Predictive Analytics for Water Quality Assurance

Mansi kharwar

Department of Computer Science and Engineering
Indira Gandhi Delhi Technical University, New Delhi, India
mansikhrwar27@gmail.com

Abstract. Water is a lifesaving factor that can found only on earth and it's a growing face of our country, so we can face a lot of changes and difficulties in the shortage of a basic necessity Safe and potable drinking water. Different region water resources should be analyzed separately. In this paper, the analysis of potability of water by using a dataset is done. Because in this dataset gap data are found, the aim of this project would be to use the best algorithm on a statistically imputed dataset and then to find an algorithm that gives the best prediction about whether water is potable or not. The goal is to determine the potability of water. Nowadays, Various algorithms have been used to predict the water potability with help of nine reliable features, obtained from Kaggle. In this set, some data is missing for three features—pH, chloramine, and trihalomethane and two column Carcinogenic and medical waste that can contain no value. The completion of missing values is done by the mean. After experimenting with various algorithms—Random Forest, Support Vector Machine (SVM), XGBoost, CatBoost, K-Nearest Neighbors (KNN), and Gradient Boosting—Gradient Boosting topped in the accuracy charts at **99.47%**, while XGBoost and CatBoost followed closely with **97%**. SVM trailed with the lowest accuracy at **62%**. This research Ensuring to provide a reliable result.

Keywords: Water potability prediction, Random Forest, SVM, XGBoost, KNN, Logistic Regression, CatBoost, Missing Values

1 INTRODUCTION

Water is life's backbone which is essential for drinking, daily needs, and food. Secure and accessible water drives public health, fuels economic growth, and fights poverty. [1] So effective Policies should be made by all nations regarding the water potability Therefore, industrial areas and pollution are some of the factors responsible for the deterioration of water in India, which results in diseases like typhoid, dysentery, polio, cholera, hepatitis, and diarrhea. Insufficient facilities for water and sanitation expose people to preventable health risks, especially in health facilities. It is estimated that of the available water in India, 70% is polluted by industrial and household waste, and 80% of the rural population and 20% of the urban population do not have access to safe drinking water and sanitation, account for 46 percent of the mortality in children below the age of five years [1]. The work is geared towards the enhancement of water quality,

filtering, normalizing, running classification algorithms, and deploying models to achieve accuracy from 75% to 82% in testing on water

China's environmental monitoring agencies are increasingly leveraging online water quality monitoring technology to gather continuous, real-time data on water quality. Technology highly uses like sensors, automatic measurement, and computer applications to analyze detection data, output maximum and minimum monitoring data, and calculate average values. [2] The system also features functions like early warning, signal output, and automatic operation.

Machine learning is a programming technique used in programming computers to enable them to learn by themselves from data without instruction. It includes different techniques of supervised, unsupervised, semi-supervised, and reinforcement learning. The algorithms of machine learning identify patterns and behaviors in the data. On the other hand, they will adjust themselves automatically to these changes to achieve the desired output. Missing data has a huge effect on the quality of the model.[2] improving model accuracy and dataset reliability, leading to better machine learning models.

For This research using different parameter have been explained below:

