Tellus A
Dynamic Meteorology and Oceanography

Comparative Analysis of Machine Learning Algorithms for Water Quality Prediction

ORIGINAL RESEARCH PAPER

MUHAMMAD AKHLAQ
ASAD ELLAHI
RIZWAN NIAZ
MOHSIN KHAN
SAAD Sh. SAMMEN
MIKLAS SCHOLZ

*Author affiliations can be found in the back matter of this article

STOCKHOLM UNIVERSITY PRESS

ABSTRACT

This study aims to identify the influential parameters and heavy metals in water and assess the water quality classification at the Alpine glacial lakes and rivers in three districts of Pakistan. For this purpose, nine water quality parameters (Cd, Cr, Pb, Ni, Fe, As, and TDS) in mg/L, pH, Ec µS/Cm are used to compute the Water Quality Index (WQI). The Boruta approach was utilized for the identification of influential parameters associated with the water quality classes. Moreover, we employed supervised machine learning models, including a decision tree, the k-nearest neighbor method, a neural network model (multi-layer perception), a support vector machine, and a random forest, to predict and validate the water quality class**.** The performance of all algorithms is assessed by an accuracy metric. The accuracy rates for the validation set were observed to be 83% for the decision tree model, 75% for the K-nearest neighbor method, 83% for the neural network, 88% for the support vector machine, and 88% for the random forest model. Water quality assessments for observed locations specify significant insights, revealing that 49% of the locations exhibit low water quality. According to the current study, the government should address problems with water quality in Pakistan's impacted areas by implementing suitable measures designed water monitoring systems and innovative technologies.

CORRESPONDING AUTHOR:
Rizwan Niaz

Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan; Department of Statistics, Kohsar University Murree, Murree, Pakistan

razwanniaz11@gmail.com; rizwanniaz@stat.qau.edu.pk

KEYWORDS:
Heavy metals; glacial lakes; river contamination; Boruta algorithm; supervised machine learning

TO CITE THIS ARTICLE:
Akhlaq, M., Ellahi, A., Niaz, R., Khan, M., Sammen, S.Sh. and Scholz, M. (2024) Comparative Analysis of Machine Learning Algorithms for Water Quality Prediction. *Tellus A: Dynamic Meteorology and Oceanography*, 76(1): 177–192. DOI: https://doi.org/10.16993/tellusa.4069

1. INTRODUCTION

Aquatic ecosystems receive multiple pollutants from diffuse and point sources and act as sinks for various pollutants. Any type of river water contamination, whether direct or indirect, could be very dangerous to people's health because they are the final consumers. Water is necessary for the metabolism of cells and correct functioning; thus, it can cause a variety of disorders in humans (Alam et al., 2017). According to Khanna and Ishaq (2013), dangerous wastewater that is discharged into rivers is a major global problem because it has a negative effect on aquatic life. The quality of water reflects environmental disturbances in their surrounding regions through the water quality (Khan et al., 2022). Due to climate change, increasing urbanization, food demand, and unrestrained use of natural resources, almost 40% of the world's population is currently experiencing water scarcity (Scanlon et al., 2023). Recently observed is a significant increase in the release of polluted wastewater into the environment due to increased urbanization, industrialization, agricultural activity, geothermal water release, and olive wastewater discharge, particularly in places where olives are grown (Wijerathna et al., 2023). According to the previous reports, approximately 80% of the diseases reported in developing countries are caused by water pollution, since approximately 4 million deaths and an estimated 2.5 billion illnesses are reported from water pollution (Ahmed et al., 2019; Rafiee, 2018). Notable diseases in Pakistan are observed, including diarrhea, typhoid, gastroenteritis, cryptosporidium infections, hepatitis, and giardiasis intestinal worms (Zhao et al., 2024; Noureen et al., 2022), which affect the country's GDP in the range of 0.6 to 1.44% (Jabbar, 2020) and become an urgent and vital issue in Pakistan. Since water contamination affects different regions and countries (Sharma et al., 2022), almost a billion individuals have faces lake access to safe drinking water, resulting in two million deaths occurring annually because of water contamination, poor sanitation, and unhygienic conditions (Kumar et al., 2022).

The presence of nitrates and heavy metals in water captured the attention of many water quality researchers worldwide (Unigwe & Egbueri, 2023). Water covers approximately 70% of the Earth's surface and plays an important role in human survival, such as being used for domestic, agriculture, and industrial activities (Khan et al., 2023). However, water quality is reduced due to heavy metal contamination (Balali-Mood et al., 2021; Rafiee, 2018; Tchounwou et al., 2012) and causes cancer in the kidneys, blood, and cardiovascular and nervous systems (Firisa, 2023; Genchi et al., 2020; Tchounwou et al., 2019; Ulapane et al., 2023). Lead exposure damages the nervous system and kidney function, which is especially hazardous to children and pregnant women (Chan & Qi, 2003; Ray et al., 2009; de Jesús Rubio et al., 2023; Yu et al., 2017). Furthermore, climate change has made it difficult to get fresh water in areas already affected by pollution through increasing temperatures, frequent floods, and drought persistence (Li et al., 2023; Ling et al., 2024). After the industrial revolution and the scientific growth of humans, additional environmental contaminants entered water, soil, and air, as well as increased entrance of these compounds into plants, particularly vegetables, through water. Continuous growth in the human population spread urbanization, industrialization work, technical improvement, and growth in agriculture, resulting in a lack of natural freshwater bodies, including lakes, wetlands, and rivers (Ibrahim and Nafi'u, 2017). It is a global issue, affecting both the environment and public health (Buckley & Casson, 2003; Stephenson, 1993). Agricultural products contaminated with heavy metals reduce their quality while also posing a major threat to human health. Some synthetic organic dyes could cause health and environmental issues because they contain toxic components (Chen et al., 2022). Therefore, environmental considerations are critical, and heavy metals, including copper, zinc, and cobalt, are critical for many biological systems.

Monitoring water quality parameters according to relevant standards is essential, but the traditional approach to water quality parameter monitoring is limited due to its lack of expansiveness and technical difficulties (Mostofi, 2018; Zhu et al., 2022). However, by using machine learning algorithms, the complexity of non-linear relationships could be easily dealt with, which can handle the underlying discovery mechanisms (Ray, 2019; Li et al., 2022). In recent years, many researchers have believed in the advancement of machine learning that highly dimensioned data can be captured successfully and evaluate the complex and large scale of water quality requirements (Zhu et al., 2022). The Boruta algorithm was employed to find significant influential factors and improve the model performance associated with water quality classification (Jamei et al., 2022). Simply put, the model's efficiency can be elevated by identifying the valuable factors and eliminating the unnecessary factors corresponding to the water quality classes.

In literature, various machine learning algorithms frequently employed for classification and regression, including the decision tree model (Saghebian et al., 2014; Sihag et al., 2021a), the k-nearest neighbor method (Naimi et al., 2022), the multiple perceptron neural network model (Abba et al., 2021; Juna et al., 2022), support vector machine model (Ehteram et al., 2021; Sakaa et al., 2022), the relevance vector machine model (Pham et al., 2021), the random forest model (Alnahit et al., 2022; Sihag et al., 2021b), the M5 three and random subspace model (Almohammed et al., 2022; Pande et al., 2023), and others (Maroufpoor et al., 2022) have been employed for estimating quantitative target variables (Dai et al., 2024; Hrnjica et al., 2021). Since we

know that water quality is important factor for living, numerous researchers and environmental scientists have utilized various methods to monitor and classify it. It gives a complete evaluation of the methodology used for monitoring water quality classification based on the existing literature. Various traditional measurements have been used for water quality when the enumerator wants to collect data about the water quality assessment, like manual data collection from different locations, laboratory analysis, calculating the water quality index, and experimental work carried out on the critical parameters that affect water quality (Jayaraman et al., 2024; Wu et al., 2021). To calculate WQI, it involves the selection of water quality class categories and computing the sub-index, along with the aggregation function. All these calculations are crucial to calculating the WQI that affects the water quality class (Hipsey et al., 2020; Uddin et al., 2022). Zhang et al. (2022) explore the utilization of the ultraviolet spectroscopy and the CNN model for the monitoring of water quality. So, the standard normal and multiplicative correction methods have high accuracy in total organic carbon and total suspended solid parameters. In 2022, Zhang et al. introduced the AT-BILSTM model for forecasting watershed water quality. Li et al. (2020) used three deep learning models, including RNN, LSTM, and GRU, to estimate the dissolved oxygen concentration in the region's fishing ponds. Furthermore, T^2-bsased principle component analysis could be affected by an outlier if there is an outlier in observation, and it gives an incorrect result in classification.

