Analyzing Water's Characteristics Health Impact with Classification Algorithms

Khadim Gueye(✉), Ndiouma Bame, and Aliou Boly

Dept. Mathématiques et Informatique, Cheikh Anta Diop University, BP 5005, Dakar Fann, Senegal
{khadim40.gueye,ndiouma.bame,aliou.boly}@ucad.edu.sn

Abstract. The water crisis is compounded by a number of factors, including population growth. In order to assess water potability, several indicators need to be taken into account during water quality evaluation. The World Health Organisation (WHO) sets concentration standards for each parameter to ensure that it is fit for drinking. The aim of this work is to take an in-depth look at these various water parameters, which have a significant impact on human health, and to understand how they influence water quality by using advanced machine learning techniques. The methodology consists on the one hand, to build a model for predicting the potability of water and on the other hand to study the impact of certain physico-chemical factors related to human health in this potability. The study was based on the use of three machine learning algorithms, namely Decision Tree, XGBoost and Random Forest, to analyze the impact of parameters such as pH, chlorine, chlorides, turbidity, nitrates, conductivity and fluoride. The results for the prediction model are promising especially for the Random Forest algorithm which gives the best performances. Regarding the impact of physico-chemical factors in the potability, all the algorithms place pH and chlorine at the top. Other parameters such as chlorides and turbidity are also significant, although their contribution is slightly lower than that of the previous characteristics.

Keywords: Anomaly detection · water quality · machine learning algorithms · health parameters · impact of physico-chemical parameters

1 Introduction

Water scarcity is a major global crisis, resulting in the death of 297,000 children under the age of five each year due to diarrheal diseases. Eighty percent of industrial and municipal activities discharge their wastewater into the environment without prior treatment[1,2]. Water quality is assessed based on several parameters to ensure its safety for human consumption. Nitrate, arsenic, chlorine, pH,

[1] Unisef rapport 2019.
[2] Unesco rapport 2021.

K. Cheikh M. F. Kebe et al. (Eds.): InterSol 2024, LNICST 610, pp. 59–67, 2025.
https://doi.org/10.1007/978-3-031-86493-3_6

lead, fluoride, and turbidity are among the substances found in water, with maximum recommended concentrations set by the World Health Organization (WHO)[3]. The emergence of machine learning algorithms presents an opportunity to establish effective systems for monitoring and ensuring the quality of drinking water [1,4]. These tools provide an innovative approach to analyzing and predicting water quality, enabling informed decision-making by policymakers and water resource managers. In this context, our study aims to explore the water parameters that most significantly impact human health and to examine how these parameters affect water quality assessment using machine learning algorithms. We will detail our exploration of essential drinking water parameters and conduct an in-depth analysis of water anomaly detection, emphasizing the specific advantages of machine learning algorithms in this domain. Finally, we will conclude by identifying future research perspectives related to this issue.

2 Drinking Water Parameters

For water to be drinkable, it must comply with more than 60 parameters grouped into five main groups: organoleptic parameters, microbiological parameters, physico-chemical parameters, toxic substance parameters and undesirable substance parameters [5,6,10]. Some of these physico-chemical parameters have a significant impact on human health. In this section, we will examine some key water parameters, looking at their impact on health and the concentrations recommended by the World Health Organisation (WHO).

2.1 Chlorine (Cl^-)

Chlorine (Cl^-) disinfects water by eliminating pathogenic microorganisms such as bacteria, viruses, and parasites, and also acts as an oxidant to reduce unpleasant odors and tastes. However, high exposure to chlorine can cause skin and eye irritations, respiratory issues, and even long-term carcinogenic risks (See footnote 3)[4]. The WHO recommends a limit of 5 mg/L of chlorine in drinking water.

2.2 Hydrogen Potential (pH)

The pH of water is crucial for human health and overall water quality. Extreme pH levels, either too acidic or too basic, can promote the growth of microorganisms such as bacteria, fungi, and algae, potentially leading to health issues (See footnotes 3 and 4). To ensure the quality of drinking water and public health, it is essential to maintain the pH within an optimal range, typically between 6.5 and 8.5.

[3] WHO: Guidelines for Drinking-Water Quality: 4th ed. Incorporating First Addendum.

[4] WHO: Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First and Second Addenda.

2.3 Turbidity (NTU)

The turbidity of water measures the amount of suspended particles, such as organic matter, bacteria, viruses, and algae (See footnotes 3 and 4). High turbidity in drinking water can indicate the presence of harmful contaminants like pathogenic microorganisms, chemicals, or heavy metals. Quality standards recommend a maximum limit of 5 nephelometric turbidity units (NTU) to ensure the safety of drinking water.

