Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 4, 2025

AN INTEGRATED MODEL OF MACHINE LEARNING AND FUZZY LOGIC FOR QUALITY ASSESSMENT OF DRINKING WATER

Muhammad Imran[*1], Sajid Ali[2], Suhaib Ahmed[3], Dr. Abdur Rashid Khan[4]

[*1,2,3]*MS Scholar Computer Science, Department of Physical and Numerical Sciences, Qurtuba University of Science and Information Technology DI Khan, Pakistan*
[4]*Professor, Department of Physical and Numerical Sciences, Qurtuba University of Science & Information Technology, DI Khan, Pakistan*

[*1]imran.bhachar@gmail.com, [2]sajidalibaloch790@gmail.com, [3]Suhaibahmad59@gmail.com, [4]rashidkh08@yahoo.com

DOI: https://doi.org/

Keywords
Supervised machine learning, classification,fuzzy system,regression , water quality monitoring

Article History
Received on 28 February 2025
Accepted on 28 March 2025
Published on 05 March 2025

Corresponding Author: *

Abstract
Ensuring access to safe drinking water is essential for environmental sustainability and public health. However, existing water quality assessment methods often encounter challenges in accurately predicting water quality parameters due to the inherent uncertainties and complexities associated with water quality data. This research addresses this gap by proposing a novel hybrid machine learning approach to enhance water quality prediction. While traditional machine learning models exhibit strong predictive capabilities, they often struggle to effectively manage the imprecise and ambiguous nature of water quality data. To address this limitation, this study investigates the integration of Random Forest with Fuzzy Logic to improve predictive performance. Specifically, Random Forest enhances the model's classification accuracy, while Fuzzy Logic enables the nuanced interpretation of qualitative parameters. The proposed hybrid model was trained and evaluated on a comprehensive water quality dataset. The experimental results indicate that the integrated Random Forest-Fuzzy Logic model achieves a high level of predictive performance, with an accuracy of 99.92%, precision of 99.47%, recall of 100%, and an F1-score of 99.73%. These findings highlight the effectiveness of the proposed approach in improving water quality monitoring and management. The integration of machine learning and fuzzy logic offers a robust framework for addressing uncertainties in water quality assessment, with significant implications for water resource management, public health protection, and evidence-based policy development.

INTRODUCTION

The drinking water quality in every Pakistani province, including Islamabad and Gilgit-Baltistan, was evaluated in a 2021 study by the Pakistan Council for Research in Water Resources (PCRWIR). The startling discovery showed that 61% of Pakistan's sources of drinking water are unfit for human use [1].Water is an indispensable element for life on Earth, sustaining microorganisms, plants, animals, and humans alike. Safe drinking water is a fundamental human right, and its provision is a crucial responsibility of governments. However, with a burgeoning population, rampant industrialization, and the looming threat of global warming, water scarcity is emerging as a formidable challenge for

humanity[2].Only a small portion of the water on Earth is easily used by humans, despite the widespread belief that 71% of the planet is covered by water. Oceans contain the great bulk of Earth's water, which is unfit for human consumption. Pakistan is one of several nations that are experiencing acute freshwater shortages [3]. Furthermore, these techniques frequently only offer a moment in time view of the water quality, missing the predictive power required for proactive water management. Intelligent systems that can efficiently evaluate and understand the data are also required due to the complexity of water quality data. By putting forth combined machine learning and fuzzy logic approach, this study tackles these issues. This method uses fuzzy logic to manage uncertainties and produce a more nuanced evaluation, while leveraging machine learning's predictive capacity to forecast water quality trends and identify possible contamination hazards. In this study these methods were combined in order to provide a quick, affordable, and trustworthy method for evaluating the quality of water that could enhance public health and water management. Therefore, evaluating drinking water quality is paramount before its consumption. Poor water quality is a significant risk factor for numerous illnesses, accounting for up to 80% of global diseases. The lack of access to safe drinking water results in thousands of preventable deaths and a multitude of waterborne diseases, including cholera, malaria, polio, and typhoid [4]

Literature Review

Despite covering a large portion of the Earth's surface, we can clearly see that freshwater is becoming an increasingly strained resource. Safe drinking water has been established as a basic human right under the law and therefore it is up to governments to ensure its supply [5].

Pakistan is among the majority of nations that are experiencing a freshwater shortage. They were very driven to use alternative resources by this concerning problem. For instance, the Gulf nations use a laborious desalination process to obtain freshwater from the sea. However, this procedure is becoming increasingly difficult due to increased coastal urbanization and the ensuing water contamination. Rainwater is being treated in other nations to produce freshwater. But recently, rainfall has been impacted by climate change, which is jeopardizing this possibility. Sadly, water-related problems still exist in nations with greater access to freshwater. Concerns about water contamination have been raised for years [6].

The seas contain around 97% of the world's water, which is too salty for plants, humans, or agriculture to utilize. The remaining 3% of the earth's water is found as freshwater, of which 30% is groundwater and 69% is trapped in polar icecaps and glaciers. As a result, only 1% of freshwater is accessible to humans. Freshwater is extremely scarce these days, and there are numerous other issues putting it under extreme strain, like increased urban consumption, extensive industrial use (primarily for agricultural purposes), and changing climate due to the global warming phenomenon that affects water quality [7].

Data Source and Context

The data set used for this research was extracted from data from National Water Quality Monitoring Program (NWQMP) by PCRWR. The NWQMP was first launched in 2001 and has been conducted as a national program annually since then and the most recent being conducted in year 2020 which included 29 cities of Pakistan. The program must be valuable as it provided detailed information about water quality, which would help to determine the main issues in bringing safe drinking water to the people. The monitoring phase of the survey in the year 2020 involved 435 water sources, which show that only 39% of the water sources provide water that has met the National Standards for Drinking Water Quality, thus the need for increased effort to ensure that the quality of water produced by the sources is improved.

Data Selection and Refinement

This study uses a more diverse dataset that draws from different sources over longer timescales to increase generalizability and reduced biasness. There were 19 water quality parameters which also included a binary classification of 'safe' or 'unsafe' which was ideal for creating a predictive model of

Water Quality Assessment. Data sources include PCRWR (Pakistan Council of Research in Water Resources) annual reports from the years 2010, 2015, and 2021 as well as data obtained from EPA (Environmental Protection Agency) for the year 2024. A broader diversity of water quality scenarios and geographic regions are included in this multifaceted dataset, improving the comprehensiveness and ultimately generalizability of the analysis. In addition, the sample size is increased with data augmentation and small diversities are added to already existing data so that the model can learn better as well as generalize on new and unseen data. Having said this, the diversified data collection approach and data augmentation technique together will be justified to provide a suitable model for water quality assessment with high reliability and prediction capability.

Selection of Water Quality Parameters

Water quality assessment involves measuring various physical, chemical, and biological parameters to assess the condition of water. The specific factors or parameters to monitor can vary depending on the context and objectives of the assessment.

Logistic Regression

Logistic Regression is a type of classifier which predicts the likelihood of yes/no based on certain inputs. To this end, it employs the logistic function (sigmoid) to map the coefficients summation of inputs into probability measures ranging from 0 and 1. The coefficients (β) in the linear combination are estimated by using Maximum likelihood estimation method, which tends to select the values that would make the occurrence of the provided data more probable. The last identification is made by comparing the probability values with the given threshold which is normally 0.5. The core equation representing the logistic regression model is:

$$P(Y=1|X) = 1/(1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn)}$$

.....Eq. 1

Where:

i. P(Y=1|X) represents the chances of obtaining the result being 1, for given input features X, such as 'unsafe'. arning is the base of natural logarithms.

ii. β0 is the intercept

iii. Coefficients calculated for each of these input features are β1, β2,... βn that has to be determined. Xn

Decision Tree

Decision Tree is a classification algorithm, where data is divided into sets based on the features the algorithm has, in the hope of achieving the 'purest' nodes in which most cases belong to a single class. Such a partitioning is done based on splitting measures which include among others Gini Impurity a measure of the probability of wrong classification within the node, Information gain which is the measure of entropy reduction within each split. The process of constructing trees goes on until this criterion is met and we get a hierarchical model where each node at the lowest level predicts a class. The core formulas for Gini Impurity and Information Gain are:

i. Gini Impurity

$$Gini(node) = 1 - \Sigma [p(i|node)]^2$$ .....

Eq. 2

ii. Information Gain

$$Information\ Gain(parent, feature) = Entropy(parent) - sum\ of\ [Weighted\ Average * Entropy(child)]$$

.....Eq. 3

iii. Entropy

$$Entropy(node) = -\sum [p(i|node) * \log p(i|node)]$$

.....Eq. 4

Random Forest

Random Forest is a supervised learning technique that generates a number of decision trees to improve the accuracy and avoid problems related with high variance. It builds thousands of decision trees each restored on different bootstrapped sample of data and features, this ensures diversification of models. Moreover, it is used the technique Random Subspace where in every node inside a tree only a random subset of features is considered to split, enlarging this diversity. The last forecasting of a new instance is the sum of the conclusions of all discrete trees; this is typically accomplished using the voting approach for the classification problem or averaging for the regression problem.

$$Prediction = Aggregate(Prediction\_1, Prediction\_2, \ldots, Prediction\_n)$$ .....Eq. 5

where **Prediction_i** represents the prediction of the i-th tree in the ensemble, and **Aggregate** denotes the aggregation function (e.g., majority vote or average).