Therefore, many modifications have been developed to solve this problem. The hotelling T^2-based principle component analysis control chart and T^2 statistic were widely used for the detection of outliers. For classification problems, T^2 and Q-statistics have been frequently used (Eide and Westad, 2018; Li et al., 2018; Ruiz et al., 2004). However, due to the complexity and time consumption of a data-driven model, it might be an uncertain prediction (Bui et al., 2022). Moreover, the advancement of machine learning and deep learning has significantly impacted a vital role in water quality assessment and has fundamentally changed our prediction methods and understanding of environmental data. This research study focuses not just on assessing and predicting water quality class but also on identifying the harmful factors in the water and rivers in Khyber Pakhtunkhwa, Pakistan. To resolve water quality issues, various machine learning algorithms were employed, including the Boruta algorithm to overcome the complexity of the model, the decision tree model, the k-nearest neighbor method, the multilayer perceptron algorithm, the support vector machine, and the random forest algorithm. The aim of this study is to protect human health from harmful heavy metal contamination and serves as a significant guideline for water quality management and resource management in the Eastern Hindukush regions of Pakistan.

2. METHODOLOGY

2.1. DESCRIPTION OF STUDY AND LOCATION

The present study examines and evaluates the water quality in various glacial Alpine lakes and glacier-fed rivers, including the Swat and Panjkora Rivers in Khyber Pakhtunkhwa, Pakistan. These glacial lakes have a critical role in maintaining the daily lives of approximately 10 million people by providing drinking water and are used for agriculture, electricity, and recreation purposes. Tourism plays a vital role in the economy of the northern areas of Pakistan but has negative impacts on the water quality of the Swat and Dir regions. (Muzamil, 2021). The interconnection of the Alpine glacial lakes with the studied rivers, surrounding landscapes, and floodplain areas highlights the significance of comprehensive management techniques to control and overcome pollution and protect these vital water resources. The dataset for this investigation was obtained from a previous study conducted by Khan et al. (2022), as shown in Figure 1 using ArcGIS. In this research study, 79 samples were collected from the Swat and Panjkora rivers. Swat is located in the northern mountain ranges. The temperate zone in the Swat region has elevations ranging from 500 to 6,500 meters above sea level (Qasim et al., 2011). Swat is located between 34 and 36 degrees north latitude and 71 and 73 degrees east longitude, with an area of 5,337 square kilometers. In Pakistan's eastern Hindukush region, the Swat River is a major flood-prone area and travels 288 kilometers from the Seri Kalam Valley to Sardaryab Charsadda (Nasir et al., 2023). The Swat River starts in Kalam with the merging Ushu and Utror rivers and runs downstream through the valley for 160 kilometers to Chakdara. It covers more than 250 kilometers from Kalam to near Charsada. Numerous large and small tributaries join this river. Among them, the Panjkora River is 220 kilometers long and flows south through Dir Upper and Dir Lower districts until joining the Swat at Bosaq. It is situated at the intersection of Dir Lower, Bajaur Agency, and Malakand, and then continuously travels downstream as the Swat River. Additionally, the Panjkora River ultimately enters the Kabul River in the area of Charsadda.

2.2. METHODS

The main stages in developing the classification of water classes are as follows. Before fitting the models, we use the Boruta technique to identify critical parameters related to water quality classes. To model water quality classification utilizing DT, KNN, MLP, SVM, and RF, we selected a collection of 79 water samples, with their respective quality classes acting as the model's inputs and outputs. In this configuration, water quality parameters serve as the model's input, while water quality classes are generated from the WQI during the models fitting in this research study. During models fitting for this study,

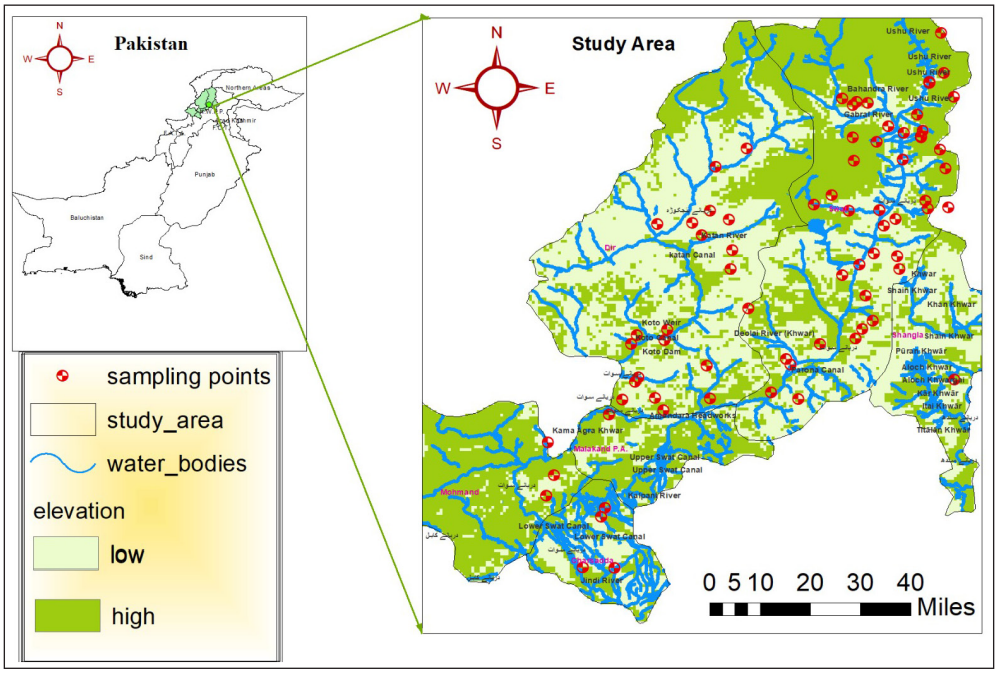

Figure 1 Location of the study area in Khyber Pakhtunkhwa, Pakistan.

Figure 2 Flowchart of the applied process for the water quality classification.

70% of the dataset was used for the training set, and the rest of the dataset was used for the testing dataset in order to generalize our result. Figure 2 provides a graphical representation of the various phases of the current investigation during this research. It starts with data collection and ends with the preprocessing of the data before calculating the WQI. The Boruta algorithm was applied to identify the significant factors for the water quality class. Furthermore, machine learning algorithms, including the decision trees model (DT), the multi-layer perceptron (MLP) algorithm, the k-nearest neighbors method (KNN), the support vector machines (SVM) model, and the random forest (RF) model, are used to develop water quality prediction.

2.3. WATER QUALITY INDEX (WQI)

The main objective of water quality is to provide an overall assessment by considering various parameters and heavy metals and simplifying the difficulties as a standard method for evaluating the water quality in the

world. The WQI was developed by Horton (1965) in the United States, and it was further modified by Brown et al. (1970). Recent studies have been extensively conducted for water quality assessment globally (Das et al., 2022; Kachroud et al., 2019; Udeshani et al., 2020). Furthermore, the weights were added in 2001 when computing WQI to determine the water quality class. For a detailed description, see Cude (2001). Over time, it was improved by different researchers (Hooshmand et al., 2011; Qasim et al., 2011; Raza et al., 2021; Sani et al., 2022; Uddin et al., 2021; Wali et al., 2020a; Wali et al., 2020b; Wali, et al., 2021; Wali et al., 2022). In this research study, the weighted arithmetic method was employed to calculate WQI because it is the most widely used method in both surface and groundwater studies (Călmuc et al., 2018; Tyagi et al., 2013). WQI was calculated for the river at the nine most used water quality parameters and screened for heavy metals such as cadmium (Cd in mg/L), chromium (Cr in mg/L), lead (Pb in mg/L), nickel (Ni in mg/L), iron (Fe in mg/L), arsenic (As in mg/L), power of hydrogen (Ph), electrical conductivity (Ec in µS/Cm), and total dissolved solids (TDS in mg/L). The weights are assigned to each physicochemical parameter of water according to their relative importance in determining water quality.

Computing WQI is given by the following formula:

$$WQI = \sum_{k-1}^{n} \beta_I \tag{1}$$

Where

$$\beta_I = \omega_i \rho_i \tag{2}$$

and

$$\rho_i = \frac{\theta_i}{S_i} \times 100 \tag{3}$$

In equation (1), β_I represents the ith parameter. In equation (2), ω_i is the weight of the ith parameter, heavy metals, and ρ_i indicates the quality rating of the ith parameter of water. In equation (3), θ_i denotes the value of each parameter in the sample of water, while S_i denotes the standard value of the ith parameter of water. In Table 1, the WQI is categorized (Khan et al., 2022; Tyagi et al., 2013).

WQI RANGE	CATEGORIES
WQI < 50	Excellent
50–100	Very good
100–150	Poor
150–200	Very poor
WQI > 200	Unsuitable for drinking

Table 1 Distribution of water quality index (WQI) categories.