1. **pH**: A measure of water's acidity or alkalinity on a scale of 0 to 14, where 7 is neutral. The World Health Organization (WHO) recommends a pH range of 6.5 to 8.5 for potable water to ensure chemical stability and minimize corrosion or scaling.
2.**Hardness**: Defined by the concentration of calcium and magnesium ions, measured in milligrams per liter (mg/L). Hardness affects water's ability to lather with soap and contributes to scaling in pipes and appliances.
3.**Total Dissolved Solids (TDS)**: The total concentration of dissolved organic and inorganic substances in water, including minerals and salts, expressed in mg/L. WHO advises a TDS limit of 500 mg/L for safe drinking water, as higher concentrations may affect taste and health.
4.**Chloramines**: Compounds formed by the combination of chlorine and ammonia, used as disinfectants in water treatment. WHO considers chloramine concentrations up to 4 parts per million (ppm) safe for consumption, as higher levels may have adverse health effects.
5. **Sulfate**: Naturally occurring sulfate in fresh water typically ranges between 3 to 30 mg/L. While sulfates are not harmful at low levels, concentrations exceeding WHO recommendations may cause health issues and affect water taste.
6.**Conductivity**: The ability of water to conduct electrical current, influenced by dissolved ions and measured in microsiemens per centimeter (μS/cm). WHO guidelines suggest a safe conductivity threshold of 400 μS/cm for drinking water, as higher levels can indicate excessive mineral content.
7.**Total Organic Carbon (TOC)**: Represents the total amount of organic matter in water, which can support microbial growth if elevated. WHO recommends maintaining TOC levels below 4 mg/L to prevent health risks and maintain water quality.
8.**Trihalomethanes (THMs)**: Byproducts formed during the chlorination process in water disinfection. WHO sets a safety threshold at 80 micrograms per liter (μg/L) for THMs in drinking water, as prolonged exposure to higher levels is associated with health risks.

9.**Turbidity**: An indicator of water clarity, measured in Nephelometric Turbidity Units (NTU), reflecting the concentration of suspended solids. WHO recommends a turbidity limit of 5 NTU to ensure adequate disinfection and aesthetic quality in drinking water.
10.**Potability**: A binary indicator denoting water's suitability for drinking, where a value of 1 signifies potable (safe) and 0 signifies non-potable (unsafe) water based on compliance with health standards.

2 RELATED WORK

Recent studies on machine learning models for water quality assessment have demonstrated the effectiveness of various techniques. Researchers have explored multiple approaches, achieving high accuracy rates in water quality prediction and providing insights into the strengths of different algorithms for handling water quality data.

Pal et al. [6] explored the use of Artificial Neural Networks (ANN) and concluded that ANN achieved the highest accuracy, with a success rate of 99.1%. This finding underscores ANN's potential in precisely predicting water quality levels.

Patel et al. [5] addressed the challenge of imbalanced data by employing the Synthetic Minority Over-sampling Technique (SMOTE). Their study indicated that Gradient Boost and Random Forest models performed well in water quality prediction, each achieving an accuracy rate of 81%.

Uddin et al. [7] evaluated a variety of models and found that Support Vector Machines (SVM) were notably effective, achieving over 95% accuracy. This result highlights SVM as a reliable model in the context of water quality analysis.

Aldhyani et al. [8] conducted a comprehensive comparison of multiple AI algorithms and found that SVM yielded an accuracy of 97.01%, slightly lower than the Nonlinear Auto-Regressive Neural Network (NARNET), which was identified as the top-performing model in their research.

Tufail et al. [9] focused on water potability prediction in Pakistan, where their findings reinforced the effectiveness of SVM, establishing it as the most reliable model for predicting water quality in that region.

3 METHOLOGY

Assuring the quality of water is a complex process because it depends on many physical, chemical, and biological factors. The machine learning models have been proven to be very useful for the prediction and assessment of the quality of water to find the assurance potable water. This will help in better water management and ensure that communities have access to clean drinking water.

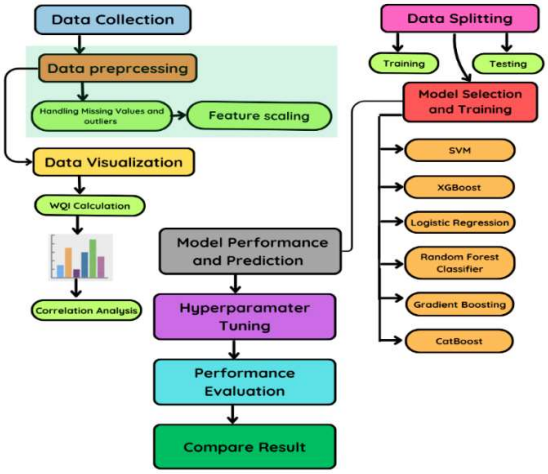


Fig.1: Workflow for water prediction model

3.1 COLLECTING DATA

The data used to do this research is collected from Kaggle, which provided 3,276 water quality samples from various locations.

