IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Water quality management using hybrid machine learning and data mining algorithms: An indexing approach

Bilal Aslam[1], Ahsen Maqsoom[2], Ali Hassan Cheema [2], Fahim Ullah [3], Abdullah Alharbi [4], and Muhammad Imran [5], Member, IEEE

[1]School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ 86011, USA
[2]Department of Civil Engineering, COMSATS University Islamabad, Wah Cantt 47040, Pakistan
[3]School of Surveying and Built Environment, University of Southern Queensland, Springfield Central 4300, Queensland Australia
[4]Department of Computer Science, Community College, King Saud University, P.O. Box 28095, Riyadh 11437, Saudi Arabia
[5]School of Engineering, Information Technology and Physical Sciences, Federation University, Brisbane, 4000, Australia.

Corresponding author: Ahsen Maqsoom (ahsen.maqsoom@ciitwah.edu.pk).

ABSTRACT One of the key functions of global water resource management authorities is river water quality (WQ) COD assessment. A water quality index (WQI) is developed for water assessments considering numerous quality-related variables. WQI assessments typically take a long time and are prone to errors during sub-indices generation. This can be tackled through the latest machine learning (ML) techniques that are renowned for superior accuracy. In this study, water samples were taken from the wells in the study area (North Pakistan) to develop WQI prediction models. Four standalone algorithms, i.e., random trees (RT), random forest (RF), M5P, and reduced error pruning tree (REPT), were used in this study. In addition, 12 hybrid data-mining algorithms (combination of standalone, bagging (BA), cross-validation parameter selection (CVPS), and randomizable filtered classification (RFC)) were also used. Using the 10-fold cross-validation technique, the data were separated into two groups (70:30) for algorithm creation. Ten random input permutations were created using Pearson correlation coefficients to identify the best possible combination of datasets for improving the algorithm prediction. The variables with very low correlations performed poorly, whereas hybrid algorithms increased the prediction capability of numerous standalone algorithms. Hybrid RT-Artificial Neural Network (RT-ANN) with RMSE = 2.319, MAE = 2.248, NSE = 0.945 and PBIAS = -0.64, outperformed all other algorithms. Most algorithms overestimated WQI values except for BA-RF, RF, BA-REPT, REPT, RFC-M5P, RFC-REPT, and ANN- Adaptive Network-Based Fuzzy Inference System (ANFIS).

INDEX TERMS Water quality index, machine learning, hybrid data-mining algorithms, cross-validation techniques, North Pakistan

I. INTRODUCTION

Water pollution is one of the critical challenges of the modern world where the goals such as the United Nations Sustainable Development Goals (UN-SDGs) and a smart and sustainable planet are being pursued. All societies, ecologies, and productions require abundant clean water supplies for farming, drinking, sanitation, and energy production. The global water crisis is among the serious threats that the human race is currently confronted with. Accordingly, the quantity and quality of groundwater are significant global concerns [1]. Many diseases occur due to polluted water, like cholera, diarrhea, typhoid, amebiasis, hepatitis, gastroenteritis, giardiasis, campylobacteriosis, scabies, and worm infections. Almost 1.6 million people died due to diarrhea in 2017 alone [2]. Water pollutants impact its conditions, which impact human health and marine life. Inadequate sewage networks, uncontrolled and improperly planned urbanization, and dumping industrial trash, pesticides, and fertilizers contribute to water pollution [3]. Such pollution is more evident in local rivers or water channels closer to urban developments.

With both non-point and point sources, river pollution is becoming a more significant problem and presents a tough challenge to global water management authorities. Such pollution seriously deteriorates water quality (WQ). WQ degradation substantially impacts aquatic life and the availability of clean water for drinking and agricultural purposes [4]. The pollution challenge is harder to tackle in developing countries which frequently go through times of economic fluctuations. Further each development action can have severe environmental consequences. For example, with an increase in the population and demand for more resources, the requirement for more agricultural production pressures soils' organic fertility, increasing the demand for artificial fertilizers to enhance yield [5]. Accordingly, surplus fertilizers are frequently dumped into rivers and waterways that pollute ground and underground water sources [1]. This increases the need for WQ assessment and surveillance.

WQ surveillance and evaluation are critical for environmental, climate, and human health protection. This can be achieved through timely, efficient, and long-term water management plans. The WQ is assessed through the water quality index (WQI). WQI helps guide policymakers' actions and decisions. However, calculating WQI is not a simple process due to the involvement of multiple sub-indices and equations. WQI is a non-dimensional index derived from defined WQ variables. It uses variables such as pH (potential of hydrogen), DO (dissolved oxygen), TSS (total suspended solids), BOD (biological oxygen demand), AN (ammoniacal-nitrogen), COD (chemical oxygen demand), and others [6]. The associated matrices enable a definite evaluation of WQ. Measurements of variables such as Ca2+, Mg2+, NO3, and others are commonly used to estimate groundwater quality indicators (GQIs) [7-9].

Several aspects of water, including physical, chemical, biological, and radiological, are included in the assessment of WQ [10]. In addition, WQI is a frequently used technique for assessing the effectiveness or failure of WQ management measures [11]. Some examples of WQIs include the Canadian WQI (CQI), United States National Sanitation Foundation WQI (NSFWQI), Interim National Water Quality Standards for Malaysia (INWQS), British Columbia WQI (BCWQI), Oregon WQI (OWQI), Florida Stream WQI (FWQI), and others. WQI is calculated through multiple methods and algorithms around the globe. However, WQI calculation is not a straightforward process, and the associated computations have many drawbacks [12]:

1. The computation algorithms are complex.
2. It is a lengthy process
3. The computations are verbose and harder to understand
4. The process is subject to inconsistencies and errors as there is no uniform WQI approach and the WQI computations frequently utilize different and varying algorithms.

Some experts have used a non-physical strategy to address these difficulties. Accordingly, they suggest using artificial intelligence (AI) to forecast WQI [13-15]. AI-based modeling eliminates the need for sub-index computations and quickly generates WQI values. Such AI algorithms are gaining popularity because of their nonlinear structures, capacity to forecast complicated events, ability to handle large datasets, and lack of sensitivity to missing data [16]. For WQI modeling, artificial neural networks (ANN) and adaptive network-based fuzzy inference system (ANFIS) based classic AI algorithms have been extensively developed. On the other hand, environmental scientists have researched more robust and trustworthy AI algorithms [17-19]. However, the methodology and quality of data gathering and analysis are critical to the predictive capability of AI systems.

Data mining is a form of AI algorithm developed to tackle nonlinear equations and reduce AI's drawbacks. It has been used to quantify suspended sediment yield [20], approximate benchmark water loss [21-23], and replicate direct sunlight [24]. New algorithms such as M5P, random tree (RT), random forest (RF), bagging (BA), reduced error pruning tree (REPT), instance-based k-nearest neighbors (IBK), random committee (RC) are currently explored in hydrological processes, climate science, and hydraulic systems [12, 20, 22, 23, 25, 26]. Another prominent solution for different environmental and hydrological issues includes the usage of tree-based algorithms such as decision trees [27] [28]. Furthermore, the known powerful machine learning (ML) tool for both linear and nonlinear regression problems is the support vector machine (SVM), which is used in a range of scientific problems with remarkable forecast accuracy [27, 29-31]. DT and SVM algorithms have been used to predict parameters of WQ, such as TDS (total dissolved solids), TSS, BOD, and COD.

Granata et al. [32] developed a regression tree (RT) algorithm and a support vector regression (SVR) algorithm for predicting wastewater quality indicators and discovered that the SVR model provided the best results. Kayaalp et al. [33] developed a hybrid SVR model using monthly WQ parameter data with the firefly algorithm (FFA) to forecast WQI. The algorithm showed a significant increase in prediction performance compared to the standalone SVR model. Kamyab-Talesh et al. [34] looked into the optimization of the SVM algorithm to investigate the factors having the highest impact on the WQI. The authors observed that nitrate is the most crucial parameter for WQI prediction. Wang et al. [35] analyzed three ML algorithms, SVR, SVR-GA (genetic algorithm), and SVR-PSO (particle swarm optimization), to predict WQI and compared their performance. Since decision tree-based algorithms (i.e., M5P, RF, RT, REPT, and others) lack hidden units and modeling clarity, they can produce superior modeling results than ANFIS and ANN [36]. Furthermore, integrated modeling gives more reliable results than using standalone algorithms.

Researchers from Iran [12] have introduced a new WQI to focus on the characteristics and conditions of the rivers and lakes because previous algorithms are time-consuming and

not accurate enough to be trusted. In addition, they added more parameters to their algorithm to improve prediction accuracy. However, its feasibility is not tested yet due to the diverse weather conditions that vary between the arid and moist seasons. As a result, applying this index to specific locations may be risky and yield variable results. But since our study area lies in a similar climate region and has the same metrological and climatic properties, we can rely on this algorithm for our study area.

Northern Pakistan has gained economic significance over the last decade because of China Pakistan Economic Corridor (CPEC) project [1, 37]. The urban areas, for instance, Gilgit city, are experiencing an economic boom because of the latest development, which has brought improved linkage and connectivity through the upgradation of the Karakoram Highway (KKH). This enhanced connectivity is also inviting an urban sprawl in the region and is expected to face many environmental issues, including WQ [1, 38]. Very recently, Maqsoom et al. [38] mapped the groundwater susceptibility of this region and found that the region has moderate to high groundwater susceptibility, particularly the region around Gilgit city. Moreover, Awais et al. [1] also conducted a study and assessed nitrate contamination in this region and found that the region has moderate to high groundwater nitrate contamination risk. Overall, the two studies discovered that the water in the extreme Northern side appears to be of good quality, with minimal contamination and protected through natural vegetation. However, as the system approaches a developed region, Gilgit city, WQ rapidly degrades because of improper and unregulated discharges, a typical trend of build-up regions [39].

The objective of this research, however, is to forecast the WQI for the region along KKH. For achieving this, the current study utilizes four standalone algorithms, M5P, RT, RF, REPT, and 12 unique hybrid data mining algorithms (randomizable filtered classifier, CV parameter selection, and BA) combined with the four standalone algorithms. It was expected that WQI could be accurately forecasted using a standalone decision tree algorithm, as its ability to predict diverse hydrological events has been proved in the literature mentioned earlier. However, by combining it with classifier algorithms, it was aimed that the precision rate could be enhanced further and the fundamental flaws of the given algorithms could be reduced. Therefore, a combination was proposed and utilized in this study.