2.4. BORUTA TECHNIQUE

The Boruta algorithm is utilized as the water quality parameter selection in this research, and it can select optimal parameters conveying relevant information for model prediction. The Boruta method, proposed by Zhang et al. (2020), is employed here to achieve this goal. The Boruta algorithm is based on R programming, which employs the importance measure provided by the original algorithm. It is based on a random forest implementation that uses a single function to identify all of the important parameters in a dataset in relation to an outcome variable. The number of estimators is predefined in this research study as the Boruta algorithm in R-programming and allows an automated number of estimator selection rather than finding a possibly compact subset of the parameters of water and heavy metals in a dataset because it is applicable to all types of parameters. The Boruta algorithm was utilized with settings such as doTrace = 2 and maxRun = 500 to identify the minimal optimal parameters and heavy metals for fitting the models.

2.5. DECISION TREE (DT) MODEL

The DT model is a key tool in the context of machine learning algorithms, depending on labeled training data to produce accurate predictions. It divides the dataset systematically into smaller subsets and performs tests at each node in the tree (Friedl and Brodley, 1997). A basic feature of decision trees is their frequent ability to segment the training set of the data and examine relevant attribute information and classification results. The primary goal is to build a decision tree that facilitates successful decision-making (Ren et al., 2022). Each route from the starting node to the leaf node in the DT model contains a set of rules, providing practical categorization rules that can be tested on real-world data. This hierarchical model uses an iterative process of categorizing the group of independent variables into homogeneous groups while highlighting essential characteristics at nodes from top to bottom (Hssina et al., 2014; Li et al., 2022). The Gini index is a crucial measure of inequality used in the DT model (Berndt et al., 2003; Ceriani and Verme, 2012). In the DT model, the Gini index helps in the selection of the most important parameter to split the data at each node, aiming to reduce impurity and increase homogeneity among the resulting subsets. Both the training and validation sets are used in the DT model to evaluate it. The DT model works similarly to a flowchart, as shown in Figure 3, based on the training data.

2.6. K- NEAREST NEIGHBOR (KNN) METHOD

Due to its popularity and robust performance, the KNN method is extensively used in data mining and machine learning applications, as observed in studies (Khamis, 2014; Wang et al., 2022; Zhang et al., 2017). Although

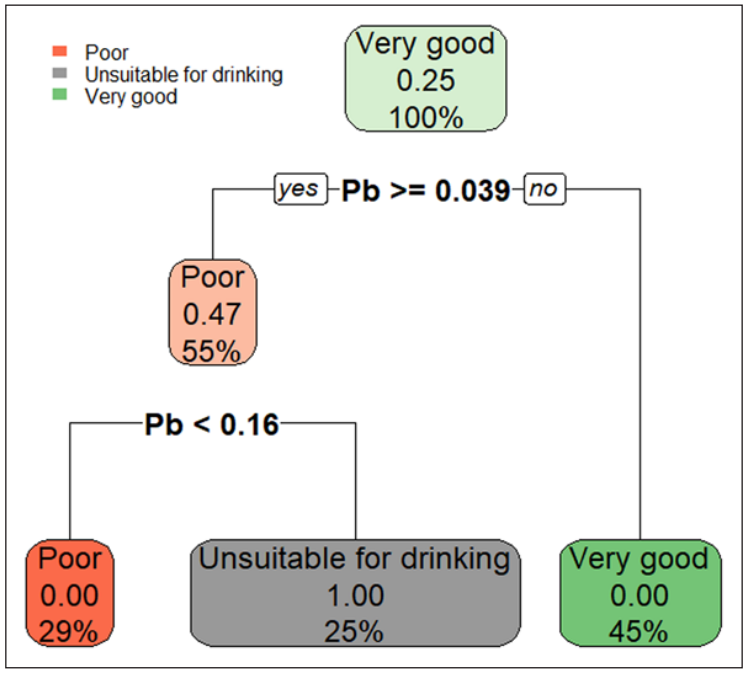

Figure 3 Decision tree visualization using the training set insight into model training.

famous for its simplicity and efficacy in classification, it has some drawbacks. Its efficiency is low, and it faces difficulties in k parameter selection. During a specific search, the KNN identifies the nearest data points from a provided training dataset. Once the k-nearest neighbor points are identified based on their distance from the search point, the algorithms employ a majority voting for the most major class and the rest for the minor class. In this research study, the KNN method was employed based on the training dataset, and then the model was trained with different values of k ranging from 1 to 10. Through this approach, the optimal value of k was determined to be 9 for this search. Furthermore, the model's performance was evaluated using the rest of the dataset.

2.7. MULTI- LAYER PERCEPTION (MLP)

Neural networks, particularly the MLP model, are widely viewed as requiring less domain-specific expertise for optimal performance. This is due to their ability to iteratively and independently tune parameters (Aitkin and Foxall, 2003; Piramuthu et al., 1994). This feature dramatically reduces the amount of time and human input required, speeding the machine learning process. It is a traditional model of neural networks mostly employed for classification problems (Sharma et al., 2022). For the classification task, nodes in the input and output layers in the MLP model are based on the number of parameters and classes. Moreover, in the MLP model, additional configuring includes, the number of hidden layers, node allocation to each hidden layer, complete iterations, and also weights learning when using MLP-NN (Azad et al., 2022; Hrnjica et al., 2021). The input layer contains the prediction parameters that affected the prediction class, and the output layer derives the prediction based on those affected factors. In this article, we used an R "neuralnet" function after training the MLP-NN to determine the setup of neurons within the neural network's hidden layers.

2.7.1 Proposed Architecture

The hidden parameter is used in R "neuralnet" functions with an activation function such as a logistic function. After training the MLP-NN to determine the setup of neurons within the neural network's hidden layers, Should not be selected randomly as the number of hidden layers and neurons when fitting the MLP model. Using too many layers and neurons can cause overfitting of the model; similarly, using a smaller number of neurons and hidden layers might be cause underfitting of the MLP model. Therefore, deciding the number of hidden layers and neurons should be done carefully. Two hidden layers are optimal for prediction studies. In difficult problems, one or two hidden layers provide optimal results in classification, as suggested in the given article (Panchal et al., 2011). There are two hidden layers used here for modeling, as shown in Figure 4. In each layer, we used hidden layer = (9,9) based on the formula (2/3 multiplied by the size of the input layer) and then added the size of the output layer (Panchal et al., 2011).

2.8. SUPPORT VECTOR MACHINE (SVM)

We used the SVM model with varied settings in our research. To build the SVM model in this research, e1071 package was used to implement the model. Different kernel functions were examined to determine how they affected the model's performance, like linear, radial, and polynomial kernels. The SVM is a strong tool for

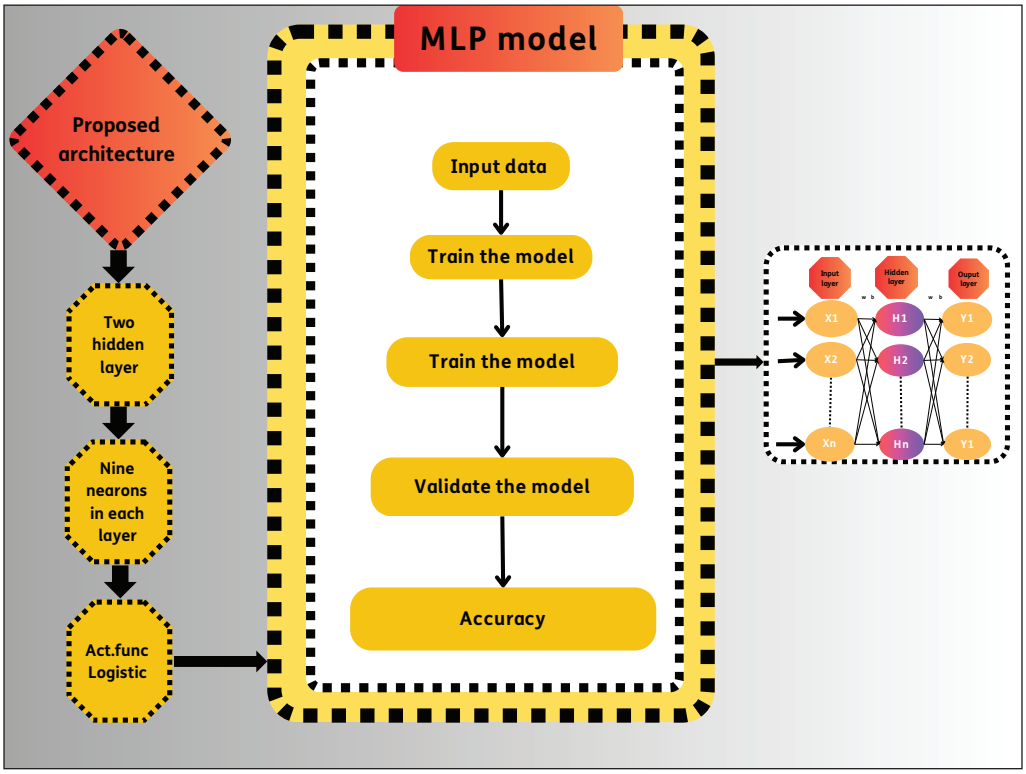

Figure 4 Multiple layer perceptron for water quality classification.

categorizing and predicting (Patle and Chouhan, 2013), and we wanted to identify the optimal parameters and kernels for our particular case. This extensive method emphasizes the significance of the SVM model prediction and model performance. The key challenge in the classification process is determining the hyperplane that separates these classes, and this hyperplane is dependent on the maximum margin (Banerjee and Mondal, 2023). The SVM can handle a variety of optimization problems, including regression and data classification (Gunn, 1998). To distinguish between two categories, which can also hold the data if they have more than two categories, a hyperplane must be identified to partition the data points into two classes. For a more detailed description of the SVM model, see Baccour et al. (2022).

