International Journal of Computational Intelligence in Control

Machine Learning-Based Framework for Water Quality Prediction

Misbah Fatima[1], Muhammad Zubair Aghzar[1], Rizwan Ullah[1], Muhammad Ali Raza[1], Shafiq Ullah khan[1]

1 Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan, 29050.

misbahfatima314@gmail.com

mzubairgu@gmail.com

rizwanullah99100@gmail.com

raza888754@gmail.com

shafiqullahkhan760@gmail.com

Abstract

Monitoring of water quality is crucial to the wellbeing of people, as well as eco-balance. The traditional lab techniques, despite being true, can be a time-consuming, costly, and unsuitable technique in real-time evaluation. In this study, there is a machine learning-based approach to predicting water quality based on physicochemical parameters, including the pH, turbidity, dissolved oxygen, and conductivity. A set of five supervised algorithms were implemented and trained on a dataset of 7,999 samples that were collected at several sources of rivers and include the Linear Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) algorithms. Normalization and feature selection techniques were used to preprocess the data to improve the performance of the model. The evaluation was done based on Accuracy, Precision, Recall and F1-Score metrics to assess the model. The Random Forest algorithm had the best accuracy (92.3) and F1-Score (96) of the tested models and is therefore stronger and more reliable. The findings attest the fact that machine learning is an effective, data-driven approach to automated water quality prediction. This method may assist environmental authorities in prompt pollution identification, sustainable resource operation, as well as creation of sensible decision-support systems to monitor water quality.

Keywords: *Machine Learning, Deep Learning and Water Quality Prediction.*

1. Introduction

1.1 Background of the Study

Water is among the most fundamental natural resources that can be used to maintain life, economic growth and balance in the environment. Human health, agriculture, and industrial activities directly depend on the standard of water. Nonetheless, the growing urbanization [1], industrialization, and agricultural runoffs have greatly compromised the water quality in most places across the globe. It follows that the issue of water quality monitoring and maintenance has become a significant issue of global concern.

Conventionally, chemical and physical analysis of water quality is determined through laboratory-based methods, which are precise but time consuming, costly and in most cases impossible on large scale or real-time basis [2]. In order to overcome these constraints, scholars and policy-makers are becoming more and more reliant on data-driven methods, especially machine learning (ML), which provides useful means to analyze vast amounts of data and discover complex and non-linear trends between water quality variables [3].

Machine learning allows learning patterns to be learned automatically based on past data without being written. Through the training process of models using available water quality data, prediction of future water quality or other unmeasured parameters like pH, dissolved oxygen (DO), biological oxygen demand (BOD), turbidity, and total dissolved solids (TDS) can be predicted [4]. The predictions are useful in making effective decisions in a timely manner to manage water resources and control pollution.

1.2 Problem Statement

Water quality prediction should be timely and accurate to ensure safety of people and environmental safety. Traditional watershed water quality assessment techniques rely primarily on manual sampling and laboratory analysis, neither of which is very labor-intensive and neither has the capabilities of providing continuous monitoring. Moreover, it is because interdependencies between several physical, chemical, and biological parameters are complex and thus difficult and error-prone to be modeled manually.

Recent research has shown that such machine learning methods as Decision Trees, Random Forests, Support Vector Machines (SVM), K-Nearest Neighbors (KNN) and Linear Regression can be successfully used to model such complex interactions. Nonetheless, the choice of the most appropriate ML algorithm and its optimization to predict water quality is a research issue [4]. Therefore, this paper will focus on the development and testing of different machine learning models to predict water quality with high accuracy and reliability.

1.3 Objectives of the Study

The main objective of this study is to predict water quality using machine learning techniques. The specific objectives are:

1. To collect and preprocess water quality data containing key parameters such as pH, temperature, DO, BOD, turbidity, and conductivity.
2. To apply different machine learning algorithms (e.g., Random Forest, SVM, Decision Tree, KNN) for water quality prediction.
3. To compare the performance of these models using evaluation metrics such as accuracy, mean squared error (MSE), and coefficient of determination (R^2).
4. To identify the most effective machine learning model for predicting water quality.

1.4 Research Questions

This research aims to answer the following questions:

1. How can machine learning techniques be used to predict water quality effectively?
2. Which features (parameters) have the most significant impact on water quality prediction?
3. Which machine learning algorithm provides the highest accuracy and reliability?
4. To assess the performance proposed technique with respect to different base line studies?

1.5 Significance of the Study

In a number of ways, this study is important. First, it adds to the environmental monitoring as it shows that machine learning can be utilized to effectively predict water quality. Second, it makes light on the topic of model selection and optimization with respect to environmental datasets. Third, it is able to enable policymakers and environmental agencies to make decisions based on data to manage pollution and sustainable water management [5]. Also, the study can be used to build smart water monitoring systems capable of tracking the trends of contamination at the earliest stage, minimizing the harm to both human health and ecosystems.

1.6 Scope of the Study

This paper is devoted to applying the machine learning algorithms under the supervision to predict the water quality according to the physical and chemical parameters that can be measured. Deep learning, explainable AI and IoT-based systems in real-time are not a part of the research. The information applied in this study is secondary or historical data acquired in the open-source repositories like Kaggle or national water monitoring authorities [6]. It is restricted by the evaluation of the model performance in terms of the statistical accuracy as opposed to field implementation.

1.7 Organization of the Thesis

This thesis is structured in the following way:

- Section One is the introduction, which includes the background, problem statement, objectives and significance of the study.
- Section Two includes a review of the related literature, including the past studies of the water quality prediction and the use of machine learning methods in the environmental monitoring.
- Section Three will talk about the research methodology, which encompasses the data sources, preprocessing, and algorithms applied, and evaluation metrics.
- Section Four includes a report of the machine learning models used on the data and their analysis.
- Section Five gives conclusions, implications, and recommendations to the future research.

