Journal of Mines, Metals and Fuels, *73(6): 1765-1770; 2025. DOI: 10.18311/jmmf/2025/49103*

Print ISSN : 0022-2755

Journal of Mines, Metals and Fuels

Contents available at: *www.informaticsjournals.co.in/index.php/jmmf*

Optimising Water Quality Assessment: A Comparative Study of Machine Learning Algorithms for Imbalanced Potability Datasets

Srinivasulu Akasam[1]*, R. Rajasekar[2], Nareddy Sudha Rani[3], Niranjan A[4], Ch. Praveen Kumar[4], Garlapati Narayana[5], G. Gopi Krishna[6], Murali Gundagani[1]

[1]Geethanjali College of Engineering and Technology, Cheeryal, Hyderabad - 501301, Telangana, India; ak.srinivasulu@gmail.com

[2]Alliance University, Bengaluru - 562106, Karnataka, India

[3]Vignan Bharathi Institute of Technology, Aushapur - 501301, Telangana, India

[4]GITAM University, Rudraram - 502329, Telangana, India

[5]Chaitanya Bharathi Institute of Technology, Hyderabad - 500075, Telangana, India

[6]Marri Laxman Reddy Institute of Technology and Management, Hyderabad - 500043, Telangana, India

Abstract

This research demonstrates the power of machine learning in predicting water potability, showcasing improvements in prediction accuracy through hyperparameter tuning. Models like Random Forest and Gradient Boosting Machines outperform others, providing reliable insights into water safety. Hyperparameter tuning enhances model performance, optimising accuracy and generalisation. The study highlights key water quality parameters, such as pH and turbidity, for predicting potability. It also compares different algorithms to determine the best fit for water quality prediction tasks. This approach could revolutionise water monitoring, enabling real-time assessments and early contamination warnings. Machine learning models can scale globally, adapting to various water sources and conditions. This research emphasises the importance of technology in improving water management practices. Predictive models can guide policymakers in addressing water contamination risks and formulating regulations. Ultimately, machine learning offers a proactive, data-driven solution to global water quality challenges.

Major Findings: *Hyperparameter tuning and dataset balancing boost ML accuracy for water potability prediction. Random Forest performed best (89%), ahead of Bagging (87%) and XGBoost (83%). Optimized parameters and class imbalance handling improve reliability, suggesting deep learning potential..*

Keywords: *Accuracy Improvement, Class Imbalance Handling, Hyperparameter Tuning, Machine Learning Models, Model Evaluation Metrics*

1.0 Introduction

The availability of clean and safe drinking water is one of the most significant global challenges, with millions of people facing the threat of contaminated water sources. The quality of potable water is increasingly jeopardised by a range of pollutants, including agricultural runoff, untreated sewage, and industrial discharge. These contaminants can lead to severe health risks and environmental damage, necessitating rigorous monitoring

**Author for correspondence*

and rapid detection of water quality issues. Heddam[1] conducted tests on various transfer and kernel functions to develop both the ANN and SVM models. The results revealed that both models strongly predict water quality components. While effective, traditional water quality testing methods are often time-consuming and labour-intensive, making it difficult to identify contamination risks in real-time.

With the rise of data science and machine learning, there is growing potential to improve the efficiency and accuracy of water quality monitoring. Machine learning offers powerful tools to analyse large and complex datasets, identify patterns, and predict water contamination risks more quickly than traditional approaches. To build predictive models that can assess water quality more accurately and provide timely alerts to prevent potential risks. Sharma[2] highlighted that the XGBoost algorithm outperforms conventional machine learning models, achieving remarkable metrics such as 99.5% accuracy, 0.99 precision, 0.99 sensitivity, 1.0 specificity, and an F1 score of 0.99. In comparison, the Random Forest algorithm also showed good performance, reaching an accuracy of 74%. Hossain[3] explored the effect of different feature selection techniques on the performance of machine learning models for predicting water potability. The study underscored the importance of choosing relevant features to improve both the accuracy and efficiency of the models.