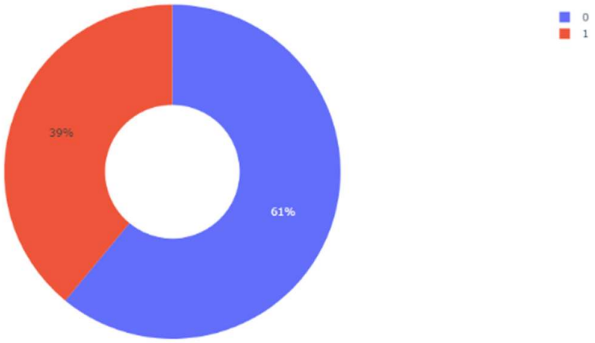


Fig. 2. Pie chart of the water potability

The dataset includes nine essential water quality indicators—such as pH, hardness, and turbidity—with two features having missing values. Alongside the potability target, these parameters were used to predict water safety through various machine learning models. The standard values recommended by world health organization shown in table

1. If the values of these parameters exceed their standard limit means water is not suitable for drinking.

Parameters	WHO limits
Ph	6.5–8.5
Hardness	200 mg/L
Solids	1000 ppm
Chloramines	4 ppm
Sulfate	1000 mg/L
Conductivity	400 μS/cm
Organic carbon	10 ppm
Trihalomethanes	80 ppm
Turbidity	5 NTU

Table 1. Standard Water Quality

3.2 DATA PREPROCESSING

It helps to increase the quality of data for analysis. It involves various steps:

a. Dealing with Missing Values: The dataset contains a lot of missing values. However, if outliers are present, they should be addressed first, as the mean may not be suitable in such cases. In Some parameter Like Carcinogens and medical water contain no value. Entire column has Null Value.

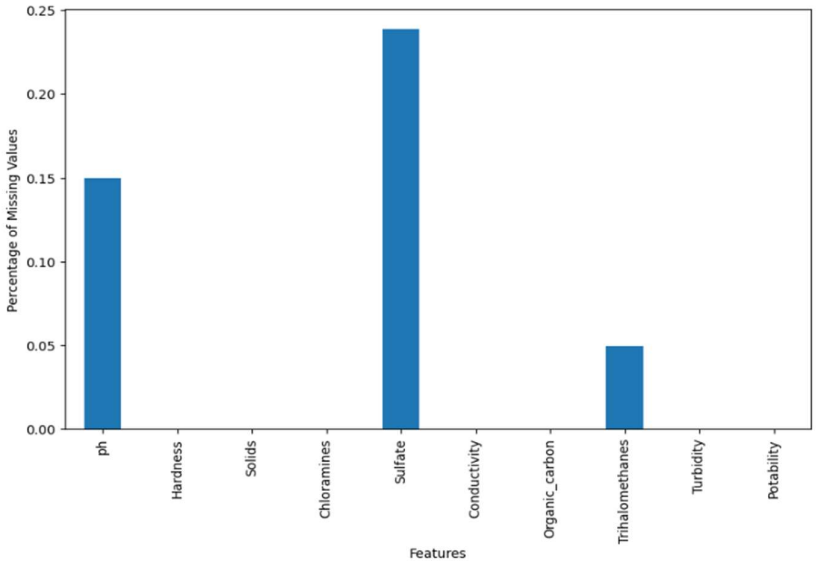


Fig. 3. Missing Value Data

b. data Normalization Using Z-Score: T Z-Score normalization is a statistical technique used to standardize data, bringing it to a common scale by expressing values in terms of their deviation from the mean in standard deviation units. This method transforms each data point to reflect its distance from the dataset's average, measured in standard deviations, typically constraining values within a range of -3 to +3. By centering the data around a mean of zero and normalizing its spread to a standard deviation of one, Z-Score normalization facilitates the comparison of data across different scales, enhancing interpretability in analyses.