Support Vector Machines

SVMs are strong classification techniques used to identify the right hyperplane that efficiently classifies classes in a high dimensional space. For linear SVM the goal is to maximize the distance between the hyperplane and the nearest data point of every class (support vectors). This optimization problem can be mathematically expressed as:

$$Maximize\ Margin = \frac{2}{(|w|)}$$ .....Eq. 6

Subject to:

$$yi(w.xi + b) \geq 1$$

for all i where 'w' is the weight vector of the hyperplane direction, 'b' is the bias term constant, 'xi' is the input feature vector and 'yi' is the output class label.

For non linearly separable data, SVMs use the kernel trick, which maps the data implicitly in to a higher dimension where a line can be fit. Some of the kernel functions used are the Radial Basis Function (RBF) and Polynomial kernel functions. The core principle of SVMs remains the same: in order to generalise LSE to transformed spaces and identify the hyperplane giving the maximum margin of difference between the classes. After finding the mathematical solution of the hyperplane, new instances are assigned to which side of the hyperplane they belong in.

K-Nearest Neighbors (KNN)

K Nearest Neighbor (KNN) is simple and widely used effective classification technique which classifies a new data point based on the major class of its 'k' closest neighbors in feature space. It works on the concept of similarity; it uses distance functions like Euclidean distance or Manhattan distance of instances. It computes the distances among the new instance and all the training instances, finds 'k' nearest neighbors and assigns a dominant class among them. The mathematical representation of the commonly used Euclidean distance is:

$$Distance(x,y) = \sqrt{(\Sigma(xi - yi)^2)}$$ .....Eq. 7

where 'x' and 'y' are two instances, and 'xi' and 'yi' are the corresponding features that those two instances possess. For classification problems, especially when decision boundaries may not be clearly observable, KNN is a basic and straightforward approach to prior problem and dependent on nearby data.

Naïve Bayes

Naive Bayes is a probabilistic classifier based on the hypothesis of the Bayes formula, adapted for case where each feature is independent from all others given the class variables. It calculates the probability of a class given a set of observed features using the following formula:

$$P(Class|Features) = \frac{[P(Features|Class) * P(Class)]}{P(Features)}$$ .....Eq. 8

The "naive" assumption states:

$$P(Feature1, Feature2, \ldots | Class) = P(Feature1|Class) * P(Feature2|Class) * \ldots$$ Eq. 9

This assumption although violated, sometimes in most part, means that we can still easily compute the probabilities even with the presence of many features. The algorithm makes an assumption of prior probabilities of each class and conditional probabilities of values of features in each class computed from the training sample. For a new instance it involves bayes' theorem of probability with the naive assumption and assigns to the class with the highest poster probability. However, sometimes Naive Bayes does not play the role of an analyst but actually outperforms itself, especially in text classification and other tasks where the independence of features is quite accurate.

Gradient Boosting

Gradient Boosting is a machine learning technique that combines several weak models most of the time being a decision tree, in a step wise manner, to form a strong model. It operates on the principle of additive modeling, where the final prediction is a weighted sum of the predictions from individual models:

$$F(x) = \Sigma fm(x)$$ Eq.10

This learning algorithm just continually introduces new models into the blend; each model in the new mixture is intended to provide a lesser, or less significant, quantity of error or residual for models in the prior mixture. This process is an optimization step known as gradient descent, where with every new tree a function is fitted to the negative gradient of the loss function with respect to the current mode's predictions. There are different categories of loss functions that depend on the type of a problem, the most common ones are mean squared for regression and log loss for classification. Since the given loss function is optimized stepwise in Gradient Boosting, the corresponding collaborative model's accuracy is gradually raised, which is why this technology can be applied to most machine learning issues.

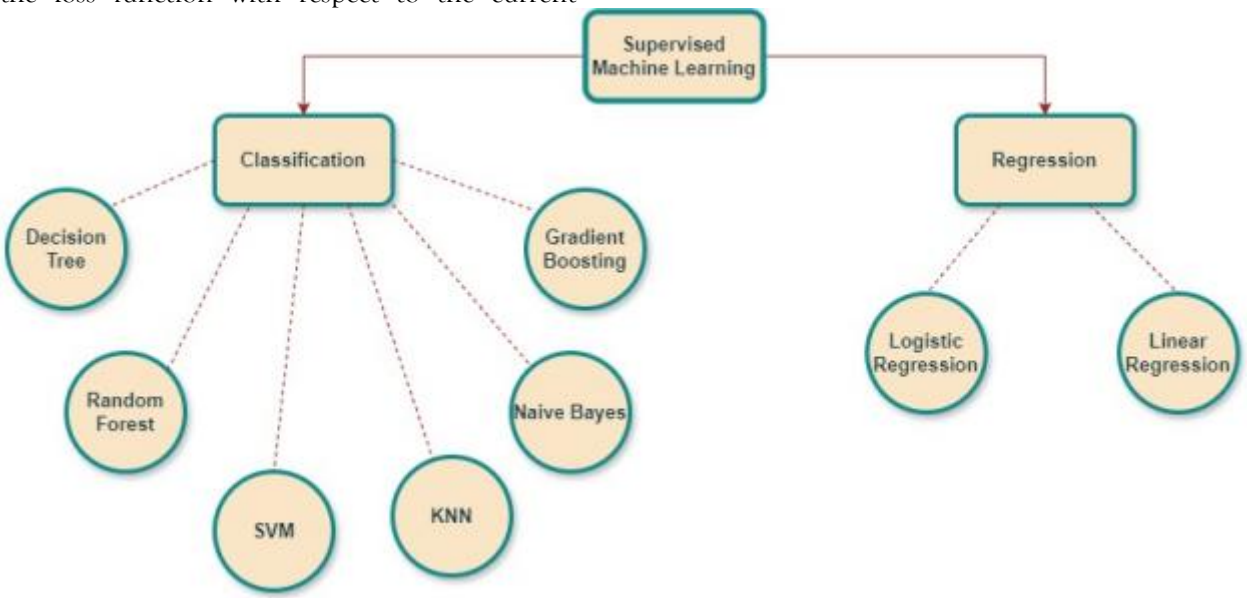

Figure 3. 2: Supervised Machine Learning Algorithms

Fuzzy Logic

Fuzzy logic is a mathematical framework that provides a way to represent and reason with uncertainty. Unlike traditional logic systems that rely on binary values of true or false, FL allows for degrees of membership, enabling a more nuanced representation of the world. This feature makes FL particularly well-suited for handling uncertainties inherent in water quality data

Fuzzy sets

Fuzzy sets are crisp sets where the characteristic function are transformed to the membership function A: X →[0, 1] fuzzy logic operators compute:

- MIN(x; y) for conjunction of two fuzzy logic values x
  - and y
- MAX(x; y) for disjunction of two fuzzy logic values x
  - and y
- 1-x for negation of a fuzzy logic value x.

Problem Definition

The objective when training the models is to be able to classify water samples in real time for analysis. Thus, in this research, seven specific machine learning algorithms derived from research have been trained with high accuracy and less false alarm rate.

Acquisition of data

The dataset employed for training the selected classifiers was derived from the publicly accessible 2021 water quality monitoring report issued by the Pakistan Council of Research in Water Resources (PCRWR), 2010 and 2015 report of PCRWR and EPA report 2024.

Data Preprocessing and Preparation

The raw data obtained from the report was preprocessed in a very stringent way to prepare it for modeling using machine learning as well as fuzzy logic. Data cleaning was done to check for and remove all the unnecessary or irrelevant information: for the cases where necessary, a simple imputation technique was used in order to make all the fields complete. To reduce the impact of outliers on the subsequent analyses, they were properly dealt with by applying the IQR(Inter Quartile Range) method. The enhanced data set was then exported into a CSV file format which is suitable for the next phases of machine learning and fuzzy logic. The preprocessing stage was significant in cleaning the raw data into a credible dataset in order to act as basis for the subsequent modeling stages.

i. Data Cleaning

Here, the PCRWR water quality data cleaning process is completed to remove null and infinite data values from the available data set.Data types were also converted into the correct format, whether numerical or categorical, so that memory space is not wasted and errors are avoided in analysis.

ii. Feature Selection

From the sci-kit-learn library, use the SelectKBest method to select the 30 most important features with a high significance level.Employed a correlation heatmap to eliminate multi-colinear features while increasing the generalized performance of the model.

iii. Class Balancing

Solved the class imbalance problem in the PCRWR data set using the SMOTE-Tomek Links method.The minority class was oversampled by synthesizing samples through the SMOTE technique. Reduced data set by taking samples of Tomek links; these are the instances in the majority class nearest to each instance in the minority class. This was done to help manage class distribution and also minimize over fitting.

Implementation Details

The project work is performed on personal systems which has the following specifications: An Intel Core i5-4200U, 1.6GHz clock-speed, 8 GB RAM, and Windows operating system. The models will be implemented in Anaconda 3 Jupyter Notebook development environment. Python V.3 was used as the main project language. Numpy, Pandas, Scipy, Matplotlib, and Scikit-learn libraries were used for the implementation and evaluation of this project.

Evaluation

The Evaluation was done on a matrix that is on confusion matrix with different attribute that contains information of the predicted and actual classes. A confusion matrix is 2-Dimensional matrix consist of the following attributes:

- **True Positive (TP):** As has been mentioned the data was correctly categorized as an Attack by the classifier.
- **False Negative (FN):** The data was wrongly predicted as Normal.
- **False Positive (FP):** This data was wrongly classified as an Attack.
- **True Negative (TN):** The Normal instances include: the instances correctly classified as Normal instances.

