RESEARCH ARTICLE

International Journal of Computer Communication and Informatics

DOI: 10.54646/ijcci.2025.01

WAPOT: Data Driven Approach for Water Potability Detection using Machine Learning

Saleem Raja Abdul Samad [a,*], Maria Rajesh Antony [b], Pradeepa Ganesan [a], Sathya Ramasamy [c], Madhubala Radhakrishnan [a], S. Sajithabanu [d]

[a] Department of Information Technology, College of Computing and Information Sciences Shinas, Sultanate of Oman, Oman

[b] Department of Civil Section, Engineering, College of Engineering and Technology Shinas, Sultanate of Oman, Oman

[c] Department of Information Technology, PSG College of Arts and Science Coimbatore, India

[d] Department of Information Technology, Mohammed Sathak Engineering College, Kilakarai, India

* Corresponding Author: saleem.abdulsamad@utas.edu.om

Received: 23-07-2025, Revised: 14-11-2025, Accepted: 21-11-2025, Published: 03-12-2025

Abstract: Water potability grading is crucial to public health and safety. It is a critical responsibility of regulatory authorities and water treatment facilities to guarantee that individuals have access to potable and secure drinking water, an inherent human right. The water potability classification is a preventative measure to detect potential impurities or contaminants that may present adverse health effects upon ingestion. This study examines a machine learning approach for classifying the potability of drinking water, utilizing ensemble learning methods (WAPOT) such as Stacking classifiers. Stacking, as a form of ensemble learning, consistently outperforms standalone classifiers and other existing research works, offering improved accuracy of 97% in potability classification. The findings underscore the capacity of machine learning to significantly contribute to the monitoring and managing of water treatment processes.

Keywords: Potability, Machine Learning, Water Quality, Ensemble Classifier, Stacking Classifier.

1. Introduction

Water pollution presents a global threat with farreaching consequences. It endangers ecosystems, public health, and socioeconomic stability [1]. The World Health Organization (WHO) emphasizes that water pollution is driven by many factors, including industrial discharges, agricultural runoff, insufficient waste management practices, and natural significantly contribute to the contamination of water supplies. Extreme weather and climate change can worsen pollution by causing erosion and flooding, which lowers the quality of the water [3]. Contaminated water sources, tainted by pathogens, toxic chemicals, and heavy metals, pose severe health risks, particularly in regions with limited access to clean water and sanitation [4]. Addressing water pollution is not only a matter of public health but also an urgent environmental imperative. To ensure that the public receives potable water and to prevent health risks associated with contamination, it is vital, therefore, to implement sophisticated methods for real-time classification of water potability. Automatic classification systems are critical to resolving the global concern of rising contamination and water quality issues [5]. These systems can detect contaminants in water sources quickly, providing an instant response in the event of contamination. This is of the utmost importance for the protection of public health, as it prevents the consumption of contaminated water and reduces exposure to hazardous substances.

Building efficient, precise, and understandable machine learning detection methods for evaluating the quality and safety of drinking water requires a solid understanding of water potability indicators. Water potability indicators are processes [2]. Human activities such as inappropriate waste disposal and pollutant emissions from industrial sites critical metrics employed to assess the appropriateness of a substance for human consumption and a wide range of applications. These indicators aid in assessing potable water quality and safety. Common indicators of water potability are presented in Table 1.

Table 1. Water Potability Indicators [6]

Indicators	Description
pH	The acidity or alkalinity of water is quantified by its pH. It is crucial for ensuring optimal taste and efficacy in water treatment procedures. pH of 7 is safe, values above or below 7 are alkaline and acidic respectively.
Hardness	The concentration of specific minerals, mainly calcium and magnesium, in water is measured as water hardness. It is quantified in parts per million (ppm) of calcium carbonate (CaCO3) equivalent or milligrams per litre (mg/L). Water with low hardness levels, typically in the range of 0-120 mg/L (CaCO3) or ppm, is considered soft. It is generally within the acceptable range for drinking and most domestic uses.
Solids	Total Dissolved Solids (TDS) represent the concentration of dissolved minerals, salts, and organic matter in water. High TDS can affect water quality and taste. Excellent (0-300 mg/L or ppm): Water with low TDS in this range is typically considered superior in taste and suitability for drinking.