The Z-Score for a value x is calculated as follows:

$$Z = \frac{x - \mu}{\sigma}$$

where:

- x represents the data point,
- μ denotes the dataset's mean,
- σ signifies the dataset's standard deviation.

3.3 DATA VISUALISATION

a. Correlation Analysis

We analyse the correlation between all features to identify potential relationships and determine which features might be more informative for predicting potability. The correlation matrix shows the values that range from -1 to 1. Values close to 1 or -1 indicate strong relationships, while values close to 0 suggest weak or no correlation. A correlation heatmap was generated to visualize these relationships clearly. This heatmap helps identify trends and interactions among features.

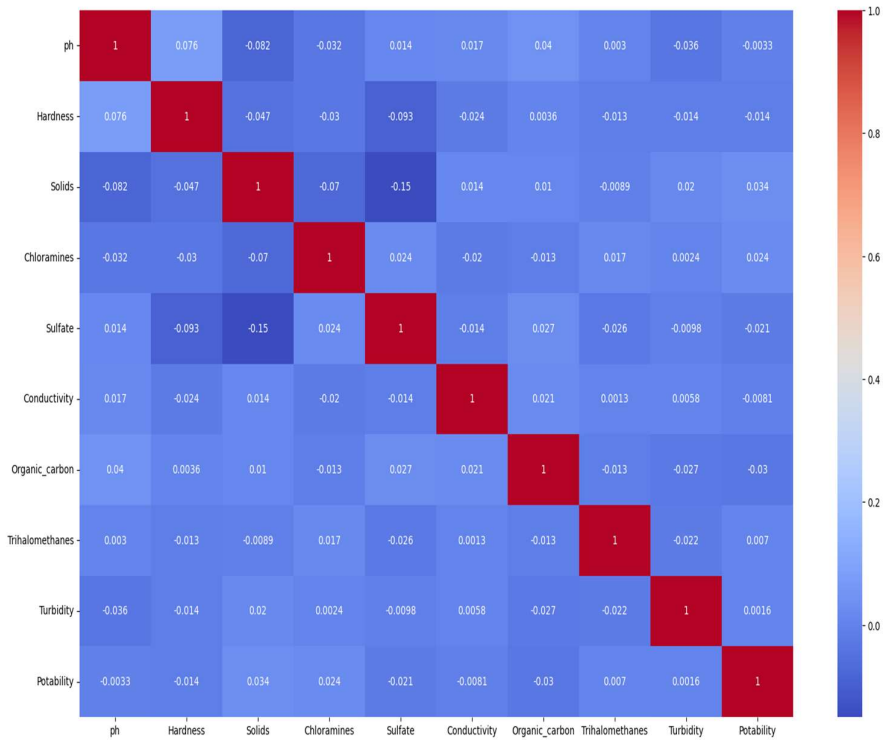


Fig. 4. Correlation Analysis

3.4 DATA SPLITTING

The dataset is comprising 3,276 samples, was divided into training and testing sets with a 70:30 ratio. In this configuration, 70% of the data (2,293 samples) was allocated for training the machine learning models, while the remaining 30% (983 samples) was reserved for testing. This approach enables a reliable evaluation of model performance, ensuring that models do not merely fit the training data but also generalize effectively when applied to unseen data. The 70:30 split ratio is widely recognized as a balanced strategy, allowing for robust model training while maintaining sufficient data for an accurate assessment of predictive accuracy and generalization capability.

3.5 MODEL SELECTION

a. XGBoost

XGBoost is an efficient and scalable implementation of gradient boosting. It is optimized for speed and performance, and is known for its accuracy in many machine learning tasks. XGBoost works well with both numerical and categorical data, using a boosting approach to combine the predictions of multiple decision trees to improve accuracy.