Diagonal elements give correct predictions while non-diagonal elements give wrong predictions of the matrix.
Based on these attributes following evaluation matrices are calculated: Various measures such as F1 score, precision, accuracy, recall and so on were used to assess the performance of the model. These metrics were

computed by various factors of the confusion matrix including the True Positive (TP), False Positive (FP), True Negative (TN) as well as the False Negative (FN). These metrics are defined as follows:

Accuracy

This is simply equal to the proportion of predictions that the model classified correctly as shown in eq.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....Eq. 11}$$

Precision

It is defined as the ratio of correctly predicted Attacks to all the samples predicted as Attacks as shown in eq .

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{.....Eq. 12}$$

Recall

It is a ratio of all samples correctly classified as Attacks to all the samples that are Attacks as shown in eq .

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{.....Eq. 13}$$

A confident result requires high recall and precision.

F1-score

It is a harmonic mean of the model's precision and recall as shown in eq.

$$F1 = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad \text{.....Eq. 14}$$

Web Application Development

The web application is developed using Streamlit Sharing Website development framework. Moreover, in Jupyter notebook best performing algorithm is saved using streamlit and Joblib library. Real-time water quality parameters are given as input and the result is shown to the user in the browser.

DataCleaning and Preprocessing

Data cleaning has immense significance during the data prepossessing phase and ensures that the dataset does not have any contradictions, missing data, or noise. In regard to the current dataset revolved around the water quality assessment, the following intervention were made:

i. Addressing the Missing Values in Data: The dataset was analyzed in a way to find out missing values for all the features in the model which would help reduce the integrity of the model. For example, using suitable approaches such as median or mean imputation for numerical variables, missing data deficiency is tackled. The summary shows the missing values regarding some major parameters such as pH, EC, TotalColiforms, etc.

a. Evaluation of Missing Data: The evaluation of each parameter was scarried out to verify that none of the parameters was either 'none' or 'empty'.

b. In such cases, however, some missing values were revealed, and measures were taken to ensure that the model prediction was not distorted.

ii. Detection of Outliers: These instances, which must also be addressed during data cleansing and modeling, are very critical features such as pH, TDS, TotalColiforms, which can achieve unordinary values because of water conditions. The effectiveness of the histogram/boxplots in assessing such values was evaluated using outlier detection models/Stata statistics and, where applicable, removed inapplicable values.

Normalization and Scaling: In bringing all parameters within the same range, indices with different scales, like EC and HCO3, were targeted to ensure no one single feature was excessive in the model when computation was taking place.

Table 4. 1: Statistical Analysis of Dataset

Parameters	Mean	Median	Mode	Standard Deviation
EC	1254.22	762.68	197.8	2248.43
pH	7.58199	7.6158	7.8	0.298312
Turbidity	8.71476	1.6743	0.3	56.4348
HCO3	260.162	240	250	140.038
CO3	0.0890564	0	0	0
Ca	65.9458	48.75	40	80.6199
Mg	40.4239	24	22	67.9408
Hard	328.162	231.35	170	459.226
Cl	172.188	53	14	672.949
Na	137.106	63.4	70	324.975
K	10.3122	3.6	1	19.4607
SO4	143.933	71.73	30	240.54
NO3	2.51352	1.3297	0	3.20103
TDS	764.785	448.47	474	1453.46
Fe	0.235689	0.1	0.02	0.505315
F	0.454237	0.3577	0.03	0.475704
As	8.42909	1.9932	0.12	15.015
TotalColiforms	12.3221	3.49	0	19.132
Ecoli	1.6668	0	0	4.66012

All these parameters contribute to water quality laws of some degree. Nevertheless, none of them is necessarily relevant for the prediction on water safety and feature selection was applied.

To know the most important features, we used feature importance techniques with Random Forest Classifier. By using the Gini importance value, Random Forest can tell us which field is more informative than othes (helps reduce uncertainty / impurity in the process of making a model) We used it to rank the features.

Furthermore, we performed correlation analysis to find the highly correlated features as multi-collinearity impacts model performance negatively. High correlated features were further evaluated to determine if they could be redundant or eliminated.

We calculated feature importance scores using the Random Forest algorithm to determine features that are important in predicting whether water is unsafe/safe. Feature importance from Random Forest Classifier is shown below.

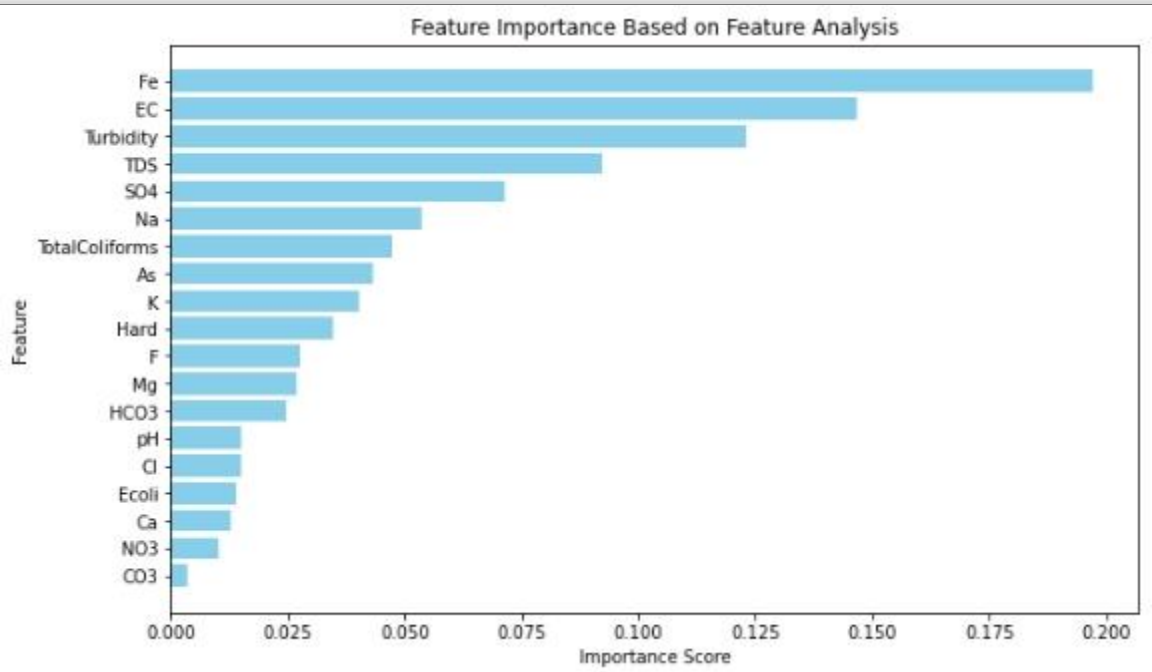

Figure 4.20: Importance of Features in quality analysis

Class Balancing

In processing imbalanced datasets, class balancing plays an indispensable role to enhance machine learning models' performance over minority classes. Results In our dataset, we had an imbalanced target variable for the water safety classification (Water_Safety_Class), with much more samples labeled as "safe" than "unsafe". This might unbalance the model and cause it to favor the majority class which would then be sensitive for false positive, voiding over detection of unsafe water samples leading to lower sensitivity in detecting unsafe conditions.

i. Imbalanced Class Problem

In a binary classification task with class imbalance, such as in this case where one of the classes is much more frequent than the other, random forest models will usually simply predict every sample to be that majority. The authors of the current study found that this was mainly due to the fact there were many more "safe" samples than "unsafe", so although initial model accuracy might be high, it often had low sensitivity for identifying where water is unsafe.

ii. Balancing Techniques

We tackled the class imbalance as follows:

- **Random Under sampling:** It reduces the majority class by choosing a subset of "safe" samples at random to equalize the classes. Since under sampling removes samples from the data, it also decreases the total size of the datasets and reduces potential bias because we get down to almost an even number of class 1 (minority) and class 0(majorities).
- **Synthetic Minority Over-sampling Technique (SMOTE) :** SMOTE, as the name suggests create a synthetic sample for minority class by interpolation between existing samples especially near each other. This balances enough data to get a fair number of "unsafe" class examples for the model, but not so much that it is almost all majority-class.

For the second aim, we applied both approaches (data set partition with stratified sampling and supervised clustering–based cross-validation) separately to assess which approach resulted in a better model performance; after seeing the models metrics for each case study in Appendix E, we selected the best performing strategy.

- **Results of Class Balancing**

SMOTE has improved detection of minority class by the model. The confusion matrix and classification report bellow show the improved sensitivity

regarding the class for "unsafe" water, e.g., more samples was identified as category to be in which were indeed labeled 'Unsafe' from both sides of NULL purified_countagrams,following table shows results before and after SMOTE result of the classification report.

Table 4. 2: Results after SMOTE analysis

Class	Precision (Before)	Recall (Before)	Precision (After)	Recall (After)
Safe	0.98	0.99	0.96	0.95
Unsafe	0.54	0.21	0.89	0.90

Machine Learning Models

A dataset consisting of various water quality indicators had to be analyzed and machine learning models were used for the prediction of water parameters. The goal was to produce well detailed and robust models that could be used for the monitoring and sound management of water resources. Several algorithms, such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), K-Nearest Neighbors(KNN) Naive Bayes and Gradient Boosting was evaluated to know which algorithm is best for this particular task.

