Optimizing Water Filtration Through Machine Learning Based Water Quality Indexing

Zaeem Farooq

zaeem.ips[at]outlook.com
marshallawcursae[at]gmail.com

Abstract: ***Water quality is a critical aspect of public health and environmental sustainability, necessitating efficient and effective water filtration systems. The selection of appropriate water filtration technology, whether Reverse Osmosis (RO), Ultraviolet (UV), or Ultrafiltration (UF), is pivotal to ensuring safe and clean water. This research explores the application of predictive analytics and machine learning techniques to optimize the selection process between these filtration methods based on various water quality parameters. This study explores the application of machine learning techniques for optimizing water filtration strategies based on water quality parameters The research aims to predict the most suitable filtration methods, including Reverse Osmosis RO, Ultraviolet UV, and Ultrafiltration UF derived from water quality indexing (WQI). We collected extensive datasets encompassing a range of water quality indicators, such as pH, turbidity, temperature, dissolved oxygen, total dissolved solids (TDS), and concentrations of various contaminants. Machine learning algorithms, including decision trees, random forests, support vector machines (SVM), and neural networks, were then applied to this dataset to develop predictive models. These models were trained to classify and recommend the most suitable filtration technology based on the input water quality parameters. We leveraged a supervised learning approach in order to design as accurate as possible predictive models from a labelled training dataset for the identification of filtration methods. A set of physiochemical and microbiological parameters as input features help represent the water's status and determine its suitability class namely safe, non-safe, moderate, or excellent. A comparative evaluation of various machine learning models is done to identify the best algorithm and classify the data into labels. These labels are water quality index (WQI) derived from the classifier algorithms for the objective of this study, K-means Clustering, kNN, AdaBoostM1, Random Forest (RF), K-means clustering, Stacking, Voting and Bagging are selected in order to establish the desired filtration approach i.e., RO, UV, UF, TDS with the greatest precision and accuracy. The results demonstrated that machine learning models could predict the optimal filtration method with high accuracy. Decision trees and random forests showed particularly robust performance, with accuracy rates exceeding 90%. These models were able to handle the complex and nonlinear relationships between water quality parameters and the effectiveness of different filtration methods. In addition to predictive accuracy, the models provided insights into the importance of various water quality parameters in the decision-making process. This research underscores the potential of machine learning and predictive analytics in transforming water quality management. Future work will focus on refining the models with larger datasets, incorporating more diverse water quality parameters, and exploring the application of deep learning techniques for even more accurate predictions. The ultimate goal is to develop a comprehensive decision support system that can be deployed in various water treatment facilities to ensure the delivery of safe and clean water to communities worldwide.***

Keywords: Water Quality Index, Machine Learning, Filtration Strategies, Predictive Analytics, Environmental Science

1. Introduction

The research begins with a comprehensive review of the existing literary works on water purification innovations and their corresponding capacities and limitations. RO systems are known for their capability to remove liquified salts and other contaminations [7], UV systems work in inactivating microorganisms without changing the chemical structure of water [8], and UF systems excel at removing put on hold solids and pathogens as an example, TDS and turbidity were found to be considerable forecasters for advising RO systems, while microbial web content was a crucial factor for UV systems. The UF systems were often chosen based upon the visibility of put-on hold solids and specific virus. The development of water top quality evaluation and therapy innovations has actually undergone a considerable improvement, proceeding from standard mechanical and chemical procedures to innovative, data-driven systems. At first, water quality was examined by manually gathering samples and performing lab evaluations, a method that was not only labor-intensive but additionally did not have real-time data [10] Early treatment techniques mostly concentrated on simple filtration and chemical therapies. Nevertheless, the arrival of digital innovations has transformed the industry. Modern advancements in sensing units and automated systems now make it possible for real-time tracking, offering much more accurate and prompt data on different water high quality parameters such as pH, temperature, turbidity, and contaminant degrees. In addition, the integration of corrosion-resistant products with high wear resistance has actually improved the effectiveness of water systems. The use of GIS and remote picking up innovations has better improved the ability to keep an eye on huge water bodies and containers effectively. The trustworthiness of AI prediction techniques is being analyzed in figure 1[4].

