International Journal of Information Systems and Computer Technologies (IJISCT) Vol. 3, No.2
ISSN: 2791-3635(Print), 2791-3643(Online). DOI: https://doi.org/10.58325/ijisct.003.02.0094

Optimized Water Quality Forecasting Using Machine Learning

Suhail Aslam Khaskheli[1], Mushtaque Ahmed Rahu[2], Saima Siraj[3] , Huma Jamshed[4], Shahid Iqbal[*5]

[1]Department of General Faculty SBBU SBA, Pakistan.
[2]Department of Electronic Engineering, QUEST Nawab Shah, Pakistan.
[3]Department of Information Technology, QUEST Nawab Shah, Pakistan.
[4]Department of Computer Science DHA Suffa University Karachi, Pakistan.
[5]Department of Computer Engineering, Bahauddin Zakariya University Multan, Pakistan

Corresponding Author: Shahid Iqbal (shahid.iqbal@bzu.edu.pk)

ABSTRACT
Water is one of most critical assets for a healthy life, covering around 70% of the Earth's surface. However, rapid industrialization and urbanization have led to an alarming rate of water quality degradation, resulting in horrible diseases. Traditional techniques for evaluating water quality are expensive, time-consuming, and require laboratory and statistical tests, making real-time monitoring impractical. Therefore, there is a need for a quicker and more workable solution to address the catastrophic consequences of poor water quality so the current trends like AI, ML and IoT are the solutions. To this end, a particular index called Water Quality Index (WQI) is defined to complete Water Quality Class (WQC) and Water Quality (WQ) and use the WQI. So, in this research work, we explore several supervised machine learning methods to calculate WQI. Suggested approach uses four input parameters: Total Dissolved Solids (TDS), pH, turbidity, and temperature. The most effective algorithms for forecasting the WQI are gradient boosting and polynomial regression with degree of 2 with learning rate of 0.1. However, these algorithms have Mean Absolute Errors (MAE) of 1.8074 and 2.8957. Conversely, a Multi-Layer Perceptron (MLP) classifies WQC most accurately, with accuracy of 0.8796. Proposed approach confirms it's prospective for use in water quality detecting systems by attaining rational accuracy with inadequate set of parameters. Therefore, it is crucial to develop more accurate methods to monitor and cope with WQ to ensure the sustainability of life on Earth.

Keywords: Dataset; Water Quality Assessment; WQ Prediction; ML Algorithms.

1. INTRODUCTION

Although water is the most incredible dynamic natural resource for people, finding clean water is becoming increasingly difficult. Two key reasons for the worldwide water calamity are growth in inhabitants and an increase in temperature. To raise volume of freshwater existing for all mankind drinking, salty water that has been desalinated and treated may come from the sea, ponds, canals and lakes, or underground aquifers. However, water storage, delivery, or quality mismanagement poses significant risks to ecosystems and human health [1]. From the 71% of water covering Earth's surface, only 2.5% or less is suitable for human use. This significantly influences the lack of water in many parts of the world and the rising pollution concentrations in freshwater sources. There is no other way to stop the uncontrollable loss of freshwater supplies than to monitor and maintain as many water bodies as is physically feasible. The twenty-first century saw significant advancements, but because of increased emissions and global warming, it is impossible to monitor WQ globally for unpolluted drinking water in real-time [2]. Water is utmost vital reserve since it is both necessary for the existence of all living forms and continuously at risk of contamination from those same lives. Moreover, water is one of the most versatile and efficient communication mediums. Unfortunately, the water's quality is quickly deteriorating due to increased industrialization. Poor water quality is acknowledged as one of the leading causes of the prevalence of terrible diseases. As per reports, 2.5 billion persons have been ill, and five million have expired due to waterborne diseases in underdeveloped nations. The most prevalent conditions in Pakistan are giardiasis, intestinal worms, gastroenteritis, typhoid, various types of hepatitis, and diarrhea [3]. Due to the presence of numerous pathogens such as viruses, bacteria, and protozoa in the contaminated ground water supply, Pakistan has around 100 million instances of diarrhea each year and 2.5 million deaths from endemic diarrheal illness, Water scarcity not only negatively impacts the nation's Gross Domestic Product (GDP) in form of decreased production in the agricultural, industrial, and service sectors, but it also decreases the population's capacity for creativity, innovation, and production. Every year, water-borne illnesses in Pakistan reduce the country's GDP by almost 0.6 to 1.44%. It becomes an urgent issue, especially in an emerging nation like Pakistan. Since water is the contagious time and medium of a spirit, if water is dirty with malady prompting discarded, WQ is presently predictable by exclusive and time-

overwhelming statistical and laboratory examines that need sample gathering, transference to laboratories, and significant quantity of calculation and time [4]. This method is also reasonably unproductive. Given the terrible effects of water pollution, there must be a faster, less expensive solution. Hence aim of this study is to propose and assess a diverse approach based on supervised ML for an accurate prediction of WQ.

1.1 Contribution of study

The study presents the Available data underwent the first exploration to clean, stabilize, and implement feature selection on WQ processes to find the smallest related subset that permits high accuracy at a reasonable price. This allows for the avoidance of time-consuming and costly lab tests using specific sensors in subsequent analyses of a similar nature. The dataset was used to test a number of supervised ML prediction (regression and classification) methods. Finally, a perfect methodology is projected in a framework of water quality numerical exploration. Gradient Boosting and Polynomial Regression predict WQI with MAE of 1.8074 and 2.8957, respectively, according to thorough testing results, although MLP categorizes WQC superlatively with accuracy of 0.8796.

Rest of this article is arranged as: Literature survey in this field is included in Section 2. Few words about water quality assessment and management in Section 3; we examined dataset and preprocess it in Section 4. We use a range of machine learning (ML) performances to evaluate water quality using only a few factors. Section 5 examines the outcomes of Regression and Classification algorithms in the form of classification precision and error rates. In Section 6, we highlight suggestions and the uniqueness of this research. Lastly, in Section 7, we accomplish the article and offer potential directions for future research.