For each algorithm a 10-fold cross-validation was used to evaluate and ensure that the model is not overfitted. 10-Fold Cross Validation — This works by splitting the dataset into 10 equal groups(Folds). This is done 10 times, for each fold as the test set once. The average performance overlooking these iterations offers a more reliable measure of the quality in which model generalizes. The table 4.3 below shows the Accuracy, Precision, Recall and F1 score for each fold with mean.

Table 4. 3: Classification models accuracy Comparison

	Logistic Regression	Decision Tree	Random Forest	SVM	KNN	Naive Bayes	Gradient Boosting
Fold							
Fold 1	0.5882	0.9346	0.9281	0.5163	0.7190	0.5229	0.8627
Fold 2	0.5882	0.9346	0.9542	0.5359	0.7124	0.5621	0.8497
Fold 3	0.6013	0.9477	0.9346	0.5294	0.7124	0.5359	0.8627
Fold 4	0.5425	0.9477	0.9739	0.5163	0.7320	0.5490	0.9085
Fold 5	0.6078	0.9346	0.9477	0.5556	0.7386	0.5556	0.8562
Fold 6	0.5948	0.9673	0.9739	0.5490	0.7059	0.5294	0.9085
Fold 7	0.5921	0.9145	0.9079	0.5066	0.6711	0.5461	0.8224
Fold 8	0.5789	0.9211	0.9605	0.5263	0.7105	0.5395	0.8684
Fold 9	0.6250	0.9013	0.9276	0.5329	0.6908	0.5526	0.8947
Fold 10	0.5724	0.9605	0.9803	0.4934	0.6974	0.5329	0.8816
Mean	0.5891	0.9364	0.9489	0.5262	0.7090	0.5426	0.8715

The table above shows the accuracy scores of a machine learning model in each fold during cross-validation. This also includes a mean accuracy over all folds. The Random Forest gets by far the best accuracy scores in all folds. The Random Forest approach has been demonstrated to perform very well, as evidenced by its consistently high accuracy percentage in our study and is therefore likely the best candidate for predicting water quality aspects on this context. Its capacity to model non-linear relationships and prevent over fitting is probably what makes it such a strong performer. Classification algorithms accuracy comparison is shown in figure 4.21.

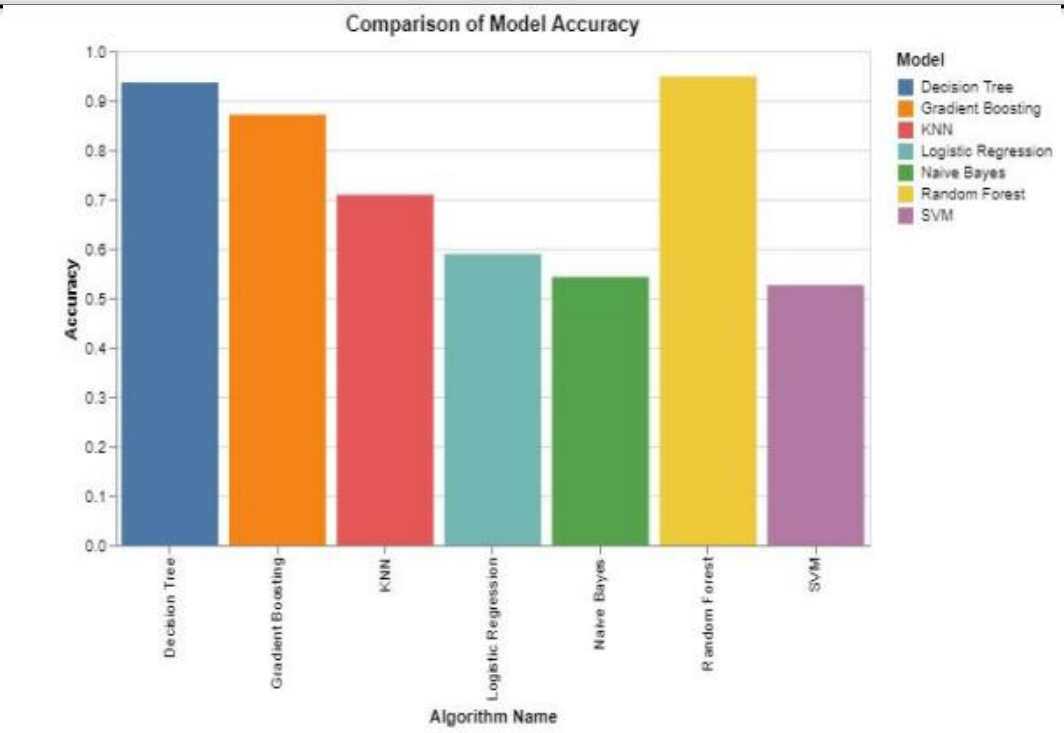

Figure 4. 21: Classification Algorithms Accuracy Chart

Table 4. 4: Classification models Precision Comparison

	Logistic Regression	Decision Tree	Random Forest	SVM	KNN	Naive Bayes	Gradient Boosting
Fold							
Fold 1	0.5892	0.9505	0.9562	0.5539	0.7307	0.5773	0.8644
Fold 2	0.5905	0.9318	0.9422	0.5604	0.7181	0.6314	0.8497
Fold 3	0.6019	0.9505	0.9242	0.5623	0.7139	0.5931	0.8644
Fold 4	0.5428	0.9543	0.9739	0.5283	0.7320	0.5911	0.9085
Fold 5	0.6088	0.9449	0.9751	0.6582	0.7491	0.6582	0.8511
Fold 6	0.5976	0.9674	0.9742	0.6305	0.7222	0.5713	0.9110
Fold 7	0.5931	0.9151	0.9277	0.5138	0.6712	0.6053	0.8224
Fold 8	0.5826	0.9313	0.9605	0.6070	0.7179	0.6180	0.8751
Fold 9	0.6261	0.8992	0.9369	0.6051	0.6911	0.6260	0.8765
Fold 10	0.5739	0.9750	0.9810	0.4815	0.7008	0.6872	0.8818
Mean	0.5907	0.9420	0.9552	0.5701	0.7147	0.6159	0.8705

The table 4.4 above represents the precision scores of each machine learning model on all 10 folds of cross-validation. Precision: Proportion of correctly predicted Positive instances out of Total Predicted as Positive. It clearly beats other models in the class validation and there is not much variance between folds, suggesting that Random Forest can classify positive instances more efficiently than others (it has a higher precision). Almost all models have been reasonably accurate, with Naïve Bayes the only model that consistently has very low accuracy. Random Forests achieved the same high precision consistently, which confirms that it is a prominent model in this context to predict water quality parameters. This is something that makes it very powerful as was shown in these analyses: their ability to handle non-linear relationships and control for over fitting.

Classification algorithm precision comparison is shown in figure 4.22

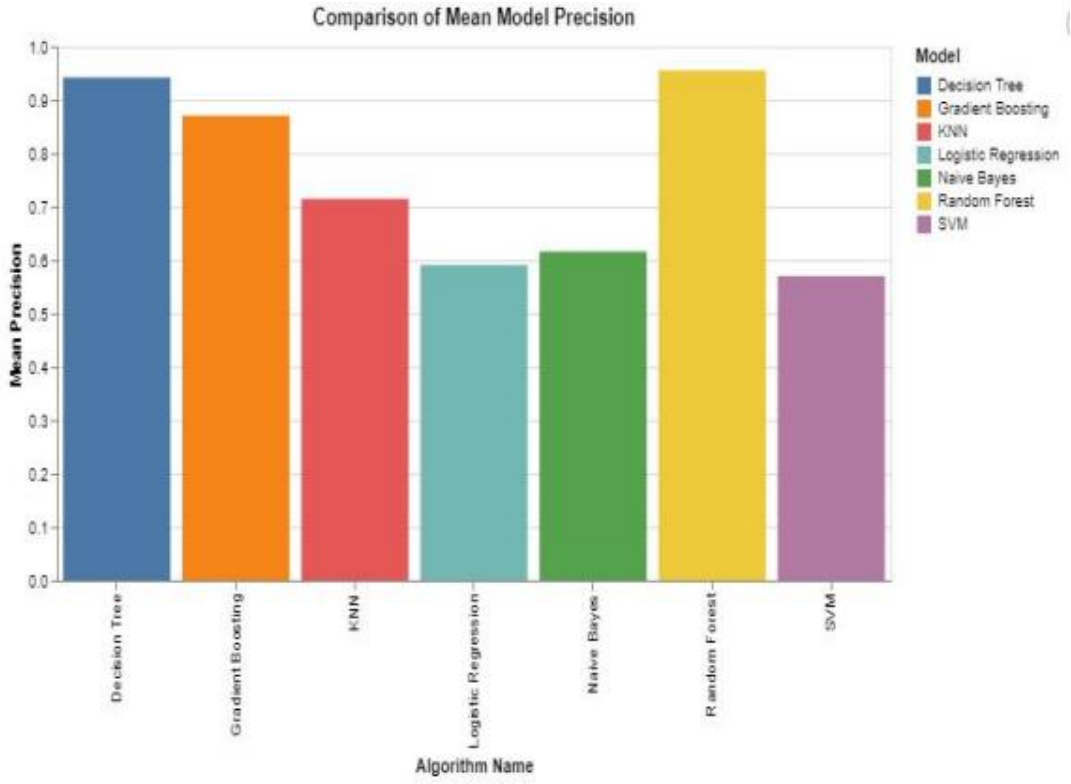

Figure 4. 22: Classification Algorithms Precision Comparison

Table 4. 5: Classification Models Recall Comparison

	Logistic Regression	Decision Tree	Random Forest	SVM	KNN	Naive Bayes	Gradient Boosting
Fold							
Fold 1	0.5882	0.9477	0.9412	0.5163	0.7190	0.5229	0.8627
Fold 2	0.5882	0.9412	0.9477	0.5359	0.7124	0.5621	0.8889
Fold 3	0.6013	0.9412	0.9412	0.5294	0.7124	0.5359	0.8627
Fold 4	0.5425	0.9281	0.9608	0.5163	0.7320	0.5490	0.8627
Fold 5	0.6078	0.9346	0.9346	0.5556	0.7386	0.5556	0.8562
Fold 6	0.5948	0.9739	0.9739	0.5490	0.7059	0.5294	0.9085
Fold 7	0.5921	0.9211	0.9211	0.5066	0.6711	0.5461	0.8224
Fold 8	0.5789	0.9342	0.9671	0.5263	0.7105	0.5395	0.8750
Fold 9	0.6250	0.8947	0.9211	0.5329	0.6908	0.5526	0.8618
Fold 10	0.5724	0.9605	0.9868	0.4934	0.6974	0.5329	0.8816
Mean	0.5891	0.9377	0.9495	0.5262	0.7090	0.5426	0.8683