DOI: https://dx.doi.org/10.21275/SR24826150426

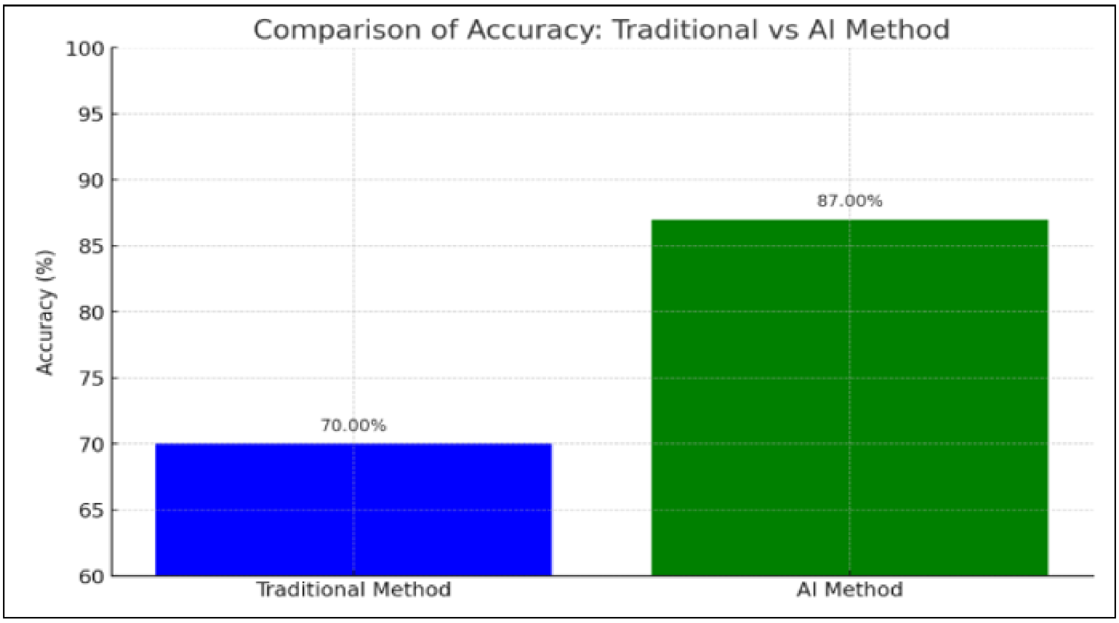


Figure 1

The advancement of artificial intelligence (AI) and expert system has actually introduced a brand-new age in water quality administration. These technological innovations have fundamentally transformed how water quality data is analyzed and predicted [4] Artificial intelligence formulas have actually confirmed very efficient in numerous applications, including forecasting contamination events, identifying air pollution resources, and maximizing treatment techniques [5,16] These algorithms can manage big datasets and spot complicated patterns, significantly progressing research study in these areas. The use of AI-driven systems in water quality monitoring gets on the surge, encompassing not just checking however additionally forecasting and replying to problems [3,4]. Water source management systems now enable proactive oversight, assisting in early discovery of prospective dangers and thereby decreasing risks to public health and the environment. A comparative evaluation of various machine learning models is done to identify the best algorithm and classify the data in to labels. These labels are water quality index (WQI) derived from the classifier algorithms For the objective of this study, Artificial neural Network (ANN), kNN, AdaBoostM1, Random forest (RF), k means classifier, Stacking, [2] Voting and Bagging are selected in order to establish the desired filtration approach i.e RO, UV, UF, TDS with the greatest precision and accuracy .Synthetic Minority Oversampling Strategy SMOTE is used to enhance the quality prediction with 10-fold cross-validation. The subject area of this article is applying machine learning techniques for water quality indexing and filtration strategies.

