Potable Drinking Water Quality prediction using Machine learning

Mansi Kharwar
Department of Computer Science and Engineering,
INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN
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 analysed separately. In this paper, the analysis of potability of water by using a dataset is done. Because in this dataset gap data are found, so the purpose of this study is to use the best algorithm that works on a statistically imputed dataset and find an algorithm which gives the best prediction about whether water is potable or not. Aim 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. Using of Random Forest, Support Vector Machine, XGBoost, CatBoost, KNN, Gradient Boosting. Gradient Boosting gives a highest accuracy with 99.47%. Additionally, XGBoost and CatBoost with 97% and therefore SVM has least accuracy with 62%. This research Ensuring to provide a reliable result.**

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

I. INTRODUCTION

Water is an absolutely necessary resource for all living beings on our planet, whether it is used for drinking, household usage, or food. Safe and accessible water is crucial for public health, and thus progress in water supplies and management is linked with economic growth, prosperity, and poverty reduction. [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 diarrhoea. 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. Infectious diseases and environmental causes, especially the water supply 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, normalising, running classification algorithms, and deploying models to achieve accuracy from 75% to 82% in testing on water. For example, the head margin in this template measures proportionately more than is customary. The U.S. Environmental Protection Agency (EPA) sets maximum contaminants limits for over 90 in drinking water, while the Nepal Government sets limits for 27 under the Water Resources Act, 2005. [3]

Environmental supervision departments in China are highly using online water quality monitoring technology to obtain continuous data on water quality. This technology uses 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] Different imputation techniques, such as imputation using stochastic regression, imputation using mean or median, and imputation using K-NN .it improves model accuracy and dataset reliability, leading to better machine learning models. For This research using different parameter.

1. **pH:** Acid-base balance, expressed on a scale of 0 to 14. The recommended limits by (World health organization) WHO are from 6.5 to 8.5.

2. **Hardness:** The concentration of calcium and magnesium salts in water, which can be described as improving its soaping action (mg/L).

3. **Total Dissolved Solids (TDS):** Total dissolved inorganic and organic minerals in the water with a desirable limit of 500 milligrams per Liter for drinking.

4. **Chloramines:** Number of chloramines present, this is a disinfectant formed by a reaction of chlorine and ammonia, which is safe up to 4 ppm.

5. **Sulfate:** Amount of naturally occurring sulfates in the water, typically between 3 to 30 milligrams per liter in freshwaters.

6. **Conductivity**: The ability of water to conduct electricity. This is directly related to its content of ions, and the WHO standard is up to 400 μS/cm.

7. **Organic Carbon**: a value showing the total organic carbon present in water, including decaying natural organic matter; levels should not be more than 4 mg/L to be safe.

8. **Trihalomethanes (THMs)**: byproducts created during chlorine disinfection; their concentration may reach 80 μg=/L in drinking water and still remain safe.

9. **Turbidity**: Expression of the clarity of the water in suspended solids; this is, the WHO recommended limit for safe drinking water is 5 NTU.

10. **Potability**: Binary indicator of whether or not water is safe to drink; 1 = potable, 0 = non-potable.

II. LITERATURE REVIEW

This research paper looks at different ways to tackle water Quality problems. It compares traditional lab tests and data analysis with newer methods like machine learning, to find the best solutions.

Poor water quality is a major health issue globally. Over 2 billion people drink water contaminated with feces. [1] There are some factors that will affect the quality of the water: wells, rivers, lakes, or reservoirs. The quality of source water is critical because it is determined by the level of contaminants and impurities checked before any treatment. These are affected by overuse of land by harmful agriculture runoff, industrial discharge, and chemical waste, which can introduce nitrates or heavy metals. Proper management is essential to prevent blooms and other issues.

Handling and distribution involve the transfer of water from the source to consumers, including storage, treatment, and distribution systems. it impacts the purity of water, which is affected by storage in poorly maintained tanks or pipes that become contaminated by debris or biofilm growth. Pipes and infrastructure through old corroded pipes harmful substances into the water, such as lead or rust particles. Treatment facilities including coagulation, sedimentation, filtration, and disinfection, are critical for removing contaminants. Ineffective treatment or failures in the treatment process can leave harmful substances in the water. Distribution system ensuring that distribution networks are well-maintained helps prevent contamination and provide safe water to consumers. Regular Maintenance to repair the water system and routine checks while filtration before reaching the consumer, System inspections help to upgrade the integrity of water supply and Sanitation of all parts helps to prevent the growth of harmful microorganisms and the buildup of biofilms. Filtration is also required to remove impurities. Such methods are activated carbon filters, reverse osmosis, and UV treatment. Regular maintenance and Filtration at homes impact drinking water quality.