2. RELATED WORK

This study examines approaches that have been used to report issues with water quality. In studies, old-style lab investigations and statistical investigations are often engaged to support define quality of water, and some analyses use ML methodologies to discover greatest promising explanation for water quality problems comfortably. We were able to realize the Pakistani water quality concern better thanks to indigenous research that used lab analyses. In one of these studies, [5] Authors evaluated the infrastructure for water delivery and WQ in Gilgit-Baltistan (GB), Northern Pakistan, as a portion of a Water and Sanitation Extension Program (WASEP). 26 chaotically chosen rural and urban WASEP schemes, as well as 12 control locations, had their water samples taken. Results demonstrate that compared to 8% of rural and 0% of urban locations, 83% of urban and 94% of rural water trials from WASEP projects complied with World Health Organization (WHO) requirements for drinking water, with no total Escherichia coli colonies. This is consistent with WASEP projects receiving higher water arrangement notches based on engineering. The study in [6] authors Eeploy manual lab analysis, researchers looked at 46 distinct water samples from Orangi town in Karachi and exposed substantial amounts of sulphates and total faecal coliform count. The research study in [7] collected water trials from various sites in Pakistan and found high levels of E. coli caused by industrial and sewage pollution through manual lab analysis. After being made aware of the research on water quality in Pakistan, we examined research using ML approaches in the field of WQ. Another study in [8] Authors examined the performance of three ML algorithms, Extreme Gradient Boosting (XGboost), Deep Neural Network (DNN), and Gradient Boosting Machine (GBM) in assessing groundwater indices in state of Haryana (India). Two water quality indices, the Entropy Water Quality Index (EWQI) and WQI, were used to assess the applicability of these models. The research found that DNN outperformed the other models in predicting both indices, with relatively lower error values. The correlation coefficient (CC), root mean square error (RMSE), Nash-Sutcliffe efficiency (NSE), and index of agreement (d) were determined for EWQI and WQI. Electrical conductivity (EC) was recognized as utmost important input parameter, while pH had most negligible significant impact on predictions of WQI and EWQI using these models. This research outcome is expected to help accurately predict WQI and EWQI of groundwater to determine its portability.

The work in [9] water main failure can be caused by pipe structural collapse, variations in WQ, or arrangement of both. To develop prediction models for key failure of water, research articles on water quality variables and sub-factors since 2000 were reviewed using a systematic procedure. Out of 4598 available articles, 304 were examined for water core disappointment prediction models. Resultant set was additionally selected for sub-factors and factors related to WQ, such as pH, temperature, etc. As a result, 18 relevant research papers were identified and thoroughly analyzed. A summary and discussion of the findings related to water quality, as well as associations with other data like kind of prediction model and prediction, are provided. The author in 10] presented the variety of pollutants that can affect water quality, predicting water quality has received much attention in recent years. By implementing the methods outlined in this article, we can reduce the risks of water contamination. This research discusses the basics of ML models and how they are applied to categorize WQC. Three ML algorithms, like MLP, J48, and Naïve Bayes,

were used for WQC prediction. The study in [11] to estimate the WQI, particular feed-forward neural networks and a mixture of several neural networks were used. They employed 25 distinct metrics for measuring water quality as input. By combining backward eradication and onward choice selected arrangement techniques, they attained MSE and R^2 or R-squared of 0.1200, 0.1158, and 0.9270, 0.9390 correspondingly. However, their method uses 25 parameters, which is too many for cheap actual system assuming a cost of parameter sensors.

The study in [12] To assess the water quality, they utilized k-Nearest Neighbors (kNN), DNN, Support Vector Machines (SVM), and Neural Networks (NN), with DNN having the greatest accuracy (93%). They only looked at three factors: pH, temperature, and turbidity, in accordance with WHO standards, to determine predictable water quality in their study. However, using only three criteria and matching them to established values when estimating water quality is rather partial. The study in [13] Authors used two established ML methods, namely ANN and multivariate linear regression, to estimate chemical oxygen demand (COD) and biochemical oxygen demand (BOD). To forecast COD and BOD, authors employed 4 variables: Total Suspended solids (TSS), pH, temperature, and total solids (TS). The study in [14] Authors used sixteen water quality metrics and ANN with Bayesian regularization to estimate WQI. Author's research formed correlation coefficients of 0.94 and 0.77 between observed and predicted values, respectively. The study in [15] Authors forecasted dissolved oxygen (DO) by means of a Feed forward Neural Network (FNN). To anticipate DO, they employed ten parameters; this could provide difficulties if utilized to calculate real-time WQI with an Internet of Things (IoT) system. whereas the work presented in [16] used models that accounted for around 99.5% of volatility of data; authors have used ANN to predict WQI. They used 23 criteria to estimate WQI, which turned out to be a bit pricey if applied to an IoT system, given the cost of sensors. the study in [17] study examined the various artificial intelligence (A.I.) methods for evaluating the quality of water, such as K-nearest neighbours, support vector machines, deep neural networks, and traditional machine learning methods. Groundwater quality measurements may be made with far more precision using machine learning and deep learning techniques than with more conventional approaches. However, a number of factors will influence the accuracy, including the quality of the training data, the complexity of the water quality measures, and the frequency of monitoring.

Several authors have proposed different structures and methods for managing water. The management of water quality, distribution, storage, and device monitoring represents the next major challenges after seawater filtration. Only after filtration over 40% of pure water goes to waste.

3. WATER QUALITY ASSESSMENT AND MANAGEMENT

Fundamental human rights are to have access to unpolluted water. Pakistan is required to work towards reaching its water and sanitation targets set by United Nations (UN) Sustainable Development Goals (SDGs) for 2030 and Pakistan Vision 2025. The 2018 National Water Policy, which has been authorized, places drinking water as the top priority [18].

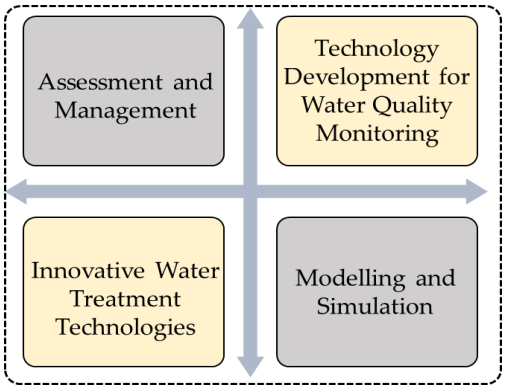


Figure 1. Water Quality Assessment and Management [16]

Monitoring the quality of water for various uses related to human and environmental health, such as drinking, irrigation, and ecological consequences, is one of the key elements of WQ. For WQ development and research, advocacy and policy inputs, proliferation, adaptation, and innovative technology development, support with measuring WQ through facilitation, socialization of innovations and best practices, networking and collaboration towards SDG 6, and sensitization and awareness are the prime motives in research. The majority of studies used manual lab analysis, which

either under or overestimated the WQI criteria, or had numerous parameters to be effectively utilized. Figure 2 illustrates the current technique and how the suggested methodology enhances these ideas.