2. Literature Review

When it comes to estimating water quality using machine learning, Ahmed, U et al. [23] estimated water quality using classical machine learning algorithms namely, Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks (Deep NN) and k Nearest Neighbors (kNN), with the highest accuracy of 93% with Deep NN. The estimated water quality in their work is based on only three parameters: turbidity, temperature and pH, which are tested according to World Health Organization (WHO) standards. Using only three parameters and comparing them to standardized values is quite a limitation when predicting water quality. Ahmad et al. [**8**] employed single feed forward neural networks and a combination of multiple neural networks to estimate the WQI. They used 25 water quality parameters as the input. Using a combination of backward elimination and forward selection selective combination methods, they achieved an R2 and MSE of 0.9270, 0.9390 and 0.1200, 0.1158, respectively. The use of 25 parameters makes their solution a little immoderate in terms of an inexpensive real time system, given the price of the parameter sensors. Sakizadeh [**9**] predicted the WQI using 16 water quality parameters and ANN with Bayesian regularization. His study yielded correlation coefficients between the observed and predicted values of 0.94 and 0.77, respectively. Abyaneh [**10**] predicted the chemical oxygen demand (COD) and the biochemical oxygen demand (BOD) using two conventional machine learning methodologies namely, ANN and multivariate linear regression. They used four parameters, namely pH, temperature, total suspended solids (TSS) and total suspended (TS) to predict the COD and BOD. Ali and Qamar [**11**] used the unsupervised technique of the average linkage (within groups) method of hierarchical clustering to classify samples into water quality classes. However, they ignored the major parameters associated with WQI during the learning process and they did not use any standardized water quality index to evaluate their predictions. Furthermore, in [**9**], four algorithms, namely RF, M5P, Random Tree (RT) and Reduced Error Pruning Tree (RepTree), and 12 hybrid data-mining algorithms (combinations of standalone with Bagging, CV Parameter Selection (CVPS) and Randomizable Filtered Classification) were used to create the Iran water quality index (IRWQIsc) predictions. Hybrid Bagging–Random Forest outperformed the other models (R2 = 0.941, RMSE = 2.71, MAE ≅ 1.87, NSE = 0.941 and PBIAS = 0.500).In addition, the basic models of the two hybrid ones in [**4**] are Extreme Gradient Boosting and RF, which, respectively, introduce an advanced data denoising technique–complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN). The results show that the prediction stability of CEEMDAN–Random Forest and CEEMDAN–Extreme Gradient Boosting is higher than other benchmark models for short-term water quality prediction. Similarly, in [3], water quality was evaluated via the detection of anomalies occurring in time series data. For this purpose, the performance of several models, such as LR, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), ANN, Deep Neural Network (DNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) is assessed using the F1 score metric. The best F1 score is achieved using the SVM model. Finally, the main purpose of the current study is to present a general methodology for water quality prediction by leveraging supervised learning models. The adopted methodology is irrespective of what features are used to capture the water status. The class variable is the water quality index with two possible states "safe" and "non-safe". We do not emphasize dataset engineering but the investigation of several classification schemes using single classifiers (such as SVM, NB, RF, etc) and Ensemble Learning (Voting, Stacking and Bagging). We consider an adequate set of labelled data with which a list of models is trained and tested (after the application of class balancing) to identify the one with the highest performance metrics. This research aims to develop predictive models using machine learning algorithms to optimize water filtration strategies and determine filtration methods based on diverse water quality parameters.

3. Methodology

The study of this paper is focused on deriving a classification based on gauging the purity of water and suggesting the corresponding filtration methods. We divide the water contamination in to selective filtration methods based on their contamination and impurity measures. The normalized data is the mean of each feature providing the centroid values determining the purification strategies.

Disadvantages of existing system:

- Water prediction is done only to determine whether the water is pure or not without recommending the filtration strategy
- Prediction is based on un-supervised data which has substantial missing values to access proportionate data.
- Accuracy is based on feature selection which effects the correlation between features
- The results are prone to underfitting where the machine is trained with bias of the parameters.