The application of machine learning to water quality prediction, focusing on optimising model performance through hyperparameter tuning. By exploring various machine learning techniques and optimising their parameters, the study aims to enhance the accuracy of water quality assessments and contribute valuable insights for managing and safeguarding water resources. The outcomes of this research could significantly improve water management practices and support decision-making in ensuring safe and accessible drinking water.

The research demonstrates how algorithms can identify intricate patterns in complex datasets, leading to better predictions about whether water is potable or not.

This research on predicting water potability using machine learning can greatly benefit human health by ensuring safer drinking water. Accurately identifying contamination risks allows for early detection of harmful substances in water, reducing health hazards like waterborne diseases. Machine learning models can help monitor water quality in real time, enabling timely interventions to prevent health crises. The research supports better resource allocation for water treatment, improving overall water safety. It can guide policymakers in enforcing stricter water quality standards to protect public health. Moreover, the technology can be applied in underserved areas, ensuring access to safe water where monitoring is limited. By analysing key water quality parameters, the models can detect patterns that may be missed by traditional methods. This research empowers water treatment plants to make more informed decisions, optimising processes for safer water. It also fosters a proactive approach to global water challenges, ensuring sustainable and healthy water resources for future generations.

Anwar[4] determined that machine learning algorithms can effectively predict water potability, with decision tree and random forest models demonstrating superior accuracy compared to other techniques. The study emphasised the necessity of choosing the most appropriate algorithm based on the dataset's characteristics. The findings underscored the potential of these models in improving water quality monitoring and protecting public health. Mahajan[5] discovered that the random forest algorithm achieved the highest accuracy in predicting water potability among various machine learning approaches. Support Vector Machine (SVM) and K-Nearest Neighbours (KNN) also performed well, but with slightly lower accuracy. This research highlighted the significance of feature selection and data preprocessing in optimising model performance. Patel *et al.*[6] conducted a comparative analysis of different machine learning methods, including SVM, Decision Tree (DT), Random Forest, Gradient Boosting, and AdaBoost, in the classification of water quality.Kouadri *et al.*[7] applied ML to predict WQI from irregular data in Illizi, Algeria. Uddin *et al.*[8] conducted a study to assess the effectiveness of the water quality index model in predicting water conditions. The research utilised several methods, such as SVM, DT, and KNN, to develop and evaluate the water quality index model. The results showed that the SVM algorithm outperformed

the others, achieving an accuracy rate of over 95%. Bilali *et al.*[9] developed and evaluated models using Adaptive Boosting, Random Forest, ANN, and SVR to predict parameters such as TDS, PS, SAR, ESP, MAR, and RSC, with physical parameters serving as the features.These studies, along with the substantial body of work detailed in the given references provide a robust foundation for the present research[10,11]. Elbeitagi *et al.*[12] proposed the AR model as an optimal approach, achieving satisfactory results due to its enhanced prediction accuracy with a minimal number of input parameters. This makes it a reliable and precise method for predicting the Water Quality Index (WQI) in the Akot basin. Salam *et al.*[13] compared these ensemble models with two widely used tree-based models, Random Forest (RF) and kernel-based models, using a daily climatic dataset from nine meteorological stations in Bangladesh spanning from 1983 to 2017. Zhu *et al.*[14] emphasised that the study's results have important implications for forecasting water levels and managing lake water resources, especially concerning the complexity of machine learning models. Valentini *et al.*[15] concluded that statistical methods are highly effective tools for monitoring water quality. These methods were instrumental in developing a WQI equation with fewer variables, making them valuable for managing water resources.

2.0 Methodology

2.1 Data Collection

The research begins by gathering a broad range of water quality data from diverse geographical locations, covering multiple water sources. The datasets contain various parameters such as pH, turbidity and other key water quality metrics. The data is sourced from public water quality databases, local environmental agencies, and academic research repositories.