The table 4.5 above lists the recall scores for every machine learning model over 10 cross-validations. Recall: It measures how many actual positive instances were correctly predicted from all the true and depicted positives. Results always show Random Forest to have the highest recall scores in each fold of cross-validation, highlighting its ability to capture a high percentage of positive instances. This is not quite net as strong as Random Forest, however again demonstrating that ensemble methods are helpful

for this task Gradient Boosting registers similar recall numbers. The regular high recall of Random Forest again reinforces it as a top performer in predicting water quality parameters here. One reason for its excellent performance is that it can handle relationships between variables in a non-linear way and also reduce overfitting. Classification algorithms recall(sensitivity) comparison is shown in figure 4.23 below.

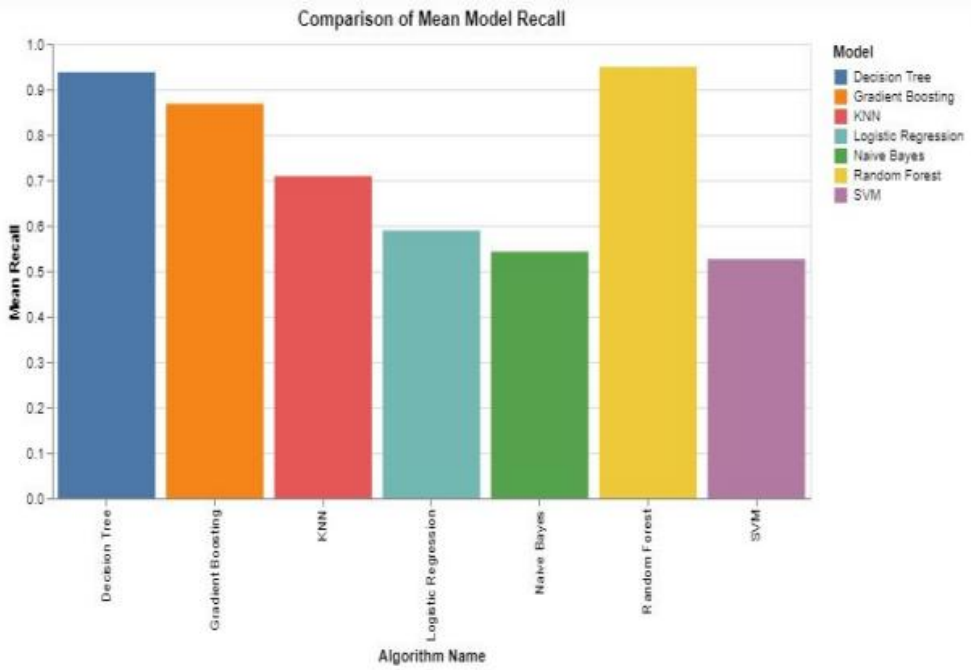

Figure 4. 23: Classification Algorithm Recall Comparison

Table 4. 6: Classification Models F1 Score Comparison

	Logistic Regression	Decision Tree	Random Forest	SVM	KNN	Naive Bayes	Gradient Boosting
Fold							
Fold 1	0.5875	0.9542	0.9476	0.4234	0.7157	0.4276	0.8626
Fold 2	0.5863	0.9345	0.9410	0.4888	0.7108	0.4986	0.8497
Fold 3	0.6010	0.9608	0.9214	0.4640	0.7120	0.4567	0.8626
Fold 4	0.5403	0.9542	0.9542	0.4459	0.7320	0.4860	0.9085
Fold 5	0.6066	0.9345	0.9608	0.4663	0.7355	0.4663	0.8560
Fold 6	0.5911	0.9608	0.9608	0.4618	0.7000	0.4419	0.9083
Fold 7	0.5910	0.9079	0.9211	0.4322	0.6710	0.4718	0.8224
Fold 8	0.5742	0.9341	0.9671	0.4163	0.7080	0.4476	0.8815
Fold 9	0.6242	0.8944	0.9208	0.4360	0.6907	0.4764	0.8605
Fold 10	0.5701	0.9671	0.9803	0.3962	0.6961	0.4117	0.8815
Mean	0.5872	0.9402	0.9475	0.4431	0.7072	0.4585	0.8694

F1-scores are shown in the provided table for every machine learning model on 10 different cross-validation folds. F1-score is the harmonic mean of precision and recall, so it can provide 2/3 quality information about model capability. Random Forest always achieves the best F1-scores for all folds, showing that it has a nice value of both precision and recall at once. For example, in some models such as Logistic Regression and Naive Bayes there is more variance of F1-score across the k-folds implying these which might indicate their performance could be

much dependent on what exact part of data they are trained on.

The consistently high F1-score of Random Forest among the best classifiers, indicates that it is a stark competitor for water quality parameters in this regard. Its superior performance can probably be attributed to its ability in dealing with non-linear relationships as well reducing over fitting. Classification algorithms F1 score comparison is shown in figure 4.24 below.

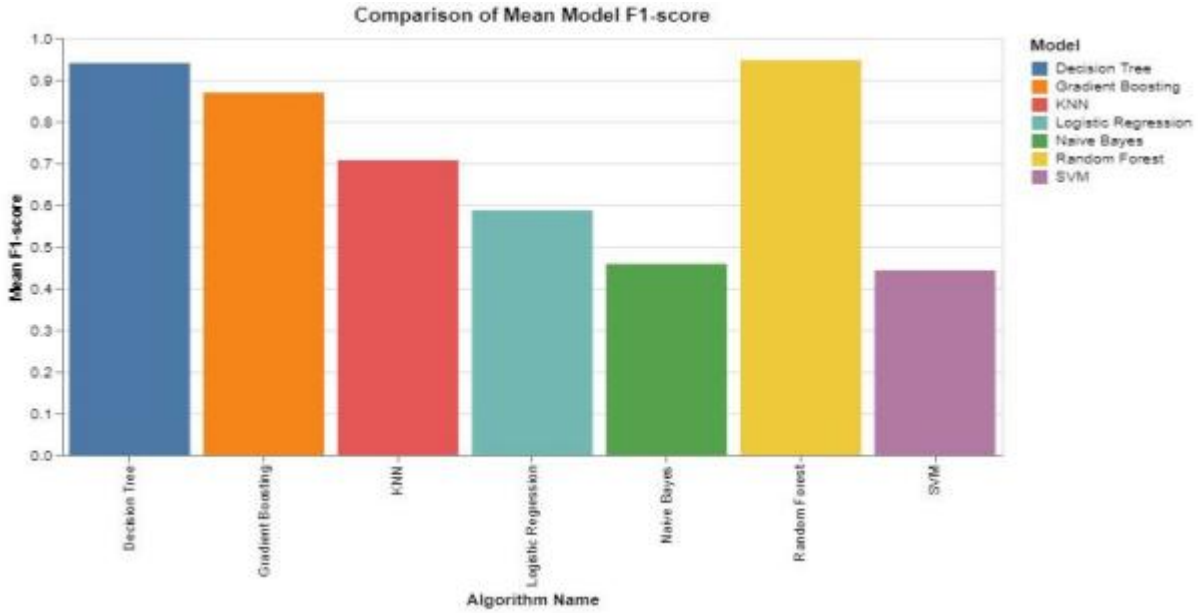

Figure 4. 24: Classification Algorithms F1 Score Comparison

Fuzzy Logic Membership functions and Development of rules

The following figures show membership function diagrams for each of the key water quality parameters used in our fuzzy logic model and outline thresholds for deeming waters safe or unsafe. The diagrams represent the membership functions that have been defined to capture different degrees of safety for parameters such as Electrical Conductivity (EC), pH, Total Coliforms, E. coli., Turbidity, Nitrate(NO_3) and Total Dissolved Solids(TDS). The model can then, according to predefined safe ranges for each parameter of the water quality indicators be able to assess more nuanced statuses about—briefly speaking—the states/crisis levels etc. by which it is bound into three intervals (intermediate / serious and extreme), as built on these membership functions that form basis for our fuzzy logic rules. What follows is a comprehensive description of the membership functions, including safe and unsafe transition points in which they are grounded within standard water quality guidelines.