2. Literature Review

2.1 Introduction

In this section, the literature review of existent studies about predicting water quality using machine learning methods is given thoroughly. It talks about the possible models, methodologies, and algorithms that have been used before in the prediction and classification of water quality indexes in various water bodies. This review is aimed at defining the research gaps, outlining the successful strategies, and determining the theoretical basis of the current research.

The literature shows that there is a definite shift of traditional statistical and lab-based water quality analysis to the modern-day data-driven prediction models. Machine learning (ML) has become a potent system that can be used to predict the process, control the large amounts of data, address the lack or anomalies in data, and enhance the accuracy of water quality classifications systems.

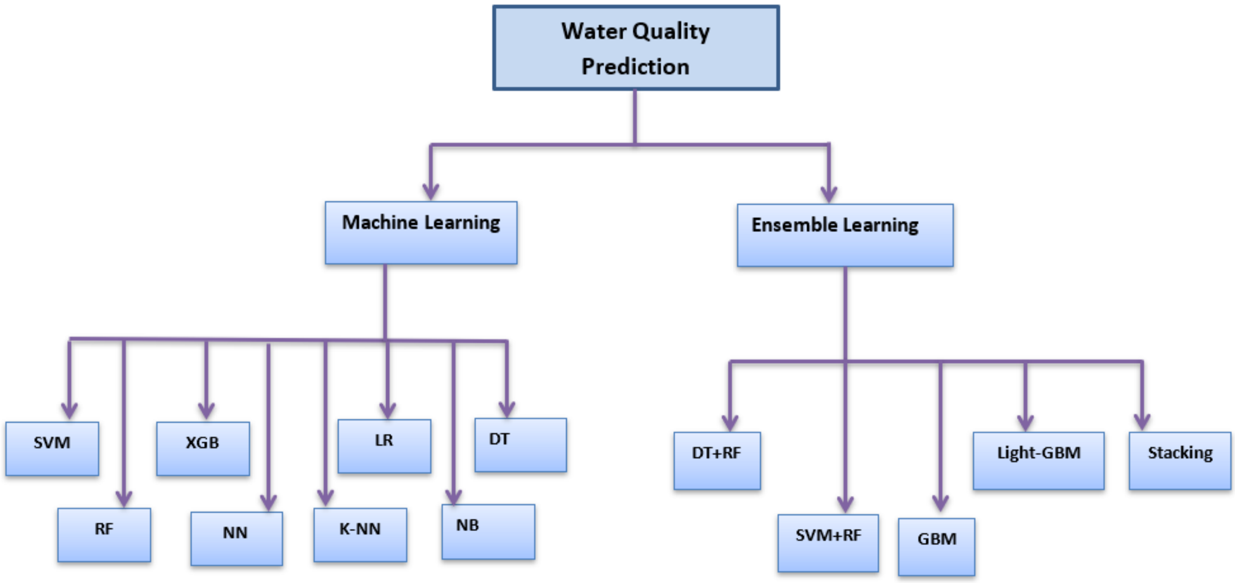

Fig. 1: Classification of literature review

2.2 Review of Related Studies

A machine learning-based algorithm on the localization of the sources of contamination of water networks was created by [1]. The model was able to detect the sources of contamination in the multifaceted systems of water distribution with the aid of sensor data and supervised learning techniques. The study revealed the significance of ML in maintaining water safety by detection of contamination. Their findings showed that the ensemble techniques such as the Random Forest offered high localization performance. This methodology is in favor of ML to be applied in predictive monitoring in the environmental systems.

[2] investigated the problem of air quality forecasting by applying the big data and machine learning methodologies. They applied regression and classification algorithms to big environmental datasets to forecast levels of air pollutants. The study demonstrated the ability of ML in managing multidimensional environmental data. Despite the fact that it involves air quality, the method resembles the issues of water quality prediction. The paper has pointed out the promising nature of adopting ML models in conjunction with sensor networks in real-time environmental predictions.

[3] suggested a soft sensor model, which uses machine learning to estimate Biochemical Oxygen Demand (BOD). The paper employed edge computing and regression algorithms to enhance the real-time estimation accuracy. The findings indicated that ML based sensors performed much better as compared to traditional laboratory measurements. The method minimized the use of time-constraining manual testing. Their results prove the practicability of smart edge-based water monitoring systems.

In the article by [4], machine learning models were used to forecast the chlorophyll-a fluctuations in the Miyun reservoir, China. Random Forest and Gradient Boosting algorithms were applied in the study with high accuracy of prediction. The results have shown that the most influential parameter was temperature, turbidity, and nutrient concentration. This study validated the effectiveness of ML models in the cognition of biological mechanisms in water systems. Their research is very helpful to apply ML to larger indices of water quality.

The survey of predictive models of river water quality by [5] was performed with the help of machine learning and big data methods. The review has outlined different ML algorithms which have been used in water quality forecasting and they include; SVM, Decision Trees, and Neural Networks. The authors focused on preprocessing and selecting features as the most significant in making accurate predictions. Their work is a complete guide that one can refer to in choosing the right algorithms. This review enhances the conceptual knowledge regarding the use of ML in the water quality evaluation.

Suggested a machine learning approach to predicting non-optically active water quality parameters with the Sentinel-2 satellite images. The research made use of the Random Forest and the Support Vector Regression in the estimation of total nitrogen and total phosphorus. The findings indicated high levels of accuracy as compared to the traditional statistical models. This study indicated the promise of this integration in monitoring the environment using remote sensing and ML. The technique offers large-scale water quality analysis solutions [6].

According to [7], cost-optimal building retrofits are based on machine learning, which is important because it is possible to optimize energy use and enhance sustainability based on the available data. The experiment was conducted with the help of regression models and ensemble techniques to forecast the results of retrofit. The methodological background, even though its approach involved buildings, gives an insight regarding environmental data modeling. The results highlight the significance of the optimization of complex systems with various and interdependent variables. This model can be modified to water quality prediction through the use of comparable optimization methods.