2.9. RANDOM FOREST (RF) MODEL

RF is a popular machine learning model widely used for both classification problems and regression problems, and it was developed by Breiman (2001) with a set of decision trees with controlled variation. The RF model is employed for both classification and regression analysis but is commonly used for classification (Jakhar et al., 2023; Li et al., 2022). It constructs several decision trees from the input data. Each decision tree generates a prediction output from each of the decision trees separately. The final prediction is determined by aggregating votes from different decision trees to select the best prediction (Pandimurugan et al., 2022). The number of decision trees in the RF model increases from the input data, and the precision of the RF model also increases, resulting in an effective model prediction (Breiman, 2001; Strobl et al., 2009). The accuracy of the RF algorithm mainly depends on the individual tree classifiers and the relationship between the classifiers (Amit et al., 1997; Galiano et al., 2012). Involving two main parameters, one is the number of trees, which is denoted by (ntree) in R language, and the second is (mtry), which is the number of variables. To optimize the model, we used various numbers of trees and numbers of variables at each node in the RF model.

2.10. MODEL PERFORMANCE EVALUATIONS

2.10.1. Accuracy

Accuracy serves as a metric that measures the overall correctness of the model. It is calculated by the number of correct predictions, classincluding the sum of the true positives and true negatives, and then divided by the total number of predictions class. In Equation (4), it is derived using TP, where TP is true positive, similarly TN (true negative), FP (false positive), and FN (false negative) (Alqahtani et al., 2022; Anđelić et al., 2023; Koranga et al., 2022).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

2.10.2. Precision (Positive predictive value)

Precision is a measure of a model's accuracy in making good predictions when we are fitting various models. In equation (5), we devide the total number of positive predictions class by the sum of true positive and false positive predictions class. False positive predictions will

be rare when precision is high (Armah et al., 2014; Miao & Zhu, 2022).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

2.10.3. Recall (True positive rate)

According to the following, in equation (6), Recall is defined as the ratio of true positives (TP) to the sum of true positives and false negatives (FN) in the model. It captures all actual positive classes in the model, and it is also called sensitivity. A greater recall value suggests that the model was successful in identifying most positive cases (Armah et al., 2014; Koranga et al., 2022; Miao & Zhu, 2022).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

2.10.4. F1-score

The F1 score provides the model performance that combines precision and recall statistics. According to the following in equation (7), using the harmonic mean of precision (also defined in equation (6)). The range of the F1 score is from 0 to 1. A lower F1 value indicates that the model performance, in terms of precision and recall, is poor fitting, and a high value of the F1 score indicates that the model performance is better as compared to other models.

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

3. RESULTS AND DISCUSSION

Water is necessary for many human activities, including agriculture, residential use, recreation, and industry. Unfortunately, agricultural and human activities frequently pollute water sources, posing a severe danger to water quality. Water quality is traditionally calculated using water quality criteria obtained through a time-consuming laboratory examination. For water quality prediction, various machine learning algorithms are used for estimating and predicting WQI based on the nine parameters and heavy metals. Except Ec in µS/Cm and Ph in number, all the parameters and heavy metals are represented using a mg/L scale. The findings of the study classified the water samples into three distinct categories: very good, poor, and unsuitable for drinking. According to the WQI classification scale proposed by Brown et al. (1970), most samples of the dataset fell into 'Very good (51.9%)', and the other two classes (24.05%) were identified as the same, unfortunately, as shown in Figure 5. All the summary results presented in Table 2 are in the units of mg/L and uS/Cm. Cadmium (Cd) levels are within an acceptable range, with a standard value of 0.05; mean, minimum, and maximum values are 0.04, 0.01, and 0.07, respectively, and showed a low standard deviation of 0.01. However, chromium (Cr) varies significantly because it exceeds the normal range of 0.03. The average concentration is 0.05 mg/L, with a wide range of 0 to 0.18 mg/L and a standard deviation of 0.04 mg/L. Lead (Pb) levels are of great concern because they surpass the standard threshold of 0.01 mg/L based on the World Health Organization's (WHO) standard value. The mean concentration of 0.19 mg/L, with a wide range of 0.01 mg/L to 0.91 mg/L, exhibits a high standard deviation of 0.29 mg/L. Nickel (Ni) and iron (Fe) levels are adequate; however, arsenic (As) levels are routinely recorded at 0 mg/L, indicating probable measurement problems. The Ph is within the acceptable range of 6.5–8.5, with a mean of 7.74. Electrical conductivity (Ec) and Total Dissolved Solids (TDS) both deviate significantly from the acceptable standard values of 400 S/Cm and 500 mg/L, suggesting excessive levels. The average Ec is 107.14 S/Cm, while the average TDS is 53.98 mg/L.

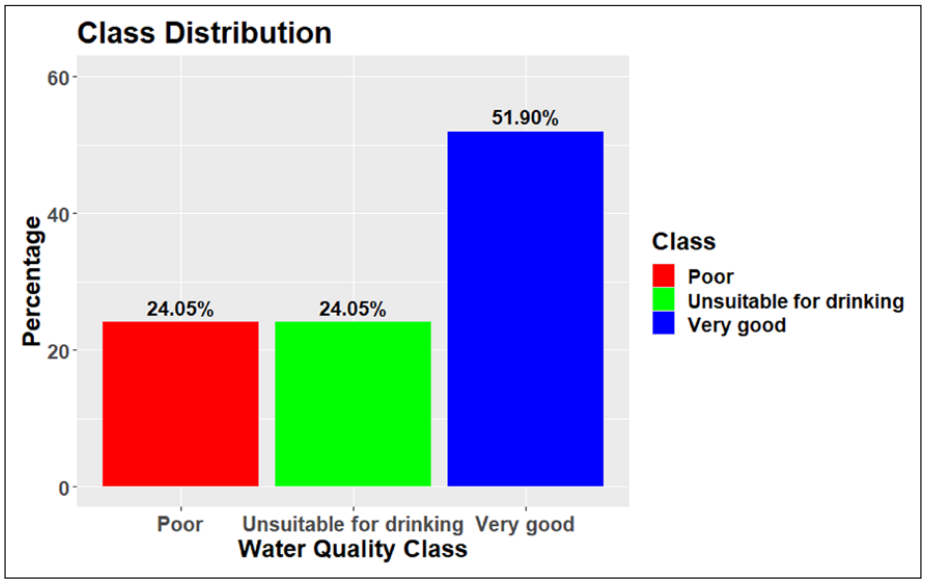

Figure 5 Visualizing the distribution of water quality classes.

S.NO	PARAMETERS AND HEAVY METALS	UNIT	WHO VALUE	ASSIGNED WEIGHTS	RELATIVE WEIGHTS	MEAN	SD	RANGE
1	Cd	mg/L	0.05	4	0.12	0.04	0.01	(0.01–0.07)
2	Cr	mg/L	0.03	4	0.12	0.05	0.04	(0–0.18)
3	Pb	mg/L	0.01	4	0.12	0.19	0.29	(0.01–0.91)
4	Ni	mg/L	0.07	3	0.09	0.05	0.02	(0–0.17)
5	Fe	mg/L	0.3	3	0.09	0.05	0.02	(0.02–0.13)
6	As	mg/L	0.01	5	0.15	0	0	(0–0)
7	Ph	Number	6.5–8.5	4	0.12	7.74	0.25	(7.3–8.6)
8	Ec	µS/Cm	400	2	0.06	107.14	74.1	(20–309)
9	TDS	mg/L	500	5	0.15	53.98	39.7	(9.5–193)

Table 2 Descriptive statistics of water quality index deviation from the WHO value (Ilaboya et al., 2014; Khwaja and Aslam, 2018).

VARIABLE	MEAN-IMP	RANGE	NORM-HITS	DECISION
Cd	12.89	(11.51–13.84)	1.00	Confirmed
Cr	15.79	(13.77–17.46)	1.00	Confirmed
Pb	24.74	(22.14–27.06)	1.00	Confirmed
Ni	10.13	(8.96–11.65)	1.00	Confirmed
Fe	11.03	(9.51–12.05)	1.00	Confirmed
As	1.02	(–0.77–2.48)	0.00	Rejected
Ph	3.59	(1.53–4.99)	0.76	Confirmed
Ec	13.27	(12.01–14.92)	1.00	Confirmed
TDS	15.14	(13.34–16.51)	1.00	Confirmed

Table 3 Boruta algorithms based relative importance of water quality parameters and heavy metals.