The current study differentiates itself from published works as two new hybrid algorithms were tested in this study for WQI analysis. Moreover, the outcomes of the hybrid algorithms were compared with the previously established algorithms and techniques to establish a more robust algorithm in terms of better accuracy. This study will benefit this fast-growing region as it is expected to be highly induced by human activities and causing many environmental issues, i.e., water pollution, and will help policymakers in the CPEC region with better water management.

The rest of the paper is organized as follows. First, the study area is explained in Section 2. Then Section 3 describes the research methodology, followed by the presentation of the algorithms used in this research in Section 4. Section 5 compares the algorithms and their performance. Section 6 presents results and pertinent discussions. Finally, Section 7 concludes the study and explains the key takeaways, limitations, and future direction for further expanding the current research.

II. Study Area

The research area extends from Gilgit to Khunjerab Pass and lies at 35.8819° N, 74.4643° E, and 36.8539° N, 75.4589° E. The study area is located in northern Pakistan and lies in the Districts Gilgit and Hunza-Nagar, located near the Pakistan-China border. The study area is 20 kilometers buffer along the 236 kilometers (146.6 miles) stretch of the traditional Silk route/ KKH from Gilgit to Khunjerab Pass, encompassing a hilly terrain. This route has tremendous importance as it connects Pakistan and China and is considered the backbone of the CPEC project [1, 38]. The study area is a part of the Himalayas, Hindukush, and Karakoram Mountain ranges, having an elevation range from 1294 meters to 7330 meters. River Hunza and River Gilgit flows from this region to provide domestic water. The area is located at a high altitude and receives lots of snow in winter, which melts in summer, thus providing freshwater [1, 37]. In the past, this region had an excellent WQ, but the local WQ is deteriorating due to the recent construction and other development due to CPEC. This calls for a WQ study for the region to better manage the groundwater and surface water in line with the global sustainability goals. Figure 1 shows the study area and locations of water wells from where the water samples were taken and analyzed for the research. Figure 1 further shows the water channels, district boundary, and elevation in the study area.

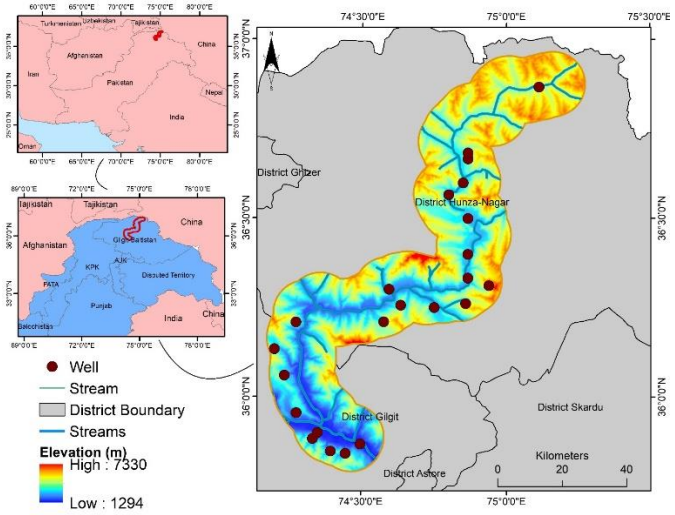


FIGURE 1. Water quality monitoring stations

III. Methodology

Figure 2 shows the methodology flow chart of this research and the associated steps.

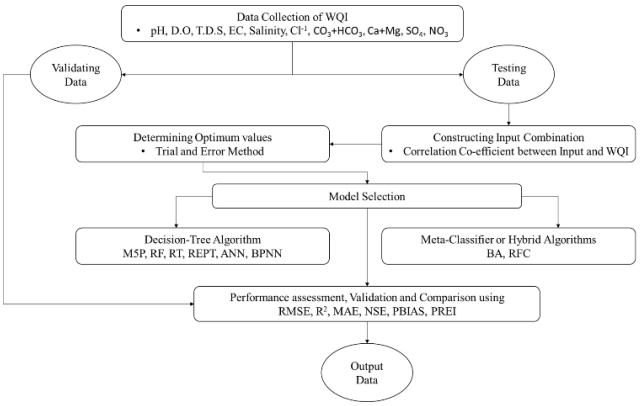


FIGURE 2. Flow chart explaining the involved steps

Figure 2 shows that the data collection was initially performed, and followingly different WQ parameters were calculated from the water samples. The data was then distributed into testing and validation datasets. From the testing datasets, the best input combination was identified. Finally, multiple algorithms were applied to the best varieties, and an algorithm assessment was conducted for the best possible algorithm selection to predict WQI. The detailed steps of the method are subsequently presented and discussed.

A. DATA COLLECTION AND PREPARATION

Water samples were collected from different random water wells in the study area so that they covered the area entirely. Overall, the data is collected from 39 locations. To minimize the seasonality impact, the samples were taken over two years, 2020 and 2021. Followingly, multiple WQ parameters were calculated. These parameters include PH, DO, TDS, conductivity, salinity, chloride, total alkalinity, total hardness, sulfate, nitrate, and WQI. Pakistan's WQ Index (PKWQI) was calculated using these datasets. The COMSATS University Islamabad, Wah Campus's laboratory was used for the WQ parameter calculations. Using the 10-fold cross-validation method, the dataset was partitioned into two subsets for algorithm training and testing (70:30). This ratio is one of the most popular modeling strategies for spatial [22, 23, 26, 40] and temporal [20, 22-25, 40] predictions. The PKWQI was created using the NSFWQI equation. In the index, the cleanness of water depends on the value of PKWQI. The river is cleaner if the PKWQI value is more significant (a WQI of 80 or higher denotes a clean river) and vice versa [15]. The PKWQI formula is used to determine the PKWQI, as shown in (1).

$$\text{PKWQI} = \sum_{i=1}^{n} W\text{i x SIi} \qquad (1)$$

where, W_i is the variable $I's$ weight (between 0 and 1), and SI_i is the sub-index resulting from the quality-index curve (0–100). The calculation techniques align with the NSFWQI [40, 41]. Table 1 shows the ranges of quality parameters for the PKWQI. It classifies the WQI into seven classes based on the ranges of WQI. For example, <15 WQI is classified as very low-quality water, while >85 WQI values are classified as very good quality water, as stated in table 1. The generic ruleset is that the higher the WQI value, the better the WQ will be.

TABLE 1.
RANGES OF PKWQI AND THEIR QUALITATIVE DESCRIPTIONS

Index	Range	Quality
PKWQI	<15	Very low
	15 - 29.9	Low
	30 - 44.9	Approximately low
	45 - 55	Moderate
	55.1 - 70	Approximately good
	70.1 - 85	Good
	>85	Very good

Table 2 shows the used WQ parameters and the results of multicollinearity analysis. These parameters were selected based on the literature [12, 41-45] and are among the standard characteristics used for WQ assessment. The variation inflation factor (VIF) value for all factors is less than 5 and satisfies the maximum threshold [46]. Thus, it can be stated that there is no multicollinearity present among the selected parameters.

According to the descriptive data (see Table 3), the WQI varies from 11.45 to 87.45 (the maximum value is 100). Thus, the WQ ranges from excellent to unsuitable for drinking in the study area [47]. The average pH for the training dataset is 7.9 and 7.5 for the testing dataset. Overall, the region has a very weak basic pH. The mean total hardness for the training dataset is 104.6 and 103.50 for the testing dataset, which means this area has moderately hard water. Also, if we notice the TDS values, the area processes hard water as the TDS for both training and testing datasets are 90.9 and 90.4, respectively.

TABLE 2.
USED INPUT VARIABLES AND THEIR VARIATION INFLATION FACTOR.

Variable	pH	DO	TDS	EC	Salinity	Cl-1	CO3+HCO3	Ca+Mg	SO4	NO3
VIF	0.93	0.43	1.34	1.67	1.44	1.51	1.68	1.71	1.76	1.77

TABLE 3.
THE TRAINING AND TESTING DATASET'S DESCRIPTIVE STATISTICS.

Variables	Training				Testing			
	Min	Max	Mean	Std. dev	Min	Max	Mean	Std. dev
pH	7	8.75	7.9	0.59	6.67	8.42	7.5	0.26
DO	0.49	2.2	1.6	0.37	0.16	1.87	1.3	0.04
TDS	19.3	277	90.9	66.24	18.97	276.67	90.6	65.91
Conductivity	20.3	513	170.8	142.13	19.97	512.67	170.5	141.80
Salinity	0	0.3	0.1	0.07	0	0.02	0.0	0.04
Chloride	6	14.89	10.6	2.04	5.67	14.56	10.3	1.71
Total Alkalinity	360	1200	773.8	203.36	359.67	1175	764.7	199.66
Total Hardness	28	240	104.6	59.16	26.95	238.95	103.50	58.11
Sulphate	9	119	36.5	28.75	7.95	117.95	35.47	27.70
Nitrate	5	42.7	23.0	11.06	3.95	41.65	21.91	10.01
WQI	11.45	87.45	48.6	9.58	19.55	91.63	51.4	11.32

The data were normalized (X_i') to a 0 to 1 range to increase prediction ability using the following relation [48]:

$$X_{i'} = (X_i - X_{min}) / (X_{max} - X_{min}) \quad [48]$$

Where $X_{i'}$ is the normalized value of a variable (i.e., BOD, COD, etc.), xi is the value at a given location, and X_{min} and X_{max} are the variable's minimum and maximum values.

B. CONSTRUCTING THE INPUT COMBINATION

Before modeling, the ideal input combination and the best value for each algorithm's operator must be identified. Ten factors were examined as potential inputs, and correlation coefficients (CCs) between input and WQI were used to determine the outcome, as presented in Table 4. CCs range from -1 to +1. Where -1 means strong negative relations and +1 means strong positive relations, and 0 means no relation among the two variables. Ca+Mg, SO4, and NO3 strongly relate to the WQI, while pH has no relation to the WQI. TDS and salinity have moderate relation. A total of 10 input combinations for this purpose as presented in Table 5.