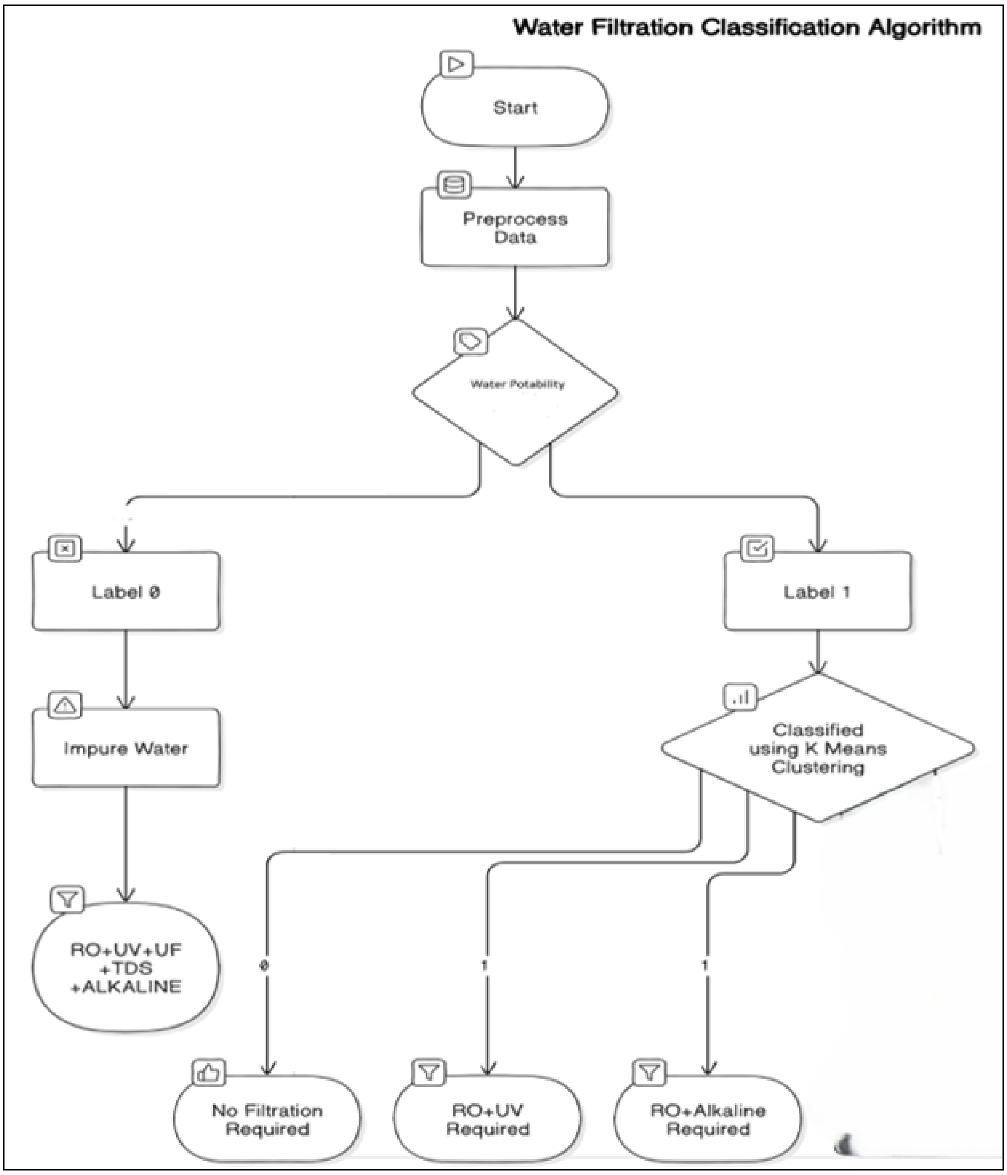
Proposed System

The proposed work is a systematic processing and training of dataset with all standard parameters which are responsible for water contamination. All such parameters are identified and corresponding filtration strategies are recommended based on the centroid and mean values of all the derived classifications. Filtration strategies include Reverse osmosis (RO), ultra-violet (UV), ultra-filtration (UF), alkaline ALK and total dissolved solids (TDS). These filtration methods are assigned after classification of dataset and standard mean values are taken to determine the filtration method.

Advantages Proposed System Advantages

- Water prediction is done to determine the filtration strategy using indexing techniques
- Prediction is based on supervised data which can evaluate and access proportionate data.
- Accuracy is based on feature extraction and information gain ratio (IGR) which enhances the correlation between features
- The results are not prone to underfitting and the machine is trained without bias of the parameters.

System Architecture



4. Implementation

K-Means Clustering classifier

The data set is labelled however to get the desired results we are using supervised learning algorithm K-means clustering algorithm on selected label 1 which is potable water while as 0 is not-potable and no clustering is done on this label. K- means is a popular unsupervised machine learning algorithm used for clustering. The goal of K-means is to partition a set of data points into KKK distinct, non-overlapping clusters.

Evaluating the data for WQI: The WQI is evaluated based on the classification done by k-means algorithm which divided the data-set in three clusters when applied on label 1 only. After getting the clusters are examined for water quality parameters which are turbidity, solids, hardness, sulphates, organic carbon, trihydrides etc. The values are compared with labels of other values obtained after classification. after getting the label data we further group the data by their centroid values by applying means all three labels. The resultant data is than processed for feature selection which we have taken as TDS and pH ranges. Comparative mean is obtained by using standard mean formula.

Centroid Mean = ∑(fi.xi)/∑fi $(x + a)^n = \sum_{k=0}^{n} \binom{n}{k} fi^k xi^{n-k}$

where x = $f1x1 + f2x2 + \frac{fix1}{f1+f2+fn} + \cdots fn$

Therefore, x is the number of observations in potable water

5. Experimental Results

The results are derived from evaluating the water quality index of each label. feature selection in done to assign the corresponding filtration method. The results are evaluated using metrics such as accuracy, precision, recall, and F1 score. The following models are applied to water quality data: logistic regression, linear discriminant analysis, support vector machines (SVM), The performance evaluation is conducted using F-score metric. A simulation study is conducted to check the performance of each algorithm using F-score.