2.4 Electrical Conductance (EC)

Electrical conductivity measures the water's ability to conduct electricity. This capacity is linked to the presence of dissolved salts in the water, such as sodium, calcium, magnesium, chloride, and sulfate ions. The presence of dissolved salts in water can affect its taste, odor, color, and turbidity (See footnotes 3 and 4). The WHO sets a maximum limit of 1000 microsiemens per centimeter (μS/cm) for the electrical conductivity of drinking water.

2.5 Nitrate (NO_3^-)

Nitrates (NO_3^-), primarily from excessive use of nitrogen fertilizers in agriculture, pose a major concern in water. At high levels, they can convert to nitrites in the body, disrupting the blood's ability to carry oxygen (See footnote 3). This is particularly concerning for infants, who may develop methemoglobinemia, commonly known as blue baby syndrome (See footnote 4). The WHO recommends a maximum limit of 50 mg per liter (mg/L) of nitrates in drinking water to ensure safety.

2.6 Fluorine (F^-)

Fluoride (F^-) is often naturally present in water, but at excessive levels it can cause dental fluorosis (stains on the teeth) and, at much higher concentrations, fluoride can cause bone problems such as bone fluorosis, which manifests itself as changes in bone structure (See footnotes 3 and 4). The WHO requires an optimum concentration of fluoride in drinking water of between 0.5 and 1.5 milligrams per litre (mg/L), with a maximum permissible concentration of 1.5 mg/L.

2.7 Trihalomethane (THM)

Trihalomethanes are a concern due to their carcinogenic potential and effects on human health. Prolonged exposure to high concentrations of trihalomethanes in drinking water has been associated with an increased risk of bladder, kidney, and colon cancer (See footnotes 3 and 4). Additionally, they can affect the respiratory system and cause skin problems. To safeguard public health, the WHO has established a maximum limit of 0.1 mg/L for trihalomethanes in drinking water.

Although the WHO issues international guidelines, each country may have its own standards and regulations regarding water quality, including maximum allowable concentrations (MAC) for contaminants. These standards can vary based on environmental conditions, agricultural practices, water resource availability, and country-specific public health considerations.

3 State of the Art on Water Anomaly Detection

Anomaly detection methods include a variety of approaches such as statistical models, supervised and unsupervised machine learning, signal processing techniques and deep learning-based methods. These techniques aim to extract unusual behaviour or aberrant patterns in data, enabling organisations to react quickly and take preventive action. Many researchers have explored machine learning algorithms to accurately assess the potability of water. This trend is widely documented in the scientific literature.

- In [11] Dorado-Guerra et al., the study employed Random Forest (RF) and XGBoost algorithms to analyze nitrate concentrations. The input data were divided into training (70%) and test (30%) sets, encompassing nineteen variables that span various climatic, hydrological, ecological, and anthropogenic aspects. The algorithms demonstrated high correlations (0.93 for RF and 0.92 for XGBoost). Predictions of nitrate concentrations were accurate with both methods, evaluated using the KGEM index ranging from 0.85 to 0.90 for RF and from 0.77 to 0.80 for XGBoost.
- In [12] Chinnappan et al., the study by presents a water quality monitoring system integrating sensors such as temperature, flow, and chlorine, connected to a Raspberry Pi 3 microcontroller. This system uses a decision tree to predict chlorine levels based on variables like temperature and pH. Data is preprocessed to remove outliers and normalize it. Fuzzy rules describe the relationship between input variables and chlorine levels, with a fuzzy algorithm adjusting the solenoid valve accordingly. The chlorine sensor measures in PPM, and the model defines ranges to interpret chlorine levels. Model performance includes a recall of 90%, precision of 92%, F-score of 89%, and an AUC of 91%.
- The study [13] by Ivan Ivanov, Borislava Toleva examines a dataset on the potability index of water with a dataset available on Kaggle. Missing values are filled in using the mean of the corresponding variable. To avoid overfitting, the data are randomly mixed. The input variables are standardised using the StandardScaler method, and the data are divided into training (72%) and test (28%) sets. Various classification models, such as SVM, DT and Random Forest (RF), are fitted. The paper evaluates the performance of the algorithms using measures such as precision, recall and f1 score. In particular, the Random Forest model has a precision of 0.81, a recall of 0.81 and an f1 score of 0.81.

In these studies, several machine learning algorithms are used. In [11] Dorado-Guerra et al. successfully employed the powerful XGBoost and Random Forest (RF) algorithms to assess nitrate concentration, obtaining satisfactory results. In [12] Chinnappan et al. opted for a decision tree to determine chlorine concentration, following predefined rules, and obtained highly conclusive accuracies. In [13] Ivan Ivanov and Borislava Toleva explored various classification models, including SVM, DT, and Random Forest, to assess the overall potability of water, achieving their best results using RF. In summary, algorithms such as Random Forest, Decision Tree, and XGBoost are frequently used in water potability assessment, providing satisfactory results.