NO3 was the initial variable included in the algorithm, having an excellent CC value, as shown in Table 5. The best estimate of the WQI is obtained using this variable alone; hence it is the known, accurate and effective variable. Until the final variable with the lowest CC was included (i.e., pH and other combinations), each variable with the next highest CC (i.e., SO4, then Ca+Mg, then CO3+HCO3, etc.) was added to the preceding variety. Each algorithm's most successful (i.e., most predictive) combination is determined by applying fixed input variable values (or default values) to all ten input combinations. The testing phase was evaluated using the root mean square error (RMSE) criterion.

TABLE 4.
EACH INPUT VARIABLE AND WQI HAVE A PEARSON CORRELATION COEFFICIENT.

Variable	pH	DO	T.D.S	EC	Salinity	Cl-1	CO3+HCO3	Ca+Mg	SO4	NO3
Correlation (r)	0.07	0.57	-0.34	-0.67	-0.44	-0.51	-0.68	-0.71	-0.76	-0.77

TABLE 5.
VARIOUS COMBINATIONS OF INPUTS

No.	Different Input Combinations
1.	NO3
2.	NO3, SO4
3.	NO3, SO4, Ca+Mg
4.	NO3, SO4, Ca+Mg, CO3+HCO3
5.	NO3, SO4, Ca+Mg, CO3+HCO3, EC
6.	NO3, SO4, Ca+Mg, CO3+HCO3, EC, DO
7.	NO3, SO4, Ca+Mg, CO3+HCO3, EC, DO, Cl-1
8.	NO3, SO4, Ca+Mg, CO3+HCO3, EC, DO, Cl-1, Sal
9.	NO3, SO4, Ca+Mg, CO3+HCO3, EC, DO, Cl-1, Sal, TDS
10.	NO3, SO4, Ca+Mg, CO3+HCO3, EC, DO, Cl-1, Sal, TDS, pH

C. DETERMINING THE OPERATOR'S OPTIMUM VALUES

After establishing the optimal input parameters, trial and error were used to obtain the optimal values for each algorithm's operator. Since operators have no universal optimum value (values vary per research), various values should be examined using the hit and trial approach to determine the most efficient value. To achieve this, each algorithm was run using default settings. Based on these findings, higher and lower numbers were randomly entered until the optimal value was found. The batch size for all the algorithms was set to 100, and the model was operated at 100 iterations. DT algorithms were used as classifiers, and random projection was used to filter all the algorithms. The minimum variance proportion was set as 0.001, and the number of decimal places for output values was 3. 15 hidden layers were used for the ANN algorithm to get a single output.

IV. Descriptions of the Algorithms

This study uses sixteen ML algorithms to predict WQI. The used algorithms are divided into two groups. Jupyter notebook was used to implement the algorithms and process the obtained data. The most essential used packages are TensorFlow, scikit-learn, ANFIS, and weka-pyscript, and the most important used libraries are NumPy, Matplotlib, and pandas [49-52]. In this study, the WQI was predicted and evaluated using the unique algorithms in group 1. Following that, ensemble algorithms based on the algorithms in Groups 1 and 2 were created to assess the accuracy of the WQI prediction. Finally, sixteen algorithms are analyzed and evaluated in 2 categories to choose the best algorithm to predict WQI. The two groups of algorithms are explained below:

D. GROUP 1 (DT ALGORITHMS)

This group contains six algorithms. These include M5P, RT, RF, REPT, ANN, and BPNN (back propagation neural network) as discussed below:

1) M5P

M5P, a machine-learning algorithm, is the first member of the DT group included in this study. M5P is a robust decision-tree algorithm used in various applications [53-57]. It works like a regression tree, with constants acting as the leaves [58]. The M5P algorithm was derived from the M5 algorithm given by Quinlan [59]. The classification and regression tree [26] algorithm modify M5P [48]. As the M5P method is centered on classification and regression analysis, it uses a divergence metric to generate a decision tree. It calculated continuous parameters using the decision tree with linear regression functions as nodes that produced numerical attributes.

2) RT

The RT algorithm is a well-known DT technique first developed in 2000 [60, 61]. In contrast to typical DTs, it builds DTs from a random selection of columns. In addition, RT offers flexible and quick training [62]. From the training dataset containing features and labels, the RT developed the DT by formulating its own set of rules and then used those rules to make the predictions.

3) RF

RF was suggested for the first time by Breiman [63]. Supervised ML, ensemble ML, and RT are some algorithms that fall within this category [27, 64]. The sample subsets from the original data are used in the RF algorithm. It creates a DT for each subgroup and summarizes the sub-decision

tree forecasts. The DT was built with around two-thirds of the dataset, and the algorithm is evaluated with the remaining data. This type of evaluation is known as "out-of-bag" (OOB) evaluation. More details are given in [65-69] about the utility of RF algorithms in natural science areas.

4) REPT

REPT can learn quickly where the DTs are created based on data enrichment or variance reduction [70]. Reduced-error pruning with back over-fitting is the primary approach used in this strategy. Pruning procedures are used to reduce the size of a DT. The REPT algorithm examines each node of the DT and lowers the number of branches until the tree's correctness is compromised [71]. The REPT considered each node for pruning and removed the subtrees at nodes. As per the REPT, the performance is compromised, making them leave by assigning weights. The REPT, by iteratively operating, continued the removal of nodes till the pruning became harmful.

5) ANN

ANNs are computer systems modeled after the biological neural networks that make up animal brains. An ANN is made up of artificial neurons, a collection of linked units or nodes that resemble the neurons in a biological brain [72]. The neurons were grouped into layers, and the best possible match was made for each input layer to form a single group. Signals went from the first layer (the input layer) to the last layer (the output layer) by going through the middle layer (the hidden layer). Neurons were assigned a threshold at which the signal was only transmitted once the aggregate signal exceeded it. The process was repeated many times till the convergence was achieved.

6) BPNN

The BPNN was created to solve the challenge of multi-layer perceptron training. The addition of a differentiable transfer function at each node of the network and using error back-propagation to adjust the internal network weights after each training period were the BPNN's key innovations. Backpropagation helped fine-tune the weights of every neural based on the error rate obtained in the previous epoch during the iterations. Proper tuning of the weights ensured lower error rates, thus, making BPNN consistent by increasing its generalization. Because of its capacity to construct complicated decision boundaries in the feature space, the BPNN was chosen as a classifier by Hornik et al. [73].

E. GROUP 2 (HYBRID ALGORITHMS)

This group includes ten algorithms: BA-RT, BA-RF, BA-REPT, RT-ANN, RFC-M5P, RFC-RF, RFC-RT, BA-M5P, RFC-REPT, and ANN-ANFIS. These are the combination of various algorithms with four key algorithms: BA, RFC, ANN, and BPNN. ANN and BPNN have already been discussed, while BA, RFC, and ANFIS are subsequently discussed.

1) BA

Breiman [74] proposed the idea of BA predictors to combine forecasters and increase single prediction accuracy. The "bootstrap aggregating" process is known as "bagging" [75-77]. BA was used to train the M5P, RF, RT, and REPT base learners to predict WQI in this research, resulting in four hybrid algorithms: BA-M5P, BA-RF, BA-RT, and BA-REPT.

2) RFC

RFC is a data-classification approach that uses randomly filtered data [78]. The filter uses the training dataset with a specific structure [79]. RFC was used to train the M5P, RF, RT, and REPT base learners to predict WQI, similar to how the bagging and CVPS algorithms were trained, resulting in four hybrid algorithms: RFC-M5P, RFC-RF, RFC-RT, and RFC-REPT. The validation dataset was run through a filter to ensure the algorithm was of good quality without affecting its structure. A random number of seeds were used to create each base classifier using the same data. The result was the average of the classifiers' predictions. Followingly, the class was utilized to construct a random classifier committee. The committee members were then categorized, and the randomizable interface was implemented.

3) ANFIS

Adaptive Neuro-Fuzzy Inference System (ANFIS) uses two sets of the algorithm as a single unit, i.e., Fuzzy Logic [18] and ANNs. Because of this combination, this algorithm handles complex large data structures very quickly and efficiently and speeds up the execution time. First, the ANFIS mapped input characteristics into input membership functions (MFs) and then input MF to a set of if-then rules. Followingly, the rules were converted to a set of output characteristics, and then the output characteristics to output MFs. Lastly, the output MFs were transformed into a single-valued output or a decision associated with the output.

V. Comparison and assessment of algorithms

Six statistical metrics were used to analyze the algorithms quantitatively. These metrics have been used in the past by several researchers to assess the performance of data mining algorithms. The used metrics include the root mean square error (RMSE) [1], coefficient of determination (R^2) [80], mean absolute error (MAE) [81], Nash- Sutcliffe efficiency [82], percentage of bias (PBIAS) [83], and percent of relative error index (PREI) [84]. RMSE is the difference between the actual and predicted value. The greater the RMSE, the higher the error in the model. MAE is the mean of errors among all actual and predicted values. The lower the mean error, the more reliable will be the prediction model.

Similarly, R^2 depicts the fitness of the model against the actual values. A higher R^2 means a high correlation between actual and predicted values, and the model generates good results. NSE calculates the relative magnitude of the residual variance compared to the measured data variance. Its values range from negative infinity to 1. Where 1 means perfect answer and prediction of values, and values close to 1 show higher accuracy. PBIAS defines whether the predicted data

is overestimated or underestimated than the actual dataset. Its optimal value is 0, which means perfect estimation, and values low or higher than 0 mean overstated or overestimated, respectively. Finally, PREI calculates the error percentage. The higher the ratio, the higher the error would be. Overall, all of these parameters give information on how accurate the model is and which model has what type of limitations, i.e., the model provides an overestimated prediction, the model is not fit, etc. These parameters are calculated by using the following relations from Breiman et al. [83] and Breiman [84].

$$RMSE = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} (WQI_{predicted} - WQI_{measured})^2 \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (WQI_{measured} - WQI_{predicted})^2}{\sum_{i=1}^{n} (WQI_{measured} - \overline{WQI_{measured}})^2} \quad (4)$$

$$MAE = \frac{1}{n} \left| WQI_{measured} - WQI_{predicted} \right| \quad (5)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n} (WQI_{measured} - WQI_{predicted})^2}{\sum_{i=1}^{n} (WQI_{measured} - \overline{WQI_{measured}})^2} \quad (6)$$

$$PBIAS = \left(\frac{\sum_{i=1}^{n} (WQI_{measured} - WQI_{predicted})}{\sum_{i=1}^{n} WQI_{predicted}} \right) x\ 100 \quad (7)$$

$$PREI = \left(\frac{WQI_{measured} - WQI_{predicted}}{WQI_{measured}} \right) x\ 100 \quad (8)$$

where, $WQI_{predicted}$ and $WQI_{measured}$ are the predicted and measured WQI mean values, respectively.