Formula used:

$$y = \sum_{m=1}^{M} \eta_m \cdot h_m(x)$$

Where:

- η_m is the learning rate for the m-th tree,
- $h_m(x)$ is the output of the m-th decision tree,
- M is the total number of trees.

In XGBoost, a regularization term is also added to prevent overfitting:

$$L(\theta) = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(h_m)$$

Where:

- $\ell(y_i, \hat{y}_i)$ is the loss function for the i-th instance,
- $\Omega(h_m)$ is the regularization term for the m-th tree.

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	617
1	0.97	0.97	0.97	366

b. Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees using random subsets of data. Each tree independently predicts, and the final output is the majority vote or average of the individual tree predictions, reducing overfitting.

Formula used:

$$y = \frac{1}{M} \sum_{m=1}^{M} h_m(x)$$

Where:

- M is the number of trees,
- $h_m(x)$ is the output of the m-th tree.

Classification Report:

	precision	recall	f1-score	support
0	0.70	0.86	0.77	617
1	0.61	0.37	0.46	366

c. Support Vector Machine (SVM)

SVM is a powerful classifier that finds the optimal hyperplane to separate classes in high-dimensional space. It works by maximizing the margin between different classes, making it effective for classification and regression tasks, especially with complex data.

Formula used:

$$w \cdot x + b = 0$$

Where:

- w is the weight vector (normal to the hyperplane),
- x is the input vector,
- b is the bias term.

Classification Report:

	precision	recall	f1-score	support
0	0.63	1.00	0.77	617
1	0.00	0.00	0.00	366

d. Logistic Regression

Logistic Regression is a statistical method used for binary classification, modeling the probability of an event occurring using the logistic function. It is widely used in problems like predicting whether water is 'safe' or 'unsafe' based on various features.

Formula used:

$$P(y = 1|x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

Where:

- w is the weight vector,
- x is the input vector,
- b is the bias term,
- $P(y = 1|x)$ is the probability of the positive class.

Classification Report:

	precision	recall	f1-score	support
0	0.63	1.00	0.77	617
1	1.00	0.02	0.03	366

e. CatBoost (Categorical Boosting)

CatBoost is an efficient gradient boosting algorithm that handles categorical features automatically. It builds an ensemble of decision trees, boosting predictions through sequential tree learning. It is known for its speed, accuracy, and robust handling of categorical data.

Formula used:

$$y = \sum_{m=1}^{M} \eta_m \cdot h_m(x)$$

Where:

- η_m is the weight of the m-th tree,
- $h_m(x)$ is the output of the m-th decision tree,
- M is the total number of trees.

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.99	0.98	617
1	0.98	0.96	0.97	366

f. Gradient Boosting

Gradient Boosting combines weak models into a strong predictor by sequentially training new models to fix errors. It minimizes loss using gradient descent which achieving high accuracy and highlighting feature importance. Careful tuning is needed to prevent overfitting.

Formula used:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$$

Where:

- $F_{m-1}(x)$ is the prediction from the previous model,
- $h_m(x)$ is the decision tree's output,
- η is the learning rate.

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	617
1	0.99	0.99	0.99	366

3.6 MODEL PERFORMANCE

To evaluate the performance of the model, there are several key metrics.

a) **Precision**: It measures true positives out of all positive predictions & how many hits were truly accurate.

$$Precision = \frac{TP}{TP + FP}$$

b) **Accuracy**: It shows the ratio of all correct predictions to total prediction & show often we got it right.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}.$$

c) **Recall**: It reflects the rate of true positives captured & how well we identified actual positives.

$$Recall = \frac{TP}{TP + FN}$$

d) **F1 Score**: It blends precision and recall into one score, ranging from 0 to 1 & the higher, the better the model's balance.

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

3.7 HYPERPARAMETER TUNNING

Model tuning involves adjusting model settings (hyperparameters) to find the best configuration that maximizes performance.

- **GridSearchCV**: This method systematically tests all possible combinations of hyperparameters to find the best setting, though it can be time-consuming as it checks every option.
- **RandomizedSearchCV**: This approach selects a fixed number of random combinations to test, making it more efficient. It may not explore every option but often finds a near-optimal solution faster.