Chloramines	Disinfection byproducts known as chloramines are produced due to the use of chlorine in the process of water disinfection. The monitoring of chloramine levels holds significant importance in ensuring the safety of drinking water. The Environmental Protection Agency (EPA) suggests a recommended range of 2.0 to 4.0 mg/L (or parts per million) of chloramines in order to provide efficient disinfection while simultaneously minimizing the production of detrimental byproducts. The water within this range is deemed healthy for consumption.
Sulfate	Sulphate is an ion that occurs naturally in water. Excessive sulphate content might have a laxative effect and alter the flavor of the water. A maximum sulphate content of 250 mg/L (or ppm) is recommended by the WHO in drinking water for aesthetic reasons (to avoid a laxative effect from high sulphate levels). This guideline is primarily based on concerns about taste and odor.
Conductivity	The capacity of water to pass electrical current is measured by conductivity, which is affected by dissolved salts and ions. It can indicate the overall mineral content of water. Low Conductivity (0-300 µS/cm) water is typically considered good quality for drinking and most domestic uses.
Organic_carbon	Organic carbon indicates the presence of organic matter in water, which can react with disinfectants and affect water quality. Organic carbon concentrations between 1 and 5 mg/L (or ppm) typically fall within the acceptable range for drinking water quality.
Trihalomethanes (THM)	Chemical reactions between chlorine and organic matter in water produce trihalomethanes. There is potential for adverse health effects associated with exposure to high levels of trihalomethanes. THM levels in drinking water are typically regulated by local water quality standards, usually measured in µg/L or ppb. The acceptable range varies by region and depends on health and water treatment. THM levels below 80 µg/L (ppb) are safe for drinking.
Turbidity	Cloudiness or haziness in water due to suspended particles like sediment, organic matter, and microorganisms is quantified by turbidity. High turbidity may indicate contaminants and affect water quality. Drinking water with turbidity levels below 15 nephelometric turbidity units (NTU) is safe. Low turbidity water is clear and has few suspended particles.
BDO (Biochemical Oxygen Demand)	BOD measures how much dissolved oxygen microorganisms need to break down water's organic material. A high BOD level may be

	indicative of pollution and other problems with the water quality. BOD values for drinking water are generally regarded as acceptable when they fall between 0 and 2 mg/L. A low biochemical oxygen demand (BOD) level indicates minimal organic pollution in the water and that it is safe for human consumption.
COD (Chemical Oxygen Demand)	The chemical oxygen demand (COD) of water is a measure of how much oxygen is required to oxidize organic and inorganic materials. It's an indicator of water pollution and the presence of substances that can affect water quality. Low COD (< 10 mg/L) values indicate that the water contains a minimal amount of organic and inorganic substances that can be easily oxidized. This water is generally considered to have low pollution levels and is likely to be good quality.

This paper explores a machine learning approach for classifying the potability of drinking water, utilizing ensemble learning methods (WAPOT) such as Stacking classifiers. Stacking, as a form of ensemble learning, consistently outperforms standalone classifiers and other existing research works, offering improved accuracy in potability classification. The findings underscore the potential of machine learning to contribute to the monitoring and managing water treatment processes significantly.

The subsequent sections of the work are structured as follows. The significance of the problem and water potability indicators are discussed in Section I. Section II exhibits the recent research studies in the problem domain. Section III, the details of the proposed method and its process flow are elucidated. Experiment results and analysis are in Section IV. Paper concludes in Section V.

2. Related Works

Researchers have used many ML algorithms to classify water as potable or non-potable based on water quality standards. Explainable AI (XAI) has also improved model predictions and public health decision-making. This segment outlines recent research conducted in the domain. Table 2 presents the summary of the existing research works. Salisu et al. [7] used the Kinta River dataset to find the best water quality classification model. The model using the KStar method had the highest accuracy at 86.67%. Theyazn et al. [8] used a dataset with seven significant parameters to forecast the index of Water quality and classify using advanced AI algorithms. The results showed that their methods accurately predicted WQI and classified water quality. In WQI prediction, the NARNET model surpassed LSTM, while in WQC prediction, the SVM algorithm was best with 97.01% accuracy. Using the Kaggle dataset, Semparudhi [9] demonstrated a machine learning based method to classify the water potability.

The dataset was normalized using variance, and the missing values were removed due to sensitivity. The best result of 84% accuracy was obtained using the XGB boost classifier. Muhammad et al. [10] presented a water potability classification technique using the kaggle dataset. They utilized both Gaussian Naïve Bayes and Decision Trees machine learning algorithms, together with k-fold crossvalidation for validation. The outcomes demonstrated apparent differences in the models' performance, with the Decision Tree model obtaining an accuracy rate of 97.22 %. Using the kaggle Water Quality Index dataset, Jinal et al. [11] evaluated machine learning techniques including SVM, DT, RF, GB, and AB, for water quality classification. The data is normalized using Z-score, and the asymmetrical dataset is balanced using SMOTE. Random Forest and Gradient Boost are most accurate at 81%. To address the absence of model transparency, XAI, and LIME in particular, are utilized to determine the importance of features. A rule-based approach for labeling water quality was proposed by Víctor et al. [12]; it utilized historical data obtained from seven monitoring stations in the Loa River. It uses expert knowledge and a Random Forest predictive model with physicochemical parameters. For model validation, accuracy, precision, and recall are 89.7, 89.73, 92.8. Tahani et al. [13] proposed a predictive system to assess water suitability for irrigation.*a.* Additionally, it can recommend crops based on crop predictions. The experiment involved crop and water potability datasets, utilizing a deep learning model that achieved a 96% accuracy in water quality prediction and 97% in crop recommendations. The SVM model predicts water quality with 92% accuracy and recommends crops with 96% accuracy. Table 2 summarizes the recent research works.