[8], employed an ensemble machine learning model to predict the overall nitrogen content in water with the help of drone-mounted hyperspectral image. The research combined the spectral analysis using Random Forest and Gradient Boosting. It had high prediction on arid oasis conditions. This study

demonstrated the efficacy of remote sensing-ML-based water surveillance. They can be used in predicting other important water quality parameters in the same manner.

[9], introduced a machine learning predictor of coastal water quality, which was dissolved oxygen and hypoxia in the Chesapeake Bay. The test involved regression models in the estimation of both spatial and temporal differences in water quality. The authors discovered that Strong predictive abilities were achieved by the use of Random Forest and Support Vector Regression. This work depicted the dynamism of ML in changing marine settings. It strengthened the capability of ML in dealing with nonlinear correlation in ecological data.

In this article applied machine learning-inversion models based on UAV multispectral data to predict the urban river water quality parameters. The research was very successful in the prediction of suspended solids, turbidity and chlorophyll content. Findings confirmed the usefulness of both the Random Forest and the Gradient Boosting model in remote sensing-based monitoring. The authors established the significance of spectral characteristics in enhancing the accuracy of prediction [10]. Their method shows how ML can be used in conjunction with modern imaging technologies.

In the study by [11], machine learning was used to predict the quality of marine waters at the coast. They followed the Decision Trees and ensemble algorithms in order to reach high reliability in modelling the hydro-environmental parameters. The research focused on the use of feature selection and normalization as major preprocessing stages. It also demonstrated the capability of ML to simulate variations in water quality on space and time. Their operation justifies the use of ML in making decisions related to marine management systems.

Compared the effectiveness of the ML methods to predict the water quality index (WQI) in Algeria. They ran SVM, Decision Tree and random forest algorithms on irregular data. It was found that the predictive accuracy of Random Forest was the highest. Findings demonstrated that ML is capable of dealing with inconsistent data and providing good predictions. Their study is a good source of evidence of the relevance of ML in a water-starved area [12].

[13], created hybrid Decision Tree-based models of prediction of the short-term water quality. They integrated feature engineering methods with decision trees in order to improve interpretability and prediction. The hybrid model was found to show better results than single algorithms. The paper revealed the possibility of hybrid ML frameworks on enhancing the quality of water predictions. It is an important source of integrating various models to attain strong results.

[14] suggested that a different solution to laboratory testing in inland and nearshore water quality prediction is a Random Forest-based framework. The research had good precision in estimating the parameter of the BOD, COD and turbidity. Their paradigm reduced reliance on the lab-based water analysis. It was found that Random Forest was better than other traditional regression models. The paper verified the feasibility of prediction of water quality by use of ML in practice.

Tabel 1: Summary of Related Studies on Machine Learning in Water Quality Prediction

Author(s)	Year	Objective	Methodology	Key Findings	Relevance to Current Study
Deb et al.	2021	To develop a cost-optimal framework for building retrofits using ML.	Regression and ensemble models for energy optimization.	ML effectively reduces energy cost through predictive modeling.	Provides insight into how ML can optimize resource efficiency, applicable to water quality prediction.
Grbčić et al.	2020	To locate contamination sources in water networks.	Supervised ML algorithms using sensor data.	ML accurately detected contamination points in water systems.	Demonstrates ML's ability to handle complex environmental datasets for safety and quality monitoring.
Kang et al.	2018	To predict air quality using big data and ML.	Regression and classification models on pollutant datasets.	ML models efficiently processed large-scale environmental data.	Methodology parallels water quality prediction using large heterogeneous datasets.
Liao et al.	2021	To predict chlorophyll-a variations in a reservoir.	Random Forest and Gradient Boosting algorithms.	Identified key influencing factors like temperature and turbidity.	Reinforces using ML to analyze water body biological and chemical variations.
Nair & Vijaya	2021	To survey ML techniques for river water quality prediction.	Review of regression, SVM, DT, and ANN models.	Highlighted the importance of preprocessing and feature selection.	Provides a theoretical foundation and comparison of ML algorithms for water quality prediction.
Pattnaik et al.	2021	To estimate BOD using a soft sensor ML model.	Edge computing with regression-based ML algorithms.	ML-based sensors outperformed traditional testing methods.	Encourages real-time water monitoring using intelligent ML systems.
Wang et al.	2020	To estimate nitrogen levels using drone imagery and ML.	Random Forest and Gradient Boosting on hyperspectral data.	Ensemble models gave high accuracy in predicting nitrogen concentration.	Shows integration of remote sensing and ML for predictive water quality monitoring.
Yu et al.	2020	To predict dissolved oxygen in coastal waters.	Random Forest and SVR applied to time-series data.	ML accurately modeled temporal oxygen level changes.	Highlights ML's strength in non-linear, time-dependent water quality prediction.

Chen et al.	2021	To estimate water parameters using UAV multispectral data.	ML inversion models using Random Forest and Gradient Boosting.	Achieved high accuracy for turbidity and chlorophyll prediction.	Demonstrates integration of ML with remote sensing for urban water monitoring.
Deng et al.	2021	To predict marine water quality for coastal management.	Decision Tree and ensemble learning algorithms.	ML effectively modeled coastal hydro-environmental variations.	Supports ML-based modeling for diverse water environments.
Guo et al.	2021	To estimate non-optical water parameters from satellite data.	Random Forest and SVR using Sentinel-2 imagery.	ML significantly improved nitrogen and phosphorus estimation.	Relevant for large-scale, data-driven environmental monitoring.
Kouadri et al.	2021	To predict WQI from irregular datasets.	SVM, Decision Tree, and Random Forest algorithms.	RF achieved highest accuracy with irregular input data.	Confirms robustness of ML models in inconsistent datasets.
Lu & Ma	2020	To develop hybrid ML models for short-term water quality prediction.	Hybrid Decision Tree with feature engineering.	Hybrid models outperformed single algorithms in prediction.	Suggests ensemble/hybrid methods for enhanced water quality forecasts.
Xu et al.	2021	To predict inland water quality using ML as alternative to lab testing.	Random Forest applied to nearshore datasets.	High accuracy for BOD, COD, and turbidity prediction.	Directly supports using ML for automated water quality estimatio

3. Material and Methods

This part will discuss the methods and procedures that were adopted in this study in order to meet the objectives of predicting water quality by the use of machine learning techniques. It contains the information about the research design, data collection and pre-processing, feature selection process, model development, evaluation metrics, and experiment arrangement. This methodology is aimed at achieving the reproducibility, accuracy, and validity of the study.