This research aims to find the several key issues related to safe drinking water. It tackles potential misinterpretations of WHO criteria for water safety and inefficiencies in current clinical methods for predicting water potability, and the lack of effective applications for water quality prediction. Additionally, it highlights the need to raise awareness among rural populations about crucial factors affecting access to safe drinking water. By comparing the effectiveness of different machine learning algorithms, the research seeks to develop a more efficient method for predicting water potability, ultimately contributing to improved public health outcomes.

III. RELATED WORK

Pal et al. [6] studied different machine learning models for predicting water quality and found that Artificial Neural Networks (ANN) were the most effective with a 99.1% accuracy.

Patel et al. [5] developed a method for predicting water potability using SMOTE to handle class imbalance. They found that Gradient Boost and Random Forest models performed best and both with 81% accuracy.

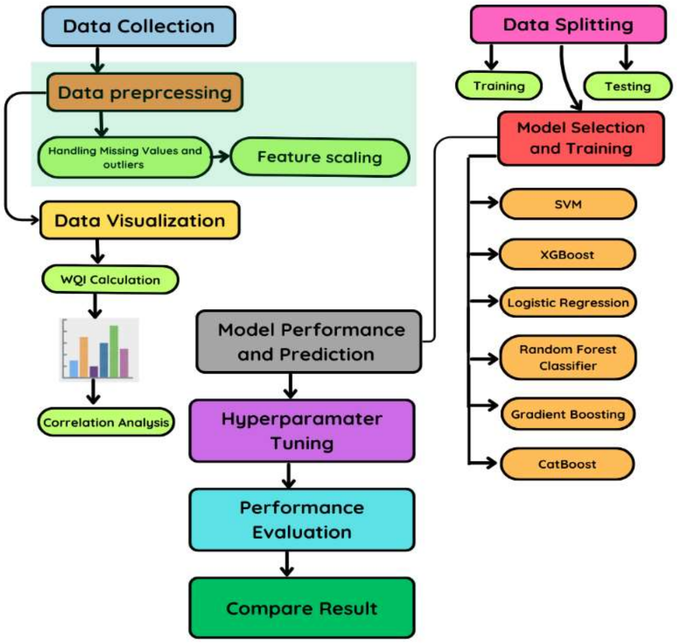
Uddin et al. [7] evaluated water quality prediction models and found that Support Vector Machines (SVM) performed well and achieving over 95% accuracy.

Aldhyani et al. [8] compared several AI and ML algorithms for predicting water quality, with the SVM achieving a high accuracy of 97.01% and just slightly behind the NARNET model.

Tufail et al. [9] studied water potability in Pakistan and found that SVM was the most effective model for making predictions.

IV. 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.



The main goal of this research is to develop accurate machine learning models that can effectively predict water potability. This will help in better water management and ensure that communities have access to clean drinking water.

A. *COLLECTING DATA*

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

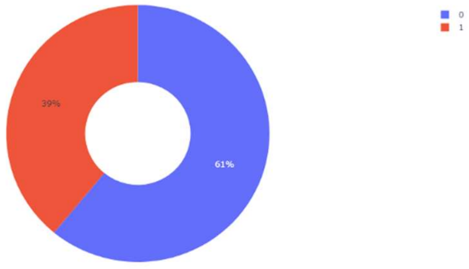
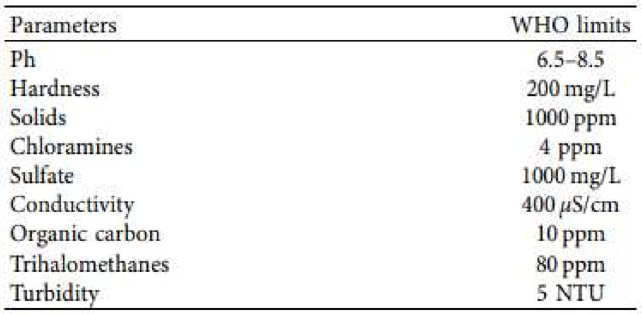


Figure 1. Pie chart of the water potability

The dataset contains nine key physicochemical parameters: pH, hardness, solids, chloramines, sulfates, trihalomethanes, organic carbon, conductivity, and turbidity but two of that features contain null values. These features, along with a target feature for potability, were used to make predictions using different machine learning algorithms.