Table 2. Recent Research Work Summary

Author (s)	ML Algorithm	Accuracy %
Salisu et al. [7]	KStar	86.67
Theyazn et al. [8]	SVM	97.01
Semparudhi [9]	XGB boost	84
Muhammad et al. [10]	Decision Trees	97.22
Jinal et al. [11]	Random Forest, Gradient Boost	81
Víctor et al. [12]	Random Forest	89.7
Tahani et al. [13]	SVM	92

3. Proposed Method (WAPOD)

The purpose of the system being proposed (WAPOT) is to use stacking ensembles to develop a robust and efficient classification model. A meta-learning algorithm is used in stacking, an effective machine learning ensemble technique, to optimize the combination of predictions made by the multiple base machine learning algorithms [14]. Stacking uses numerous best classification models to combine their strengths, often outperforming individual models. Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbours (KNN), Naive Bayes (NB), Random Forest (RF), and Decision Trees are used in the proposed method. The meta-classifier (LR) combines the base classifier's predictions to make the final decision. To facilitate the experiment, the dataset is partitioned into two segments: training is allocated 80% of the data, and testing is allocated 20%. The complete process flow of the proposed method is depicted in Figure 1.

Integrating biometric authentication technology (such as fingerprints, facial recognition, or iris scanning) to safely authenticate people attempting to access their pension accounts online is known as a Web Biometric Authentication System for Pension Retrieval. To provide safe authentication, pension retrieval, and data management, the system architecture for this type of solution would be composed of several parts, each of which would play a distinct role as shown in Figure 1.

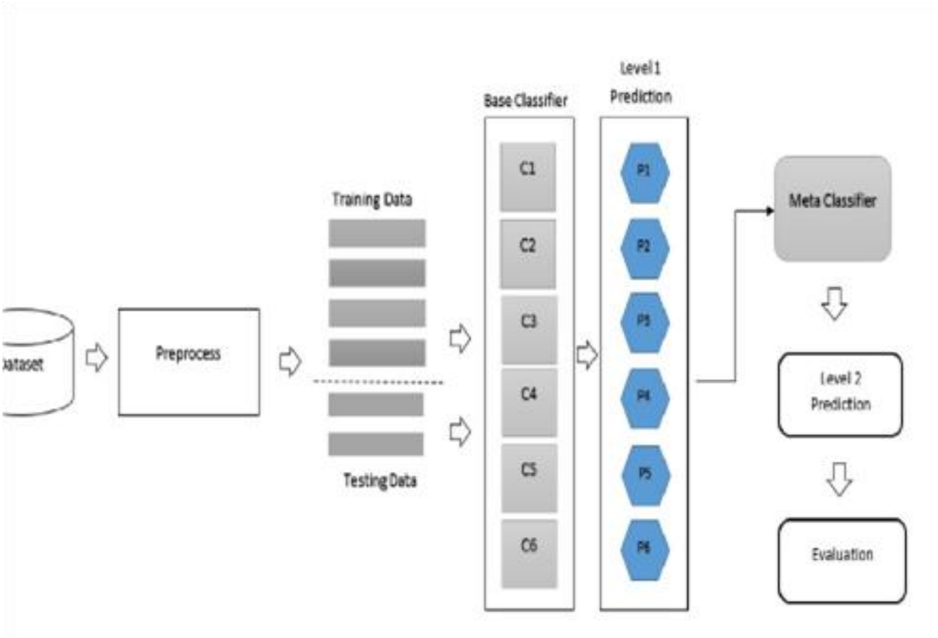

Figure 1. Process flow of the proposed method (WAPOD)

3.1 Dataset

The proposed method (WAPOD) utilizes the dataset from Kaggle [15], which contains water quality metrics for 3276 different water bodies. The water quality indicators are listed in Table 1. The potability attribute in the dataset indicates 1- Potable and 0-Not potable. Table 3 shows the dataset details.

Table 3. Dataset Details

Class	Number of samples
Potable (1)	1278
Not potable (0)	1998

The density (distribution or the frequency of values) within each attribute (feature) of a dataset [16] is depicted in Figure 2.