Both techniques help boost the accuracy and effectiveness of a model by finding the best hyperparameter values.

4 EXPERIMENTAL OUTCOME

This study explores the power of artificial intelligence to enhance quality prediction and ultimately protect public health. Using a dataset of 3,276 samples split into a 70:30 training-to-testing ratio, the experiments revealed standout results. Gradient Boosting achieved the highest accuracy at 99%, while XGBoost and CatBoost closely followed with approximately 97%. These ensemble models demonstrated exceptional ability to identify complex patterns, underscoring their potential for real-world water monitoring. In comparison, Support Vector Machine and Logistic Regression, with lower accuracies of 62% and 63%, did not perform as effectively.

These results highlight the potential of advanced models like Gradient Boosting and XGBoost, which skillfully capture important relationships among water quality indicators. This accuracy and adaptability suggest that it could transform water quality assessment by enabling proactive monitoring and early insights. However, the study's scope was limited by a relatively small dataset and fewer quality parameters, signaling the need for future research to include larger datasets and additional factors. By building on these foundations, future work can bring us closer to reliable, real-time solutions for safe and clean drinking water.

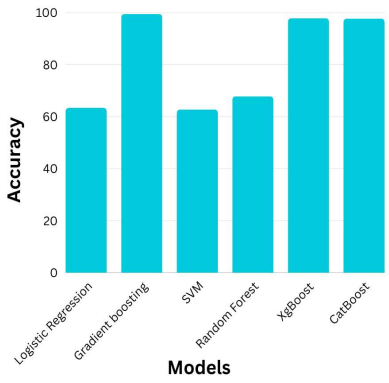
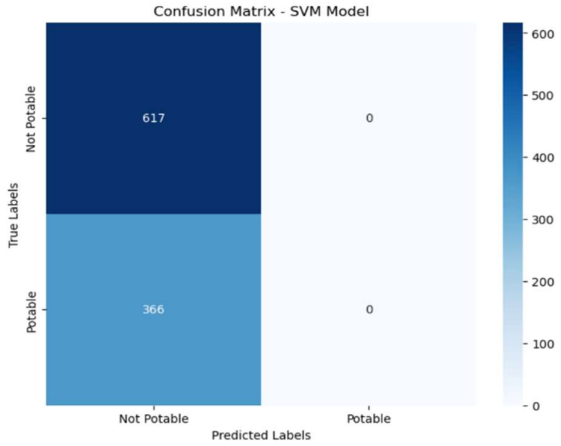
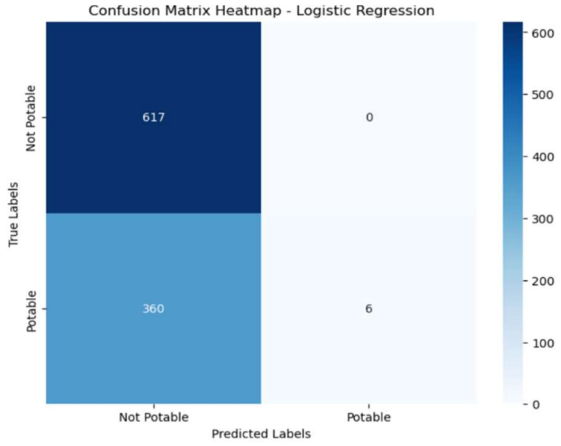