2.2 Data Preprocessing

In this step, the collected data undergoes several preprocessing stages to improve the quality and usability for machine learning models. Missing data is handled using imputation methods like mean, median, or forward-fill techniques, depending on the nature of the variable.

2.3 Feature Engineering

The next step involves transforming raw water quality data into meaningful features that can enhance model performance. Feature selection is carried out using statistical tests and correlation analysis to identify which variables are most relevant to predicting contamination. Additional derived features, such as rolling averages or seasonal trends, are created to capture temporal patterns.

2.4 Model Selection

The study explores several machine learning algorithms to build predictive models for water quality.

2.5 Decision Trees

These are employed due to their interpretability and ability to model non-linear relationships.

2.6 Support Vector Machines (SVM)

SVM is used for its strength in binary classification tasks (safe vs. unsafe water) and its ability to handle complex decision boundaries.

2.7 Random Forests

This ensemble method is utilised to reduce overfitting and improve generalisation.

2.8 Hyperparameter Tuning

Grid search exhaustively searches over a manually specified hyperparameter space, testing combinations of values for key parameters like learning rate, max depth (for trees), and the number of estimators. Random search is employed to explore a broader range of hyperparameters by selecting random combinations, often yielding faster results with high accuracy. Cross-validation is used during this process to ensure that the tuning results are robust and not subject to overfitting on any specific subset of the data.

2.9 Prediction and Analysis

After training and hyperparameter tuning, the models are used to predict water quality for unseen data. The predictions are compared against actual water quality readings, and various analysis techniques, such as

confusion matrices for classification or regression error analysis for continuous predictions, are employed. Precision-recall curves and decision trees (in the case of decision tree models) are used to interpret and visualise model outcomes.

2.10 Deployment and Recommendations

The final optimised models are integrated into a decision-support system designed for water authorities, environmental agencies, and policymakers. This system provides real-time water quality predictions, alerts for potential contamination risks, and recommendations for interventions. The study also explores the possibility of future improvements, such as integrating additional data sources like satellite imagery, remote sensing, or IoT sensor networks, to enhance model accuracy and provide more timely assessments.

This methodology combines machine learning, hyperparameter optimisation, and interpretability to provide a comprehensive approach to monitoring and improving potable water quality, ensuring the ability to predict contamination risks and take preventive actions in a timely manner

3.0 Review of Results

The water potability dataset is intended to evaluate the safety of water for consumption based on various chemical properties. It contains attributes such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. The dataset is imbalanced, with a greater number of non-potable water samples than potable ones. Managing this imbalance is crucial to improving the accuracy and reliability of predictive models. Machine learning techniques are applied to analyse these key features and predict water potability.

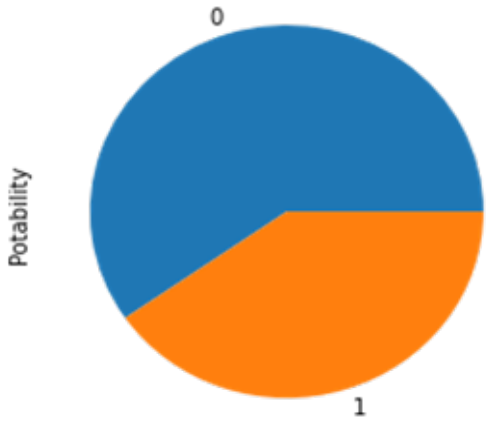
Table 1. Missing values in different variables

ph	491
Hardness	0
Solids	0
Chloramines	0
Sulfate	781
Conductivity	0
Organic_carbon	0
Trihalomethanes	162
Turbidity	0
Potability	0

Several columns in the dataset have missing values shown in Table 1, including pH, sulfate, and Trihalomethanes, while the other columns are fully populated. This indicates that the dataset is largely complete, but the missing values in pH, sulfate, and trihalomethanes may require preprocessing or handling. Since water quality data is sensitive, it is crucial not to alter the data by imputing values such as the mean, median, or mode.