3.2 Preprocess

After data collection, preprocessing is a crucial step that is the base for reliable data analysis. Handling missing values and standardizing attributes are the two primary functions that preprocessing executes in the experiment. Lost data can result in skewed or incorrect outcomes, as well as the loss of • essential information. To preserve data integrity, the experiment utilizes the mean imputation [17]. This • straightforward and efficient technique substitutes the absent values in a given feature with the average of the available data. It guarantees that the dataset is complete, enabling more thorough analysis and precise model training.

$$\text{Imputed value } = \frac{\Sigma_{i=1}^{n} xi}{n} \tag{1}$$

Imputed value is a Missing Value

X_i • Represents the observed values of the variable Xi.

n • Is the total observed values

Furthermore, to establish uniformity and guarantee that the features fall within a consistent range, the min-max normalization method was implemented [18]. By converting each feature to a standard interval, usually between 0 and 1, this transformation facilitates an equitable comparison and optimizes the model's training.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

- x_{norm} is the normalized value of the data point.
- x is the original data point that you want to normalize.

x_{max}, x_{min} are the max and min value of the variable in the dataset.

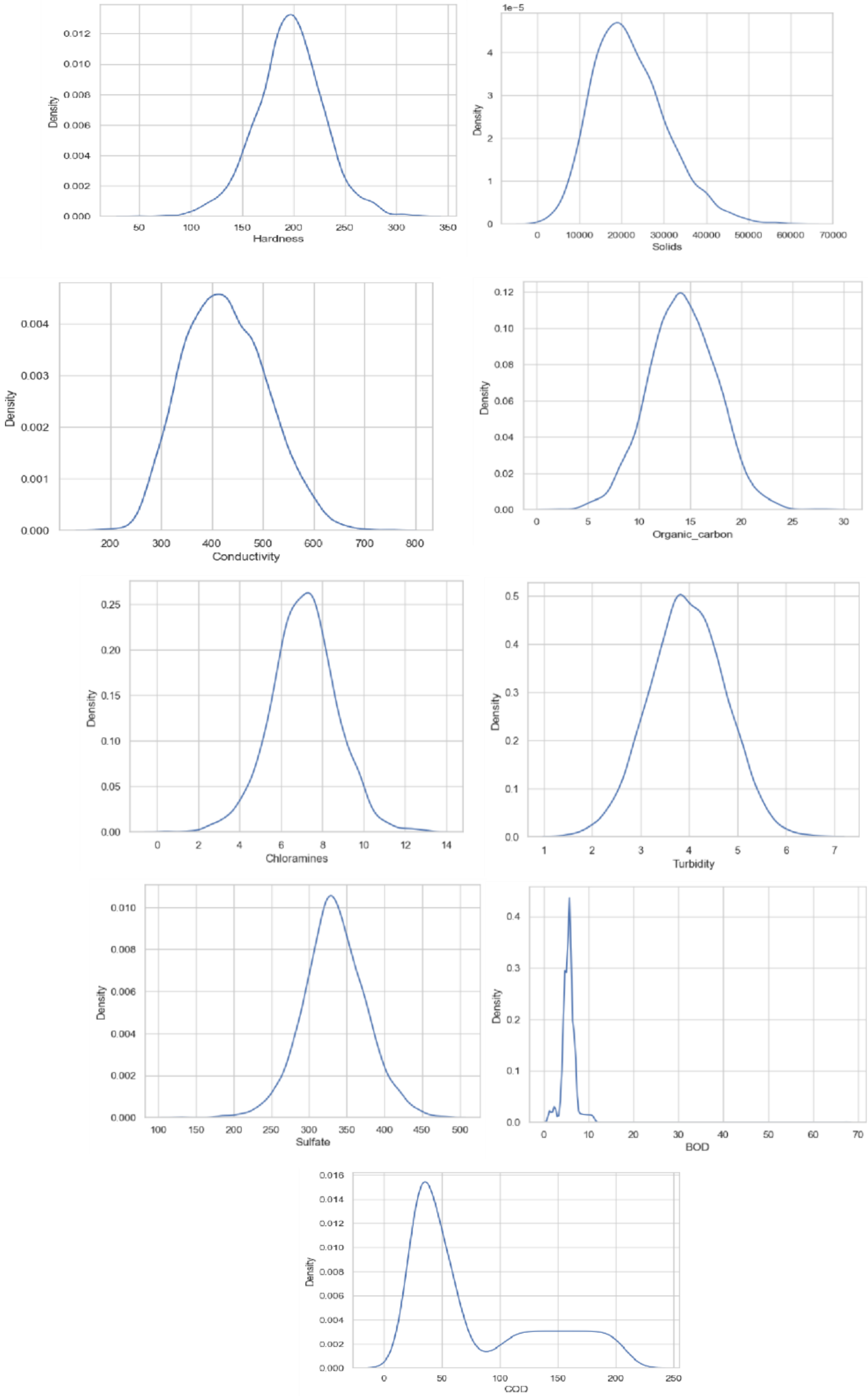

Figure 2. Density of a dataset

3.3 Train and Test Split

A dataset with 80:20 training and testing subsets is 80% usually utilized for training and 20% for model testing [19]. In the training phase, a subset of the data, commonly called the training dataset, is utilized to train the base classifiers. The base classifiers in this data subset capture numerous patterns and correlations. Subsequently, the base classifier predictions on the testing data set serve as the Level 1 predictions, representing the diverse insights from individual models. The performance of the meta-classifier, which aggregates and improves the Level 1 prediction, is also assessed using the testing data set.