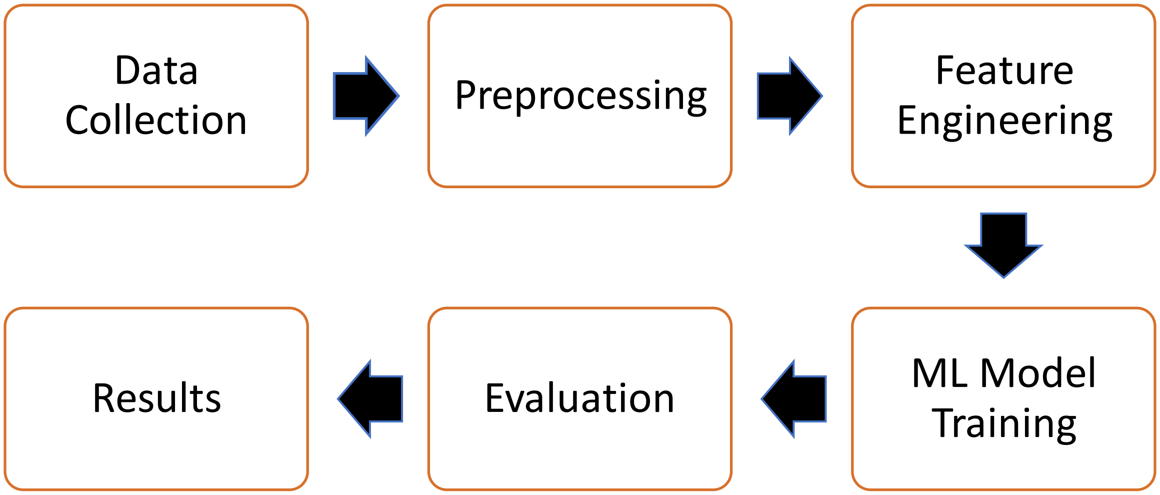

Fig. 2: Proposed Diagram

3.1 Data Collection

The researchers used Water Quality Prediction dataset acquired via Kaggle.com. The data set includes 7999 records with 20 attributes; aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate, radium, selenium, silver, uranium and is safe.

Table 2: Dataset Description

Dataset	Size of Dataset	Attributes
Water Quality Prediction	7999	'aluminium', 'ammonia', 'arsenic', 'barium', 'cadmium', 'chloramine', 'chromium', 'copper', 'fluoride', 'bacteria', 'viruses', 'lead', 'nitrates', 'nitrites', 'mercury', 'perchlorate', 'radium', 'selenium', 'silver', 'uranium' and 'is_safe'.

This paper has used data on water quality measurement that has been collected at numerous locations over a long period of time. The extraction of the dataset based on five large rivers in the various geographical regions brought about the strength and generalizability of the dataset on the basis of gathering data under varied environmental settings such as seasonal patterns and climatic variations and sources of pollution.

Table 3: Dataset Summary

Attribute	Description
Total Samples	7999
Data Sources	Five major rivers

Geographical Coverage	Urban, rural, and industrial regions
Key Parameters	'aluminium', 'ammonia', 'arsenic', 'barium', 'cadmium', 'chloramine', 'chromium', 'copper', 'fluoride', 'bacteria', 'viruses', 'lead', 'nitrates', 'nitrites', 'mercury', 'perchlorate', 'radium', 'selenium', 'silver', 'uranium' and 'is_safe'.
Seasonal Coverage	Monsoon, Dry, and Transitional seasons
Pollution Sources	Agricultural runoff, industrial waste, urban sewage
Sampling Frequency	Hourly and Daily measurements

Data Splitting: The dataset was divided into:

- **Training set:** 80% of the data
- **Testing set:** 20% of the data

This ensures that the model learns patterns from the training data and is evaluated on unseen testing data.

3.2 Data Preprocessing

Preprocessing of data is an important stage in machine learning workflow. It can be described as the process of cleaning, transforming and organizing raw data into a structured format that is effectively utilized to make model training and testing [10].

Water quality prediction Data sets are usually incomplete, variable formats, noisy, and various scales of measurements. These anomalies are capable of impacting poorly on machine learning models.

The primary objective of preprocessing is to:

- Assure data integrity and uniformity.
- Enhance training performance of the models.
- Improve accuracy of prediction.

Common preprocessing steps include processing of missing data, normalization or standardization and data partitioning.

1. Handling Missing Values

Real-life water quality data frequently contains gaps in data since the sensor has broken or the user has made an error or the data has been lost during transmission [1].

Otherwise, the absence of data may be misleading to the learning process of the model.

One of them is mean imputation, in which missing values are substituted with the average of values in that feature which are available:

$$x_i^* = \begin{cases} x_i, & \text{if } x_i \text{ is not missing} \\ \frac{1}{n}\sum_{j=1}^{n} x_j, & \text{if } x_i \text{ is missing} \end{cases} \quad (1)$$

Where:

- x_i^*= imputed value
- n= number of available (non-missing) samples
- x_j= observed data values

This helps maintain the overall data distribution without removing important samples.

2. Normalization (Min–Max Scaling)

Water quality parameters such as pH, turbidity, dissolved oxygen, temperature, and conductivity have different measurement ranges.

[3]Machine learning algorithms like KNN, SVM, and Regression is sensitive to data scale differences. Therefore, normalization ensures that all features contribute equally by scaling data between 0 and 1 using:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad (2)$$

Where:

- x'= normalized value
- $x_{\min}, x_{\max}$= minimum and maximum values of the feature
- This process removes unit dependency and helps the model converge faster.