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. Drinkable Water Quality Standards

The standard values of each important quality parameter recommended by WHO and the EPA are represented in Table 1. If the values of these parameters exceed their standard limit means water is not suitable for drinking.

B. *DATA PREPROCESSING*

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

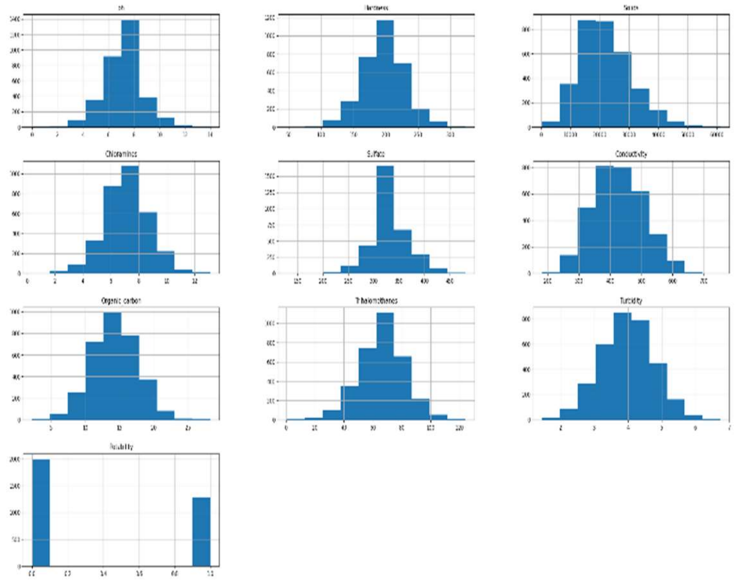


Figure 2. Distribution of all water parameters

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 wate contain no value. Entire column has Null Value.

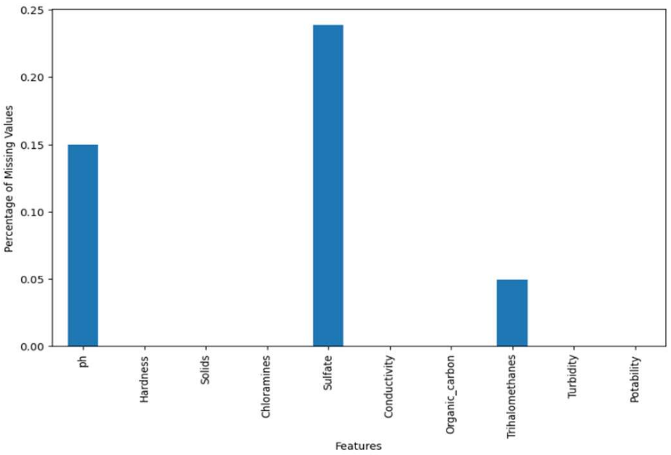


Figure 3. Missing Value Data

b. *Data Normalization Using Z-Score:* The Z-Score normalisation method standardises data by converting values into a common scale. Values are transformed to show how many standard deviations they will obtain from mean ideally within the range of -3 to +3.

C. *DATA VISUALISATION*

D. *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, shown in Figure 11, was generated to visualize these relationships clearly. This heatmap helps identify trends and interactions among features.

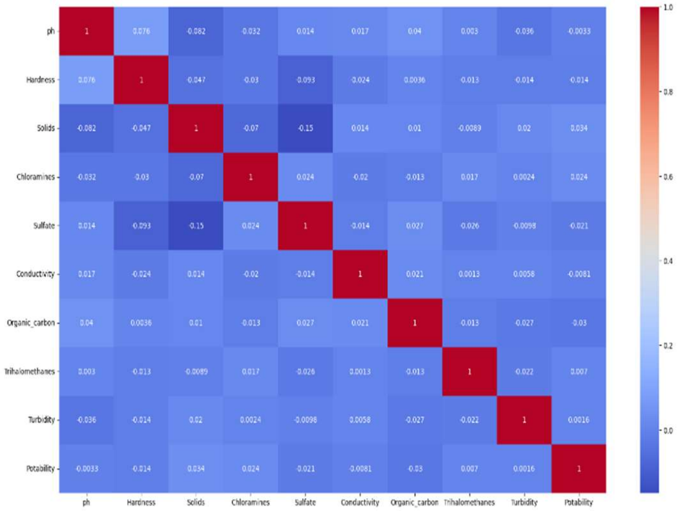


Figure 4. Correlation Analysis

E. *DATA SPLITTING*

The dataset consisting of 3,276 samples was split into training and testing sets using a 70:30 ratio. This means 70% of the data (2,293 samples) was used for training the machine learning models, while the remaining 30% (983 samples) was set aside for testing. This approach allows for robust model performance assessment, ensuring that the models are not just overfitting to the training data but can also generalize effectively on unseen data. The split ratio is a standard practice to balance between model training and evaluation accuracy.