3.4. Base Classifiers

The proposed system incorporates six distinct classifiers (C1, C2, C3, C4, C5, C6) as base classifiers, as depicted in Figure 1.

C1-Logistic Regression (LR): LR estimates the probability of an instance belonging to one of two classes.

C2-Support Vector Machine (SVM): By optimizing the hyperplane for maximal class margin, SVM excels in linear and non-linear binary and multi-class classification.

C3-k-Nearest Neighbors (kNN): kNN is a simple and effective instance-based learning method for regression and classification that labels data points by their k-nearest neighbor's feature space majority class.

C4-Naive Bayes (NB): Based on Bayes theorem, the Gaussian Naive Bayes classifier assumes feature conditional independence and Gaussian distribution.

C5- Random Forest (RF): RF a decision tree ensemble method, combines several trees to improve accuracy and reduce overfitting.

C6- Decision Trees (DT): DT, which are simple and interpretable, recursively partition data by important features to produce a tree with leaves reflecting the final decision.

3.5. Level1 and Level2 prediction

Level 1 predictions (P1, P2, P3, P4, P5, P6) are produced by individual base classifiers, each of which is trained on the same dataset but provides different perspectives on the problem. The Level 1 predictions are preliminary assessments from the base classifiers, whose performance may differ regarding their strengths and weaknesses. In contrast, Level 2 predictions are produced when a meta-classifier is trained to integrate and enhance the Level 1 predictions. This process results in the generation of a novel set of predictions to optimize the overall accuracy of predictions. This two-tiered strategy maximizes the ensemble's use of the base classifiers' diversified knowledge to provide more accurate and robust predictions, increasing machine learning model performance in our experiment. The Level 2 prediction of the logistic regression meta-classifier can be mathematically represented as

- $y_{stacked=} \quad \sigma(w1.y1 + w2.y2 + w3.y3 + w4.y4 + w5.y5 + w6.y6 + b)$ (3)
  $y_{stacked}$ represents the final prediction of the stacking ensemble.
- $y1, y2, y3, y4, y5, y6$ are the predictions of the base classifiers.
- w1 to w6 base classifier prediction weights. These weights are learned during the training process.
- b is the bias term.
- σ represents the logistic function that maps the weighted sum to the [0, 1] range to obtain the final classification probability.

3.6 Performance Evaluation

The metrics employed for the detection of potability encompass accuracy, which assesses the ratio of accurately classified instances, precision, is the ratio of true positive predictions to all positive predictions, recall, the percentage of true positives identified, and F1-score, which is the harmonic mean of precision and recall. AUC-ROC, which stands for the area under the Receiver Operating Characteristic (ROC) curve, is an additional metric utilized to evaluate the discriminatory power of the model [20]. These metrics are used to assess the performance of the proposed model.

4. Result and Discussion

Python and its ecosystem of libraries, such as Scikitlearn, Pandas, and Matplotlib, enabled flexible and efficient data analysis, model development, and evaluation. Imputation is an essential part of data preprocessing that improves dataset reliability and value for data analysis and machine learning. Because it can handle missing or partial data, which is common in real-world datasets. Imputation approaches can prevent data loss, preserve dataset integrity, and enable statistical and machine learning model applications by filling gaps. In this context, the experiment is carried out in two separate methods: one involving imputation (using mean values) and the other without imputation. The dataset size for each way is presented in Table 4.

Table4. Dataset Size For Experiment Methods

Experiment	Potable	Not Potable
With imputation	1278	1998
Without imputation	811	1200

The experimental dataset is unbalanced, with 40% of the data being potable and 60% being non-potable class. As a result, the F1-Score, Confusion Matrix, and ROC-AUC Curve are indispensable instruments for assessing the performance of a binary classification model along with accuracy. The results of the stacking classifier with imputing are displayed in Table 5 and yield 97% accuracy and 96% F1-Score.