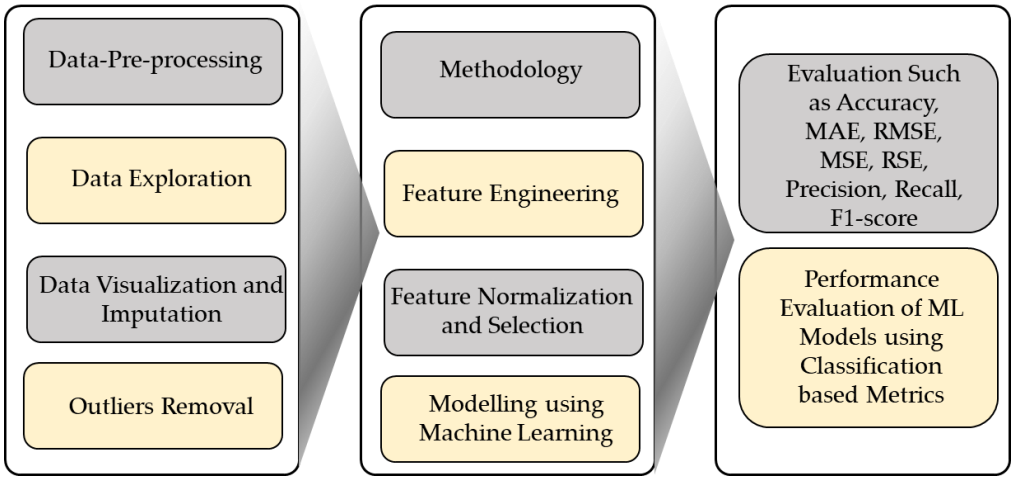


Figure 2. Design Methodology

4. DATA PREPROCESSING

This study made use of the Kaggle dataset, which was then cleaned using the box plot technique discussed in this section. Following data cleaning, the range was changed to 0-100 using q-value normalization, allowing WQI to be calculated using 4 available parameters. Initial values were normalized by z-score after WQI was calculated so that they were all on same scale. A comprehensive process is explained here.

4.1 Collection of Data

For this study dataset was downloaded from: https://www.kaggle.com/datasets/parnianmalekian/water-quality and contained 2000 samples and 11 parameters are given below in Table 1.

Table 1. Parameters as per "WHO" standard limits [17]

Parameter	WHO Limits
Temperature	^{O}C
pH	6.5 – 8.5
TDS	1000mg/L
Turbidity	5 NTU
Nitrite as NO2−	<1 mg/L
Hardness as CaCO3	500 mg/L
Conductance	2000 µS/cm
Chlorides	200 mg/L
Calcium	200 mg/L
Look	Perfect
Alkalinity	500 mg/L

4.2 Outlier Detection and Boxplot Analysis

As the majority of parameters speckled considerably and were on the upper end of the value spectrum, we used box plot analysis for outlier detection. This method provides informative visuals to choose outlier detection threshold values depending on problem domain. We revised an upper cap procedure to filter out outliers after box plot explora-

tion exposed that bulk of parameters were outside box, constructing outliers common. Parameter values that significantly differed from other values were identified, and they were exchanged with concentrated threshold value. Concentrated threshold value was determined based on parameter value that was just below an outlier value. For turbidity, we set a threshold value of 800 for the sample value and applied it to all values above 800, giving them a value of 800. Given our tiny dataset, we physically removed outliers using the same approach across all parameters [15]. We were also very sympathetic when determining an upper threshold of the parameters to avoid distorting the dataset and only slightly penalizing values that were far outside the range and suspect to happen.

4.3 Water Quality Index (WQI)

The only display of WQ is the WQI, which is produced by applying a number of factors that are essentially illustrative of water quality. 9 water quality parameters are often used to create a WQI, but even without all of them, we can still estimate a WQI using approximately six established parameters. In our dataset, there were five different variables: pH, temperature, turbidity, nitrates, and TDS. Although nitrites are claimed to have the same weight and virtual relevance as nitrates in a WQI calculation in many WQI studies [19-20], we also considered them as the sixth parameter. We calculated WQI of every sample using these parameters and their corresponding weights, as shown in Equation (1), where 'value' denotes a parameter's value in a range of 0 to 100, and 'weight' denotes weight of the specific parameter as stated in Table 2. The basic formula for calculating the WQI is to first multiply each parameter's q value by its consistent weight, add them all up, and divide total by sum of weights of used parameters [20-21]

$$\mathrm{WQI} = \frac{\sum \mathrm{q_{value}} \times \mathrm{w_factor}}{\sum \mathrm{w - factor}} \quad (1)$$

Table 2. Parameter weights for calculating WQI [20-21]

Weighing Factor	Weight
Temperature	0.10
pH	0.12
TDS	0.09
Turbidity	0.08
Nitrates	0.12

4.4 Water Quality Class (WQC)

After estimating the WQI, we used the WQI in classification algorithms to identify WQC of every trial [20] as given in Table 3.

Table 3. WQI Ranges [20-21].

WQI Range	Class
0–30	Very Poor
31–50	Poor
51–70	Normal
71–90	Good
91–100	Excellent

4.5 Q-Value Normalization

Values, especially WQ parameters, were normalized using Q-value normalization to adequate them in a range of 0 to 100 for at ease index calculation. Five of these parameters were converted to the 0 to 100 range (20- 21). Due to inaccessibility of q-value ranges for the fifth parameter, nitrites, By applying a set of thresholds, such as allocating

100 if the value is below 1, 80 if the value is below 2, 50 if the value is below 3, and 0 if the value is more than 3, which replicates harsh punishment, we clearly changed them to a 0-100 range following WHO standards. Equation (1) was used to determine the WQI of a dataset when values were Q-normalized and in range of 0-100.

4.6 Z-Score Normalization

A rare data point is one that is either above or below the population means, according to the traditional standardization and normalization approach known as the z-score. The ideal range for a data point is between -3 and +3. To transform data with dissimilar scales to a default scale, dataset is normalized to an aforementioned range. The population mean is subtracted from each data point, and the result is divided by the standard deviation to standardize data using z-score. Equation (2), where x is value of the specific trial, is mean and is standard deviation, provides a score that should fall between -3 and +3, indicating how many standard deviations the point is exceeding or lower mean [19].

$$z_{score} = \frac{(x - \mu)}{\sigma} \quad (2)$$

4.7 Exploration of Data

Once all of data had been processed, different machine learning (ML) procedures was employed to analyze the data and forecast the WQI and WQC using the parameters with the lowest probability. Before utilizing the ML algorithms, some prior steps were carried out to prepare the data for input, such as data separation and correlation analysis.

4.7.1 Correlation Analysis

We used correlation analysis to uncover prospective correlations between parameters in order to categorize reliant variables and predict difficult-to-estimate variables using freely achievable parameters. Pearson correlation, which is the most extensively used and reliable correlation approach, was employed. We used the q-value normalization method to normalize results before applying Pearson correlation to rare values of parameters mentioned in Table 4, as described below.