Employing the Boruta algorithm, all parameters of the water, with the exception of arsenic heavy metal, were identified as more important and played a significant role in this study for the specific region of Khyber Pakhtunkhwa, Pakistan. Table 3 summarizes the key findings of the confirmed and rejected parameters and heavy metals with different descriptive statistics: 'Mean-Imp', 'Range', 'Norm-Hits', and 'Decision'. Furthermore, we used different metrics for assessing the applied machine learning algorithms. Moreover, to ensure robust performance in this analysis, the dataset was divided into two parts, with 70% allocated to the training set and the rest to the testing or validation set. Subsequently, five supervised machine learning models were employed, including the DT model, the KNN method, neural network or the MLP algorithm, SVM, and the RF model to predict and classify the water quality class with multiple parameters and heavy metals. The overall model performance is checked by the final metric accuracy of the model, extracted from confusion metrics for each model. The DT model achieved an accuracy of 96% and 83% based on the training and testing sets, respectively. The KNN method achieved an accuracy of 75% with both the training and testing sets. The MLP achieved an accuracy of 96% and 83% based on the training and testing sets, respectively. SVM achieved an accuracy of 96% based on the training set and 88% with the testing set. Similarly, in our fifth effort, we employed the RF model, and the relative importance of each parameter or heavy metals was calculated, as shown in Figure 6. The RF model outperformed the previous models except for the SVM based on the testing set. The second comparable model is the SVM model according to the validation set. The RF model achieved 100% and 88% accuracy based on the training and testing sets, respectively, as shown in Figure 7. Furthermore, the summarization of various metrics like precision, recall, and F1-score are included in Table 4, providing insights through multiple comparison metrics for each model.

4. CONCLUSION

Water is essential for survival, so ensuring its quality is critical for our well-being. Traditional lab analyses, alternatively, may require more time and resources.

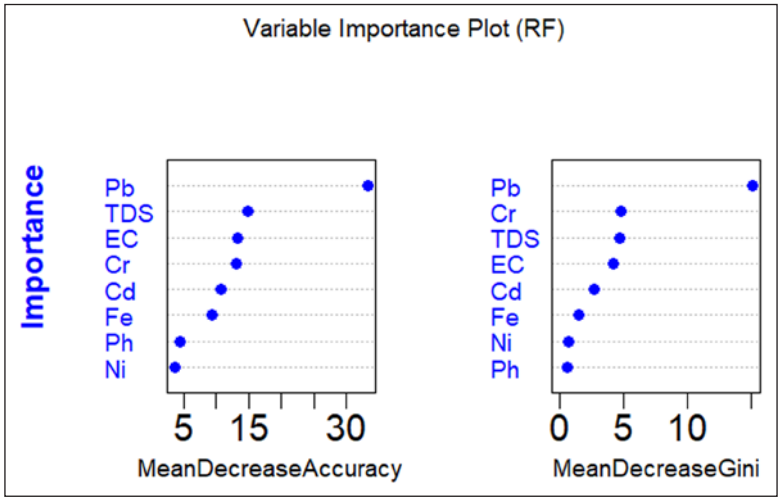

Figure 6 Ranking important parameters provides insights from the random forest model.

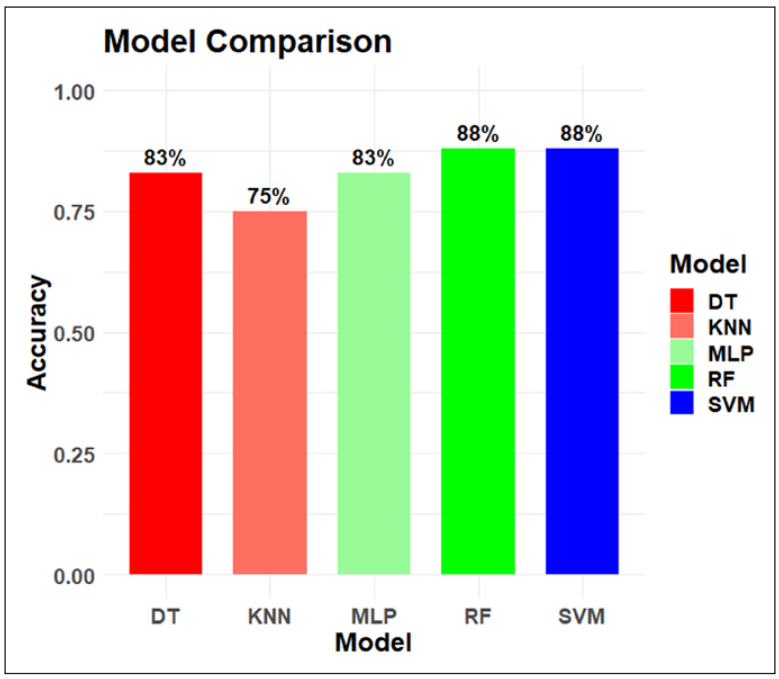

Figure 7 Evaluating accuracy metrics across various supervised machine learning models.

MODEL	PRECISION	RECALL	F1-SCORE
DT	0.91	0.88	0.89
KNN	0.86	0.59	0.70
MLP	0.86	0.83	0.85
SVM	0.87	0.89	0.88
RF	0.90	0.88	0.89

Table 4 Models performance with weighted precision, recall, and F1-score.

Under these scenarios, establishing practical and effective ways to monitor water quality for this research study, such as using the WQI, becomes essential. The present study was conducted for the water quality evaluation in the three districts of Khyber Pakhtunkhwa, Pakistan, and employed supervised machine learning algorithms to predict and validate the WQI based on the various categories. This shows that most samples fall into the 'Very good (51.9%)' category, with 24.5% falling into the 'Poor' category and 24.5% falling into 'Unsuitable for drinking water' category. Based on the present study, we can conclude that half of the water in this area is fit for drinking, but we still need to be aware of the water quality management for the other parts of the remaining percentage. In a set of supervised machine learning models, all models perform well, but the RF outperforms in predicting the WQC based on both the training and validation sets, and the second comparable model is the SVM model. Based on the findings, we should consider integrating data from multiple locations in Khyber Pakhtunkhwa, Pakistan, to improve the capabilities of machine learning models for future studies. These

methods can provide a fundamental understanding of the heavy metal parameters that affect water quality class. To ensure the long-term development of water quality in these locations, monitoring and protection actions must be prioritized, as well as actions to prevent further deterioration. The current study recommends that the government use appropriate measures, such as developed water monitoring systems and innovative technology, to address water quality issues in the influential areas of Pakistan.

DATA ACCESSIBILITY STATEMENT

The data and codes used for the preparation of the manuscript are available with the corresponding author and can be provided upon request.

ETHICS AND CONSENT

All procedures followed were in accordance with the ethical standards with the Helsinki Declaration of 1975, as revised in 2000.

All authors voluntarily agreed to participate in this research study.

All authors are agreed to for publication. There is no legal constraint in publishing the data used in the manuscript.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

All authors contributed equally.

AUTHOR AFFILIATIONS

Muhammad Akhlaq
Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan

Asad Ellahi
Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan; Department of Community Medicine, Wah Medical College, National University of Medical Sciences, Rawalpindi, Pakistan

Rizwan Niaz orcid.org/0000-0001-8959-7680
Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan; Department of Statistics, Kohsar University Murree, Murree, Pakistan

Mohsin Khan
Department of biological sciences, Quaid-I-Azam University, Islamabad, Pakistan

Saad Sh. Sammen
Department of Civil Engineering, College of Engineering, Diyala University, Diyala Governorate, Iraq

Miklas Scholz
Department of Civil Engineering Science, School of Civil Engineering, and the Built Environment, Faculty of Engineering and the Built Environment, University of Johannesburg, Kingsway Campus, PO Box 524, Aukland Park 2006, Johannesburg, South Africa; Department of Urban Drainage, Bau & Service Oberursel (BSO), Postfach 1280, 61402 Oberursel (Taunus), Germany; Kunststoff-Technik Adams, Specialist Company According to Water Law, Schulstraße 7, 26931 Elsfleth, Germany; Nexus by Sweden, Skepparbacken 5, 722 11 Västerås, Sweden; Department of Town Planning, Engineering Networks and Systems, South Ural State University (National Research University), 76, Lenin prospekt, Chelyabinsk 454080, The Russian Federation

REFERENCES

Abba, S.I., et al. (2021) Comparative implementation between neuro-emotional genetic algorithm and novel ensemble computing techniques for modelling dissolved oxygen concentration. *Hydrological Sciences Journal*, 66(10): 1584–1596. DOI: https://doi.org/10.1080/02626667.2021.1937179

Ahmed, U., et al. (2019) Efficient water quality prediction using supervised machine learning. *Water*, 11(11): 2210. DOI: https://doi.org/10.3390/w11112210

Alam, M.T., et al. (2017) The self-inhibitory nature of metabolic networks and its alleviation through compartmentalization. *Nature Communications*, 8(1): 16018. DOI: https://doi.org/10.1038/ncomms16018

