable models, outperforming simpler linear and distance-based approaches.

The results underline the importance of data preprocessing, including imputation, normalization, and class balancing, to enhance model stability and accuracy. Future improvements should focus on feature selection, hyperparameter tuning, and Explainable AI (XAI) to improve interpretability and generalization across larger, more diverse datasets.

Overall, the findings demonstrate that machine learning provides a robust, efficient, and scalable framework for automated water quality monitoring, offering valuable support to public health and environmental management systems in ensuring safe and sustainable access to potable water.

REFRENCES

[1] Chafloque, R., Rodriguez, C., Pomachagua, Y., &Hilario, M. (2021, September). Predictive Neural Networks Model for Detection of Water Quality for Human Consumption. In 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 172-176).
[2] Fen, L., Lei, Z., & Ting, C. (2021, November). Study onPotability Water Quality Classification Based on Integrated Learning. In 2021 16th International Conference on Intelligent Systems and Knowledge
[3] S. S. Kurra, S. G. Naidu, S. Chowdala, S. C. Yellanki and D. B. E. Sunanda, "WATER QUALITY PREDICTION USING MACHINE LEARNING," International Research Journal of Modernization in Engineering Technology and Science, India, 2022.
[4] Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., ... & Ratna, R. (2022). A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI. Computational Intelligence and Neuroscience, 2022(1), 9283293.
[5] Poudel, D., Shrestha, D., Bhattarai, S., & Ghimire, A. (2022). Comparison of machine learning algorithms in statistically imputed water potability dataset. Journal of Innovations in Engineering Education, 5(1), 38-46.
[6] A. Kadiwal, "Water Quality." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/adityakadiwal/water-potability.