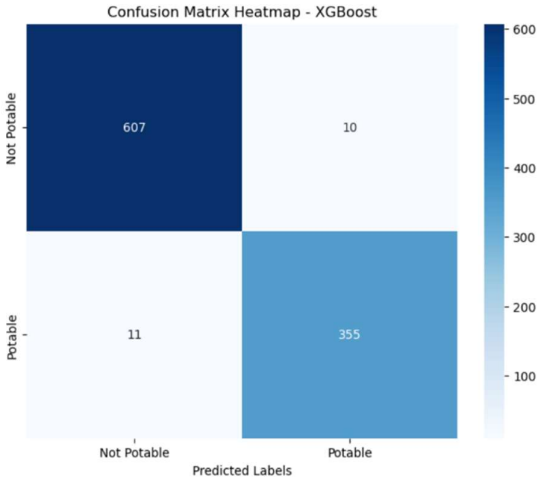
Fig 5. Bar Chart of Accuracy Comparison

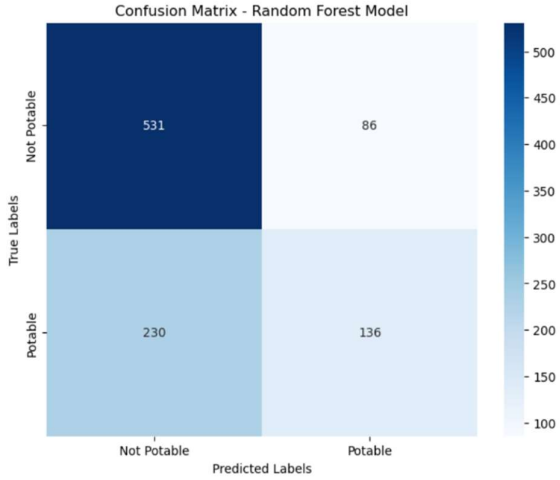
Model	Accuracy (%)
Gradient Boosting	99.49
XGBoost	97.86
CatBoost	97.76
SVM	62.77
Random Forest	67.85
Logistic Regression	63.38

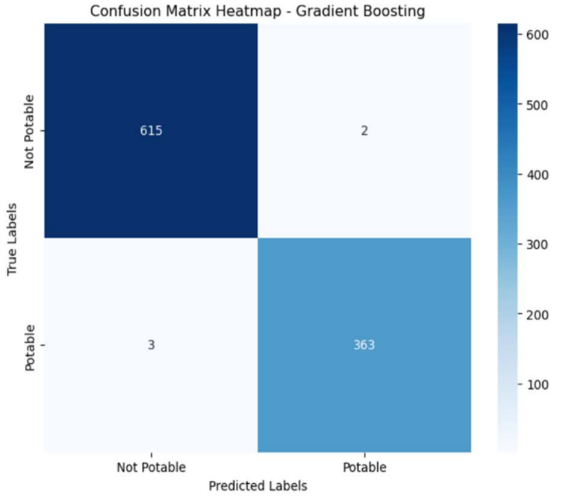
Table 2. Proposed algorithms Performance Analysis

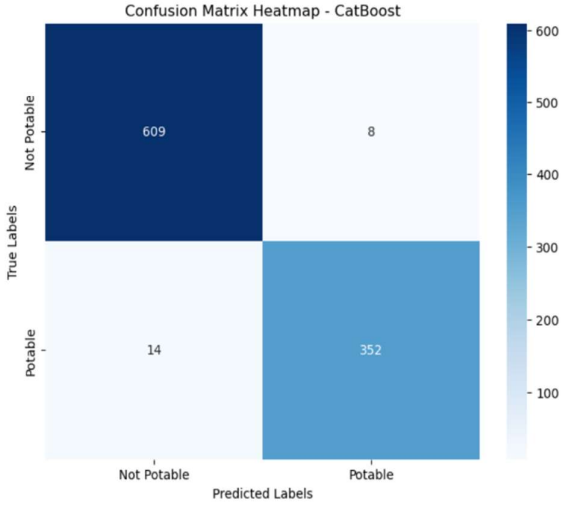

Confusion Matrix - SVM Model
Not Potable
Potable
True Labels
617
0
366
0
Not Potable
Potable
Predicted Labels
600
500
400
300
200
100
0


Confusion Matrix Heatmap - Logistic Regression
Not Potable
Potable
True Labels
617
0
360
6
Not Potable
Potable
Predicted Labels
600
500
400
300
200
100
0