An indicator of chemical acidity or alkalinity is the pH ion concentration; a range of 6 to 9 is regarded as typical. Hardness (CaCO3) and calcium (Ca) are closely related to alkalinity (Alk), and vice versa. Water's capacity to purify itself is determined by its oxidation-reduction potential. Higher temperature affects oxygen levels in the water. Alkalinity and Ca have strong relationships with hardness, while pH has Just a weak relationship. Conductance is extremely interrelated with TDS and chlorides and lightly interrelated with calcium and temperature. Using water with more nitrate content may increase chance of emerging colon cancer. Nitrate may make other mixtures more likely to reason cancer or alteration into body-produced cancer-causing molecules. Turbidity, which is defined as cloudiness or opacity and should not be less than 1NTU. Alkalinity and hardness are strongly connected with Calcium, whereas TDS, Chlorides, conductance, and pH are only weakly correlated. Conductivity is the capability to conduct an electrical current; it should be nearby 200 Micromhos/cm. TDS has a strong relationship with conductance and chlorides and only a weak relationship with temperature and Calcium. Chlorides are weakly associated with temperature and Ca, strongly correlated with conductance and TDS. Our prediction parameter, WQI, is correlated with several factors, including total dissolved solids, pH, hardness as CaCO3, temperature, turbidity, and conductance count, as shown by the list of correlation analysis observations. To keep the system's cost low, we need to select the fewest possible parameters to predict WQI. pH, turbidity, and temperature are three characteristics that have widely available and inexpensive sensors and contribute significantly to the WQI, making them natural choices. TDS is a useful statistic that is also readily available and correlates with conductance count; therefore, by selecting TDS, we may disregard the other two factors. The final quality, CaCO3's hardness, is not included because it is similarly rare and difficult to acquire. We chose four criteria for predicting WQI: temperature, turbidity, pH, and TDS, to wrap up the correlation study. Initially, given their cheap cost, we only considered the first three variables. TDS will be added if required afterward, and its effect on accuracy will be examined.

Table 4. Chart for Correlation Analysis

	Temp	pH	TDS	Turb	NO2	Caco3	Cond	Cl	Ca	Alk	WQI
Temp	1.000	0.006	0.261	0.102	-0.162	-0.243	0.261	0.284	-0.161	-0.182	**-0.463**
pH	0.003	1.000	-0.059	-0.079	0.169	-0.278	-0.068	-0.152	-0.239	-0.180	**-0.437**
TDS	0.276	-0.059	1.000	0.047	0.090	0.067	0.968	0.754	0.214	0.019	**-0.388**
Turb	0.127	-0.0768	0.042	1.000	0.0002	-0.146	0.049	0.039	-0.129	-0.098	**-0.357**
NO2	-0.166	0.159	0.089	0.0005	1.000	0.078	0.100	0.035	0.129	0.049	-0.211
CaCo 3	-0.276	-0.280	0.063	-0.146	0.079	1.000	0.067	0.137	0.637	0.468	0.368
Cond	0.253	-0.064	0.0968	0.049	0.100	0.064	1.00	0.784	0.229	0.019	-0.375
Cl	0.289	-0.149	0.758	0.039	0.039	0.139	0.784	1.00	0.268	0.069	-0.279
Ca	-0.166	-0.240	0.219	-0.122	0.124	0.639	0.231	0.267	1.00	0.495	0.178
Alk	-0.188	-0.169	0.016	-0.092	0.048	0.467	0.020	0.068	0.449	1.00	0.227
WQI	-0.479	-0.438	-0.385	-0.0359	-0.211	0.364	-0.376	-0.279	0.190	0.229	1.00

4.7.2 Data Splitting–Cross Validation

The final stage before using the ML model is to split the supplied data into specific subsets to train model, test it, and determine accuracy metrics to assess the model's effectiveness. We looked into the cross-validation data splitting strategy in this research. For cross-validation, data is divided into k subsets, with k-1 subsets serving as training dataset and one subset as testing dataset. It guarantees use of true and reliable data for testing and training. While this approach can be computationally costly with repetitions, it is suitable for WQ datasets as they are typically short. We performed cross-validation after segmenting data into k = 6 subsets. With 663 samples in the entire training set, every fold subset, including a test set, had approximately a hundred samples.

4.7.3 Machine Learning Algorithms

Both classification and regression techniques were applied. We classified samples into predetermined WQC using classification algorithms and used regression procedures to predict WQI. The following supervised machine learning algorithms were employed in our investigation for both classification and regression: these are shown in Figure 3.

A. Support Vector Machines (SVM)

SVMs are typically employed for classification, but they can also be used for regression. SVMs use data points plotted on the plane to imagine data, describe hyper plane between classes, and spread sideline to exploit difference between two classes, leading to littler scheming inaccuracies [23-24].

B. Random Forest

The random forest (RF) model applies different base models to distinct subsets of supplied data and bases its conclusions on each model. The RF model, which combines its benefits with an effectiveness of integrating numerous models, is built on decision trees [25-26].

C. Decision Trees

Decision trees are a straightforward method that can be utilized to address classification and regression problems. After training, the values of all the pertinent input parameters form the basis for decision tree's findings. The root variable is determined by entropy, and then the other parameter values are considered. It projects the choice depending on multiple parameter decisions and is structured in a top-down tree with all the parameter decisions [27-28].

D. K-Nearest Neighbor (KNN)

KNN algorithm sort's data by finding provided points' closest N neighbors and classifying them according to the majority of those neighbors. If there is a tie, there are several ways to settle it, like growing N or bias addition in favour of one class. Since total handling happens during testing and KNN recapitulates through the whole training

dataset to compute nearest neighbors every time, it is not advised for large datasets [29-30]. For our model, we chose an N of 5.

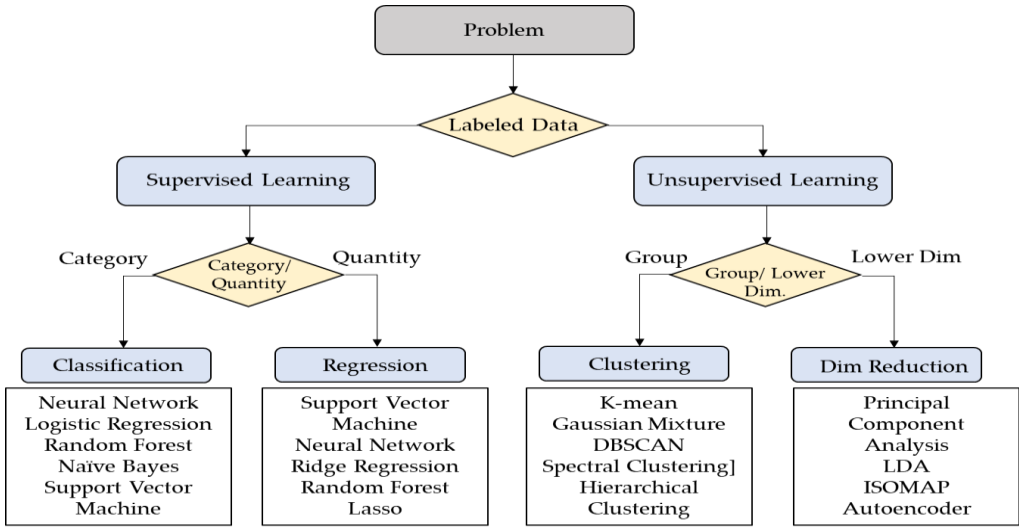


Figure 3. Machine learning Algorithms [22]

E. Gaussian Naïve Bayes

Gaussian Naïve Bayes is extension of naïve Bayes, the assumption that the likelihood of the occurrence of one feature is different from the likelihood of the occurrence of another feature underlies the use of the Bayes theorem in the simple and speedy method known as naive Bayes, Although there are different functions used to guess data distribution, Gaussian or normal distribution is most straightforward to implement because we have simply need to figure out mean and standard deviation for training data [31–32].