Visual comparisons were also performed to evaluate the algorithms. Scatter plots and box plots were two approaches used for visual comparisons. Scatter plots are frequently used to assess algorithm performance and study the distribution of datasets employed [85, 86]. For example, scatter plots are used to study the data organization and density. Box plots are a standard tool for assessing the density and distribution of datasets and findings. The datasets or results are separated into four data quartiles. It is possible to look at extreme values (minimums and maximums), medians, and first (upper) and third (lower) quartile projections. Such a boxplot helps to understand how all these models are calculating the WQI and their ranges, which are used to compare the accuracy and overall results among all models.

VI. Results and discussions

Following the holistic method adopted in this study, the results of pertinent analyses are present as follows:

A. THE IDEAL INPUT COMBINATION

Different input combinations based on CCs were constructed using a variety of WQ characteristics, as presented in Table 5. pH emerged as the least relevant predictor of WQI when using the previously submitted equations to calculate it. The same has been indicated by [87] and [13]. On the other hand, pH was the most critical predictor of WQI in research by Mohammadpour et al. [88], which is the opposite of this study's findings. The 16 algorithms were trained using the ten input combinations discussed previously. A testing dataset to assess these combinations, as presented in Table 6, and the most effective was chosen for modeling and further study. The results reveal how well the algorithms fit with the training dataset. These data points were not utilized in the algorithm's evaluation. The best possible combination was identified based on the testing data RMSE value (Table 6). Since all models were built on a training dataset, this table only specifies how the models fit with the training dataset. From the testing dataset, it can be seen that, on average, five combinations have the lowest RMSE. Still, combinations of more than 5 algorithms are also close to the lowest RMSE, which indicates the higher dataset does not contribute much to error. However, lower than five combinations have relatively high RMSE, which is understood because a lower dataset would have more errors because of the non-availability of data.

As evident from Table 6, combination 4 gives the lowest RSME value for M5P; combination 7 gives the lowest RSME value for RT, and combination 4 gives the lowest RSME value for BA-RT. Similarly, combination 4 gives the lowest RSME value for BA-RF, RF, BA-REPT, RT-ANN, REPT, RFC-RT, RFC-M5P, BA-M5P, ANN, RFC-REPT, ANN-ANFIS, and BPNN, and combination 6 gives the lowest RSME value for RFC-RF. Hence, combination 4 (NO3, SO4, Ca+Mg, CO3+HCO3), combination 6 (NO3, SO4, Ca+Mg, CO3+HCO3, EC, DO), and combination 7 (NO3, SO4, Ca+Mg, CO3+HCO3, EC, DO, Cl-1) are the best to estimate WQI and obtain the lowest RMSE values during testing.

B. ALGORITHM'S PERFORMANCE

The 16 algorithms were tested (Figures. 3 and 4). As per the observations, all of the algorithms functioned well. Of all the algorithms, RT-ANN, BA-RT, RF, BA-RF, and BA-M5P have the highest prediction power. All algorithms were validated as the predicted WQI was compared with measured WQI for each model at each testing dataset. It can be seen that all models performed well, but RT-ANN, BA-RT, RF, and BA-RF models predicted the best prediction.

Figure 3 shows how measuring and predicting WQI differ at each testing datapoint and how big the difference is among them. Again, all models performed well; no significant deviation between measured and predicted can be seen. Also, no pattern can be identified among all models identified as an error, so overall, all models gave reliable results.

TABLE 6.
RMSE OF VARIOUS INPUTS BASED ON ALGORITHMS.

Algorithms	Phase	RMSE of Various Input Combinations									
		1	2	3	4	5	6	7	8	9	10
M5P	Train	9.25	4.62	4.83	2.67	3.14	2.43	2.43	2.37	2.56	2.82
	Test	10.19	5.63	5.48	3.57	4.82	3.96	4.15	4.18	4.23	4.2
RT	Train	6.52	1.47	2.35	0.82	1.92	1.24	1.26	1.34	1.54	1.41
	Test	11.79	5.54	5.43	3.15	3.74	3.12	3.01	3.16	3.23	3.13
BA-RT	Train	6.018	0.65	0.86	0.27	1.16	0.45	0.43	0.36	0.29	0.31
	Test	12.98	7.13	7.15	4.33	4.88	4.74	5.07	5.12	5.32	5.18

BA-RF	Train	9.01	3.98	4.35	2.69	3.93	3.26	3.38	3.45	3.55	3.54
	Test	10.24	5.59	6.02	4.11	5.26	4.73	4.79	4.74	4.82	4.72
RF	Train	9.32	4.54	4.64	2.53	3.21	2.57	2.77	2.92	2.95	2.82
	Test	10.11	5.34	5.39	3.95	5.21	4.31	4.44	4.45	4.52	4.38
BA-REPT	Train	7.12	2.39	3.06	1.34	2.33	1.67	1.66	1.65	1.62	1.47
	Test	11.52	5.46	5.37	3.12	3.76	3.15	3.17	3.15	3.24	3.25
RT-ANN	Train	6.72	1.78	2.62	1.04	2.19	1.56	1.58	1.64	1.6	1.47
	Test	11.97	5.78	5.71	3.28	3.94	3.56	3.55	3.48	3.49	4.14
REPT	Train	8.18	3.62	3.95	1.92	3.08	2.45	2.56	2.54	2.57	2.56
	Test	10.53	5.34	5.54	3.29	4.13	3.65	3.65	3.63	3.69	3.56
RCF-M5P	Train	9.25	4.62	4.73	2.68	3.14	2.43	2.56	2.56	2.56	2.53
	Test	10.19	5.63	5.47	3.97	5.04	4.44	4.59	4.58	4.62	4.51
RFC-RF	Train	6.51	1.46	2.28	0.79	1.87	1.21	1.27	1.34	1.4	1.3
	Test	11.86	5.58	5.37	3.12	3.7	3.07	3.16	3.23	3.32	3.34
RFC-RT	Train	6.01	0.28	0.87	0.17	1.13	0.47	0.47	0.44	0.42	0.27
	Test	12.98	7.13	7.09	4.38	5.99	5.04	5.17	5.31	5.4	5.45
BA-M5P	Train	8.95	4.45	5.07	3.07	4.02	3.49	3.59	3.61	3.66	3.67
	Test	10.27	6.16	6.23	4.33	5.42	4.8	4.89	4.92	4.95	4.84
ANN	Train	9.26	4.62	4.78	2.77	3.95	3.87	4.15	4.54	4.89	4.99
	Test	10.19	5.63	5.74	3.57	4.67	4.06	4.11	4.03	4.06	4.02
RFC-REPT	Train	6.49	1.29	2.38	0.82	2.05	1.85	2.11	2.17	2.45	2.25
	Test	11.72	5.64	5.64	3.31	4.23	3.75	3.85	3.95	4.05	4.25
ANN-ANFIS	Train	6.01	0.51	0.86	0.59	1.11	0.44	0.44	0.38	0.36	0.2
	Test	12.95	7.26	7.42	4.77	8.21	6.54	6.59	7.12	7.32	7.25
BPNN	Train	8.93	4.36	4.52	2.71	5.92	4.63	5.49	5.75	6.03	6.02
	Test	10.74	5.36	6.52	4.71	7.92	6.63	7.49	7.75	7.03	7.02

Similar to Figure 3 a-p, Figure 4 a-p show how to fit the models by plotting the measured WQI and predicted WQI for all the models. It is another representation of predicted and measured values.

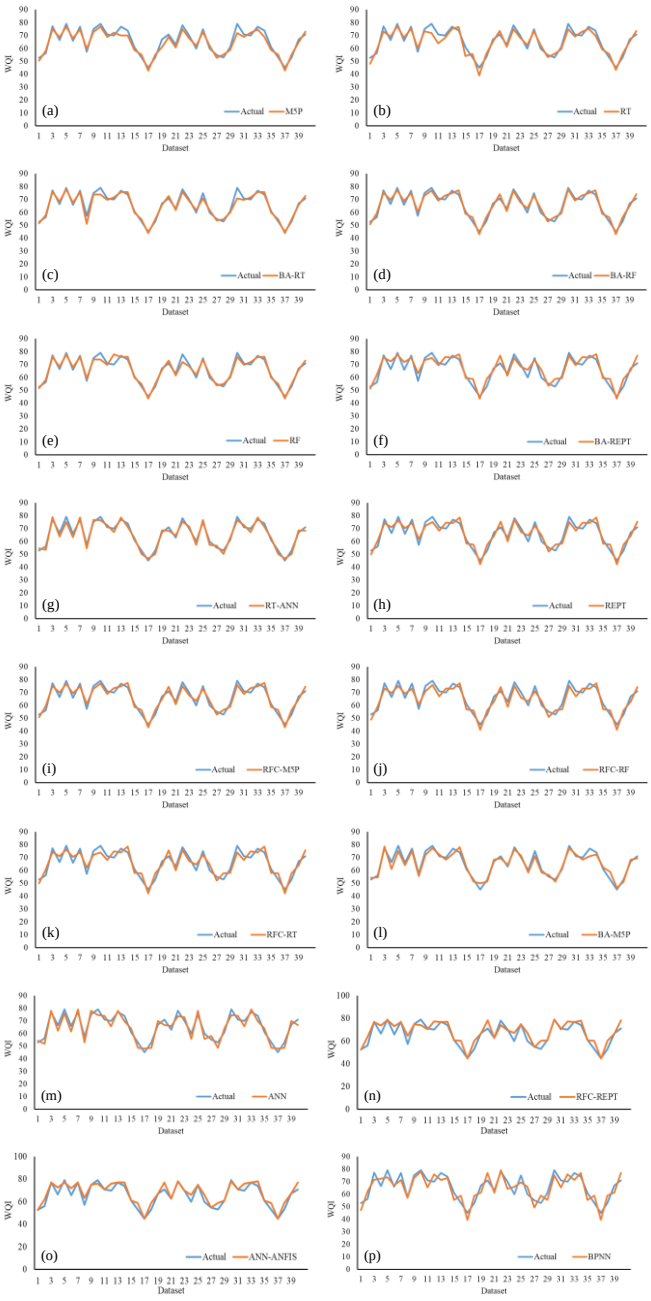


FIGURE 3. Time variation graph of predicted and measured value during the testing phase.