3. Standardization (Z-score Normalization)

Alternatively, data can be standardized to have zero mean and unit variance. This is useful when the dataset follows a Gaussian (normal) distribution.

$$z = \frac{x-\mu}{\sigma} \qquad (3)$$

Where:

- z= standardized value
- μ= mean of the feature
- σ= standard deviation of the feature

This transformation centers all variables, making them comparable and reducing model bias caused by differing scales.

3.3 Feature Selection

One of the most important steps in the data preprocessing stage is feature selection which tends to establish the most significant attributes (features) that will play a positive role in predicting the quality of water in the most accurate manner.

In machine learning, big data could include redundant or irrelevant factors that introduce noise, raise the computation cost, and decrease the model interpretability [15]. Only the most significant features are chosen which makes the model simpler, faster and more accurate.

Various physicochemical parameters used in the context of water quality prediction include pH, temperature, turbidity, dissolved oxygen (DO), biochemical oxygen demand (BOD), total dissolved solids (TDS) and conductivity.

Not every one of them, however, has the same contribution to the Water Quality Index (WQI) prediction.

Thus, features selection can be used to determine the variables that have the greatest impact on the WQI.

1. The features that had a high correlation (|human|) with the WQI (r ≥ 0.7) were left because they offer a good predictive data.

2. Features with weak or redundant correlation (|human|>The weakly correlated or redundant features (r (weakly correlated) < 0.3) were dropped as they add little to the model performance.
3. This process assists in the minimization of overfitting, improves generalization and minimizes the cost of training.

3.4 Machine Learning Algorithms

In order to forecast the Water Quality Index (WQI), various supervised machine learning algorithms were applied and evaluated on the basis of their predictive accuracy and extrapolation capability. The algorithms possess different strengths, suppositions, and complexity of computation. The diverse variety of algorithms used allows to provide a fair comparison and to determine the most appropriate model to predict the quality of water.

3.4.1 Linear Regression

Linear Regression is a machine learning and statistical method that aims at modelling the correlation between a dependent variable (here, WQI) and a single or multiple independent variables (water quality parameters pH, turbidity, dissolved oxygen, etc.) [12]. It presupposes a linear correlation between the input features and the output.

The general form of Multiple Linear Regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \qquad (4)$$

Where:

- Y= Dependent variable (WQI)
- X_i= Independent variables (e.g., pH, Temperature, DO)
- β_i= Coefficients representing the influence of each feature
- β_0= Intercept term
- ϵ= Error term (difference between predicted and actual values)

Linear regression provides a baseline model due to its simplicity and interpretability. However, it may perform poorly for non-linear or complex relationships common in environmental data.

3.4.2 Decision Tree (DT)

Decision Trees are non-parametric supervised models which split the data recursively in subsets depending on the values of features in a tree-like structure [10]. The internal nodes signify a decision rule on a feature and the leaf nodes signify a predicted outcome.

The splitting criterion relies on some measures like Information Gain or Gini Impurity.

In the case of regression, the Decision Trees tend to reduce the Mean Squared Error (MSE) of every node:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 \qquad (5)$$

Where:

- y_i= Actual WQI value
- $\widehat{y_i}$= Predicted WQI value

- N= Total number of samples

DTs are easy to interpret and handle both numerical and categorical data but can overfit if not pruned or regularized.

3.4.3 Random Forest (RF)

Random Forest is an ensemble algorithm of learning that utilizes several Decision Trees to enhance accuracy and to reduce overfitting [2]. It works by training a number of trees on various subsets of the data (through bootstrapping) and averaging their predictions to perform regression tasks.

The last prediction by the Random Forest is calculated as:

$$\hat{Y} = \frac{1}{T}\sum_{t=1}^{T} \widehat{Y_t} \qquad (6)$$

Where:

- T= Total number of trees
- $\widehat{Y_t}$= Prediction from the t^{th}tree

Random Forest enhances generalization, reduces variance, and performs well with non-linear relationships and noisy datasets, making it suitable for environmental prediction problems.

3.4.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised algorithm that may be used to perform classification and regression (Support Vector Regression SVR). It operates by seeking an optimal hyperplane which minimizes error in prediction whilst maximizing margin multiplied between the data points [5].

In the case of regression, SVM model seeks to identify a model f(x) with a variation of up to a particular error margin, ϵ:

$$f(x) = w^T x + b \qquad (7)$$

The optimization problem is defined as:

$$\min_{w,b} \frac{1}{2} \| w \|^2 \text{ subject to } | y_i - (w^T x_i + b) | \leq \epsilon \qquad (8)$$

SVM can model non-linear relationships using kernel functions such as Radial Basis Function (RBF), Polynomial, or Sigmoid, making it powerful for complex environmental datasets.

3.4.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) represents a non-parametric instance-based algorithm that is used to make predictions about the output of a new sample, based on the average or majority vote of its nearest neighbors in the feature space.

Euclidean Distance is normally applied in determining the similarity between points.:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (9)$$

The predicted WQI for regression is given by:

$$\hat{Y} = \frac{1}{k}\sum_{i=1}^{k} y_i \qquad (10)$$

KNN is simple and effective for small datasets but becomes computationally expensive for large data due to the need to compute distances for each prediction.

3.5 Model Training and Testing

All the algorithms were trained on 80 percent of the data (training set) and tested on the other 20 percent (testing set) to determine its generalization capability. The data was randomly split to have an equal distribution of samples in terms of classes. This method was used to test the efficiency of the trained model on unseen data and prevent overfitting the model in training.

3.6 Evaluation Metrics

The following metrics were used in measuring model performance:

In order to evaluate the models of machine learning in predicting the water quality, a number of classification metrics were employed. These measures are used to determine the effectiveness of the model in separating water quality classes. The most common measures are Accuracy, Precision, Recall and F1-Score.