Almohammed, F., et al. (2022) Assessment of soft computing techniques for the prediction of compressive strength of bacterial concrete. *Materials*, 15(2): 489. DOI: https://doi.org/10.3390/ma15020489

Alnahit, A.O., Mishra, A.K. and **Khan, A.A.** (2022) Stream water quality prediction using boosted regression tree and random forest models. *Stochastic Environmental Research and Risk Assessment*, 36(9): 2661–2680. DOI: https://doi.org/10.1007/s00477-021-02152-4

Alqahtani, Y., et al. (2022) Breast cancer pathological image classification based on the multiscale CNN squeeze model. *Computational Intelligence and Neuroscience*, 2022. DOI: https://doi.org/10.1155/2022/7075408

Amit, Y., Geman, D. and **Wilder, K.** (1997) Joint induction of shape features and tree classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 19(11): 1300–1305. DOI: https://doi.org/10.1109/34.632990

Anđelić, N., Baressi Šegota, S. and **Car, Z.** (2023) Improvement of malicious software detection accuracy through genetic programming symbolic classifier with application of dataset oversampling techniques. *Computers*, 12(12): 242. DOI: https://doi.org/10.3390/computers12120242

Armah, G.K., Luo, G. and **Qin, K.** (2014) A deep analysis of the precision formula for imbalanced class distribution.

International Journal of Machine Learning and Computing, 4(5): 417–422. DOI: https://doi.org/10.7763/IJMLC.2014.V4.447

Aitkin, M. and **Foxall, R.** (2003) Statistical modelling of artificial neural networks using the multi-layer perceptron. *Statistics and Computing*, 13: 227–239. DOI: https://doi.org/10.1023/A:1024218716736

Baccour, M.H., Driewer, F., Schäck, T. and **Kasneci, E.** (2022) Comparative analysis of vehicle-based and driver-based features for driver drowsiness monitoring by support vector machines. *IEEE transactions on intelligent transportation systems*, 23(12): 23164–23178. DOI: https://doi.org/10.1109/TITS.2022.3207965

Balali-Mood, M., Naseri, K., Tahergorabi, Z., Khazdair, M.R. and **Sadeghi, M.** (2021) Toxic mechanisms of five heavy metals: mercury, lead, chromium, cadmium, and arsenic. *Frontiers in pharmacology*, 12: 643972. DOI: https://doi.org/10.3389/fphar.2021.643972

Banerjee, S. and **Mondal, A.C.** (2023) A region-wise weather data-based crop recommendation system using different machine learning algorithms. *International Journal of Intelligent Systems and Applications in Engineering*, 11(3): 283–297. DOI: https://doi.org/10.17762/ijritcc.v11i1.6084

Berndt, D.J., Fisher, J.W., Rajendrababu, R.V. and **Studnicki, J.** (2003, January). Measuring healthcare inequities using the Gini index. In: *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, 2003. IEEE. pp. 1–10. DOI: https://doi.org/10.1109/HICSS.2003.1174353

Breiman, L. (2001) Random forests. *Machine learning*, 45: 5–32. DOI: https://doi.org/10.1023/A:1010933404324

Brown, R.M., et al. (1970) A water quality index-do we dare. *Water and Sewage Works*, 117(10).

Buckley, P.J. and **Casson, M.C.** (2003) Models of the multinational enterprise. *The New Economic Analysis of Multinationals*, 17. DOI: https://doi.org/10.4337/9781843766995.00010

Bui, N., Nguyen, D. and **Nguyen, V.A.** (2022) Counterfactual plans under distributional ambiguity. *arXiv preprint*. arXiv:2201.12487.

Călmuc, V.A., et al. (2018) Various methods for calculating the water quality index. *Annals of the "Dunarea de Jos" University of Galati. Fascicle II, Mathematics, Physics, Theoretical Mechanics*, 41(2): 171–178. DOI: https://doi.org/10.35219/ann-ugal-math-phys-mec.2018.2.09

Ceriani, L. and **Verme, P.** (2012) The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality*, 10: 421–443. DOI: https://doi.org/10.1007/s10888-011-9188-x

Chan, F.T. and **Qi, H.J.** (2003) An innovative performance measurement method for supply chain management. *Supply chain management: An international Journal*, 8(3): 209–223. DOI: https://doi.org/10.1108/13598540310484618

Chen, B., et al. (2022) Magnetic porous carbons derived from iron-based metal-organic framework loaded with glucose for effective extraction of synthetic organic dyes in drinks. *Journal of Chromatography A*, 1661: 462716. DOI: https://doi.org/10.1016/j.chroma.2021.462716

Cude, C.G. (2001) Oregon water quality index a tool for evaluating water quality management effectiveness 1. *JAWRA Journal of the American Water Resources Association*, 37(1): 125–137. DOI: https://doi.org/10.1111/j.1752-1688.2001.tb05480.x

Dai, H., et al. (2024) A two-step Bayesian network-based process sensitivity analysis for complex nitrogen reactive transport modeling. *Journal of Hydrology*, 632: 130903. DOI: https://doi.org/10.1016/j.jhydrol.2024.130903

Das, C.R., Das, S. and **Panda, S.** (2022) Groundwater quality monitoring by correlation, regression and hierarchical clustering analyses using WQI and PAST tools. *Groundwater for Sustainable Development*, 16: 100708. DOI: https://doi.org/10.1016/j.gsd.2021.100708

de Jesús Rubio, J., Garcia, D., Sossa, H., Garcia, I., Zacarias, A. and **Mujica-Vargas, D.** (2023) Energy processes prediction by a convolutional radial basis function network. *Energy*, 284: 128470. DOI: https://doi.org/10.1016/j.energy.2023.128470

Ehteram, M., et al. (2021) A hybrid novel SVM model for predicting CO_2 emissions using Multiobjective Seagull Optimization. *Environmental Science and Pollution Research*, 28: 66171–66192. DOI: https://doi.org/10.1007/s11356-021-15223-4

Eide, I. and **Westad, F.** (2018) Automated multivariate analysis of multi-sensor data submitted online: real-time environmental monitoring. *PLoS One*, 13(1): e0189443. DOI: https://doi.org/10.1371/journal.pone.0189443

Firisa, T.G., Geletu, A.K., Wondimu, K.T. and **Kedir, W.M.** (2023) Proximate composition, levels of heavy metals and their associated risk assessment in ginger (Zingiber officinale roscoe). *International Journal of Sustainable Energy and Environmental Research*, 12(2): 46–57. DOI: https://doi.org/10.18488/13.v12i2.3585

Friedl, M.A. and **Brodley, C.E.** (1997) Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61(3): 399–409. DOI: https://doi.org/10.1016/S0034-4257(97)00049-7

Genchi, S.A., Vitale, A.J., Perillo, G.M., Seitz, C. and **Delrieux, C.A.** (2020) Mapping topobathymetry in a shallow tidal environment using low-cost technology. *Remote Sensing*, 12(9): 1394. DOI: https://doi.org/10.3390/rs12091394

Gunn, S.R. (1998) Support vector machines for classification and regression. *ISIS Technical Report*, 14(1): 5–16.

Hipsey, M.R., Gal, G., Arhonditsis, G.B., Carey, C.C., Elliott, J.A., Frassl, M.A., et al. (2020) A system of metrics for the assessment and improvement of aquatic ecosystem models. *Environmental Modelling & Software*, 128: 104697. DOI: https://doi.org/10.1016/j.envsoft.2020.104697

Hooshmand, A., Delghandi, M., Izadi, A. and **Aali, K.A.** (2011) Application of kriging and cokriging in spatial estimation of groundwater quality parameters. *African Journal of Agricultural Research*, 6(14): 3402–3408.

Horton, R.K. (1965) An index number system for rating water quality. *J Water Pollut Control Fed*, 37(3): 300–306.

Hrnjica, B., et al. (2021) Application of deep learning neural networks for nitrate prediction in the Klokot River, Bosnia and Herzegovina. *2021 7th International Conference on Control, Instrumentation and Automation*, 1–6. DOI: https://doi.org/10.1109/ICCIA52082.2021.9403565

Hssina, B., et al. (2014) A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications*, 4(2): 13–19. DOI: https://doi.org/10.14569/SpecialIssue.2014.040203

Ibrahim, S. and **Nafi'u, S.A.** (2017) Macroinvertebrates as Indicators of Water Quality in Thomas Dam, Dambatta, Kano State, Nigeria. *UMYU Journal of Microbiology Research (UJMR)*, 2(1): 61–71. DOI: https://doi.org/10.47430/ujmr.1721.010

Ilaboya, I.R., et al. (2014) Assessment of water quality index of some selected boreholes around dump sites in Nigeria. *International Journal of Environmental Monitoring and Protection*, 1(2): 47–55.

Jabbar, M. (2020) Spatial analysis of the factors responsible for waterborne diseases in rural communities located along the Hudiara Drain, Lahore. *Pakistan Geographical Review*, 75: 84–94.