Figure 3 a-p depicts the variation between the predicted and measured values. Similarly, Figure 4 a-p depicts the fitness of the model. Figure 4 a-p shows that all models have good fitness as most data points fall near the straight line, which is nearly perfect for model reliability. The RF algorithm has the minimum error among the standalone algorithms. The error ranges for RT and REPT were also between ±10; these algorithms failed to estimate the results accurately. The predictive value of standalone algorithms was improved by hybrid algorithms, notably the bagging algorithm (compare Figure 4a with e, c with g, and d and h). The RFC-RT, RFC-

M5P, CVPS-M5P, M5P, and BAM5P algorithms are highly accurate at predicting the maximum WQI values, as shown by the box plots of measured and estimated WQI values. Only RFC-RF correctly estimated the lower values (see Figure 5).

PREI, which evaluates the efficiency of algorithms on the potential to over-or underestimates the WQI, was used to analyze the results, as shown in Figure 5. Though it has been established that all models predict reliable WQI, one factor still needs to be addressed. It must be checked if the model overestimates or underestimates the outcome. Only then can model accuracy be judged (when it predicts nearly to actual values, i.e., having a lower RMSE value). However, as shown, all the values are overestimated or underestimated, which means there is something wrong with the model, and it needs some refinement or model tuning. This overestimation of underestimation can be estimated by PREI calculation. Figure 5 shows that all the models have close to zero PREI values.

Further, all models have different PREI values for each testing dataset, which means that the model performed reasonably accurately. The model has not predicted biased values, i.e., overestimated or underestimated. It can be seen that RT-ANN, BA-RT, and BA-RF models performed well in PREI analysis as they have close to zero PREI value because usually, the ±10 PREI range is considered to be acceptable. Nevertheless, directly analyzing the algorithms' predictions to compare their effectiveness has drawbacks. Those with stronger prediction powers are easier to spot but determining the optimal algorithm and success ranking is complex.

As a result, quantitative data that gives more substantial evidence of each algorithm's performance is required, as presented in Table 7. Boxplots access the dataset's mean, range, and overall distributions. Hence to compare how our models are predicting among all dataset's boxplot was used, as shown in Figure 6.

The boxplot shows that all models range almost equally and have similar distribution except RT and BA-RF, which have higher ranges and distribution. Since the difference is minimal, it cannot be declared an outlier. Furthermore, the boxplots show that the best models are RT-ANN, BA-RT, RF, BA-RF, BA-M5P, and M5P, while RCF-RT, RCF-REPT BPNN predicted the lowest values and have relatively lower accuracy. The hybrid RT-ANN (R^2 = 0.951) had the highest prediction success (R^2 = 0.75), while the BPNN (R^2 = 0.752) had the lowest. The RT-ANN algorithm had the best MAE (2.284) and the lowest RMSE (2.319). An algorithm has excellent prediction ability when the NSE is 0.75 to 1 [89].

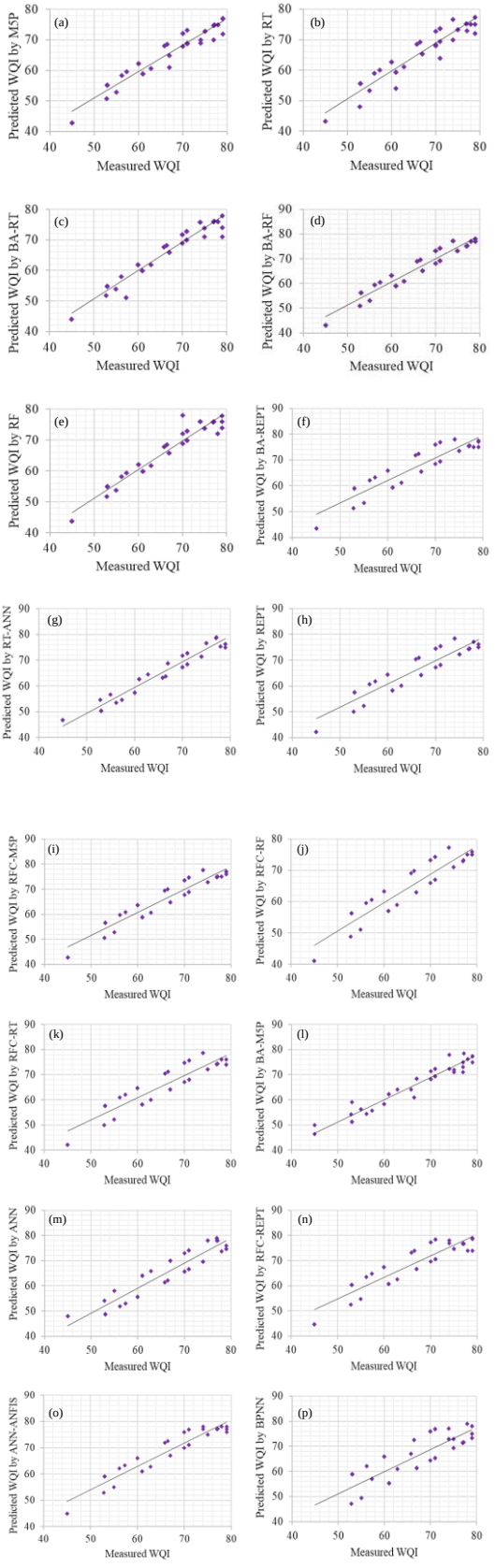


FIGURE 4. Scatter plots of predicted measured values produced by all algorithms during testing

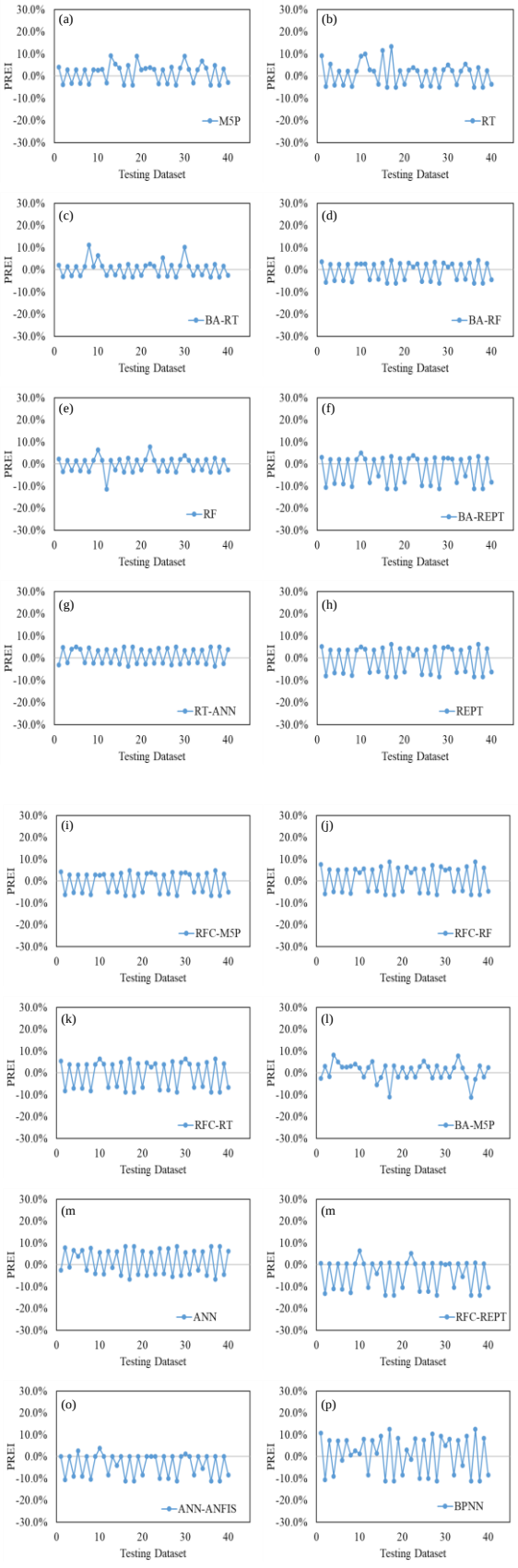


FIGURE 5. Error graph of estimated value compared to measured data in the testing phase.

TABLE 7.
FINDING THE BEST ALGORITHM FOR PREDICTING THE WQI

Algorithm	RSQ	RMSE	MAE	NSE	PBIAS	Rank
M5P	0.929	2.919	2.625	0.913	-1.127	6
RT	0.9	3.241	2.867	0.893	-0.85	12
BA-RT	0.945	2.362	1.868	0.943	-0.43	2
BA-RF	0.936	2.506	2.406	0.936	0.3	4
RF	0.942	2.376	1.943	0.942	0.05	3
BA-REPT	0.858	3.988	3.43	0.838	1.41	13
RT-ANN	0.951	2.319	2.248	0.945	-0.64	1
REPT	0.867	3.636	3.519	0.865	0.28	11
RFC-M5P	0.915	2.893	2.812	0.915	0.23	7
RFC-RF	0.872	3.651	3.633	0.864	-0.83	10
RFC-RT	0.85	3.848	3.729	0.849	0.24	14
BA-M5P	0.935	2.625	2.237	0.93	-0.62	5
ANN	0.888	3.647	3.501	0.864	-1.02	9
RFC-REPT	0.84	4.699	3.344	0.775	2.53	15
ANN-ANFIS	0.906	3.792	2.568	0.853	2.26	8
BPNN	0.752	5.192	4.875	0.725	-0.72	16

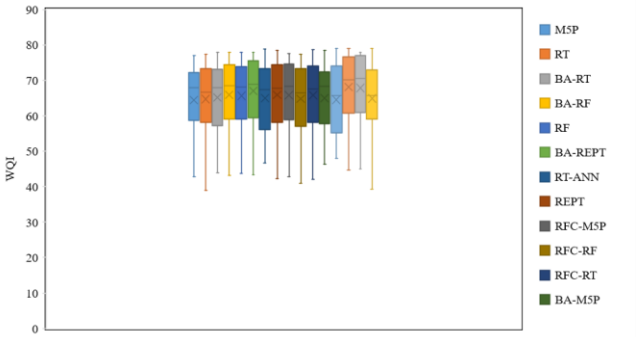


FIGURE 6. Box plot of applied algorithms for algorithm performance.