F. *MODEL SELECTION*

i. XGBoost

XGBoost is a powerful machine learning algorithm widely used for prediction, handles complex data and missing values well. It combines multiple decision trees to make accurate predictions and uses regularization to avoid overfitting. Its speed and ability to handle large datasets make it ideal for water quality analysis.

ii. Random Forest

Random forest is widely used in machine learning algorithms and is effective in both regression and classification tasks. It combines multiple decision trees which can each built from random subsets of data to improve accuracy and handle complex datasets. The final prediction is made based on a "voting" system, where the majority vote from all trees determines the result. This approach makes Random Forest effective at managing noise, outliers, and high-dimensional data.

iii. Support Vector Machine (SVM)

Support Vector Machines (SVM) are used in supervised learning to analyse and classify water samples as either potable or non-potable by analysing their chemical and microbial properties. It works with complex data that is in non-linear form. The algorithm Function by creating a hyperplane that separates the two potable and non-potable classes of water samples. It works well with complex and large datasets which is reducing overfitting by maximizing the margin between classes.

iv. Logistic Regression

Logistic regression is mostly used in prediction systems as a supervised learning algorithm operating for main particular categorical target features. Basically, it estimates the probability of a target outcome based on input features. Primarily used for binary classification tasks in the water quality domain, logistic regression can also be adapted for multiclass classification problems. Its straightforward yet effective approach to modelling the relationship between dependent and independent variables makes it particularly useful for predicting the potability of water samples.

v. CatBoost

CatBoost is a gradient boosting algorithm designed to handle categorical data without extra steps. It simplifying workflows and improving performance, especially on datasets with many categories. CatBoost also reduces overfitting, making it effective for small datasets, and supports fast training on both CPUs and GPUs while natively handling missing values. It's a powerful choice for both classification and regression tasks.

vi. Gradient Boosting

Gradient Boosting is an ensemble learning technique that builds a strong predictive model by combining multiple weak models with usually decision trees. It works by training models sequentially where each new model tries to correct the errors made by the previous ones. This method uses gradient descent to minimize the loss function, making it flexible and effective for various tasks. Gradient Boosting often delivers high accuracy and provides insights into feature importance, helping to identify which features are most impactful. However, it can be prone to overfitting, so careful tuning of parameters like learning rate and tree depth is necessary.

G. MODEL PERFORMANCE

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

a) **Precision**: Precision is calculated as the number of true positives divided by the total number of true positives and false positives.

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

b) **Accuracy**: represents the proportion of correctly predicted outcomes.

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

c) **Recall**: indicates the proportion of true positives correctly identified.

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

d) **F1 Score**: Combines precision and recall into a single metric with a score range from 0 to 1, where a higher score indicates better performance.

$$F1\,Score = \frac{2 \times Precision \times Recall}{Precision \times Recall}.$$

Hyperparameter Tuning: It involves adjusting model settings to find the best result and optimize performance. Models often have numerous hyperparameters and making the process of finding the best configuration of challenging search task. Two common methods for hyperparameter tuning include:

GridSearchCV systematically searches for all possible hyperparameter combinations to find the best set while **RandomizedSearchCV** randomly samples a fixed number of combinations for better efficiency. Both methods were used to tune five classifiers, leading to improved accuracy and performance

V. EXPERIMENT OUTCOME

This research contain dataset comprising 3,276 samples was used, with each sample analysed across nine water quality parameters: pH, Organic Carbon, Chloramines, Turbidity, Trihalomethanes, Sulphate, Hardness, Conductivity, and Solids. The dataset was split into a 70:30 ratio for training and testing. The aim of the study was to access the performance of various algorithms—Gradient Boosting, Logistic Regression, Random Forest (RF), XGBoost, CatBoost and Support vector Machine (SVM). The results visualised in a bar graph, revealed that Gradient Boosting achieved the highest accuracy with 99% and with these SVM scoring the lowest 62% and Logistic Regression holds the 63% accuracy there might be small difference.