Jakhar, J., Jakhar, J. and **Chugh, R.** (2023) Fuzzy stability of mixed type functional equations in Modular spaces. *Mathematical Foundations of Computing*. DOI: https://doi.org/10.3934/mfc.2023019

Jamei, M., et al. (2022) Developing hybrid data-intelligent method using Boruta-random forest optimizer for simulation of nitrate distribution pattern. *Agricultural Water Management*, 270: 107715. DOI: https://doi.org/10.1016/j.agwat.2022.107715

Jayaraman, P., et al. (2024) Critical review on water quality analysis using IoT and machine learning models. *International Journal of Information Management Data Insights*, 4(1): 100210. DOI: https://doi.org/10.1016/j.jjimei.2023.100210

Juna, A., et al. (2022) Water quality prediction using KNN imputer and multilayer perceptron. *Water*, 14(17): 2592. DOI: https://doi.org/10.3390/w14172592

Kachroud, M., et al. (2019) Water quality indices: Challenges and application limits in the literature. *Water*, 11(2): 361. DOI: https://doi.org/10.3390/w11020361

Khamis, H.S. (2014) Application of k-nearest neighbour classification in medical data mining in the context of Kenya. *International Journal of Information and Communication Technology Research*, 4(4): 121–128.

Khan, M., Almazah, M.M., EIlahi, A., Niaz, R., Al-Rezami, A.Y. and **Zaman, B.** (2023) Spatial interpolation of water quality index based on Ordinary kriging and Universal kriging. *Geomatics, Natural Hazards and Risk*, 14(1): 2190853. DOI: https://doi.org/10.1080/19475705.2023.2190853

Khan, M., et al. (2022) Water quality assessment of Alpine glacial blue water lakes and glacial-fed rivers. *Geomatics, Natural Hazards and Risk*, 13(1): 2597–2617. DOI: https://doi.org/10.1080/19475705.2022.2126800

Khanna, D.R. and **Ishaq, F.** (2013) Impact of water quality attributes and comparative study of icthyofaunal diversity of Asan Lake and River Asan. *Journal of Applied and Natural Science*, 5(1): 200–206. DOI: https://doi.org/10.31018/jans.v5i1.306

Khwaja, M.A. and **Aslam, A.** (2018) Comparative assessment of Pakistan national drinking water quality standards with selected Asian countries and World Health Organization. Islamabad: Sustainable Development Policy Institute.

Koranga, M., et al. (2022) Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand. *Materials Today: Proceedings*, 57: 1706–1712. DOI: https://doi.org/10.1016/j.matpr.2021.12.334

Kumar, P., et al. (2022) Prevalence and predictors of water-borne diseases among elderly people in India: Evidence from Longitudinal Ageing Study in India, 2017–18. *BMC Public Health*, 22(1): 993. DOI: https://doi.org/10.1186/s12889-022-13376-6

Li, Q., et al. (2023) Impact of Inorganic Solutes' Release in Groundwater during Oil Shale In Situ Exploitation. *Water*, 15(1): 172. DOI: https://doi.org/10.3390/w15010172

Li, W., et al. (2018) Condition monitoring of sensors in a NPP using optimized PCA. *Science and Technology of Nuclear Installations*, 2018(2): 1–16. DOI: https://doi.org/10.1155/2018/7689305

Li, P., Zhou, K., Lu, X. and **Yang, S.** (2020) A hybrid deep learning model for short-term PV power forecasting. *Applied Energy*, 259: 114216. DOI: https://doi.org/10.1016/j.apenergy.2019.114216

Li, X., et al. (2022) Sustainable decision-making for contaminated site risk management: A decision tree model using machine learning algorithms. *Journal of Cleaner Production*, 371: 133612. DOI: https://doi.org/10.1016/j.jclepro.2022.133612

Ling, X., et al. (2024) The novel application of polyoxometalates for achieving sludge deep dewatering using low-temperature thermal hydrolysis pretreatment. *Journal of Cleaner Production*, 444: 141125. DOI: https://doi.org/10.1016/j.jclepro.2024.141125

Maroufpoor, S., et al. (2022) A novel hybridized neuro-fuzzy model with an optimal input combination for dissolved oxygen estimation. *Frontiers in Environmental Science*, 10: 929707. DOI: https://doi.org/10.3389/fenvs.2022.929707

Miao, J. and **Zhu, W.** (2022) Precision–recall curve (PRC) classification trees. *Evolutionary Intelligence*, 15(3): 1545–1569. DOI: https://doi.org/10.1007/s12065-021-00565-2

Mostofi, F. (2018) Heavy metal contamination of zinc and lead in Region 1 and 2 of the main city of Ardabil. *Journal of Research in Science, Engineering and Technology*, 6(4): 14–20.

Muzamil, M.R. (2021) Climate-related disasters, conflict and development: Reflections about the past and insights into the future from the Khyber-Pakhtunkhwa Province of Pakistan. Unpublished thesis (PhD), University of Western Australia.

Naimi, A., et al. (2022) Fault detection and isolation of a pressurized water reactor based on neural network and k-nearest neighbor. *IEEE Access*, 10: 17113–17121. DOI: https://doi.org/10.1109/ACCESS.2022.3149772

Nasir, M.J., et al. (2023) Soil erosion susceptibility assessment of Swat River sub-watersheds using the morphometry-based compound factor approach and GIS. *Environmental Earth Sciences*, 82(12): 315. DOI: https://doi.org/10.1007/s12665-023-10982-4

Noureen, A., et al. (2022) The impact of climate change on waterborne diseases in Pakistan. *Sustainability and Climate Change*, 15(2): 138–152. DOI: https://doi.org/10.1089/scc.2021.0070

Panchal, G., et al. (2011) Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. *International Journal of Computer Theory and Engineering*, 3(2): 332–337. DOI: https://doi.org/10.7763/IJCTE.2011.V3.328

Pande, C.B., et al. (2023) Combination of data-driven models and best subset regression for predicting the standardized precipitation index (SPI) at the Upper Godavari Basin in India. *Theoretical and Applied Climatology*, 152: 535–558. DOI: https://doi.org/10.1007/s00704-023-04426-z

Pandimurugan, V., et al. (2022) Random forest tree classification algorithm for predicating loan. *Materials Today: Proceedings*, 57(5): 2216–2222. DOI: https://doi.org/10.1016/j.matpr.2021.12.322

Patle, A. and **Chouhan, D.S.** (2013) SVM kernel functions for classification. *2013 International Conference on Advances in Technology and Engineering*, 1–9. DOI: https://doi.org/10.1109/ICAdTE.2013.6524743

Pham, Q.B., et al. (2021) A new hybrid model based on relevance vector machine with flower pollination algorithm for phycocyanin pigment concentration estimation. *Environmental Science and Pollution Research*, 28: 32564–32579. DOI: https://doi.org/10.1007/s11356-021-12792-2

Piramuthu, S., Shaw, M.J. and **Gentry, J.A.** (1994) A classification approach using multi-layered neural networks. *Decision Support Systems*, 11(5): 509–525. DOI: https://doi.org/10.1016/0167-9236(94)90022-1

Qasim, M., et al. (2011) Spatial and temporal dynamics of land use pattern in District Swat, Hindu Kush Himalayan region of Pakistan. *Applied Geography*, 31(2): 820–828. DOI: https://doi.org/10.1016/j.apgeog.2010.08.008

Rafiee, P. (2018) Determination of heavy metal pollution products, vegetable gardens Ardabil. *Journal of Research in Science, Engineering, and Technology*, 6(4): 6–13.

Ray, S. (2019) A quick review of machine learning algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing*, 35–39. DOI: https://doi.org/10.1109/COMITCon.2019.8862451

Ray, W.A., Chung, C.P., Murray, K.T., Hall, K. and **Stein, C.M.** (2009) Atypical antipsychotic drugs and the risk of sudden cardiac death. *New England journal of medicine*, 360(3): 225–235. DOI: https://doi.org/10.1056/NEJMoa0806994

Raza, A., Hussain, I., Ali, Z., Faisal, M., Elashkar, E.E., Shoukry, A.M., et al. (2021) A seasonally blended and regionally integrated drought index using Bayesian network theory. *Meteorological Applications*, 28(3): e1992. DOI: https://doi.org/10.1002/met.1992

Ren, Q. (2022) A novel hybrid method of lithology identification based on k-means++ algorithm and fuzzy decision tree. *Journal of Petroleum Science and Engineering*, 208: 109681. DOI: https://doi.org/10.1016/j.petrol.2021.109681

Rodriguez-Galiano, V.F., et al. (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104. DOI: https://doi.org/10.1016/j.isprsjprs.2011.11.002

Ruiz, L.A., Fdez-Sarría, A. and **Recio, J.A.** (2004, July) Texture feature extraction for classification of remote sensing data using wavelet decomposition: A comparative study. In: *20th ISPRS Congress*, Vol. 35, No. part B. pp. 1109–1114.