F. Gradient Boosting Algorithm

The majority of contests employs a Gradient Boosting Algorithm and is a most recent. It makes use of an additive model that enables differentiable loss function optimization. We applied it with the learning rate of 0.1, the min samples fragmented of 2, and the loss function of ls [33–34].

G. Logistic Regression

Logistic regression is an algorithm for classifying data. Its foundation is a sigmoid or logistic function, hence the name. While it is an utmost famous method for binary classification, we chose MLR in our situation because there were more than two classes [35]. It was utilized with a "caution" solver and the L2 penalty.

H. Polynomial Regression

Polynomial regression is used when there is a nonlinear and slightly complex correlation between input and output variables. In order to effectively depict nonlinear relationship between input and output variables, higher order variables are used. We used a number of two for the degree of the polynomial equation. Equation (3) shows the relationship between y, the output value predicted by ML, x, the observed value, Beta, an appropriate value, i, No. of parameters reflected, k, order of polynomial equation, and i, an error term or residuals of ith predictor. However, using higher order variables can lead to over fitting. Therefore, we used degree 2 polynomials with an order of C.

$$y = \beta_0 + \beta_1 x_i + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_k x^k + \in_i, \; for \; i = 1, 2 \ldots n \qquad (3)$$

I. Multiple Linear Regression (MLR)

When dealing with multiple predictor variables, MLR is the type of linear regression that is used. Equation (4), where y is the outcome for which ML is used to predict value, x is an observed value, b is slope on detected value, and e is

error term, illustrates how we employ MLR to estimate the impact of each variable on the outcome when there are multiple input variables [37].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \in \quad (4)$$

J. Multi-Layer Perceptron (MLP)

Structure of neurons works as loose model for neural networks. There are numerous layers and nodes that connect between them. Hidden layers between input and output layers are required. An output layer displays predictions based on an input layer's input of the predicting parameters. By iteratively going over each training data point and adjusting weights on every node of every layer, a model is simplified. The trained model then makes decisions about which units to stimulate depending on an input using those weights. A common neural network model, an MLP, is mostly utilized for classification but can also be used for regression [38]. We utilized it for classification with the configuration (3, 7) for almost 200 iterations, utilizing the 'lbfgs' solver. Except for MLP, we utilized default values for the algorithms.

5 RESULTS AND DISCUSSIONS

5.1 Accuracy Measures

In this section, we describe several metrics used to estimate an accuracy of ML algorithms used in this study, as previously noted, including classification and regression. The outputs of both types of algorithms were calculated differently, and we used following metrics for regression.

5.1.2 Mean Absolute Error (MAE)

Regression accuracy was measured using the MAE metric. Absolute values of the errors were summed and divided by total number of values, equally considering each mistake value. Equation (5) displays the formula for computing MAE, where xobs represents an actual value, xpred represents predicted value, and 'n' represents total number of trials taken into consideration [39].

$$\text{MAE} = \frac{\sum(||x_{obs} - x_{pred}\,||)}{n} \quad (5)$$

5.1.3 Root Mean Squared Error (RMSE)

Additionally, we used RMSE metric to scale the MSE values to be closer to the range of observed values. RMSE is simply square root of MSE. It provides a more comprehensive assessment of the model's accuracy by considering the magnitude of the errors. It is calculated using Equation (6) [40].

$$\text{RMSE} = \sqrt{\frac{\sum(x_{obs} - x_{pred}\,)^2}{n}} \quad (6)$$

5.1.4 Mean Square Error (MSE)

The MSE is calculated by dividing sum of squared errors by a total number of predicted values. This method gives larger errors more significance, which is particularly useful in situations where a heavier weight for larger faults is required. Equation (7) is used to calculate it [41].

$$\text{MSE} = \frac{\sum(x_{obs} - x_{pred}\,)^2}{n} \quad (7)$$

5.1.5 R Squared Error (RSE)

A Coefficient of determination, commonly known as R^2 or R-squared, is used to evaluate how well model fits the data. This specifically measures proportion of a dependent variable's variance that can be explained by an independent variable, as given in Equation (8). Higher R-squared values indicate that an independent variable(s) significantly accounts for variance in dependent variable [42].

$$\text{RSE or } R^2 = 1 - \frac{\text{Explained variation}}{\text{Total variation}} \quad (8)$$

5.1.6 Metrics Utilized for Classification

Various metrics including accuracy, precision, F1 score and Recall have been utilized for classification. These are briefly discussed below.

i. Accuracy

The accuracy of model is measured by the number of observed values that it correctly predicted. Equation (9) measures accuracy, where TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative [43].

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (9)$$

ii. Precision

Precision is proportion of correctly classified positive examples of given class out of all positive occurrences of that class. Formula used to determine precision is Equation (10), where TP stands for true positive and FP stands for false positive [29].

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (10)$$

iii. F1 Score

As recall and precision do not capture the complete picture of accuracy on their own, we used their harmonic mean to calculate F1 score, as given in Equation (11). This score takes into account both characteristics and provides a more accurate representation of the overall accuracy. It has a range of 0 to 1, and the accuracy improves with a higher score [29].

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

iv. Recall

Recall is a proportion of examples in a given positive class that were correctly categorized. A recall is computed using Equation (12), where TP stands for true positive and FN for false negative [29].

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (12)$$

v. Regression Algorithm Results

Due to the high cost of water quality parameter sensors, this research attempted to predict water quality using a small number of features. Our initial analysis considered four variables: TDS, pH, turbidity, and temperature. As shown in Table 5 and Figure 4, when using the regression algorithm, we initiate gradient boosting to be best active technique, with MAE of 1.8074, MSE of 2.6549, RMSE of 7.1028, and R^2 or R-squared score of 1.8657.

Table 5. Results of four-parameter regression.

Algorithm	R Squared	RMSE	MSE	MAE
SVM	0.3569	3.2510	11.7456	3.5264
Random Forest	0.6818	3.1821	10.6548	2.4042
Logistic Regression	0.7062	3.5275	10.6470	2.7220
Polynomial Regression	0.7029	2.7283	7.8549	2.1142
Gradient Boosting	**1.8657**	**7.1028**	**2.6549**	**1.8074**

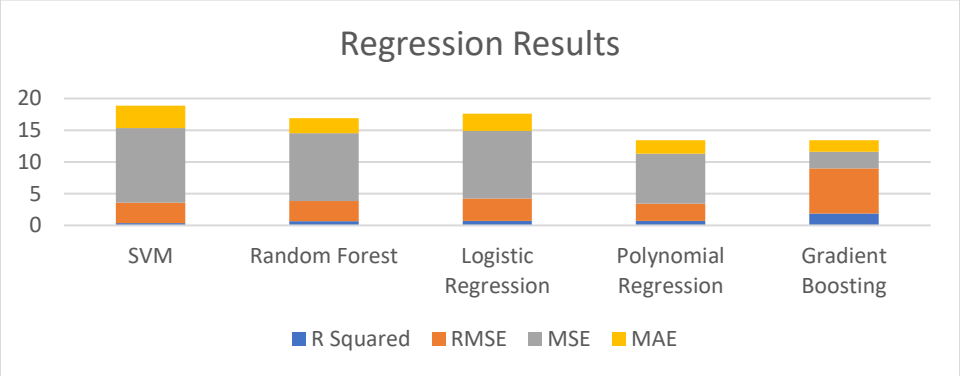


Figure 4. Regression results using four parameters

Then, in an effort to reduce the number of parameters, we decided not to include TDS as it is more difficult to obtain than other parameters. Although logistic regression and gradient boosting had better RSE values, with 0.6213 and 0.6014 correspondingly, as depicted in Table 6 and Figure 5, we found that polynomial regression was utmost effective algorithm, with MAE of 2.8957, MSE of 13.9742, and RMSE of 3.9812. Although overall error rate increased, considering cost, the increase was not concerning and the performance was satisfactory within acceptable limits.