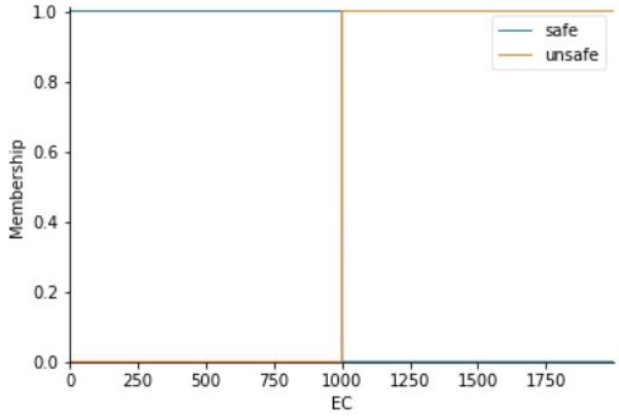

Figure 0- 1: Membership Fucntion of Electrical

Conductivity

The graph pertaining to Electrical Conductivity (EC) presents the boundary between the safe and unsafe water quality echelons based on EC levels. The plot illustrates that those EC values which are indicated to be less than 1000 are entirely safe with a membership level of one. This range undoubtedly indicates that the water assessed within this column is safe. However, at the EC level of 1000, the membership for safe status declines to 0 and from there onwards it means that the water is unsafe. In the case of values above 1000, the only function left which is safe is unsafe with a membership value of one meaning that the water is unsafe. This suggests that there is a clear delineation of what constitutes safe and unsafe levels of EC in water having a safe threshold.

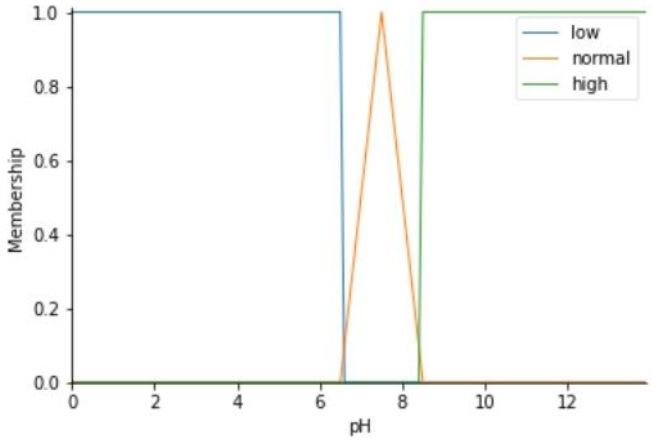

Figure 4. 25: Membership Function of pH

The blotted pH membership function depicts three levels or categories to assess water quality: low, normal and high. It has been noted that safe or normal pH levels (neutral range) are on the borderline of 6.5 and 8.5; a scoring of 1 is granted within the range which is consistent with drinking water standards. Low values of less than 6.5 suggest acidity; therefore, a low membership function of 1 is accorded as pH continues to fall below 6.5. On the contrary, pH readings greater than 8.5 suggest an alkaline condition, which means that individual is referred to as high, and once pH is greater than 8.5, that membership function of high gets a value of 1. This explains why serious attention should be given with regard to the pH level of the water consumed, as both extremes are detrimental to health.

.

Performance Evaluation

Following figures shows the confusion matrix of each classification algorithm and integrated random forest and fuzzy logic confusion matrix.

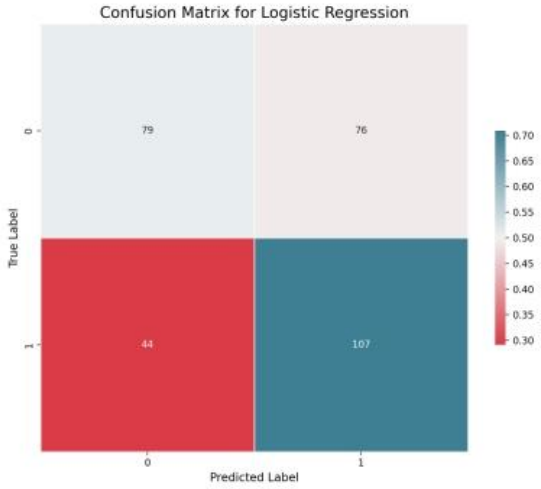

Figure 4. 42: Logistic Regression Model Prediction

The model's ability to categorize examples into two classes (0 and 1) is revealed by the Logistic Regression confusion matrix. According to the matrix, the model accurately identified 107 cases as negative (true negatives) and 79 instances as positive (true positives). Nevertheless, it misclassified 44 cases as negative when they were actually positive (false negatives) and 76 cases as positive when they were actually negative (false positives). The model classified 58% of the cases correctly, indicating a decent degree of accuracy. Only 51% of the positive predictions were successfully identified, indicating its accuracy issues. The model missed a sizable portion of real positive occurrences, as evidenced by the recall of 64%. The recall, at 64%, indicates that the model missed a significant number of actual positive cases. This suggests that the model might be better at identifying negative instances than positive ones.

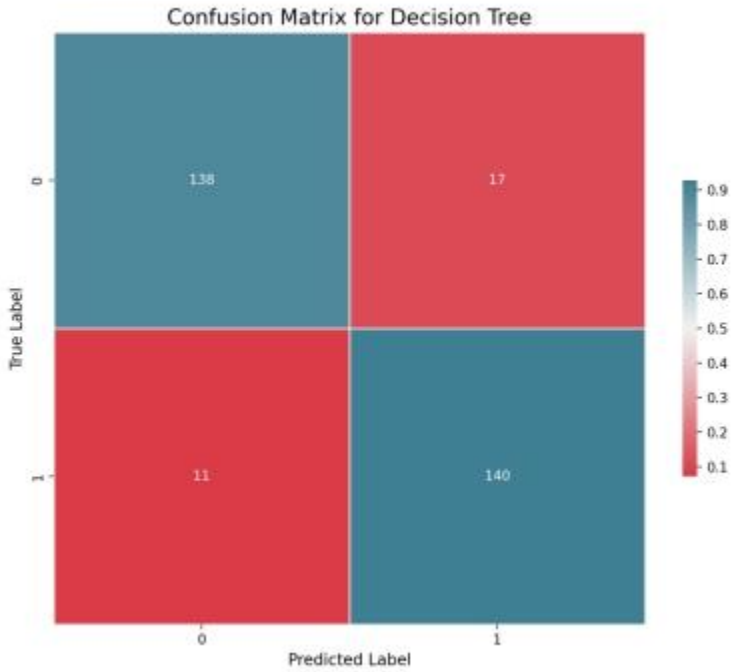

Figure 4. 43: Decision Tree Model Prediction

The Decision Tree model's confusion matrix sheds light on how well it performs when dividing cases into two groups (0 and 1). The model properly identified 140 cases as positive (true positives) and 138 instances as negative (true negatives), according to the matrix. On the other hand, it misclassified 17 cases as negative when they were actually positive (false negatives) and just 11 cases as positive when they were actually negative (false positives).

With an accuracy of almost 93%, this shows that the Decision Tree model has done quite well. The model has a recall of about 89% and an accuracy of about 93%. These high values imply that both positive and negative situations may be successfully identified by the model. The Decision Tree's capacity to identify intricate non-linear correlations in the data may be the reason for its better performance when compared to Logistic Regression.

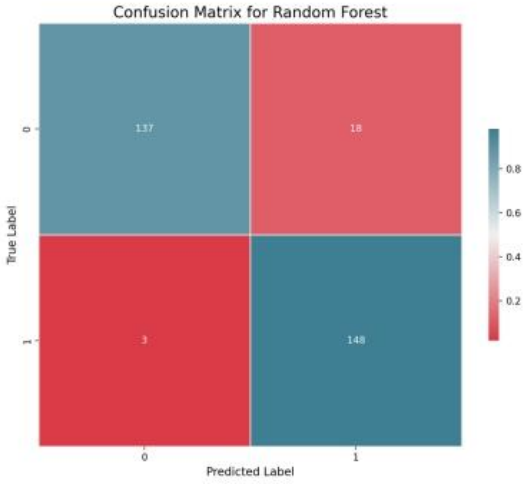

Figure 4. 44: Random Forest Model Prediction

With an accuracy of around 97%, Random Forest's confusion matrix performs quite well. The model has a recall of about 89% and an accuracy of about 98%. These high numbers demonstrate how well the model detects both positive and negative situations. The Random Forest's higher performance over the other models is probably a result of its capacity to generate several decision trees and aggregate their predictions.

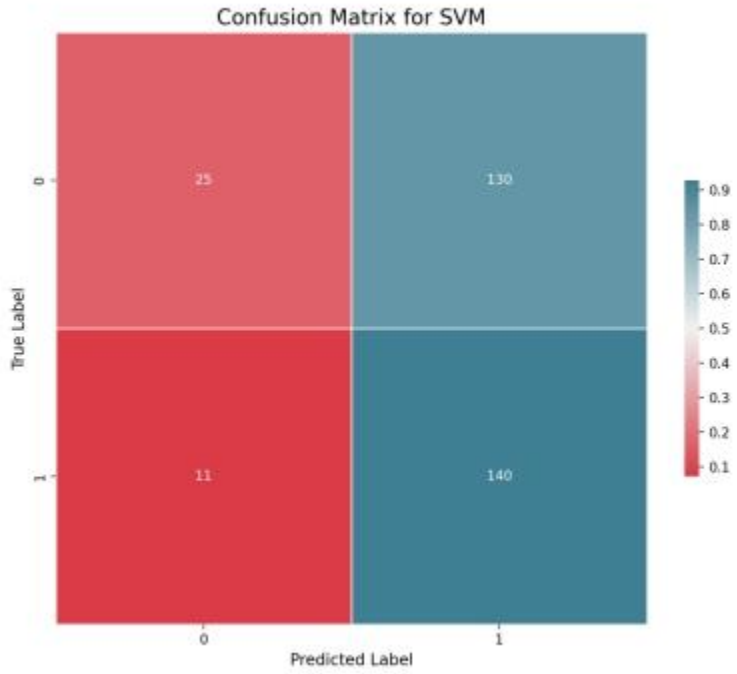

Figure 4. 45: SVM Model Prediction

Classifying examples into two classes (0 and 1) yields a reasonable performance, according to the SVM confusion matrix. About 84% accuracy was attained by the model. The model's recall is just 93%, despite its excellent accuracy of 85%. This suggests that while the SVM model is good at detecting actual positive situations, it has trouble detecting all positive cases, which results in some false negatives. The intricacy of the decision boundary and the SVM's sensitivity to hyperparameter adjustment might be the cause of this.