WQI INDEX	Feature Selection	Filtration Method
0 Not safe	All features	RO+UV+UF+Alkaline
1 (safe) Label 0	TDS and pH	No filtration required
1 (Safe) Label 1	TDS and pH	RO+UV+ Alkaline
1(safe) label 2	TDS pH	RO+UV only

6. Conclusion

Our research highlighted the importance of specific water quality indicators in determining the appropriate filtration technology. Parameters such as total dissolved solids (TDS) and turbidity were critical for recommending RO systems, while microbial content was essential for UV systems. UF systems were often selected based on the presence of suspended solids and particular pathogens. These insights are invaluable for tailoring water treatment processes to specific contamination profiles, ensuring more effective and efficient filtration. This study successfully demonstrates the application of machine learning algorithms in optimizing water filtration methods. The findings suggest that decision trees and random forests provide high accuracy in predicting the most suitable filtration strategies based on water quality parameters. Future work will expand the dataset and explore deep learning techniques to further enhance predictive accuracy and reliability. This research paves the way for smarter, more adaptive water treatment solutions, reflecting a crucial step forward in the intersection of environmental science and artificial intelligence.

7. Future Enhancement

- Feature addition can be done to WQI to enhance the filtration strategies
- Predictions have high accuracies can be used for domestic as well as commercial use.
- Further scope of improvement can be introduction of new filtration method if available
- Integration with ioT devices for real-time monitoring.

References

[1] Kang G, Gao JZ, Xie G. Data-driven water quality analysis and prediction: A survey.IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService). IEEE; 2017.

[2] Ortiz-Lopez C, Bouchard C, Rodriguez M. Machine learning models with potential application to predict source water quality for treatment purposes: a critical review. Environ Technol Rev. 2022;11(1):118–47. Available from: http://dx.doi.org/10.1080/21622515.2022.2118084

[3] Akula R, Aravinda K, Nagpal A, Kalra R, Maan P, Kumar A, et al. Machine Learning and AI-Driven Water Quality Monitoring and Treatment. International Journal of Environmental Science and Technology. 2024;15(2):123–45.

[4] Zhou J, Chu F, Li X, Ma H, Xiao F, Sun L. Water quality prediction approach based on t-SNE and SA-BiLSTM. In: 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE; 2020.

[5] Lu M, Lv F. Water quality prediction model based on GRA and CNN-GRU. In: 2022 2nd International Conference on Computational Modeling, Simulation and Data Analysis (CMSDA). IEEE; 2022.

[6] Liu P, Wang J, Sangaiah AK, Xie Y, Yin X. Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. Sustainability. 2019;11(7):2058. Available from: http://dx.doi.org/10.3390/su11072058.

[7] Gong Y, Zhang P. Research and implementation of water quality grade prediction based on neural network. In: 2021 International Conference on Networking Systems of AI (INSAI). IEEE; 2021.

[8] Muhammad SY, Makhtar M, Rozaimee A, Aziz AA, Jamal AA. Classification model for water quality using machine learning techniques. Int J Softw Eng Appl. 2015;9(6):45–52. Available from: http://dx.doi.org/10.14257/ijseia.2015.9.6.05.

[9] International Conference on Computer Communication and Informatics (ICCCI). 2021 International Conference on Computer Communication and Informatics (ICCCI). IEEE;

[10] Sakizadeh, M. Artificial intelligence for the prediction of water quality index in groundwater systems. Model. Earth Syst. Environ. 2016, 2, 8

[11] Khashab F, Moubarak J, Feghali A, Bassil C. DDoS attack detection and mitigation in SDN using machine learning. In: 2021 IEEE 7th International Conference on Network Softwarization (NetSoft). IEEE; 2021.

[12] Zazouli MA, Kalankesh LR. Removal of precursors and disinfection by-products (DBPs) by membrane filtration from water; a review. J Environ Health Sci Eng. 2017;15(1). Available from: http://dx.doi.org/10.1186/s40201-017-0285-z

[13] Fatemah S. Natural filtration unit for removal of heavy metals from water. In: IOP Conference Series: Materials Science and Engineering.

[14] Singh J, Saharan V, Kumar S, Gulati P, Kapoor RK. Laccase grafted membranes for advanced water filtration systems: a green approach to water purification technology. Crit Rev Biotechnol [Internet]. 2018;38(6):883–901. Available from: http://dx.doi.org/10.1080/07388551.2017.1417234

[15] Water Quality. Available online https:/www.kaggle.com/datasets/mssmartypants/water-quality (accessed on 9 December 2022).

[16] Dritsas, E.; Trigka, M. Efficient Data-Driven Machine Learning Models for Water Quality Prediction. *Computation* **2023**, *11*, 16. https://doi.org/10.3390/computation11020016

Paper ID: SR24826150426 DOI: https://dx.doi.org/10.21275/SR24826150426