Table 5. Experimental Result Of The Proposed Method

Classifier	Accuracy %	Precision%	Recall %	F1-Score%
Stacking without imputing	96	96	96	96
Stacking with imputing	97	96.5	96.5	96

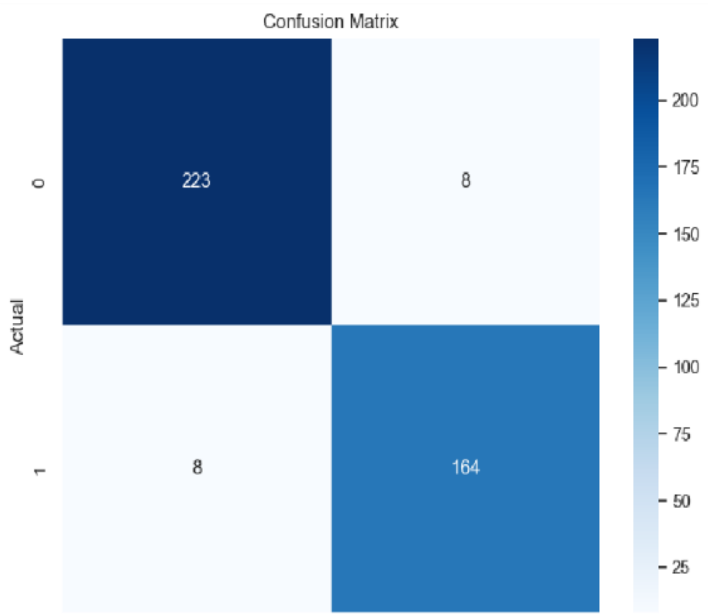

Figure 3. Confusion Matrix for Stacking without Imputing

The confusion matrix offers a comprehensive description of the model's predictions, which is essential for comprehending the model's behavior, mainly when working with datasets that are not equally balanced. The confusion matrix for the experiment, which is depicted in Figures 3 and 4, demonstrates a performance improvement. It also highlights how crucial imputation is. The ROC-AUC Curve is useful for assessing a model's discriminatory power regardless of class distribution. When evaluating a model's performance in the absence of a threshold, it's crucial. When a model has a high ROC-AUC, it is good at differentiating across classes, which is helpful when there is an imbalance. Figures 5 and 6 show the experiment's ROC-AUC curve, which more precisely classifies the class.