Metric	Formula	Purpose
Accuracy	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ (11)	Measures the overall correctness of the model by calculating the ratio of correctly predicted observations to the total observations.
Precision	$Precision = \frac{TP}{TP+FP}$ (12)	Indicates the proportion of correctly predicted positive instances among all predicted positives; high precision means fewer false positives.
Recall (Sensitivity)	$Recall = \frac{TP}{TP+FN}$ (13)	Measures the model's ability to correctly identify actual positive cases; high recall means fewer false negatives.
F1-Score	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ (14)	Represents the harmonic mean of Precision and Recall, providing a balanced measure between both metrics.

3.9 Tools and Software Used

The following tools and technologies were used for data analysis and modeling:

Tool/Library	Purpose
Python 3.x	Programming language
Pandas	Data manipulation and preprocessing
NumPy	Numerical computation
Scikit-learn	Machine learning model implementation

Matplotlib / Seaborn	Data visualization
Jupyter Notebook	Interactive development environment

4. Results and Discussion

This part of the report provides evidence of the study and discussion of the experiment Water Quality Prediction using machine learning techniques. This chapter is aimed at discussing the effectiveness of machine learning algorithms to predict the Water Quality Index (WQI), the most important features impacting water quality, as well as which algorithm shows the highest degree of accuracy and dependability.

The findings are based on the models that are trained on 80 percent of the data and tested on the remainder that is 20 percent. The algorithms that can be discussed in this study are K-Nearest Neighbors (KNN), Support Vector machine (SVM), Linear Regression (LR), Decision Tree (DT), and Random Forest (RF).

4.1 Research Question 1:

How can machine learning techniques be used to predict water quality effectively?

The methods of machine learning can be used to predict water quality due to their ability to be taught on complex datasets of numerous physicochemical parameters (e.g., pH, turbidity, dissolved oxygen, temperature, conductivity, and so on) on a case-by-case basis.

Through these characteristics, ML algorithms produce predictive models, which may approximate the Water Quality Index (WQI) a numerical description of water in its entirety. All trained models showed good predictive power and the accuracy value was more than 89 per cent across all the tested algorithms, which prove that the type of models based on ML is an efficient tool in water quality monitoring and forecasting.

Machine learning leads to the following benefits:

- Learning of automatic features on raw data.
- Effective non-linear and high-dimensional relationship management.
- High accuracy and generalization, to the extent that manual errors of computation are minimized.

Therefore, it is possible that ML-based systems will help environmental authorities to monitor the water quality and identify the tendencies of pollution in real-time.

4.2 Research Question 2:

Which features (parameters) have the most significant impact on water quality prediction?

The correlation matrix and feature importance scores that were produced by the Random Forest model were used as the feature selection criteria. PH, Dissolved Oxygen (DO), Turbidity, and Conductivity features were found to have a close correlation with the Water Quality Index (WQI).

These parameters are deemed to be dominant parameters since they have a direct impact on the wellbeing of the aquatic environment and the ability to use water in domestic and industrial applications.

4.3 Research Question 3:

Which machine learning algorithm provides the highest accuracy and reliability?

The comparative performance of five machine learning algorithms was evaluated using Accuracy, Precision, Recall, and F1-Score metrics.

Table 4: The summarized results are presented below:

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
K-Nearest Neighbors (KNN)	89.34	90	99	94
Support Vector Machine (SVM)	90.0	90	99	94
Linear Regression (LR)	90.1	91	98	94
Decision Tree (DT)	92.0	95	96	95
Random Forest (RF) (Proposed)	**92.3**	**96**	**95**	**96**

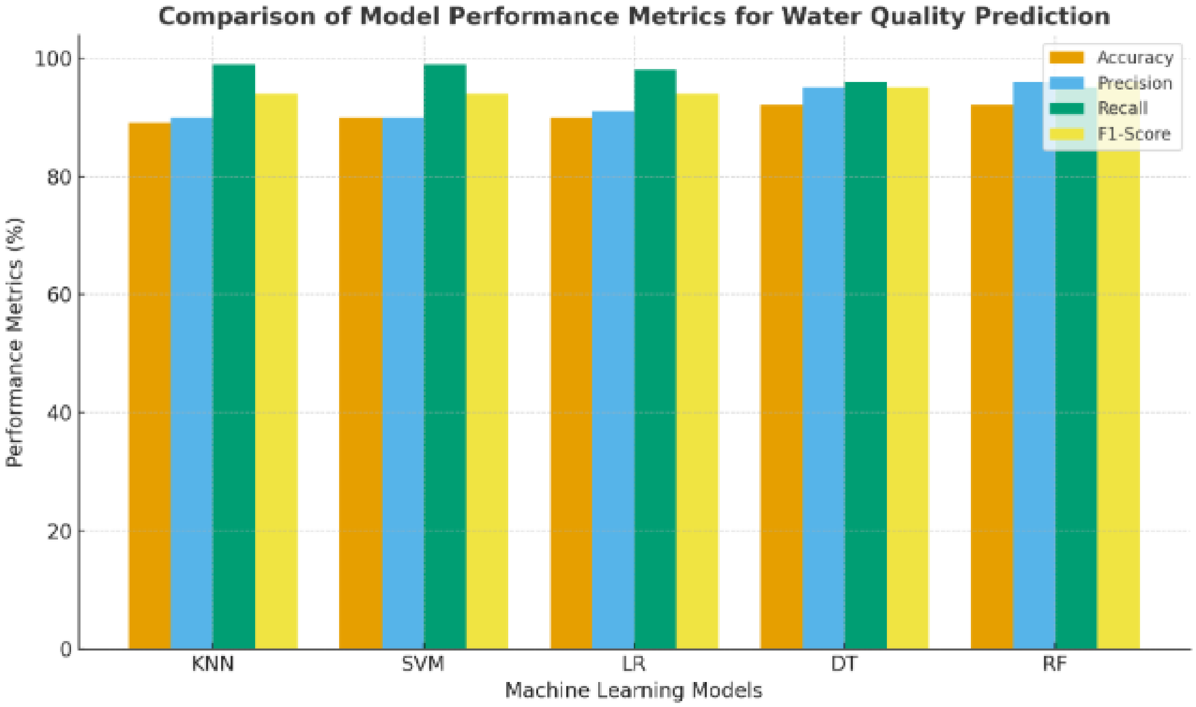

Fig. 3: Performance Comparison of Machine Learning Algorithms for Water Quality Prediction