However, although determination and detection are general, the impact of each individual parameter is not specified. Knowing this information would enable managers to focus more sharply on the parameters that have the greatest impact on water potability.

4 Our Proposition Methodology

In this section, we present the objective of our proposition and we detail the adopted methodology, starting with a study of the algorithms. We will then present the dataset used and the experimentation, before concluding with a discussion of the obtained results.

4.1 Objective

Our research is committed to taking an in-depth look at the various water parameters that have a significant impact on human health. We aim to understand how these parameters influence water quality using advanced machine learning techniques. The objective is to identify the influence of these parameters on the accurate assessment of water potability.

4.2 Machine Learning Algorithms

There are several machine learning algorithms in the literature, but some stand out from the rest in terms of anomaly detection based on the frequency, rarity and proportion of anomalies in the dataset. These algorithms include the Decision Tree (DT), XGBoost and Random Forest [7–9], which are widely used for anomaly detection.

Decision Tree (DT): A decision tree is a supervised learning method that progressively divides the dataset into smaller subsets based on feature values. The goal is to create a model that predicts the value of the target variable by learning simple decision rules derived from the data's features.

eXtreme Gradient Boosting (XGBoost): XGBoost is an advanced implementation of the gradient boosting algorithm used to build prediction models by sequentially adding base models, typically weak decision trees. The primary objective is to optimize the loss function at each step to minimize overall prediction error. This method works by sequentially adding trees to adjust the residuals of the existing model, thereby minimizing the loss function. XGBoost employs efficient optimization techniques such as gradient computation and variance reduction to accelerate the learning process.

Random Forest (RF): Random Forests are decision tree-based algorithms used to detect anomalies in water. They create multiple trees from random data samples, with variable randomization at each node to prevent overfitting. While effective for large datasets and robust against outliers, interpreting them can be challenging due to the numerous trees and parameters to adjust, and they may be sensitive to imbalanced data.

4.3 Dataset

In order to determine the impact of physico-chemical parameters of water on the determination of its potability and having dangerous consequences on human health, we use a dataset available on the Kaggle platform. The dataset contains 2140 records with 15 input parameters including pH, nitrate, fluoride, conductivity, chlorine and chloride and a binary output parameter which is potability. Missing values can compromise the quality of a model's results. We opted to use the SimpleImputer function to replace missing values with the mean, an approach commonly used in data pre-processing. Next, the SMOTE method is applied to manage class imbalance. Finally, we split the dataset into a training set (80%) and a test set (20%) before moving on to implementation.

4.4 Results and Discussion

To evaluate the proposition, metrics such as precision, recall and F-score are used to study the performance of classification models.

Results: Table 1 contains the performance of three classification algorithms: Decision Tree, XGBoost and Random Forest in determining the potability of water. Each algorithm is evaluated in terms of precision, recall and F-score. Overall, it can be seen that the Random Forest and XGBoost algorithms outperform the Decision Tree in all metrics. This suggests that these two algorithms have a greater ability to classify data correctly than to detect anomalies compared with the Decision Tree.

The importance of the different water parameters in the dataset as a function of the different types of algorithm is presented in Table 2.

Table 1. Algorithms performances

Classifier	Precision	Recall	F-score
Decision Tree	80.00	69.86	74.59
XGBoost	87.14	79.91	83.37
Random Forest	86.69	82.53	84.56

Table 2. Importance of parameters on the potability of water (%)

Classifier	DT	RF	XGBoost
pH	23.98	20.82	21.61
Chlorine	15.93	15.98	15.72
Chloride	14.98	15.41	14.80
Turbidity	15.19	14.53	14.50
Nitrate	12.24	14.05	12.96
Fluoride	10.35	11.57	10.81
Conductivity	07.29	08.23	08.96

The results obtained from the importance of the characteristics for each model (Random Forest, XGBoost and Decision Tree) provide valuable information on the relative contribution of each characteristic to the prediction of water potability.

Discussion: In all three models, pH is the most important characteristic, with importance values ranging from 20.82% to 23.98%. This suggests that the pH level of the water is a crucial factor in determining its potability. High or low pH values can indicate potentially undrinkable water due to its extreme acidity or alkalinity. Chlorine is also an important characteristic in all models, with importance values of 15.93%, 15.72% and 15.98% for the Decision Tree, XGBoost and random Forest respectively. This is consistent with the fact that chlorine is commonly used as a disinfectant in drinking water treatment to remove microbial contaminants. Chloride is also considered important in all three models with importance values ranging from 14.80% to 15.41%, although its contribution is slightly less than that of chlorine. Chloride levels in water can come from natural sources or from water treatment processes, and high levels can indicate contamination by soluble salts. Water turbidity, a measure of water clarity based on the presence of suspended particles, is also important in all models. High levels of turbidity can indicate undrinkable water due to the presence of unwanted particles. Nitrate, Fluoride and Conductivity are also important but have a slightly lower contribution than the previous characteristics. High levels of nitrate can come from agricultural or industrial sources, while the presence of fluoride can be controlled to avoid excessive concentration. Conductivity can be related to the amount of dissolved salts in the water, which may indicate the presence of certain minerals. These results underline the importance of monitoring a varied set of characteristics to assess the potability of water. pH, chlorine, chloride, turbidity and other characteristics play critical roles in determining the quality of drinking water, and monitoring them regularly is essential to ensure public health. Different machine learning methods provide consistent insights into the factors important to water potability, boosting confidence in water treatment recommendations and decisions.