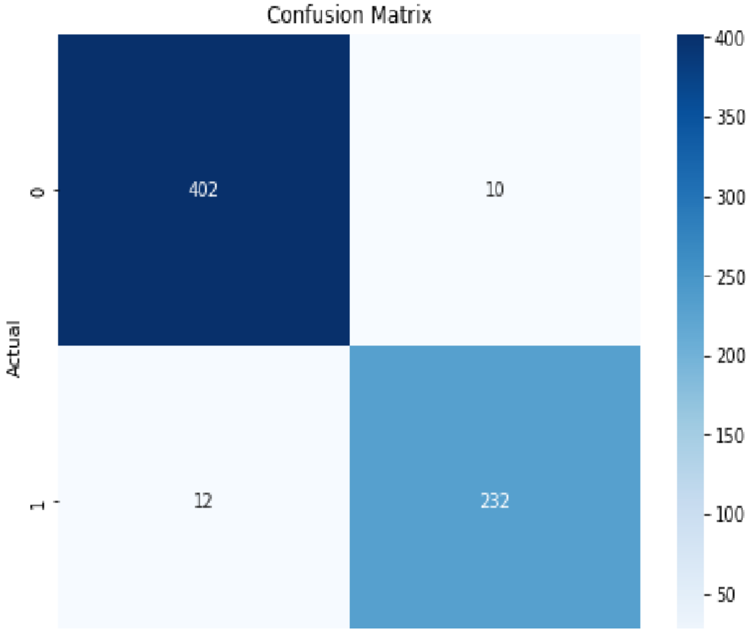

Figure 4. Confusion Matrix for Stacking with Imputing

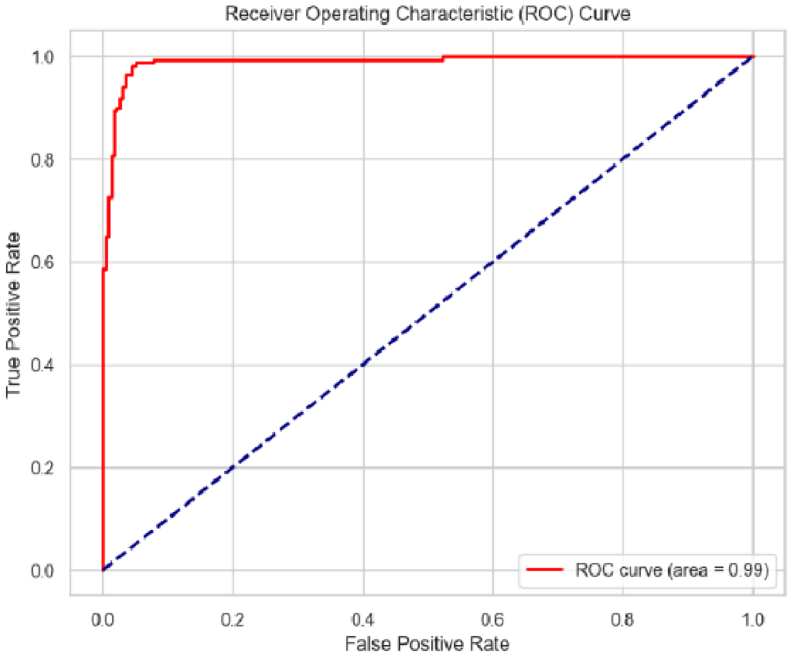

Figure 5. ROC-AUC Curve for Stacking without Imputing

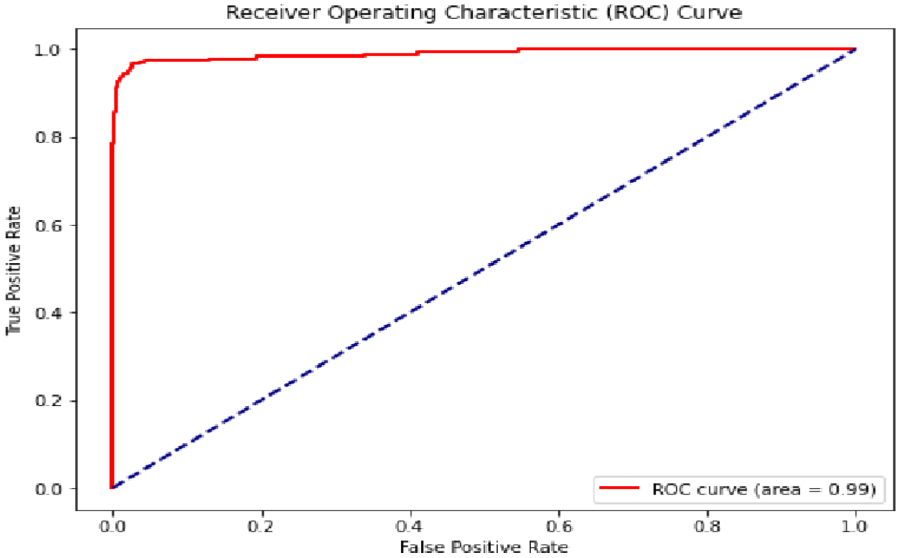

Figure 6. ROC-AUC Curve for Stacking with Imputing

The results presented in Figure 7 demonstrate that the accuracy of the proposed methods is significantly better than that of the existing methods

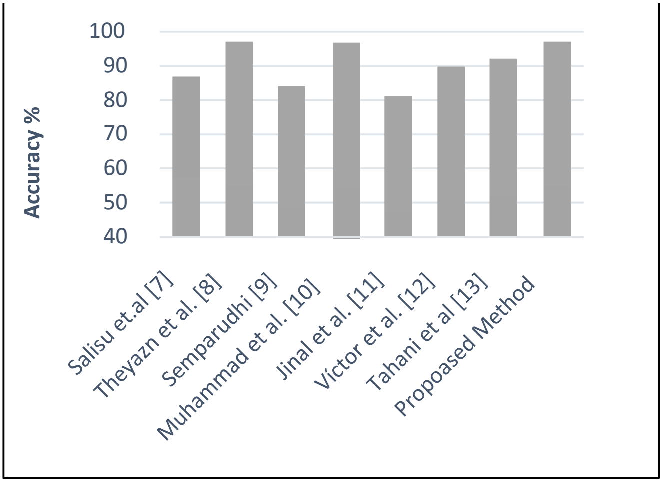

Figure 7. Performance Comparison

5. Conclusion

The introduction emphasizes the significance of classifying potable water according to its potability. It also highlights how crucial accurate and consistent potability measures are. The literature review summarizes recent field research. Potability classification accuracy is increased by our suggested methodology, which is based on ensemble learning techniques (Stacking classifiers, or WAPOD). The experiment's results perform better than individual classifiers and earlier studies and achieve 97% accuracy. This offers an excellent chance to use the machine learning model to improve and guarantee the potability of drinking water

References

[1] L. Lin, H. Yang, X., Xu, (2022) Effects of Water Pollution on Human Health and Disease Heterogeneity: A Review. *Frontiers in Environmental Science*, *10*, 880246. https://doi.org/10.3389/fenvs.2022.880246

[2] N. Akhtar, M.I. Syakir Ishak, S.A. Bhawani, K. Umar, (2021) Various Natural and Anthropogenic Factors Responsible for Water Quality Degradation: A Review. *Water*, *13*(19), 2660. https://doi.org/10.3390/w13192660