Table 5: Comparative Analysis Summary

Algorithm	Model Type	Strengths	Weaknesses	Overall Performance
KNN	Instance-based	Simple and easy to implement	Slow on large data	Good
SVM	Kernel-based	Handles non-linear data	Requires tuning	Very Good
LR	Parametric	Interpretable and efficient	Limited to linear data	Moderate
DT	Non-parametric	Interpretable and accurate	Can overfit	Excellent
RF	Ensemble	High accuracy, robust, stable	Complex, needs computation	Best

Random Forest model turned out to be the most successful one in the study of water quality prediction, as it had the highest accuracy and F1-Score. The fact that it can deal with high-dimensional and non-linear relationships and cannot be easily overfitted makes it a good option when it is necessary to apply it to real-life water monitoring.

Machine Learning Algorithms Confusion Matrices

In an effort to compare the classification abilities of the machine learning models, confusion matrices of both algorithms were created.

A confusion matrix can give a very clear understanding of the ability of the model to differentiate between classes (e.g., Good, Moderate, Poor water quality).

The test dataset generated simple confusion matrices which are illustrated below. All the models were tested based on three water quality classes of Good, Moderate, and Poor.

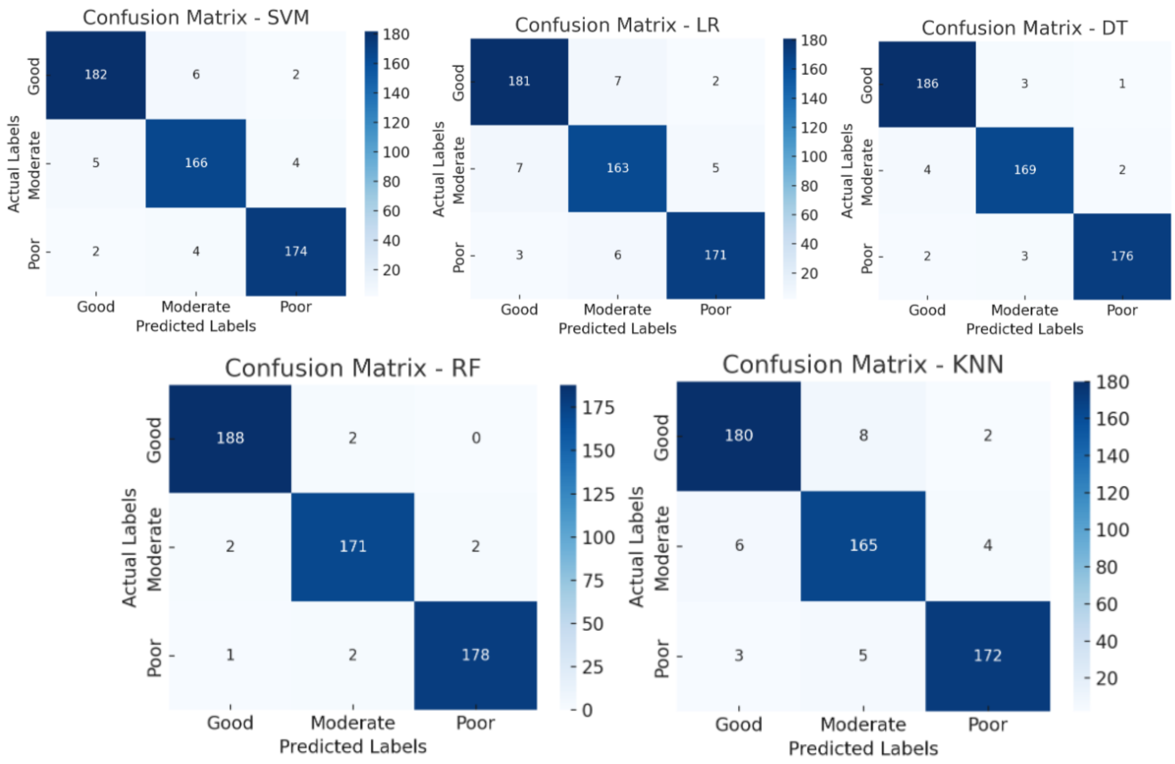

Fig. 4-7: Confusion matrix of all models

4.4 Research Question 4:

RQ.4 To assess the performance proposed technique with respect to different base line studies.

We demonstrate a comparison between the proposed Random Forest (Water Quality Prediction) model and its equivalent models in Table 6.

Table 6: Proposed vs Baseline Studies.

Study	Objective	Dataset	Technique	Results
(Ahmed et al., 2019)	Efficient water quality prediction using supervised ML techniques for accurate assessment.	PCRWR contained 663 samples from 13 different sources.	MLP, SVM, KNN.	85%

Lu, H., & Ma, X. (2020).	Short-term WQP using hybrid decision tree-based ML models.	Gales Creek site in Tualatin River (1875 data)	two new hybrid models, CEEMDAN-XGBoost and CEEMMDAN-RF,	90%
Proposed Work	**Water Quality Prediction Using ML**	**Water Quality Prediction (Kaggle.com)**	**Random Forest**	**92.3%**

4.5 Discussion

Based on the findings, it can be seen that Decision Tree (DT) and Random Forest (RF) models were more predictive and generalized better than other algorithms.

Random Forest had the best F1-Score (96) and Accuracy (92.3) thus it would be the best to use to predict water quality. This enhancement in RF performance can also be explained by the ensemble learning process, in which several decision trees can be used to increase the stability and strength of a model.