As a result, all algorithms performed admirably, but RT-ANN outperformed the competitors (NSE = 0.945). All algorithms except BA-RF, RF, BA-REPT, REPT, RFC-M5P, RFC -REPT, and ANN-ANFIS overestimated WQI, according to the PBIAS metric. Based on their performance results, the algorithm's ranking from best to worst is RT-ANN, BA-RF, RF, BA-RF, BA-M5P, M5P, RFC-M5P ANN-ANFIS, RT, RFC-RF, REPT, ANN, BA-REPT, RFC-RT, RFC-REPT, BPNN.

C. DISCUSSION

To forecast WQI along the KKH stretch from Gilgit to Khunjerab Pass, six standalone tree-based algorithms (M5P, RF, RT, REPT, ANN, and BPNN) were used in this study. In addition, ten new hybrid algorithms were created by merging the standalone algorithm with BA and RFC algorithms. The sixteen algorithms were compared in terms of performance. Previously researchers [12, 21, 23] have examined the predictive power of several independent tree-based algorithms using the neuron-based algorithm (ANFIS). These were hybridized with meta-heuristic optimization techniques. The findings of the previously conducted studies [12, 21, 23] show that isolated neuron-based algorithms have low prediction capacities due to significant flaws. Hybridization can considerably improve their forecasts. Their results also show that standalone tree-based algorithms perform similarly to ANFIS hybridized with meta-heuristic

optimization, which outperforms tree-based algorithms in prediction power.

In the current study, hybrid algorithms improved the performance of specific independent tree-based algorithms, but not all. On their own, tree-based algorithms offer high predictive potential. For example, the best algorithm was BA-RT, with an R^2 value of 0.941 in a relevant study, while in this research, the best algorithm is RT-ANN having an R^2 of 0.951. Overall, the comparison shows the improvement in algorithms, e.g., published M5P has an R^2 value of 0.923, and this research has 0.929, etc.

Apart from the structure of an algorithm, determining the appropriate mix of variables to be inputted into the algorithm is one of the most critical influences on performance. Because of the variety of point and non-point sources of pollution that generate nonlinear interactions between factors and WQ, the impact of combining variables on the result varies from catchment to catchment. Some studies failed to consider alternative variable combinations while determining the optimum set. Other researchers added all factors at the same time [6]. Similarly, some researchers used different approaches to pick the optimal input variables, such as multiple linear regression (dependent on CC) [90].

The current study shows that various input combinations have distinct outcomes. Therefore, different variable input combinations should be tried to increase performance and select the most effective set. Each algorithm may have its own "best" combo. The outcomes are determined by the structure of each algorithm and the dataset's fit to the algorithm's structure (data structure and distribution). As mentioned earlier, new proposed hybrid algorithms performed better than the existing algorithms by at least 2%. To simulate WQI, Sahoo et al. [91] utilized ANFIS, and Yaseen et al. [13] employed a hybrid ANFIS. According to our findings, all standalone and hybrid algorithms produced superior WQI predictions than any previous algorithm examined for WQI prediction. Hence, based on the results, these algorithms can be used in any part of the world for WQI estimations and prediction. These algorithms can handle large long-term datasets and lower the cost of WQI estimation as just the WQ parameters for the algorithms to predict the WQI. Modifying the inputs for the algorithms used in this research can be done to adopt the divergent effects of modeling in other regions, or perhaps it can be done with the same variable combinations.

VII. Conclusion

This study investigated the performance of six standalone (RT, EPTR, RF, and M5P) and ten hybrid data-mining algorithms (hybrids of the standalone with CVPS, RFC, and BA) algorithms for forecasting the WQI in Northern Pakistan. The goal was to develop algorithms for WQI prediction and assess the WQ in the study area. According to the modeling procedure, the essential factor of the WQI was fecal coliform concentration. BOD, NO3, DO, EC, COD, PO24, turbidity, TS, and pH were then listed in relevance. It was discovered that multiple variable combinations led to varying degrees of algorithm performance. The predicting power was the best when the algorithms' variables with the highest CCs were utilized. Low-CCs variables have a detrimental impact on predictive power. Compared to the standalone algorithms, the hybrids demonstrated an enhanced prediction accuracy rate (i.e., adequate than the standalone algorithms) as they have > 0.9 RSQ and NSE, respectively. The RT-ANN algorithm outperformed all other algorithms in terms of accuracy, with the highest RSQ value of 0.951. RF, BA-RF, BA-RT, BA-REPT, RFC-RF, RT, M5P, CVPS-M5P, RFC-M5P, BA-M5P, REPT, CVPS-REPT, CVPS-RT, RFC-REPT, and RFC-RT are in order of decreasing performance after RT-ANN. Despite having the best performance, the RT-ANN hybrid could not effectively predict severe WQI values. WQI values were overestimated by nearly all algorithms, except BA-RF, RF, BA-REPT, REPT, RFC-M5P, RFC -REPT, and ANN-ANFIS.

A. PRACTICAL AND RESEARCH IMPLICATIONS

This research compares the implementation of new and existing algorithms for WQI assessment. This is important to mention that these algorithms can give stable outputs with a short-term dataset. The stability can, however, be increased with the longer-term dataset. As a result, these algorithms may be highly efficient in emerging areas with minimal measuring networks or when gauging networks have only recently been constructed. According to our results, the recommended RT-ANN algorithm appears practical and cost-effective for assessing WQI in Northern Pakistan. In the future, relevant research can be conducted using the proposed algorithms in developed and developing countries. The proposed algorithms can become more beneficial in underdeveloped nations since the costs of testing various WQ parameters are large and may be unaffordable generally. However, local climatic modifications need to be considered before applying this algorithm.

However, the research outcomes can be valuable for the water management authorities in a way that they can take preventative measures to safeguard against the leaching of different detrimental pollutants and chemicals into the water resources, thus ensuring a relatively better WQ.

B. LIMITATIONS AND FUTURE PROSPECTS

This research has some limitations that can be potential future research areas. Firstly, the datasets used in this research were based on two years of sampling, making it a comparatively smaller sample, so the long-term analysis was impossible. The performance of these algorithms on long-term datasets can be investigated in the future. Secondly, the important WQ parameters, namely COD and BOD, were not considered in the present due to some practical limitations. In future, data over multiple years such as the last decade can be used for similar purposes.

Consequently, as the statistical and ML algorithms were used in this research, providing highly accurate results, it will be

beneficial to use deep learning algorithms, for instance, convolution neural network, to cross-check the results and compare them with this study to yield holistic results. Further, in addition to the correlation tests, other tests such as the PCA should be conducted in the future. Moreover, it would also be valuable to consider the WQ variables of COD and BOD for future research.

ACKNOWLEDGMENT

Research Supporting Project number (RSP2022R444), King Saud University, Riyadh, Saudi Arabia.

REFERENCES

[1] M. Awais *et al.*, "Assessing nitrate contamination risks in groundwater: a machine learning approach," *Applied Sciences*, vol. 11, no. 21, p. 10034, 2021.

[2] T. H. Tulchinsky and E. A. Varavikova, "Communicable Diseases," *The New Public Health*, p. 149, 2014.

[3] S. Khalid, M. Shahid, I. Bibi, T. Sarwar, A. H. Shah, and N. K. Niazi, "A review of environmental contamination and health risk assessment of wastewater use for crop irrigation with a focus on low and high-income countries," *International journal of environmental research and public health*, vol. 15, no. 5, p. 895, 2018.

[4] E. Chu and J. Karr, "Environmental impact: Concept, consequences, measurement," *Reference Module in Life Sciences*, 2017.

[5] P. M. Kopittke, N. W. Menzies, P. Wang, B. A. McKenna, and E. Lombi, "Soil and the intensification of agriculture for global food security," *Environment international*, vol. 132, p. 105078, 2019.

[6] M. Hameed *et al.*, "Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia," vol. 28, no. 1, pp. 893-905, 2017.

[7] T. Bournaris, J. Papathanasiou, B. Manos, N. Kazakis, and K. J. O. R. Voudouris, "Support of irrigation water use and eco-friendly decision process in agricultural production planning," vol. 15, no. 2, pp. 289-306, 2015.

[8] F. Rufino, G. Busico, E. Cuoco, T. H. Darrah, D. J. E. m. Tedesco, and assessment, "Evaluating the suitability of urban groundwater resources for drinking water and irrigation purposes: an integrated approach in the Agro-Aversano area of Southern Italy," vol. 191, no. 12, pp. 1-17, 2019.

[9] M. Vadiati, A. Asghari-Moghaddam, M. Nakhaei, J. Adamowski, and A. J. J. o. E. M. Akbarzadeh, "A fuzzy-logic based decision-making approach for identification of groundwater quality based on groundwater quality indices," vol. 184, pp. 255-270, 2016.

[10] A. Shalby, M. Elshemy, B. A. J. E. S. Zeidan, and P. Research, "Assessment of climate change impacts on water quality parameters of Lake Burullus, Egypt," vol. 27, no. 26, pp. 32157-32178, 2020.

[11] D. Sharma and A. Kansal, "Water quality analysis of River Yamuna using water quality index in the national capital territory, India (2000–2009). Appl Water Sci 1: 147–157," ed, 2011.

[12] D. T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, and N. Kazakis, "Improving prediction of water quality indices using novel hybrid machine-learning algorithms," *Science of the Total Environment*, vol. 721, p. 137612, 2020.

[13] Z. M. Yaseen, M. M. Ramal, L. Diop, O. Jaafar, V. Demir, and O. J. W. R. M. Kisi, "Hybrid adaptive neuro-fuzzy models for water quality index estimation," vol. 32, no. 7, pp. 2227-2245, 2018.

[14] C. Iticescu *et al.*, "Lower Danube water quality quantified through WQI and multivariate analysis," vol. 11, no. 6, p. 1305, 2019.

[15] W. C. Leong, A. Bahadori, J. Zhang, and Z. J. I. J. o. R. B. M. Ahmad, "Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM)," vol. 19, no. 2, pp. 149-156, 2021.

[16] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, pp. 1-21, 2021.

[17] M. J. Alizadeh, M. R. Kavianpour, M. Danesh, J. Adolf, S. Shamshirband, and K.-W. Chau, "Effect of river flow on the quality of estuarine and coastal waters using machine learning models," *Engineering Applications of Computational Fluid Mechanics*, vol. 12, no. 1, pp. 810-823, 2018/01/01 2018.