Confusion Matrix Heatmap - XGBoost
Not Potable
Potable
607
10
11
355
Not Potable
Potable
Predicted Labels
600
500
400
300
200
100


Confusion Matrix - Random Forest Model
Not Potable
Potable
True Labels
531
86
230
136
Not Potable
Potable
Predicted Labels
500
450
400
350
300
250
200
150
100


Confusion Matrix Heatmap - Gradient Boosting
Not Potable
Potable
True Labels
615
2
3
363
Not Potable
Potable
Predicted Labels
600
500
400
300
200
100


Confusion Matrix Heatmap - CatBoost
Not Potable
Potable
True Labels
609
8
14
352
Not Potable
Potable
Predicted Labels
600
500
400
300
200
100

6 CONCLUSIONS

Ensuring access to safe, clean drinking water is crucial for public health, especially as population growth and environmental pollution intensify pressures on water resources. This study highlights the potential of machine learning techniques to accurately predict water potability, enabling proactive monitoring and management of water quality. By classifying water samples as potable or non-potable, machine learning models provide valuable insights that support public health initiatives, allowing for early detection of contamination and aiding in better-informed decisions for water treatment and resource allocation.

However, this research faced limitations, particularly due to a relatively small dataset of 3,276 samples and a restricted set of water quality parameters, which may limit the findings' generalizability across diverse geographic and environmental conditions. Expanding future studies to incorporate additional variables—such as microbial contaminants, chemical pollutants, and seasonal shifts—could enhance the models' predictive accuracy and broaden their applicability. By integrating more comprehensive datasets, future research can further refine machine learning applications in water quality management, contributing to global efforts in safeguarding public health and ensuring access to clean drinking water.

References

1. Patel, J.: A Machine Learning-Based Water Potability Prediction Model. *Computational Intelligence and Neuroscience*, 2(5), 99–110 (2022).
2. Patel, S., Shah, K., Vaghela, S., Aglodiya, M. A., Bhattad, R.: Water Potability Prediction Using Machine Learning. (Publication Year Not Provided).
3. Poudel, D., Shrestha, D., Bhattarai, S., Ghimire, A.: Comparison of Machine Learning Algorithms in Statistically Imputed Water Potability Dataset. *Journal of Innovations in Engineering Education*, 2(4), 120-134 (2023).
4. Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., Althobaiti, Y. S., Ratna, R.: A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI. *Computational Intelligence and Neuroscience*, 2(7), 135-145 (2022).
5. Pal, O. K.: The Quality of Drinkable Water Using Machine Learning Techniques. *International Journal of Advanced Engineering Research and Science*, 8(5), 87–97 (2021).
6. Uddin, M. G., Nash, S., Rahman, A., Olbert, A. I.: Performance Analysis of the Water Quality Index Model for Predicting Water State Using Machine Learning Techniques. *Process Safety and Environmental Protection*, 169, 808–828 (2023).
7. Aldhyani, T. H., Al-Yaari, M., Alkahtani, H., Maashi, M.: Water Quality Prediction Using Artificial Intelligence Algorithms. *Applied Bionics and Biomechanics*, 4(3), 211-219 (2020).
8. Addisie, M. B.: Evaluating Drinking Water Quality Using Water Quality Parameters and Esthetic Attributes. *Air, Soil and Water Research*, 15, 11786221221075005 (2022).

9. World Health Organization. pH in Drinking-Water. *WHO Guidelines for Drinking-Water Quality*, (2007).
10. Kent RO Systems. Kent Health Care Products. *Kent RO Systems*, (2020).
11. World Health Organization. Chloramine in Drinking Water. *WHO Guidelines for Drinking-Water Quality*, (1998).
12. United States Environmental Protection Agency. Sulfate Health Effects. *EPA*, (2003).
13. Whitehead, P.: Total Organic Carbon. *Elga Veolia*, (2021).
14. Health Canada. Guidelines for Canadian Drinking Water Quality: Trihalomethanes. *Health Canada*, (2018).