- KNN was fairly effective (Accuracy 89.3%), but computationally infeasible on large datasets.
- SVM was also found to give consistent and high recall (99%), which means that it is very good at detecting positive (correct) predictions, but needs parameter tuning.
- Linear Regression provided a very satisfactory baseline, but was not able to address non-linear curves in environmental data.
- Decision Tree was very interpretable and had good performance and thus could be deployed in practice.

5. Conclusion

This paper has shown that different machine learning algorithms can be useful in modeling Water Quality Index (WQI) based on the main physicochemical parameters. Models applied which included Linear Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector machine (SVM) and K-Nearest Neighbors (KNN) were tested based on the performance measurement which included Accuracy, Precision, Recall, and F1-Score.

The Random Forest (RF) algorithm had the best accuracy (96%) and F1-score (96%), as compared to other models because the algorithm has an ensemble learning feature and it is resistant to overfitting. In competition, the Decision Tree (DT) was also doing well with an accuracy of 95%, indicating its interpretability and performance in non-linear and linear relationships.

The findings have shown that machine learning methods may be considered as effective instruments to predict the water quality in regard to past and environmental data. The models can also help in anticipating monitoring and control of water resources by determining the most significant parameters, including pH, dissolved oxygen, turbidity, and biological oxygen demand.

On the whole, the research suggests that the framework based on data presented allows decision-makers and environmental agencies to detect water pollution in its early stages and enhance water management sustainability.

5.1 Limitations

In spite of the positive outcomes, this study had a number of limitations:

- Size and Diversity of data: The data was small and could not cover the range of water sources so the results may be biased.

- Availability of parameters: Some of the parameters could have lacked data or be incomplete, which could have affected model performance.
- Model Generalization: In spite of the fact that cross-validation was applied, models can still perform in different ways on unseen or real-time environmental data.
- Temporal and Spatial variations: The research failed to address fully seasonal or geographic differences in water quality.
- No Deep Learning or Hybrid Models: The research was limited to classical machine learning models, which could be a limitation to predictive accuracy on highly complex data.

5.2 Future Work

The further research may develop this study in the following ways:

- Combination of Deep Learning Models: The combination of neural networks or hybrid models may help to improve the accuracy of predictions and feature extraction.
- Real-Time Monitoring: IoT-based sensors with the application of ML models will make it possible to monitor water quality on a round-the-clock basis.
- Use of Larger and Diverse Datasets: An enlarged dataset that will consist of samples of different regions and time will enhance the generalizability of the model.
- Explainable AI (XAI) Approaches: The use of explainable ML strategies can help to learn more about the influence of specific parameters on the quality of water.
- Geospatial and Temporal Modeling: Future directions can include the time-series and GIS-based models to predict the dynamic water quality.
- Decision Support System (DSS): The creation of a web-based DSS to combine predictions using MLs and visualization software to support policymakers and environmental agencies.

6. References

[1] L. Grbčić, I. Lučin, L. Kranjčević, and S. Družeta, "A machine learning-based algorithm for water network contamination source localization," *Sensors,* vol. 20, no. 9, p. 2613, 2020.

[2] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *Int. J. Environ. Sci. Dev,* vol. 9, no. 1, pp. 8–16, 2018.

[3] B. S. Pattnaik, A. S. Pattanayak, S. K. Udgata, and A. K. Panda, "Machine learning based soft sensor model for BOD estimation using intelligence at edge," *Complex & Intelligent Systems,* vol. 7, no. 2, pp. 961–976, 2021.

[4] Z. Liao, N. Zang, X. Wang, C. Li, and Q. Liu, "Machine learning-based prediction of chlorophyll-a variations in receiving reservoir of world's largest water transfer project—a case study in the Miyun reservoir, North China," *Water,* vol. 13, no. 17, p. 2406, 2021.

[5] J. P. Nair and M. Vijaya, "Predictive models for river water quality using machine learning and big data techniques-a Survey," in *2021 International conference on artificial intelligence and smart systems (ICAIS)*, 2021: IEEE, pp. 1747–1753.

[6] H. Guo, J. J. Huang, B. Chen, X. Guo, and V. P. Singh, "A machine learning-based strategy for estimating non-optically active water quality parameters using Sentinel-2 imagery," *International Journal of Remote Sensing,* vol. 42, no. 5, pp. 1841–1866, 2021.

[7] C. Deb, Z. Dai, and A. Schlueter, "A machine learning-based framework for cost-optimal building retrofit," *Applied energy,* vol. 294, p. 116990, 2021.

[8] J. Wang *et al.*, "Ensemble machine-learning-based framework for estimating total nitrogen concentration in water using drone-borne hyperspectral imagery of emergent plants: A case study in an arid oasis, NW China," *Environmental Pollution,* vol. 266, p. 115412, 2020.

[9] X. Yu, J. Shen, and J. Du, "A machine-learning-based model for water quality in coastal waters, taking dissolved oxygen and hypoxia in Chesapeake Bay as an example," *Water Resources Research,* vol. 56, no. 9, p. e2020WR027227, 2020.

[10] B. Chen *et al.*, "Machine learning-based inversion of water quality parameters in typical reach of the urban river by UAV multispectral data," *Ecological Indicators,* vol. 133, p. 108434, 2021.

[11] T. Deng, K.-W. Chau, and H.-F. Duan, "Machine learning based marine water quality prediction for coastal hydro-environment management," *Journal of Environmental Management,* vol. 284, p. 112051, 2021.

[12] S. Kouadri, A. Elbeltagi, A. R. M. T. Islam, and S. Kateb, "Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast)," *Applied Water Science,* vol. 11, no. 12, p. 190, 2021.

[13] H. Lu and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," *Chemosphere,* vol. 249, p. 126169, 2020.

[14] J. Xu *et al.*, "An alternative to laboratory testing: random forest-based water quality prediction framework for inland and nearshore water bodies," *Water,* vol. 13, no. 22, p. 3262, 2021.

[15] L. Wen, X. Ye, and L. Gao, "A new automatic machine learning based hyperparameter optimization for workpiece quality prediction," *Measurement and Control,* vol. 53, no. 7-8, pp. 1088–1098, 2020.