3.1 Pie Diagram (Unbalanced)

The pie chart represents potable water availability, where the blue section indicates non-potable water, and the orange section represents potable water.



As shown in the Table 2, the accuracy scores of different machine learning models vary based on their ability to classify data correctly. Support Vector Machine (SVM) achieves the highest accuracy at 68.85%, followed by XGBoost at 67.09%. K-Nearest Neighbours (KNeighbours) and Decision Tree models perform slightly worse at 65.34% and 64.51%, respectively.

Table 2. Accuracy table for different machine learning models

Model	Accuracy score
SVM	0.688540
Xgboost	0.670980
knn	0.653420
DT	0.645102
AdaBoost	0.634011
Logistic Regression	0.628466
Random Forest	0.628466

AdaBoost, Logistic Regression, and Random Forest show lower accuracy, with the latter two having identical scores of 62.85%.

The above pie chart illustrates an imbalanced dataset, where Class 0 has significantly more instances than Class 1. Such an imbalance can lead to biased model predictions, favouring the majority class. To improve accuracy and fairness, techniques like resampling should be used to balance the dataset.

3.2 Pie Diagram (Balanced)

As shown in the Table 3, the provided values reflect the performance scores of various machine learning models. Logistic Regression has the lowest score of 0.51, suggesting its predictive capability is weaker. K-Nearest Neighbours, Decision Tree, and XGBoost exhibit

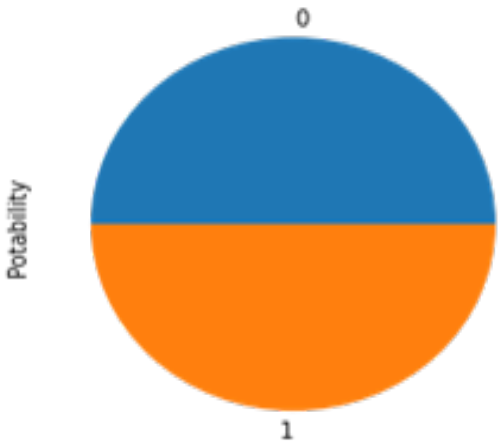


Table 3. Accuracy table for machine learning models (hyperparameter tuning)

Model	Accuracy
Logistic Regression	0.51
K Nearest Neighbours	0.79
Decision Tree	0.83
Random Forest	0.89
AdaBoost	0.69
Bagging Classifier	0.87
XGBoost : 0.83	0.83

moderate performance, and Random Forest attains the highest accuracy of 0.89. Bagging Classifier and AdaBoost also demonstrate strong performance, emphasising the effectiveness of ensemble learning techniques in improving model accuracy. In the Random Forest model, the precision for class 0 is 89%, while for class 1, it is 90%, indicating that the model correctly predicts 89% of class 0 and 90% of class 1. The first dataset, which is imbalanced, likely led to biased model predictions, favouring the majority class. In contrast, the second dataset, balanced using hyperparameter tuning, improved overall model performance. Logistic Regression showed minimal improvement, while models like Random Forest, XGBoost, and Bagging Classifier demonstrated higher accuracy. The balanced dataset allowed ensemble models to achieve better generalisation, reducing bias and improving classification results.

In summary, this research provides a data-driven, scalable solution for predicting water potability and enhances our understanding of how machine learning models can be integrated into water management systems to improve access to safe drinking water. The insights gained can guide decisions around water safety, treatment, and policy planning to mitigate water-related health risks.