Table 6. Results of regression with three parameters.

Algorithm	R Squared	RMSE	MSE	MAE
SVM	0.2634	3.8964	13.9437	2.7069
Random Forest	0.5017	3.8139	14.9876	3.1421
Logistic Regression	**0.6213**	3.8761	14.9457	3.2167
Polynomial Regression	0.4796	**3.9812**	**13.9742**	**2.8957**
Gradient Boosting	0.6014	3.5497	12.9851	2.7162

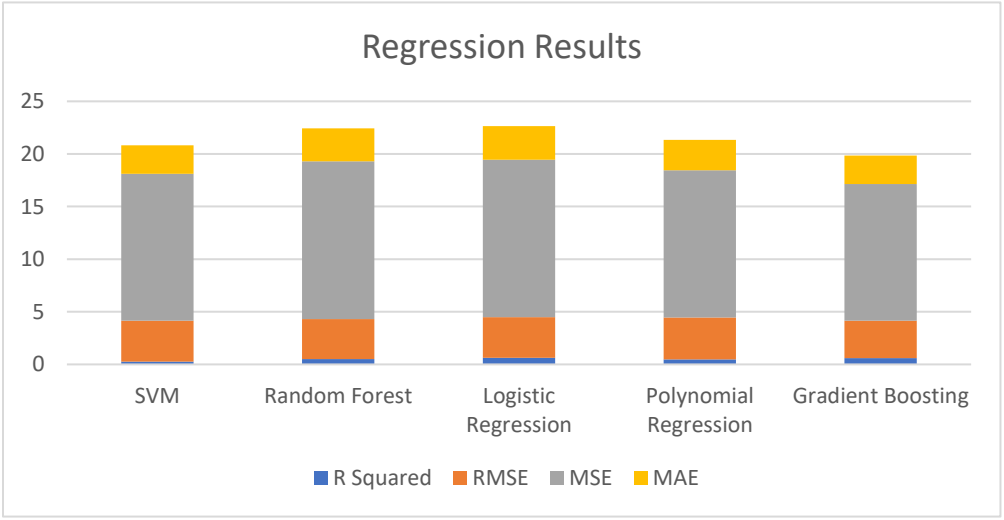


Figure 5. Regression results using three parameters.

5.2 Results for different Classifications of Algorithms

We used classification algorithms to predict WQC based on the previously determined WQI of samples, using the same four criteria as in the previous section. As shown in Table 7 and Figure 6, we found that MLP outperformed other algorithms in this scenario, with an accuracy of 0.8796, precision of 0.5589, recall of 0.5788, and F1 score of 0.5584.

Table 7. Results for different classifications of Algorithms using four parameters.

Algorithm	Accuracy	Precision	Recall	F1 Score
SVM	0.7887	0.5378	0.5237	0.5043
RF	0.78610	0.5163	0.5211	0.5136
Logistic Regression	0.8201	0.5430	0.5504	0.5288
Decision Tree	0.7894	0.5189	0.5055	0.5274
KNN	0.7182	0.4943	0.4698	0.4865
Gaussian Naïve Bayes	0.7953	0.4846	0.5382	0.5152
Gradient Boosting Classifier	0.8061	0.5400	0.5283	0.5068
Multi-Layer Perceptron	**0.8796**	**0.5589**	**0.5788**	**0.5584**

In this section, we reviewed the findings of our investigation and concluded that MLP was more effective in predicting WQC than WQI, while gradient boosting and polynomial regression were more effective in predicting WQI.

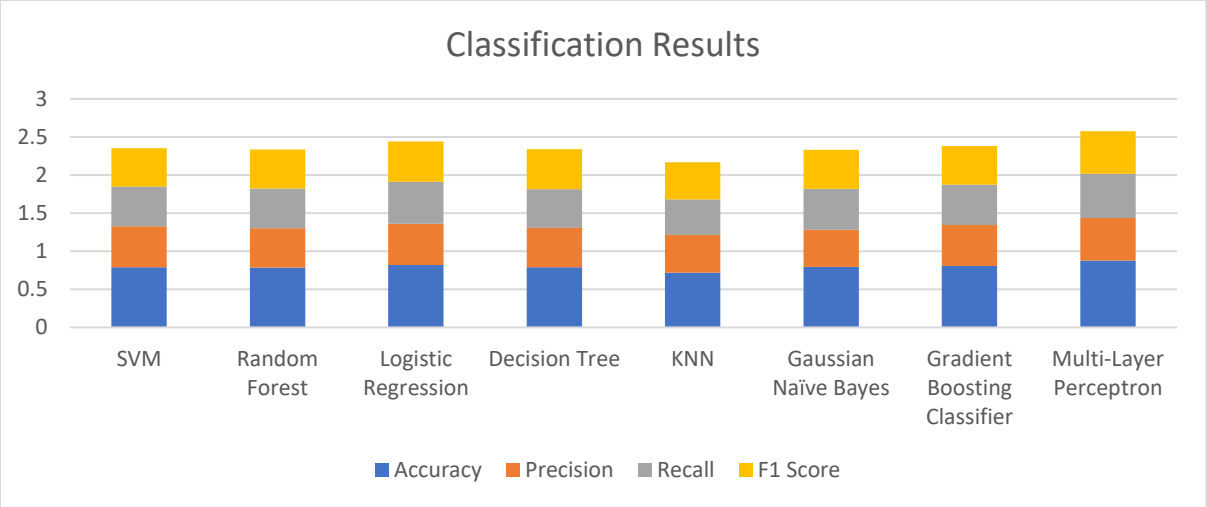


Figure 6. Classification results using four parameters

5.3 Comparison with Existing work

We have compared this study with existing studies this is quite different work from existing studies, different algorithms and models have been used by different authors. This study performed well as compare to existing models [46-47].

6 DISCUSSION

WQ metrics, obtained through laborious lab examination, have traditionally been used to calculate water quality. We reviewed various ML techniques to evaluate them and examined numerous studies that used more than ten WQ measures to predict WQI. However, these studies used up to twenty-five input parameters, making them inappropriate for low-cost real-time systems. In contrast, our method used only four WQ factors to predict WQC with an accuracy of 87% and an MAE of 1.80 for WQI. While prior studies have used ML, our results provide a foundation for a low-cost WQ detection system that uses fewer parameters and can be used in real-time systems.