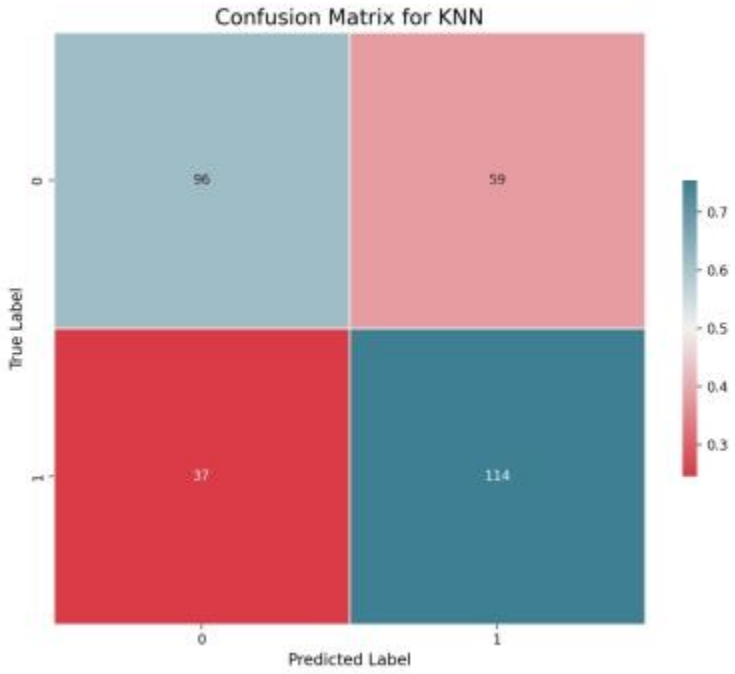

Figure 4. 46: KNN Model Prediction

Classifying cases into two classes (0 and 1) yields a reasonable performance, according to the confusion matrix for KNN. About 75% accuracy was attained by the model. The model's recall is just 76%, but its accuracy is a respectable 66%. This suggests that some false negatives result from the KNN model's inability to correctly recognize every good event. This might be because KNN is sensitive to the distance measure and k-value selection.

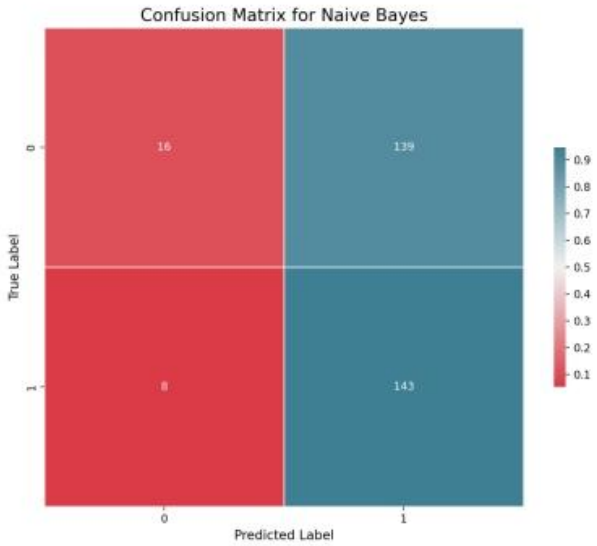

Figure 4. 47: Naive Bayes Model Prediction

Classifying cases into two classes (0 and 1) yields a reasonable performance, according to the confusion matrix for KNN. About 75% accuracy was attained by the model. The model's recall is just 76%, but its accuracy is a respectable 66%. This suggests that some false negatives result from the KNN model's inability to correctly recognize every good event. This might be because KNN is sensitive to the distance measure and k-value selection.

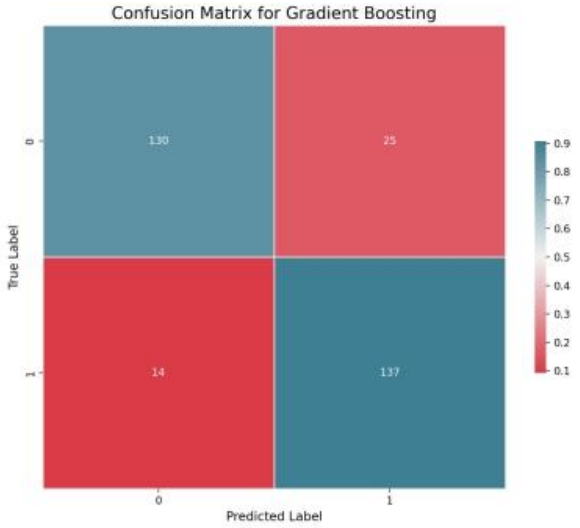

Figure 4. 48: Gradient Boost Model Prediction

Strong performance in dividing occurrences into two classes (0 and 1) is shown by the Gradient Boosting confusion matrix. The accuracy of the model was high, at about 89%. It was successful in recognizing both positive and negative cases, as evidenced by its high accuracy and recall scores. The model's outstanding performance is probably a result of its capacity to integrate several weak learners into a powerful ensemble.

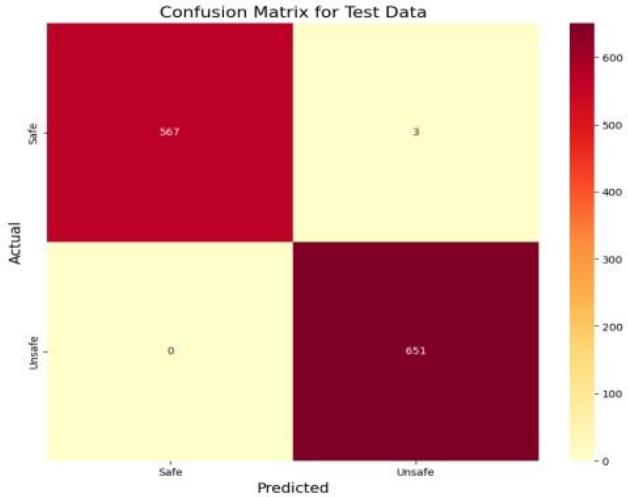

Figure 4. 49: ML and FL Integrated Model Predictions

Based on the confusion matrix, the combined machine learning and fuzzy logic model performs very well in dividing cases into "Safe" and "Unsafe." With just three cases misclassified, the model's accuracy was 99%. The model's high ability to distinguish between "Safe" and "Unsafe" cases is demonstrated by its exceptional accuracy and recall. This implies that the model is very dependable in producing precise forecasts, which is essential for

evaluating and making decisions about the quality of water.

Confusion matrix of machine learning algorithms as well as integrated model of machine learning and fuzzy logic is shown in subplot in figure 4.50 below.