Saghebian, S.M., et al. (2014) Ground water quality classification by decision tree method in Ardebil region, Iran. *Arabian Journal of Geosciences*, 7: 4767–4777. DOI: https://doi.org/10.1007/s12517-013-1042-y

Sakaa, B., et al. (2022) Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin. *Environmental Science and Pollution Research*, 29(32): 48491–48508. DOI: https://doi.org/10.1007/s11356-022-18644-x

Sani, A., Idris, K.M., Abdullahi, B.A. and **Darma, A.I.** (2022) Bioaccumulation and health risks of some heavy metals in Oreochromis niloticus, sediment and water of Challawa river, Kano, Northwestern Nigeria. *Environmental Advances*, 7: 100172. DOI: https://doi.org/10.1016/j.envadv.2022.100172

Scanlon, B.R., et al. (2023) Global water resources and the role of groundwater in a resilient water future. *Nature Reviews Earth & Environment*, 4(2): 87–101. DOI: https://doi.org/10.1038/s43017-022-00378-6

Sharma, R., Kim, M. and **Gupta, A.** (2022) Motor imagery classification in brain-machine interface with machine learning algorithms: Classical approach to multi-layer perceptron model. *Biomedical Signal Processing and Control*, 71: 103101. DOI: https://doi.org/10.1016/j.bspc.2021.103101

Sihag, P., Dursun, O.F., Sammen, S.Sh., Malik, A. and **Chauhan, A.** (2021b) Prediction of aeration efficiency of Parshall and Modified Venturi flumes: application of soft computing versus regression models. *Water Supply*, 21(8): 4068–4085. DOI: https://doi.org/10.2166/ws.2021.161

Sihag, P., Kumar, M. and **Sammen, S.Sh.** (2021a) Predicting the infiltration characteristics for semi-arid regions using regression trees. *Water Supply*, 21(6): 2583–2595. DOI: https://doi.org/10.2166/ws.2021.047

Stephenson, R.M. (1993) Mutual solubilities: water-glycol ethers and water-glycol esters. *Journal of Chemical and Engineering Data*, 38(1): 134–138. DOI: https://doi.org/10.1021/je00009a033

Strobl, C., Malley, J. and **Tutz, G.** (2009) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4): 323. DOI: https://doi.org/10.1037/a0016973

Talatian Azad, S., Ahmadi, G. and **Rezaeipanah, A.** (2022) An intelligent ensemble classification method based on multi-layer perceptron neural network and evolutionary algorithms for breast cancer diagnosis. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(6): 949–969. DOI: https://doi.org/10.1080/0952813X.2021.1938698

Tchounwou, P.B., Yedjou, C.G., Patlolla, A.K. and **Sutton, D.J.** (2012) Heavy metal toxicity and the environment. In: *Molecular, clinical and environmental toxicology: volume 3: environmental toxicology*, 133–164. DOI: https://doi.org/10.1007/978-3-7643-8340-4_6

Tchounwou, P.B., Yedjou, C.G., Udensi, U.K., Pacurari, M., Stevens, J.J., Patlolla, A.K., et al. (2019) State of the science review of the health effects of inorganic arsenic: perspectives for future research. *Environmental toxicology*, 34(2): 188–202. DOI: https://doi.org/10.1002/tox.22673

Tyagi, S., Sharma, B., Singh, P. and **Dobhal, R.** (2013) Water quality assessment in terms of water quality index. *American Journal of Water Resources*, 1(3): 34–38. DOI: https://doi.org/10.12691/ajwr-1-3-3

Uddin, M.G., Nash, S. and **Olbert, A.I.** (2021) A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators*, 122: 107218. DOI: https://doi.org/10.1016/j.ecolind.2020.107218

Uddin, M.G., Nash, S., Rahman, A. and **Olbert, A.I.** (2022) A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. *Water Research*, 219: 118532. DOI: https://doi.org/10.1016/j.watres.2022.118532

Udeshani, W.A.C., et al. (2020) Assessment of groundwater quality using water quality index (WQI): A case study of a hard rock terrain in Sri Lanka. *Groundwater for Sustainable Development*, 11: 100421. DOI: https://doi.org/10.1016/j.gsd.2020.100421

Ulapane, D., et al. (2023) Nutritional qualities and heavy metals accumulation in grains: A study on lowland irrigated rice with different fertilizer inputs and growing seasons. *International Journal of Sustainable Agricultural Research*, 10(3): 70–84. DOI: https://doi.org/10.18488/ijsar.v10i3.3531

Unigwe, C.O. and **Egbueri, J.C.** (2023) Drinking water quality assessment based on statistical analysis and three water quality indices (MWQI, IWQI and EWQI): A case study. *Environment, Development and Sustainability*, 25(1): 686–707. DOI: https://doi.org/10.1007/s10668-021-02076-7

Wali, S.U., Alias, N. and **Harun, S.B.** (2021) Reevaluating the hydrochemistry of groundwater in basement complex aquifers of Kaduna Basin, NW Nigeria using multivariate statistical analysis. *Environmental Earth Sciences*, 80: 1–25. DOI: https://doi.org/10.1007/s12665-021-09421-z

Wali, S.U., Alias, N.B., Harun, S.B., Umar, K.J., Gada, M.A., Dankani, I.M., et al. (2022) Water quality indices and multivariate statistical analysis of urban groundwater in semi-arid Sokoto Basin, Northwestern Nigeria. *Groundwater for Sustainable Development*, 18: 100779. DOI: https://doi.org/10.1016/j.gsd.2022.100779

Wali, S.U., Dankani, I.M., Abubakar, S.D., Gada, M.A., Umar, K.J., Usman, A.A. and **Shera, I.M.** (2020a) Re-Examination of hydrochemistry and groundwater potentials of Cross River and imo-kwa-ibo intersecting tropical basins of SouthSouth Nigeria. *Journal of Geological Research*, 2(3): 25–42. DOI: https://doi.org/10.30564/jgr.v2i3.2142

Wali, S.U., Dankani, I.M., Abubakar, S.D., Gada, M.A., Umar, K.J., Usman, A.A. and **Shera, I.M.** (2020b) Reassessing groundwater potentials and subsurface water hydrochemistry in a Tropical Anambra Basin, Southeastern Nigeria. *Journal of Geological Research*, 2(3): 1–24. DOI: https://doi.org/10.30564/jgr.v2i3.2141

Wang, A.X., Chukova, S.S. and **Nguyen, B.P.** (2022) Implementation and analysis of centroid displacement-based k-nearest neighbors. *International Conference on Advanced Data Mining and Applications*, 431–443. DOI: https://doi.org/10.1007/978-3-031-22064-7_31

Wijerathna, W.S.M.S.K., et al. (2023) Imperative assessment on the current status of rubber wastewater treatment: Research development and future perspectives. *Chemosphere*, 139512. DOI: https://doi.org/10.1016/j.chemosphere.2023.139512

Wu, Z., Lai, X. and **Li, K.** (2021) Water quality assessment of rivers in Lake Chaohu Basin (China) using water quality index. *Ecological Indicators*, 121: 107021. DOI: https://doi.org/10.1016/j.ecolind.2020.107021

Yu, Y., Jia, T. and **Chen, X.** (2017) The 'how'and 'where'of plant micro RNA s. *New Phytologist*, 216(4): 1002–1017. DOI: https://doi.org/10.1111/nph.14834

Zhang, Y., Ban, X., Li, E., Wang, Z. and **Xiao, F.** (2020) Evaluating ecological health in the middle-lower reaches of the Hanjiang River with cascade reservoirs using the Planktonic index of biotic integrity (P-IBI). *Ecological Indicators*, 114: 106282. DOI: https://doi.org/10.1016/j.ecolind.2020.106282

Zhang, H., et al. (2022) Online water quality monitoring based on UV–Vis spectrometry and artificial 588 neural networks in a river confluence near Sherfield-on-Loddon. *Environmental Monitoring and Assessment*, 194(9): 630. DOI: https://doi.org/10.1007/s10661-022-10118-4

Zhang, S., et al. (2017) Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(3), 1–19. DOI: https://doi.org/10.1145/2990508

Zhao, Y., et al. (2024) Release pattern of light aromatic hydrocarbons during the biomass roasting process. *Molecules*, 29(6): 1188. DOI: https://doi.org/10.3390/molecules29061188

Zhu, M., et al. (2022) A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*, 1(2): 107–116. DOI: https://doi.org/10.1016/j.eehl.2022.06.001

TO CITE THIS ARTICLE:
Akhlaq, M., Ellahi, A., Niaz, R., Khan, M., Sammen, S.Sh. and Scholz, M. (2024) Comparative Analysis of Machine Learning Algorithms for Water Quality Prediction. *Tellus A: Dynamic Meteorology and Oceanography*, 76(1): 177–192. DOI: https://doi.org/10.16993/tellusa.4069

Submitted: 26 March 2024 **Accepted:** 12 July 2024 **Published:** 30 July 2024