7 CONCLUSIONS AND FUTURE WORK

Water is one of leading necessary assets for existence, and WQ is often tested by complete and time-consuming lab analysis. To predict water quality, this study looked into a different method of ML that relies on simple and easily accessible water quality criteria. The dataset obtained from Kaggle consisted of 2000 samples and 11 parameters. The selection of illustrative supervised ML techniques was engaged to evaluate WQI. This exposed that MLP with the configuration of (3, 7) outperformed other classification algorithms by most accurately classifying WQC, while polynomial regression with degree of 2 and gradient boosting, with learning rate of 0.1, outperformed other regression algorithms by predicting WQI much more accurately. When using the regression algorithm, we initiate gradient boosting to be best active technique, with MAE of 1.8074, MSE of 2.6549, RMSE of 7.1028, and R^2 or R-squared score of 0.8657. We found that polynomial regression was utmost effective algorithm, with MAE of 2.8957, MSE of 13.9742, and RMSE of 3.9812. We found that MLP outperformed other algorithms in this scenario, with an accuracy of 0.8796, precision of 0.5589, recall of 0.5788, and F1 score of 0.5584. We highly recommend implementing this study's conclusions into significant IoT-based online monitoring system or protocol and platform-based system that solely uses sensors for requisite metrics in subsequent efforts. Tested algorithms would provide a quick prediction of WQ based on real-time data provided by an IoT device. pH, turbidity, temperature, and TDS parameter sensors would be used in proposed Internet of Things system, and values from those sensors would be communicated using an Arduino microcontroller and ZigBee transceiver. Before water is made available for ingestion, contaminated water would be detected, and appropriate authorities would be notified. Hope that fewer people will consume low-WQ, which will lessen the severity of terrible diseases such that typhoid and diarrhea. So like that, development of impending tools to assist policy and decision-makers would result from the use of prescriptive analysis based on expected values.

Conflicts of Interest: Authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] A. Manna and D. Biswas, "Assessment of Drinking Water Quality Using Water Quality Index : A Review," Water Conserv. Sci. Eng., vol. 0, pp. 23–25, 2023, doi: 10.1007/s41101-023-00185-0.
[2] K. V Modi, H. K. Jani, and I. D. Gamit, "Impact of orientation and water depth on productivity of single - basin dual - slope solar still with - Al2O3 and CuO nanoparticles," J. Therm. Anal. Calorim., vol. 143, no. 2, pp. 899–913, 2021, doi: 10.1007/s10973-020-09351-1.
[3] Pakistan Council of Research in Water Resources (PCRWR) Islamabad: Water quality reports. ; Available online: https://pcrwr.gov.pk/water-quality-reports/.
[4] D. Xu, K. Raza, K. Hussain, A. Albaker, A. I. Almulhim, and R. Alvarado, "Analyzing the factors contribute to achieving sustainable development goals in Pakistan : A novel policy framework," Energy Strateg. Rev., vol. 45, no. July 2022, p. 101050, 2023, doi: 10.1016/j.esr.2022.101050.
[5] A. Shah, "Assessing water and water infrastructure quality in community-managed water supply systems in northern Pakistan," Water Practice and Tehnology, vol. 17, no. 5, pp. 1046–1057, 2022, doi: 10.2166/wpt.2022.053.
[6] M. K. Daud et al., "Drinking Water Quality Status and Contamination in Pakistan," vol. 2017.
[7] H. Raheja, "Prediction of groundwater quality indices using machine learning algorithms," Water practice & technologt, vol. 17, no. 1, pp. 336–351, 2022, doi: 10.2166/wpt.2021.120.
[8] M. M. Nojumi and A. Bayat, "A review of water quality factors in water main failure prediction models," Water practice & Technology, vol. 17, no. 1, 2022, doi: 10.2166/wpt.2021.094.
[9] S. Y. Abuzir and Y. S. Abuzir, "Machine learning for water quality classi fi cation," Water quality research journal, vol. 57, no. 3, pp. 152–164, 2022, doi: 10.2166/wqrj.2022.004.
[10] Z. Ahmad, N. A. Rahim, A. Bahadori, and J. Zhang, "Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks Improving water quality index prediction in Perak

River basin Malaysia through a combination of multiple neural networks," Intl. J. River basin management, vol. 5124, 2017, doi: 10.1080/15715124.2016.1256297.
[11] M. M. Rahman, T. Haque, A. Mahmud, and A. Amin, "Drinking water quality assessment based on index values incorporating WHO guidelines and Bangladesh standards," Phys. Chem. Earth, vol. 129, no. August 2022, p. 103353, 2023, doi: 10.1016/j.pce.2022.103353.
[12] H. Z. Abyaneh, "Environmental Health Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters," Journal of environmental health
Science & Engineering, pp. 6–13, 2014.
[13] M. Sakizadeh, "Artificial intelligence for the prediction of water quality index in groundwater systems," Model. Earth Syst. Environ., vol. 2, no. 1, pp. 1–9, 2016, doi: 10.1007/s40808-015-0063-9.
[14] V. Ranković, J. Radulović, I. Radojević, A. Ostojić, and L. Čomić, "Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia," Ecol. Modell., vol. 221, no. 8, pp. 1239–1244, 2010, doi: 10.1016/j.ecolmodel.2009.12.023.
[15] N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, and M. F. Ramli, "Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors," Mar. Pollut. Bull., vol. 64, no. 11, pp. 2409–2420, 2012, doi: 10.1016/j.marpolbul.2012.08.005.
[16] A. K. Makarigakis and B. E. Jimenez-cisneros, "Water Challenges," Water, 2019, doi: 10.3390/w11020388.
[17] C. O. Unigwe and J. C. Egbueri, "Drinking water quality assessment based on statistical analysis and three water quality indices (MWQI , IWQI and EWQI): a case study," Environ. Dev. Sustain., vol. 25, no. 1, pp. 686–707, 2023, doi: 10.1007/s10668-021-02076-7.
[18] G. Marinho, D. F. Campos, A. Teixeira, M. Tremblay, E. M. Mendiondo, and F. Ghiglieno, "Advances in Technological Research for Online and In Situ Water Quality Monitoring — A Review," Sustainability, pp. 1–28, 2022.
[19] D. A. Method, "Assessment of Surface Water Quality Using Water Quality," Water, pp. 1–14, 2023.
[20] U. Alexandru and I. Cuza, "water quality index an instrument for water resources water quality index an instrument for water," water, 2014, doi: 10.13140/2.1.3736.3203.
[21] Yuan W, Liu Q, Song S, Lu Y, Yang S, Fang Z, Shi Z. A climate-water quality assessment framework for quantifying the contributions of climate change and human activities to water quality variations. Journal of Environmental Management. 2023 May 1;333:117441.
[22] S. Sahour, M. Khanbeyki, V. Gholami, H. Sahour, I. Kahvazade, and H. Karimi, "Evaluation of machine learning algorithms for groundwater quality modeling," Environ. Sci. Pollut. Res., no. 0123456789, 2023, doi: 10.1007/s11356-023-25596-3.
[23] P. Ditthakit, S. Pinthong, N. Salaeh, J. Weekaew, and T. Thanh, "Comparative study of machine learning methods and GR2M model for monthly runoff prediction," Ain Shams Eng. J., vol. 14, no. 4, p. 101941, 2023, doi: 10.1016/j.asej.2022.101941.
[24] M. Gupta, S. K. B. V, B. Kavyashree, H. R. Narapureddy, N. Surapaneni, and K. Varma, "Various Crop Yield Prediction Techniques Using Machine Learning Algorithms," Conference on smart structures and systems IEEE, pp. 273–279, 2022.
[25] F. Ghobadi, "Application of Machine Learning in Water Resources Management : A Systematic Literature Review," Water, 14(4),2023.
[26] O. K. Pal, "The Quality of Drinkable Water using Machine Learning Techniques," Int. J. Adv. Res Sci. Eng. no. October, 2022, doi: 10.22161/ijaers.96.2.
[27] B. Jeong, M. Renee, M. Kim, J. Kim, and J. Shin, "Uncorrected Proof Machine learning-based prediction of harmful algal blooms in water supply reservoirs Uncorrected Proof," water quality research journal, vol. 00, no. 0, pp. 1–15, 2022, doi: 10.2166/wqrj.2022.019.
[28] G. Uddin, S. Nash, A. Rahman, and A. I. Olbert, "A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches," Water Res., vol. 229, no. July 2022, p. 119422, 2023, doi: 10.1016/j.watres.2022.119422.
[29] N. Yahyaoui, M. Neji, M. Kallel, A. Wali, and S. Hajji, "A comparative approach of ML algorithms to rank irrigation water quality : case of Oriental - Coast shallow aquifer in Cap - Bon , northeastern of Tunisia," Model. Earth Syst. Environ., no. 0123456789, 2023, doi: 10.1007/s40808-023-01694-6
[30] Alamgir A, Khan MA, Hany OE, Shaukat SS, Mehmood K, Ahmed A, Ali S, Riaz K, Abidi H, Ahmed S, Ghori M. Public health quality of drinking water supply in Orangi town, Karachi, Pakistan. Bulletin of Environment, Pharmacology and Life Sciences. 2015 Oct 11;4(11):88-94.
[31] B. Das, S. M. Ali, M. Z. Shaikh, A. F. Chandio, M. A. Rahu, J. K. Pabani, and M. R. Khalil. "Linear Regression Based Crop Suggestive System for Local Pakistani Farmers." Global Conference on Wireless and Optical