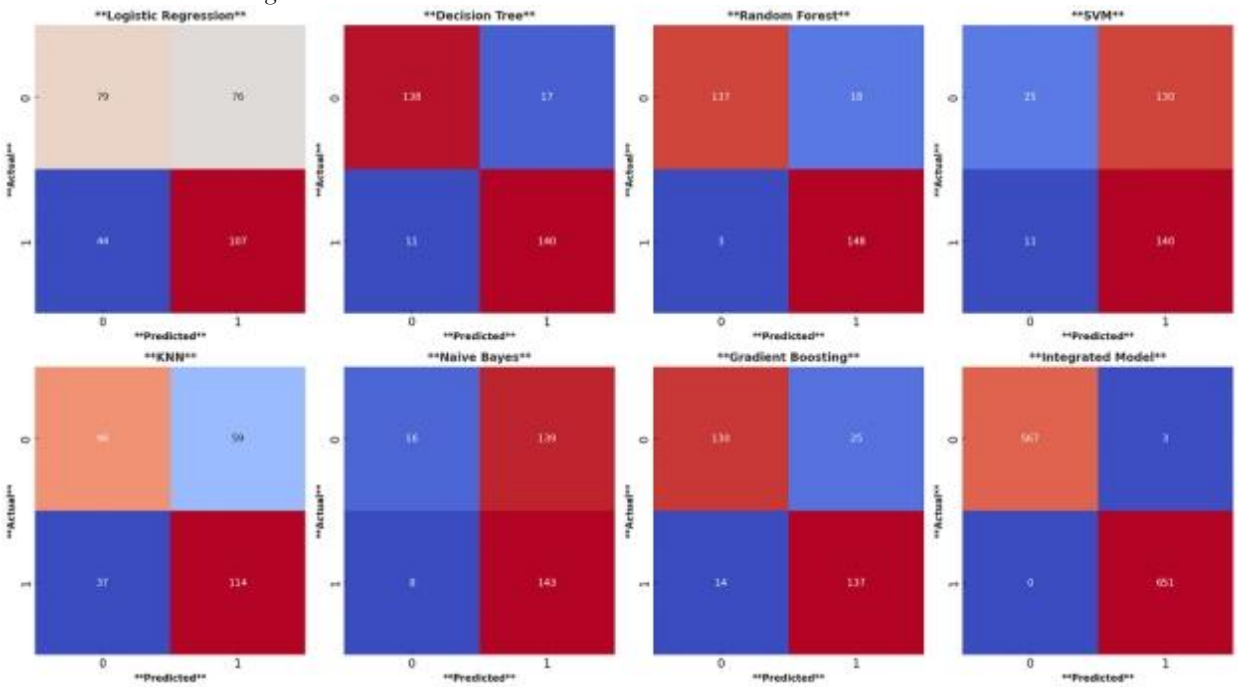

Figure 4. 50 : Subplot of Confusion Matrix of Machine Leaning Models

A comparison of machine learning methods for evaluating water quality is shown in the table 4.7 that is supplied. The Integrated Model, which combined Random Forest with Fuzzy Logic, outperformed the other examined algorithms. This model outperformed other algorithms such as Random Forest (96%), SVM (84%), Decision Tree (93%), and others, achieving a remarkable accuracy of almost 99%. Fuzzy logic improves the model's capacity to manage the imprecision and uncertainty present in water quality data, resulting in forecasts that are more reliable and accurate. A web application has been created to support real-time water quality evaluation and decision-making, based on the Integrated Model's outstanding performance.

Table 4. 7: Models Prediction Comparison

Algorithm	True Positive (TP)	False Positive (FP)	False Negative (FN)	True Negative (TN)	Accuracy	Precision	Recall	F1-Score
Logistic Regression	79	76	44	107	58%	51%	64%	57%
Decision Tree	140	11	17	138	93%	93%	89%	91%
Random Forest	148	3	18	137	97%	98%	89%	93%
SVM	140	25	11	130	84%	85%	93%	89%
KNN	96	59	37	114	75%	66%	76%	70%
Naive Bayes	16	139	8	143	55%	10%	67%	17%
Gradient Boosting	137	25	14	130	89%	85%	91%	88%
Integrated	567	3	0	651	**99.92%**	**99.47%**	**100%**	**99.73%**

Model								

Web Application

To implement the integrated model, a web application was created utilizing the Streamlit framework. Streamlit is a great option for quick prototyping and deployment since it offers a simple method for creating interactive web apps in Python.

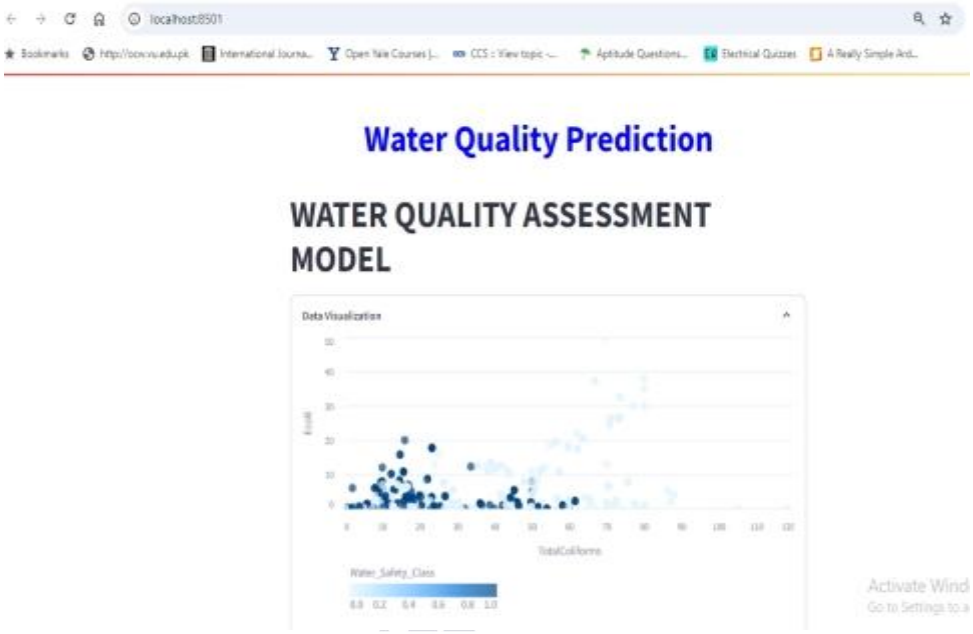

Figure 4. 51: Web Application Interface

Web Interface Working

Users may enter different water quality criteria using the online application's user-friendly interface. After submission, the model analyzes the input data and produces a water quality estimate with a confidence level. Because of its user-friendly interface and clear visualizations, the program may be used by users with different levels of technical proficiency.

Figure 4. 52: Web Application Input Parameters Interface

The integrated model, which combines the power of Fuzzy Logic and Random Forest, is deployed using the Joblib library. This library enables the serialization of Python objects, including machine learning models, allowing for efficient deployment and execution in a web environment.

Figure 4. 51: Web Application Output

Conclusion

This study aimed to develop an integrated machine learning and fuzzy logic model for assessing drinking water quality. To achieve this, various machine learning algorithms—including Logistic Regression, Decision Trees, Random Forest, SVM, KNN, Naïve Bayes, and Gradient Boosting—were explored. The findings demonstrated that combining Fuzzy Logic and Random Forest provided superior classification accuracy compared to standalone models. The research successfully addressed its objectives by developing an integrated model where Fuzzy Logic handled uncertainties in water quality parameters, while Random Forest enhanced classification accuracy by learning complex patterns from the dataset. The model was rigorously compared with existing classification approaches, and results showed that the integrated approach outperformed traditional models in terms of accuracy, precision, recall, and F1-score.

The model was further tested and evaluated, demonstrating high reliability in distinguishing "safe" and "unsafe" water. The incorporation of fuzzy logic improved interpretability, while Random Forest strengthened predictive performance, making the model both robust and effective. Additionally, a web-based application was developed to enhance the model's usability, allowing users to input water quality parameters and instantly receive a classification result, enabling informed decision-making regarding water usage and management.

In summary, this research successfully developed, compared, and validated an advanced water quality assessment model, demonstrating the potential of integrating machine learning with fuzzy logic to improve water quality monitoring systems for researchers, policymakers, and the general public.

Significance of Findings and Future Work

The findings of this study have important significance for both practical applications and future research in water quality evaluation. By combining fuzzy logic with machine learning, notably Random Forest, the model successfully manages uncertainty in water quality metrics while maintaining high classification accuracy. This hybrid method improves decision-making by offering a more interpretable and consistent framework for categorizing drinking water as "safe" or "unsafe." The creation of a user-friendly online application expands accessibility by allowing real-time evaluations for customers, regulatory agencies, and water management groups.In the future, this model might be used for larger environmental monitoring systems such as wastewater treatment, industrial effluent analysis, and agricultural water quality evaluation. Furthermore, including real-time sensor data into the model might boost its responsiveness and flexibility for large-scale deployment. Future study might look at the use of deep learning approaches to increase feature extraction and classification accuracy,

particularly with complex water quality datasets. Furthermore, developing the model to assess water quality in response to regional or climatic fluctuations might aid in tailoring water safety recommendations to specific geographic locations.

Overall, this work adds to the improvement of intelligent water quality assessment systems by providing a scalable and effective method of assuring safe drinking water. By bridging the gap between old water testing methods and current AI-driven technologies, this study lays the groundwork for future advances in water quality monitoring and management.

References

[1] H. Rasheed, F. Altaf, K. Anwaar, and M. Ashraf, *Drinking water quality in Pakistan: current status and challenges*. Islamabad: Pakistan Council of Research in Water Resources, Ministry of Science and Technology, 2021.

[2] L. Godo-Pla, P. Emiliano, J. Suquet, F. Valero, M. Poch, and H. Monclús, "Integrating data-based and knowledge-based models in an Environmental Decision Support System for the management of a Drinking Water Treatment Plant".

[3] A. Mahmood, W. Muqbool, M. W. Mumtaz, and F. Ahmad, "Application of Multivariate Statistical Techniques for the Characterization of Ground Water Quality of Lahore, Gujranwala and Sialkot (Pakistan)," vol. 12, no. 1, 2011.

[4] A. M. P. B. Alahakoon, M. M. Nibraz, P. M. S. S. B. Gunarathna, S. Thenuja, K. A. D. C. P. Kahandawaarchchi, and N. D. U. Gamage, "Water Quality Index Based Prediction of Ground Water Properties for Safe Consumption," in *2020 2nd International Conference on Advancements in Computing (ICAC)*, Malabe, Sri Lanka: IEEE, Dec. 2020, pp. 55–60. doi: 10.1109/ICAC51239.2020.9357146.

[5] "Assessment of Ground Water Quality of Dera Ismail Khan, Pakistan, Using Multivariate Statistical Approach," 2018.

[6] "The Commonwealth Scientific and Industrial Research Organisation," *Curr. Biol.*, vol. 7, no. 3, p. R126, Mar. 1997, doi: 10.1016/S0960-9822(97)70976-X.

[7] F. Jan, N. Min-Allah, and D. Düştegör, "IoT Based Smart Water Quality Monitoring: Recent Techniques, Trends and Challenges for Domestic Applications," *Water*, vol. 13, no. 13, p. 1729, Jun. 2021, doi: 10.3390/w13131729.