5 Conclusion and Perspectives

The study focused on analysing the influence of water characteristics on human health, with particular emphasis on predicting the potability of water by assessing the parameters that have a direct impact on human health. Through the analysis of many works dealing with the potability of water and the use of machine learning algorithms such as random forest, XGBoost and decision tree, we evaluate the effectiveness of these models and determine the importance of each parameter. The obtained results, expressed in terms of precision, recall and F1-score, showed that the XGBoost and Random Forest algorithms outperformed the Decision Tree in terms of predictive performance. In addition, analysis of the specific results for different parameters such as pH, chlorine, chloride, turbidity, nitrates, fluoride and conductivity revealed significant variations in the contribution of these factors to the prediction of water potability. By analysing the results, we observed that pH, chlorine and turbidity have a predominant influence on the prediction of water potability, while other parameters such as nitrate and fluoride also play a significant role. These results can inform decision makers and health professionals in their efforts to ensure a safe, high-quality water supply for all. Particular attention should be paid to identifying the most relevant combinations of parameters for predicting drinking water quality. This may require in-depth analyses, such as correlation studies and data mining techniques, to determine the most significant relationships between different parameters. By developing models that take into account the conjunction of several parameters, it will be possible to gain a better understanding of the mechanisms underlying water potability and identify effective strategies for mitigating the associated risks. This combination of parameters could play a crucial role in water quality management, helping to ensure universal access to safe drinking water.

References

1. Ainapure, B., Baheti, N., Buch, J., Appasani, B., Vidyakant, J., Srinivasulu, A.: Drinking water potability prediction using machine learning approaches: a case study of Indian rivers. Water Pract. Technol. (2023)
2. Mondal, A., Dubey, S.: Machine learning-based water potability prediction: model evaluation, and hyperparameter optimization. In: Tripathi, A.K., Shrivastava, V. (eds.) Advancements in Communication and Systems, SCRS, India, pp. 37–54 (2024)
3. Gao, H., Li, Y., Lu, H., Zhu, S.: Water Potability Analysis and Prediction. Highlights in Science, Engineering and Technology, pp. 70–77 (2022)
4. El Fadel, D.: Water resources evaluation and potability in north-east of Algeria. Alger J. Eng. Archit. Urban **5**(4), 192–198 (2021)
5. Soulounganga, P., Ndjeri-Ndjouhou, M., Ngohang, F.: Drinking water consumption habits and perception of the organoleptic quality of tap water by the population of Greater Libreville (Gabon). Int. J. Biol. Chem. Sci. 1117–1130 (2023)
6. Josiane, C., et al.: Physico-chemical and bacteriological characteristics of river waters. Afrique Sci. **23**(2), 50–64 (2023)

7. Didavi, K., Agbokpanzo, R., Agbomahena, M.: Comparative study of decision tree, random forest and XGBoost performance in forecasting the power output of a photovoltaic system. In: Proceedings of a Conference, pp. 1–5 (2021)
8. Meric, E., Ozer, C.: Symptom Based Health Status Prediction via Decision Tree, KNN, XGBoost, LDA, SVM, and Random Forest, pp. 193–207 (2023)
9. Helmud, E., Fitriyani, F., Romadiana, P.: Classification comparison performance of supervised machine learning random forest and decision tree algorithms using confusion matrix. Jurnal Sisfokom (Sistem Informasi dan Komputer) **13**, 92–97 (2024)
10. Melin, J.: Occurrence and fate of metabolites of four pesticide families (neonicotinoids, carbamates, organophosphates, phenylpyrazoles) in drinking water resources and drinking water treatment plants (PhD thesis) (2020)
11. Dorado-Guerra, D.Y., Corzo-Pérez, G., Paredes-Arquiola, J., Pérez-Martín, M.Á.: Machine learning models to predict nitrate concentration in a river basin. Environ. Res. Commun. **4**(12), 125012 (2023)
12. Chinnappan, C.V., et al.: IoT-enabled chlorine level assessment and prediction in water monitoring system using machine learning. Electronics **12**, 1458 (2023)
13. Ivanov, I., Toleva, B., Taylor, G.: Predicting the water potability index using machine learning. Environ. Ecol. Res. **11**, 537–542 (2023)