Technologies (GCWOT), 2023, pp. 1-6. IEEE
[32] A. A. Mirani, M. S. Memon, M. A. Rahu, M. N. Bhatti, and U. R. Shaikh. "A review of agro-industry in IoT: applications and challenges." Quaid-E-Awam University Research Journal of Engineering, Science & Technology, Nawabshah. 17, no. 01, 2019, 28-33.
[33] M. A. Rahu, S. Karim, R. Shams, A. A. Soomro, and A. F. Chandio. "Wireless Sensor Networks-based Smart Agriculture: Sensing Technologies, Application, and Future Directions." Sukkur IBA Journal of Emerging Technologies Vol 5, no. 2, 2022, pp: 13-32.
[34] M. A. Rahu, P. Kumar, S. Karim, and A. A. Mirani. "Agricultural Environmental Monitoring: A WSN Perspective." University of Sindh Journal of Information and Communication Technology 2, no. 1, 2018, pp: 17-24.
[35] M. A. Rahu, "Energy Harvesting for Water Quality Monitoring using Floating Sensor Networks: A Generic Framework." Sukkur IBA Journal of Emerging Technologies 1, no. 2, 2018, pp: 19-32.
[36] A. A. Mirani, M. S. Memon, R. Chohan, I. N. Sodhar, and M. A. Rahu. "Irrigation scheduling, water pollution monitoring in IoT: A Review." 2021.
[37] M. A. Rahu, A. F. Chandio, K. Aurangzeb, S. Karim, M. Alhussein, and M. S. Anwar. "Towards design of Internet of Things and machine learning-enabled frameworks for analysis and prediction of water quality." IEEE Access, 2023.
[38] A. A. Mirani, M. S. Memon, M. N. Bhati, M. A. Soomro, and M. A. Rahu. "Taxonomy of ubiquitous computing: Applications and challenges." in 2017 International Conference on Information and Communication Technologies (ICICT), pp. 202-208. IEEE, 2017.
[39] G. M. Jatoi, M. A. Rahu, S. Karim, S. M. Ali, and N. D. Sohu, "Water quality monitoring in agriculture: Applications, challenges and future prospectus with IoT and machine learning," . Appl. Eng. Technol., vol. 7, no. 2, pp. 46-54, 2023.
[40] M. A. Rahu, M. M. Shaikh, S. Karim, A. F. Chandio, S. A. Dahri, S. A. Soomro, and S. M. Ali, "An IoT and machine learning solutions for monitoring agricultural water quality: A robust framework," Mehran Univ. Res. J. Eng. Technol., vol. 43, no. 1, pp. 192-205, 2024.
[41] M. A. Rahu, M. M. Shaikh, S. Karim, et al., "Water quality monitoring and assessment for efficient water resource management through Internet of Things and machine learning approaches for agricultural irrigation," Water Resour. Manag., 2024. [Online]. Available: https://doi.org/10.1007/s11269-024-03899-5. Received: Mar. 22, 2024; Accepted: May 20, 2024; Published: Jun. 3, 2024.
[42] S. M. A. Shah, A. Soomro, K. Hussain, M. A. Rahu, and S. Karim, "Innovative machine learning solutions: Investigating algorithms and applications," J. Appl. Eng. Technol., vol. 8, no. 1, pp. 32-51, 2024.
[43] X. Wang, Y. Li, Q. Qiao, A. Tavares, and Y. Liang, ‘‘Water quality prediction based on machine learning and comprehensive weighting methods,’’ Entropy, vol. 25, no. 8, p. 1186, Aug. 2023.
[44] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, ‘‘Efficient water quality prediction using supervised machine learning,’’ Water, vol. 11, no. 11, p. 2210, 2019.
[45] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, ‘‘Water quality prediction using artificial intelligence algorithms,’’ Appl. Bionics Biomech., vol. 2020, pp. 1–12, Dec. 2020.
[46] R. Gai and H. Zhang, ‘‘Prediction model of agricultural water quality based on optimized logistic regression algorithm,’’ EURASIP J. Adv. Signal Process., vol. 2023, no. 1, p. 21, Feb. 2023.
[47] P. Jayaraman, K. K. Nagarajan, P. Partheeban, and V. Krishnamurthy, "Critical review on water quality analysis using IoT and machine learning models," Int. J. Inf. Manag. Data Insights, vol. 4, no. 1, p. 100210, 2024.