[18] K. Kargar *et al.*, "Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms," vol. 14, no. 1, pp. 311-322, 2020.

[19] T. M. Tung and Z. M. J. J. o. H. Yaseen, "A survey on river water quality modelling using artificial intelligence models: 2000–2020," vol. 585, p. 124670, 2020.

[20] K. Khosravi, L. Mao, O. Kisi, Z. M. Yaseen, and S. J. J. o. H. Shahid, "Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile," vol. 567, pp. 165-179, 2018.

[21] K. Khosravi *et al.*, "Meteorological data mining and hybrid data-intelligence models for reference evaporation simulation: A case study in Iraq," vol. 167, p. 105041, 2019.

[22] K. Khosravi *et al.*, "Stochastic modeling of groundwater fluoride contamination: Introducing lazy learners," vol. 58, no. 5, pp. 723-734, 2020.

[23] K. Khosravi *et al.*, "A comparative assessment of flood susceptibility modeling using multi-criteria decision-making analysis and machine learning methods," vol. 573, pp. 311-323, 2019.

[24] A. Sharafati *et al.*, "The potential of novel data mining models for global solar radiation prediction," vol. 16, no. 11, pp. 7147-7164, 2019.

[25] Z. S. Khozani, K. Khosravi, B. T. Pham, B. Kløve, W. H. M. Wan Mohtar, and Z. M. J. J. o. H. Yaseen, "Determination of compound channel apparent shear stress: application of novel data mining models," vol. 21, no. 5, pp. 798-811, 2019.

[26] B. T. Pham *et al.*, "A comparison of Support Vector Machines and Bayesian algorithms for landslide susceptibility modelling," vol. 34, no. 13, pp. 1385-1407, 2019.

[27] H. Nguyen, C. Drebenstedt, X.-N. Bui, and D. T. J. N. R. R. Bui, "Prediction of blast-induced ground vibration in an open-pit mine by a novel hybrid model based on clustering and artificial neural network," vol. 29, no. 2, pp. 691-709, 2020.

[28] B. Xiang, C. Zeng, X. Dong, and J. J. A. Wang, "The application of a decision tree and stochastic forest model in summer precipitation prediction in Chongqing," vol. 11, no. 5, p. 508, 2020.

[29] P. Aghelpour, B. Mohammadi, S. M. J. T. Biazar, and A. Climatology, "Long-term monthly average temperature forecasting in some climate types of Iran, using the models SARIMA, SVR, and SVR-FA," vol. 138, no. 3, pp. 1471-1480, 2019.

[30] K. F. Fung, Y. F. Huang, C. H. Koo, M. J. J. o. W. Mirzaei, and C. Change, "Improved SVR machine learning models for agricultural drought prediction at downstream of Langat River Basin, Malaysia," vol. 11, no. 4, pp. 1383-1398, 2020.

[31] B. B. Hazarika, D. Gupta, and M. Berlin, "A comparative analysis of artificial neural network and support vector regression for river suspended sediment load prediction," in *First international conference on sustainable technologies for computational intelligence*, 2020, pp. 339-349: Springer.

[32] F. Granata, S. Papirio, G. Esposito, R. Gargano, and G. J. W. De Marinis, "Machine learning algorithms for the forecasting of wastewater quality indicators," vol. 9, no. 2, p. 105, 2017.

[33] F. Kayaalp, A. Zengin, R. Kara, S. J. N. C. Zavrak, and Applications, "Leakage detection and localization on water transportation pipelines: a multi-label classification approach," vol. 28, no. 10, pp. 2905-2914, 2017.

[34] F. Kamyab-Talesh, S.-F. Mousavi, M. Khaledian, O. Yousefi-Falakdehi, and M. J. W. R. Norouzi-Masir, "Prediction of Water Quality Index by Support Vector Machine: a Case Study in the Sefidrud Basin, Northern Iran," vol. 46, no. 1, 2019.

[35] Y. Wang *et al.*, "Rapid removal of Pb (II) from aqueous solution using branched polyethylenimine enhanced magnetic carboxymethyl chitosan optimized with response surface methodology," vol. 7, no. 1, pp. 1-11, 2017.

[36] O. Kisi, A. H. Dailr, M. Cimen, and J. J. J. o. H. Shiri, "Suspended sediment modeling using genetic programming and soft computing techniques," vol. 450, pp. 48-58, 2012.

[37] A. Maqsoom *et al.*, "Landslide susceptibility mapping along the China Pakistan Economic Corridor (CPEC) route using multi-criteria decision-making method," *Modeling Earth Systems and Environment*, pp. 1-15, 2021.

[38] A. Maqsoom *et al.*, "A GIS-based DRASTIC model and an adjusted DRASTIC model (DRASTICA) for groundwater susceptibility assessment along the China–Pakistan Economic Corridor (CPEC) route," *ISPRS International Journal of Geo-Information*, vol. 9, no. 5, p. 332, 2020.

[39] F. Othman, A. E. ME, and I. J. J. o. E. M. Mohamed, "Trend analysis of a tropical urban river water quality in Malaysia," vol. 14, no. 12, pp. 3164-3173, 2012.

[40] M. Gholami, E. N. Ghachkanlu, K. Khosravi, and S. J. J. o. E. S. S. Pirasteh, "Landslide prediction capability by comparison of frequency ratio, fuzzy gamma and landslide index method," vol. 128, no. 2, pp. 1-22, 2019.

[41] F. Kamyab-Talesh, S.-F. Mousavi, M. Khaledian, O. Yousefi-Falakdehi, and M. Norouzi-Masir, "Prediction of water quality index by support vector machine: a case study in the Sefidrud Basin, Northern Iran," *Water Resources*, vol. 46, no. 1, pp. 112-116, 2019.

[42] C. Iticescu *et al.*, "Lower Danube water quality quantified through WQI and multivariate analysis," *Water*, vol. 11, no. 6, p. 1305, 2019.

[43] M. Hameed, S. S. Sharqi, Z. M. Yaseen, H. A. Afan, A. Hussain, and A. Elshafie, "Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia," *Neural Computing and Applications*, vol. 28, no. 1, pp. 893-905, 2017.

[44] F. Granata, S. Papirio, G. Esposito, R. Gargano, and G. De Marinis, "Machine learning algorithms for the forecasting of wastewater quality indicators," *Water*, vol. 9, no. 2, p. 105, 2017.

[45] R. Barzegar, A. Asghari Moghaddam, J. Adamowski, and E. Fijani, "Comparison of machine learning models for predicting fluoride contamination in groundwater," *Stochastic Environmental Research and Risk Assessment*, vol. 31, no. 10, pp. 2705-2718, 2017.

[46] B. Aslam *et al.*, "Evaluation of different landslide susceptibility models for a local scale in the Chitral District, Northern Pakistan," *Sensors*, vol. 22, no. 9, p. 3107, 2022.

[47] W. H. M. W. Mohtar, K. N. A. Maulud, N. S. Muhammad, S. Sharil, and Z. M. J. E. P. Yaseen, "Spatial and temporal risk quotient based river assessment for water resources management," vol. 248, pp. 133-144, 2019.

[48] S. Shifath, M. F. Khan, and M. J. a. p. a. Islam, "A transformer based approach for fighting COVID-19 fake news," 2021.

[49] C. J. Beckham, M. A. Hall, and E. Frank, "WekaPyScript: Classification, regression, and filter schemes for WEKA implemented in Python," 2016.

[50] A. A. Chowdhury, K. T. Hasan, and K. K. S. Hoque, "Analysis and prediction of COVID-19 pandemic in Bangladesh by using ANFIS and LSTM network," *Cognitive Computation*, vol. 13, no. 3, pp. 761-770, 2021.

[51] M. Kumar, "Surface Roughness Prediction Using ANFIS and Validation with Advanced Regression Algorithms," in *International Conference on Intelligent and Fuzzy Systems*, 2020, pp. 238-245: Springer.

[52] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825-2830, 2011.

[53] S. N. Almasi, R. Bagherpour, R. Mikaeil, Y. Ozcelik, H. J. G. Kalhori, and G. Engineering, "Predicting the building stone cutting rate based on rock properties and device pullback amperage in quarries using M5P model tree," vol. 35, no. 4, pp. 1311-1326, 2017.

[54] A. Behnood, V. Behnood, M. M. Gharehveran, K. E. J. C. Alyamac, and B. Materials, "Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm," vol. 142, pp. 199-207, 2017.

[55] L. Lin, Q. Wang, A. W. J. A. A. Sadek, and Prevention, "A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations," vol. 91, pp. 114-126, 2016.

[56] E. K. Onyari, F. J. I. J. o. I. Ilunga, Management, and Technology, "Application of MLP neural network and M5P model tree in predicting streamflow: A case study of Luvuvhu catchment, South Africa," vol. 4, no. 1, p. 11, 2013.

[57] C. Zhan, A. Gan, and M. J. I. T. o. I. T. S. Hadi, "Prediction of lane clearance time of freeway incidents using the M5P tree algorithm," vol. 12, no. 4, pp. 1549-1557, 2011.

[58] Y. Wang and I. H. Witten, "Induction of model trees for predicting continuous classes," 1996.

[59] J. R. Quinlan, "Learning with continuous classes," in *5th Australian joint conference on artificial intelligence*, 1992, vol. 92, pp. 343-348: World Scientific.

[60] D. J. S. a. Aldous, "The continuum random tree. II. An overview," vol. 167, pp. 23-70, 1991.

[61] D. J. T. A. o. P. Aldous, "The continuum random tree III," pp. 248-289, 1993.

[62] S. M. LaValle, "Rapidly-exploring random trees: A new tool for path planning," 1998.

[63] L. J. M. l. Breiman, "Random forests," vol. 45, no. 1, pp. 5-32, 2001.

[64] X.-N. Bui, H. Nguyen, H.-A. Le, H.-B. Bui, and N.-H. J. N. R. R. Do, "Prediction of blast-induced air over-pressure in open-pit mine: assessment of different artificial intelligence techniques," vol. 29, no. 2, pp. 571-591, 2020.

[65] E. J. M. Carranza and A. G. J. N. R. R. Laborte, "Data-driven predictive modeling of mineral prospectivity using random forests: a case study in Catanduanes Island (Philippines)," vol. 25, no. 1, pp. 35-50, 2016.