[3] UN-Water. (n.d.). *Water and climate change*. Retrieved 2023, from https://www.unwater.org/water-facts/water-and-climate-change

[4] A. Singh, A. Sharma, R.K. Verma, R.L. Chopade, P.P. Pandit, V. Nagar, V. Aseri, S.K. Choudhary, G. Awasthi, K.K. Awasthi, M.S. Sankhla, (2022) Heavy Metal Contamination of Water and Their Toxic Effect on Living Organisms. In *The Toxicity of Environmental Pollutants*. Intechopen. https://doi.org/10.5772/intechopen.105075

[5] M. Zhu, J. Wang, X. Yang, Y. Zhang, L. Zhang, H. Ren, B. Wu, L. Ye, (2022) A Review of the Application of Machine Learning in Water Quality Evaluation. *Eco-Environment & Health*, *1*(2), 107-116. https://doi.org/10.1016/j.eehl.2022.06.001

[6] Water Quality Indicators: Available at: https://www.rampalberta.org/river/water+sediment+quality/chemical.aspx

[7] S.Y. Muhammad, M. Makhtar, A. Rozaimee, A.A. Aziz, A.A. Jamal, (2015). Classification Model for Water Quality Using Machine Learning Techniques. *International Journal of Software Engineering and its Applications*, *9*(6), 45-52. http://dx.doi.org/10.14257/ijseia.2015.9.6.05

[8] T.H.H. Aldhyani, M. Al-Yaari, H. Alkahtani, M. Maashi, (2020) [Retracted] Water Quality Prediction Using Artificial Intelligence Algorithms. *Applied Bionics and Biomechanics*, *2020*(1), 6659314. https://doi.org/10.1155/2020/6659314

[9] S. Selvaraj, (2021). Water Potability Prediction with Machine Learning. *Medium*. https://parudhi.medium.com/water-potability-prediction-with-machine-learning-8eea74c29708

[10] M.I.T.K. Haq, F.D. Ramadhan, F. Az-Zahra, L. Kurniawati, A. Helen, (2021) Classification of Water Potability Using Machine Learning Algorithms. In *2021 International Conference on Artificial Intelligence and Big Data Analytics* IEEE. 1-5. https://doi.org/10.1109/ICAIBDA53487.2021.9689727

[11] J. Patel, C. Amipara, T.A. Ahanger, K. Ladhva, R.K. Gupta, H.O. Alsaab, Y.S. Althobaiti, R. Ratna, (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. https://doi.org/10.1155/2022/9283293

[12] V. Flores, I. Bravo, M. Saavedra, (2023) Water Quality Classification and Machine Learning Model for Predicting Water Quality Status—A Study on Loa River Located in an Extremely Arid Environment: Atacama Desert. *Water*, 15(16), 2868. https://doi.org/10.3390/w15162868

[13] T., Alkhudaydi, M.Q. Albalawi, J.S. Alanazi, W. Al-Anazi, R.M. Alfarshouti, (2023) Deep Learning for Combined Water Quality Testing and Crop Recommendation. *International Journal of Advanced Computer Science and Applications*, 14(4). https://dx.doi.org/10.14569/IJACSA.2023.0140450

[14] M. Lu, Q. Hou, S. Qin, L. Zhou, D. Hua, X. Wang, L. Cheng, (2023) A Stacking Ensemble Model of Various Machine Learning Models for Daily Runoff Forecasting. *Water*, 15(7), 1265. https://doi.org/10.3390/w15071265

[15] A.Patel, (2021). Water Quality Prediction: 7Model Input. *Kaggle*. https://www.kaggle.com/code/imakash3011/water-quality-prediction-7_model/input

[16] P. Khurana, (2021). Distribution of data — Histogram. *Medium*. https://prvnk10.medium.com/distribution-of-data-histogram-a88c7e97728b

[17] S. Alam, M.S. Ayub, S. Arora, M.A. Khan, (2023) An Investigation of The Imputation Techniques for Missing Values in Ordinal Data Enhancing Clustering and Classification Analysis Validity. *Decision Analytics Journal*, 9, 100341. https://doi.org/10.1016/j.dajour.2023.100341

[18] Loukas, S. (2020). Everything you need to know about Min-Max normalization: A Python tutorial. *Towards Data Science*, 5.https://medium.com/data-science/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79

[19] Brownlee, J. (2020). Train-Test Split for Evaluating Machine-Learning Algorithms. *Machine learning mastery*, 23(7).https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

[20] Brownlee, J. (2020). How to Calculate Precision, Recall, and F-measure for Imbalanced Classification. *Machine learning mastery*, 1. https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/

Funding

No funding was received for conducting this study.

Conflict of interest

The Author's have no conflicts of interest to declare that they are relevant to the content of this article.

About The License