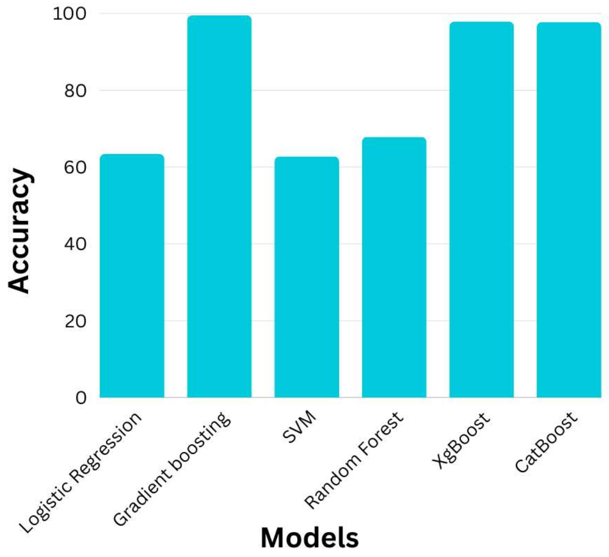


Figure 5. Bar Chart of Accuracy Comparison

Model	Accuracy (%)	Precision	Recall	F1-score
Gradient Boosting	99.49	1.00	1.00	1.00
XGBoost	97.86	0.98	0.98	0.98
CatBoost	97.76	0.98	0.99	0.98
SVM	62.77	0.63	1.00	0.77
Random Forest	67.85	0.70	0.86	0.77
Logistic Regression	63.38	0.63	1.00	0.77

Table 2. Proposed algorithms Performance Analysis

VI. Conclusion

Assuring safe and clean drinking water is vital for human health, especially with the growing global population and increasing pollution. Machine learning techniques offer a promising way to predict water potability and improve water quality. A recent study examined different machine learning algorithms to predict water potability based on specific water quality factors. The research shows the potential of these methods to help monitor and manage water quality benefiting public health. However, the study was limited by a small dataset of 3,276 samples and only a few water quality parameters, making it harder to apply the findings broadly. Future research should explore more factors that affect water safety.

References

[1] Computational Intelligence and Neuroscience - 2022 - Patel - A Machine Learning-Based Water Potability Prediction Model

[2] WATER POTABILITY PREDICTION USING MACHINE LEARNING Samir Patel, Khushi Shah, Sakshi Vaghela, Mohmmad Ali Aglodiya, Rashmi Bhattad

[3] Comparison of machine learning algorithms in statistically imputed water potability dataset Diwash Poudela, Dhadkan Shresthaa, Sulove Bhattaraia and Abhishek Ghimirea

[4] Cabral, João PS. "Water Microbiology. Bacterial Pathogens and Water – PMC." NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996186/.

[5] Patel, J., Amipara, C., Ahanger, T.A., Ladhva, K., Gupta, R.K., Alsaab, H.O., Althobaiti, Y.S. and Ratna, R., "A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI", Computational Intelligence and Neuroscience: CIN, 2022.

[6] Pal, O.K., "The Quality of Drinkable Water using Machine Learning Techniques", Int. J. Adv. Eng. Res. Sci., 8, p.5. 2021.

[7] Uddin, M.G., Nash, S., Rahman, A. and 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, pp.808-828, 2023.

[8] Aldhyani, T.H., Al-Yaari, M., Alkahtani, H. and Maashi, M., "Water quality prediction using artificial intelligence algorithms", Applied Bionics and Biomechanics, 2020

[9] Addisie, M.B., "Evaluating Drinking Water Quality Using Water Quality Parameters and Esthetic Attributes", Air, Soil and Water Research, 15, p.11786221221075005, 2022.

[10] pH in drinking-water[M]. World Health Organization, 2007.

[11] Kent health care products[M/OL]. Kent RO Systems, 2020. https://www.kent.co.in/blog/what-are-total-dissolvedsolids-tds-how-to-reduce-them/.

[12] Chloramine in drinking water[M]. World Health Organization, 1998.

[13] United states environmental protection agency [EB/OL]. 2003. https://www.epa.gov/sites/default/files/2014-09/documents/ support_cc1_sulfate_healtheffects.pdf.

[14] United states environmental protection agency [EB/OL]. https: //archive.epa.gov/water/archive/web/html/vms59.html.

[15] Whitehead P. Elga veolia[J/OL]. 2021. https:// www.elgalabwater.com/blog/total-organic-carbon-toc. [11] Guidelines for canadian drinking water quality: Trihalomethanes[M].