4.0 Conclusion

The study highlights the significance of data preprocessing and hyperparameter tuning in improving the accuracy of machine learning models for water potability prediction. Initially, models trained on an imbalanced dataset without

tuning showed moderate accuracy, with SVM performing the best at 68.8%. However, after balancing the dataset and optimising hyperparameters, all models demonstrated substantial improvement. Random Forest achieved the highest accuracy at 89%, followed by Bagging Classifier at 87% and XGBoost at 83%. These results emphasise that addressing class imbalance enhances model reliability. Moreover, hyperparameter tuning optimises feature learning, leading to better generalisation. The comparison between pre- and post-tuning models illustrates the importance of refining machine learning techniques for critical applications. By employing proper data balancing techniques and parameter optimisation, predictive accuracy can be significantly enhanced. This study contributes to improving water quality assessments using machine learning. Future work can explore additional techniques, such as deep learning, for further advancements.

5.0 References

1. Heddam M. Water quality prediction using machine learning methods. Water Qual Res J. 2016; 53(1):38171.
2. Sharma A, Kumar P. Water potability prediction using machine learning algorithms. Proc Int Conf Comput Intell Data Sci; 2020. p. 1-6.
3. Hossain MS, Akhtar F. Feature selection techniques for machine learning-based water potability prediction. Proc Int Conf Comp Comm Signal Proc; 2021. p. 51-55.
4. Anwar A, Pratama M. Comparative study of machine learning algorithms for water potability prediction. Proc Int Conf Data Software Eng; 2021. p. 1-6.
5. Mahajan N, Narula GS. Comparative study of machine learning techniques for water potability prediction. J Water Environ Technol. 2022; 20(3):115-127.
6. Patel J, Amipara C, Ahanger TA, Ladhva K, Gupta RK, Alsaab HO, *et al.* A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable AI. Comput Intell Neurosci. 2022; 2022:e9283293. https://doi.org/10.1155/2022/9283293 PMid:36177311 PMCid:PMC9514946
7. Kouadri S, Elbeltagi A, Islam ARMdT, Kateb S. Performance of machine learning methods in predicting water quality index based on irregular data set: Application on Illizi region (Algerian southeast). Appl Water Sci. 2021; 11(12). https://doi.org/10.1007/s13201-021-01528-9
8. Uddin MG, Nash S, Rahman A, Olbert AI. Performance analysis of the water quality index model for predicting water state using machine learning techniques. Process Saf Environ Prot. 2023; 169:808-828. https://doi.org/10.1016/j.psep.2022.11.073
9. El Bilali A, Taleb A, Brouziyne Y. Groundwater quality forecasting using machine learning algorithms for irrigation purposes. Agricult Water Manag. 2021; 245:106625. https://doi.org/10.1016/j.agwat.2020.106625
10. Singh N, Kaur J, Thakur P, Murali G. Structural behaviour of annular isotropic disk made of steel/copper material with gradually varying thickness subjected to internal pressure, structural integrity and life. 2023; 23(3):293-297.
11. Bhati SM, Murali G, Sanjay Ch. Law of electric fields via Bianchi identities. J Appl Sci Eng. 2022; 25(4):759-762. https://doi.org/10.6180/jase.202208_25(4).0011
12. Elbeltagi A, Pande CB, Kouadri S, Islam ARM. Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra, India. Environ Sci Pollut Res. 2021; 29: 17591–17605.
13. Salam R, Islam ARMT. Potential of RT, Bagging and RS ensemble learning algorithms for reference evapotranspiration prediction using climatic data-limited humid region in Bangladesh. J Hydrol. 2020; 590:125241. https://doi.org/10.1016/j.jhydrol.2020.125241
14. Zhu S, Hrnjica B, Ptak M, Choinski A, Sivakumar B. Forecasting of water level in multiple temperate lakes using machine learning models. J Hydrol. 2020; 585:124819.
15. Valentini M, dos Santos GB, Muller Vieira B. Multiple Linear Regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul—Brazil. SN Appl Sci. 2021; 3:70. https://doi.org/10.1007/s42452-020-04005-1