[66] P. O. Gislason, J. A. Benediktsson, and J. R. J. P. r. l. Sveinsson, "Random forests for land cover classification," vol. 27, no. 4, pp. 294-300, 2006.

[67] J. Wang, R. Zuo, and Y. J. N. R. R. Xiong, "Mapping mineral prospectivity via semi-supervised random forest," vol. 29, no. 1, pp. 189-202, 2020.

[68] Z. Wang, R. Zuo, and Y. J. N. R. R. Dong, "Mapping geochemical anomalies through integrating random forest and metric learning methods," vol. 28, no. 4, pp. 1285-1298, 2019.

[69] S. Zhang, K. Xiao, E. J. M. Carranza, and F. J. N. R. R. Yang, "Maximum entropy and random forest modeling of mineral potential: Analysis of gold prospectivity in the Hezuo–Meiwu district, west Qinling Orogen, China," vol. 28, no. 3, pp. 645-664, 2019.

[70] W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar, "A comparative study of reduced error pruning method in decision tree algorithms," in *2012 IEEE International conference on control system, computing and engineering*, 2012, pp. 392-397: IEEE.

[71] P. Kapoor, R. Rani, and R. J. I. J. E. R. G. S. JMIT, "Efficient decision tree algorithm using J48 and reduced error pruning," vol. 3, no. 3, pp. 1613-1621, 2015.

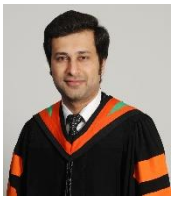
[72] J. Zou, Y. Han, and S.-S. So, "Overview of artificial neural networks," *Artificial Neural Networks*, pp. 14-22, 2008.

[73] K. Hornik, M. Stinchcombe, and H. J. N. n. White, "Multilayer feedforward networks are universal approximators," vol. 2, no. 5, pp. 359-366, 1989.

[74] L. J. M. l. Breiman, "Bagging predictors," vol. 24, no. 2, pp. 123-140, 1996.

[75] W. Chen *et al.*, "Landslide susceptibility modeling based on gis and novel bagging-based kernel logistic regression," vol. 8, no. 12, p. 2540, 2018.
[76] G. Collell, D. Prelec, and K. R. J. N. Patil, "A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data," vol. 275, pp. 330-340, 2018.
[77] F. Moretti, S. Pizzuti, S. Panzieri, and M. J. N. Annunziato, "Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling," vol. 167, pp. 3-7, 2015.
[78] L. a. B. Asaju, P. B. Shola, N. Franklin, H. M. J. F. T. i. S. Abiola, and T. Journal, "Intrusion detection system on a computer network using an ensemble of randomizable filtered classifier, K-nearest neighbor algorithm," vol. 2, no. 1, pp. 550-553, 2017.
[79] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. J. A. S. e. n. Witten, "The WEKA data mining software: an update," vol. 11, no. 1, pp. 10-18, 2009.
[80] M. Koopialipoor, H. Tootoonchi, D. Jahed Armaghani, E. Tonnizam Mohamad, and A. Hedayat, "Application of deep neural networks in predicting the penetration rate of tunnel boring machines," *Bulletin of Engineering Geology and the Environment*, vol. 78, no. 8, pp. 6347-6360, 2019.
[81] U. Khalil, B. Aslam, U. Azam, and H. M. D. Khalid, "Time Series Analysis of Land Surface Temperature and Drivers of Urban Heat Island Effect Based on Remotely Sensed Data to Develop a Prediction Model," *Applied Artificial Intelligence*, vol. 35, no. 15, pp. 1803-1828, 2021.
[82] P. H. Gude, A. J. Hansen, R. Rasker, and B. Maxwell, "Rates and drivers of rural residential development in the Greater Yellowstone," *Landscape and Urban Planning*, vol. 77, no. 1-2, pp. 131-151, 2006.
[83] L. Breiman, J. Friedman, R. Olshen, and C. J. I.-. Stone, "Classification and regression trees (Wadsworth, Belmont, CA)," pp. 978-0412048418, 1984.
[84] L. J. U. B. T. Breiman, "Random forests," 1999.
[85] S. Kahng *et al.*, "Temporal distributions of problem behavior based on scatter plot analysis," vol. 31, no. 4, pp. 593-604, 1998.
[86] P. E. Touchette, R. F. MacDonald, and S. N. J. J. o. a. b. a. Langer, "A scatter plot for identifying stimulus control of problem behavior," vol. 18, no. 4, pp. 343-351, 1985.
[87] L. Y. Khuan, N. Hamzah, and R. Jailani, "Prediction of water quality index (WQI) based on artificial neural network (ANN)," in *Student Conference on Research and Development*, 2002, pp. 157-161: IEEE.
[88] R. Mohammadpour *et al.*, "Prediction of water quality index in constructed wetlands using support vector machine," vol. 22, no. 8, pp. 6208-6219, 2015.
[89] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. J. T. o. t. A. Veith, "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations," vol. 50, no. 3, pp. 885-900, 2007.
[90] R. Barzegar, A. A. Moghaddam, J. Adamowski, E. J. S. E. R. Fijani, and R. Assessment, "Comparison of machine learning models for predicting fluoride contamination in groundwater," vol. 31, no. 10, pp. 2705-2718, 2017.
[91] M. M. Sahoo, K. Patra, and K. J. A. P. Khatua, "Inference of water quality index using ANFIA and PCA," vol. 4, pp. 1099-1106, 2015.

Bilal Aslam (ba924@nau.edu) received a B.S. degree in Earth Sciences from the COMSATS University Islamabad, Pakistan, in 2012, then an M.S. degree in Geophysics from the Quaid-I-Azam University, Pakistan, in 2014, and an M.S. degree in Data Science from the Riphah International University, Pakistan in 2021. He served the National Space Agency of Pakistan for over five years and worked on several national and international projects for urban sustainability. He is currently pursuing a Ph.D. degree in Informatics and Computing at the Northern Arizona University, USA. He is studying the deep learning tools on geospatial datasets for urban sustainability.



Ahsen Maqsoom (ahsen.maqsoom@ciitwah.edu.pk) received his M.S. and Ph.D. degrees from the Asian Institute of Technology, Bangkok, Thailand. He is currently working as an Associate Professor at COMSATS University Islamabad, Wah Cantonment, Pakistan. He has published more than 90 research articles in peer-reviewed international journals and conferences and has authored two book chapters.

Ali Hassan Cheema (fa18-cve-071@cuiwah.edu.pk) received a B.E. degree in Civil Engineering from the COMSATS University Islamabad, Pakistan, in 2022. He worked on several projects related to town planning at the local level. His core interest areas are data analytics tools, data modeling, and town planning.

FAHIM ULLAH (fahim.ullah@usq.edu.au) has been a lecturer in construction project management at the University of Southern Queensland since 2021. He holds a Ph.D. degree from the School of Built Environment, University of New South Wales (UNSW) in Sydney, Australia, and has been a casual lecturer at the same institute for four years. He also taught various courses in Project Management at the University of Sydney as a lead lecturer. Previously he worked for three years as a lecturer at the National University of Sciences and Technology (NUST) Pakistan, where he taught the courses of Construction Engineering and Management and Project Management at three schools. Further, he has more than two years of the industry as an Assistant Manager (Planning) and Planning Engineer. His research interests are in Construction Management, Project Management, Smart Cities, Digital Technologies, and Disruptive Innovation. He has been awarded multiple research grants and best paper awards. Fahim has published more than 70 high-quality research articles on construction, project, smart cities, real estate, and property management. In addition, Fahim has edited multiple special issues in Q1 journals related to digital disruptions in the built environment and industry 5.0 technologies.

Abdullah Alharbi (arharbi@ksu.edu.sa) is an assistant professor of Computer Science at King Saud University (KSU), Riyadh, Saudi Arabia. Alharbi is currently the Dean of the College of Applied Computer Sciences at KSU, Muzahmiyah Branch. Alharbi is also the CEO of the Information Security Association (Hemaya), a non-profit organization. He is also a Research Fellow at the Center of Excellence for Information Assurance, KSU. He was previously the Department of Administrative Sciences Chair at Community College at KSU. He received his Ph.D. in Computer Science from the Florida Institute of Technology, Melbourne, Florida, USA. Alharbi received his Master of Science in Information Technology from Rochester Institute of Technology, Rochester, NY, USA. Alharbi has a second Master's degree in Information Assurance and Cybersecurity from the Florida Institute of Technology, Melbourne, Florida, USA, where he also got an Information Assurance and Cybersecurity Graduate Certificate. Alharbi's research interests are Wearable Devices Security, Transparent and Continuous Security, Alternative Authentication, Usable Security, and Behavioral Biometrics.

Muhammad Imran (dr.m.imran@ieee.org) is working as a Senior Lecturer in the School of Science, Engineering and Information Technology, Federation University Australia. Previously, he served King Saud University (KSU), Saudi Arabia, as an Associate Professor. His research interest includes Mobile and Wireless Networks, the Internet of Things, Big Data Analytics, Cloud/edge computing, and Information Security. He is the founding leader of the Wireless Networks and Security (WINS) research group in KSU from 2013-to 2021. His research is financially supported by several national and international grants. He has completed several international collaborative research projects with reputable universities.

Imran has published more than 300 research articles in peer-reviewed, highly-reputable international conferences (90), journals (198), editorials (15), book chapters (1), and two edited books. Many of his research articles are among the highly cited and most downloaded. His research has been cited more than 11,500 with an h-index of 55, and an i-10 index of 175 (Google Scholar). Imran has received a number of awards and fellowships. He served as an Editor in Chief for European Alliance for Innovation (EAI) Transactions on Pervasive Health and Technology and associate editor for IEEE Communications Magazine. He is serving as an associate editor for top-ranked international journals such as IEEE Network, Future Generation Computer Systems, and IEEE Access. He served/serving as a guest editor for about two dozen special issues in journals such as IEEE Communications Magazine, IEEE Wireless Communications Magazine, Future Generation Computer Systems, IEEE Access, and Computer Networks. He has been involved in about one hundred peer-reviewed international conferences and workshops in various capacities, such as a chair, co-chair, and technical program committee member. He has been consecutively awarded Outstanding Associate Editor of IEEE Access in 2018 and 2019, besides many others.