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

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Proposition of New Ensemble Data-Intelligence Models for Surface Water Quality Prediction

Ali Omran Al-Sulttani[1], Mustafa Al-Mukhtar[2], Ali B Roomi[3,4], Aitazaz Farooque[5], Khaled Mohamed Khedher[6,7], Zaher Mundher Yaseen[8,*]

[1]Department of Water resources Engineering, College of Engineering, University of Baghdad, Baghdad, Iraq ali2003ena@gmail.com
[2]Civil Engineering Department, University of Technology, Baghdad, Iraq mmalmukhtar@gmail.com
[3]Ministry of Education, Directorate of education Thi-Qar, Thi-Qar, 64001, Iraq dr.ali_bader@alayen.edu.iq
[4]College of Health and Medical Technology, Al-Ayen University, Thi-Qar, 64001 Iraq
[5]Faculty of Sustainable Design Engineering, University of Prince Edward Island, Charlottetown, PE C1A4P3, Canada; afarooque@upei.ca
[6]Department of Civil Engineering, College of Engineering, King Khalid University, Abha 61421, Saudi Arabia; kkhedher@kku.edu.sa
[7]Department of Civil Engineering, High Institute of Technological Studies, Mrezgua University Campus, Nabeul 8000, Tunisia
[8]New era and development in civil engineering research group, Scientific Research Center, Al-Ayen University, Thi-Qar, 64001, Iraq.

Corresponding author: The Last Author (e-mail: zaheryaseen88@gmail.com).

ABSTRACT An accurate prediction of water quality (WQ) related parameters is considered as pivotal decisive tool in sustainable water resources management. In this study, five different ensemble machine learning (ML) models including Quantile regression forest (QRF), Random Forest (RF), radial support vector machine (SVM), Stochastic Gradient Boosting (GBM) and Gradient Boosting Machines (GBM_H2O) were developed to predict the monthly biochemical oxygen demand (BOD) values of the Euphrates River, Iraq. For this aim, monthly average data of water temperature (T), Turbidity, pH, Electrical Conductivity (EC), Alkalinity (Alk), Calcium (Ca), chemical oxygen demand (COD), Sulfate (SO_4), total dissolved solids (TDS), total suspended solids (TSS), and BOD measured for ten years period were used in this study. The performances of these standalone models were compared with integrative models developed by coupling the applied ML models with two different feature extraction algorithms i.e., Genetic Algorithm (GA) and Principal Components Analysis (PCA). The reliability of the applied models was evaluated based on the statistical performance criteria of determination coefficient (R^2), root mean square error (RMSE), mean absolute error (MAE), Nash-Sutcliffe model efficiency coefficient (NSE), Willmott index (d), and percent bias (PBIAS). Results showed that among the developed models, QRF model attained the superior performance. The performance of the evaluated models presented in this study proved that the developed integrative PCA-QRF model presented much better performance compared with the standalone ones and with those integrated with GA. The statistical criteria of R^2, RMSE, MAE, NSE, d, and PBIAS of PCA-QRF were 0.94, 0.12, 0.05, 0.93, 0.98, and 0.3, respectively.

INDEX TERMS Semi-arid region; river water quality; biochemical oxygen demand; principal component analysis

I. INTRODUCTION

A. THE IMPORTANCE OF SURFACE WATER QUALITY MONITORING AND DETECTION

Human life is significantly reliant on the availability of water because humans depend on water for many activities such as for drinking, cooking, farming, personal hygiene, industrial and manufacturing purposes [1], [2]. Water is also important in other activities like biotransformation, electric power generation, etc. [3]. Owing to the reliance of human life on water availability, both surface and groundwater bodies are exposed to various levels of contamination from different contaminants [4], [5]. This has made the prediction of WQ a difficult task in recent times and many scholars have dedicated much effort to WQ assessment due to its importance to human life [6], [7].

A high level of stress has been experienced over the last two decades in the area of water resources in the Iraqi region due to several reasons, such as the damming of Tigris and Euphrates Rivers, variations in global climate, and the decrease in the local annual rainfall precipitation rates [8]–[10]. Water salinity is a critical issue in Iraq that affects WQ for domestic, agricultural, and industrial purposes [11], [12]. Poor drainage and irrigation practices have brought about low water table and soil salinization in the region; agricultural developments and other human activities have affected the quality of water in the Euphrates Basin. However, these impacts are not obvious at the point of water source for irrigation. Therefore, WQ management is necessary for the effective management of all water-related resources [13].

B. MACHINE LEARNING MODELS LITERATURE REVIEW

The need for effective, dependable, accurate, and flexible prediction models has increased recently due to the acknowledgment of the issue of surface water pollution, coupled with the increasing interest in WQ assessment [14]. It is expected that these models can precisely describe the mechanisms of WQ deterioration [15]. Researchers have developed the idea of surface and underground WQ modeling using soft computing tools, such as ML models owing to their reliability and accuracy [16], [17]. However, the ML models demonstrated an inability of the generalization to handle the complicated and highly nonlinear relationship among the modeling parameters [18]. Based on the reported literature (2014-2021), Scopus database indicated that there is a substantial attention on the BOD simulation using the feasibility of ML models. Figure 1 reported the major keywords occurrence clusters and the time span, used over the literature. Over 144 keywords were presented indicating the significant of this topic on modeling river water quality. The idea of the exploration of new ML models that are capable to solve environmental engineering problems is always going on and the research domain of modeling WQ using new sophisticated models are of interest of researchers and scientists [19]–[21]. Although the literature revealed different version of ML models applied for surface WQ modeling such as artificial neural network, kernel models, fuzzy logic, genetic programming, adaptive neuro-inference system models and several others [7]; however, there are several new versions of ML models are yet to be explored for modeling surface WQ phenomena. The efficiency of integrative intelligence models in WQ modeling has also been noted [9, 36-39]. Further, although ML models are the commonly used predictive models in surface WQ prediction, they are still facing several limitations, such as the need to tune their internal parameters, the need for time-consuming algorithms, poor generalization capability, and the need for human intervention during the modeling process. Hence, there is a need for models that are flexible enough to address the complicated nature of most environmental engineering problems [22].

C. THE SIGNIFICANT OF THE SELECTED CASE STUDY

The accurate determination of BOD is necessary for water pollution control because it is an important index of good quality water [23]. This parameter is delicate and tedious to analyze, especially BOD analysis. BOD presents the approximation of the bio-degradable organic matter in the water and defines an essential indicator for water pollution. In addition, BOD is presented as the foremost parameter for the aquatic system health presentation and its proper quantification can contribute to development of strategic water resources protection and safety. Furthermore, for instance, the DO parameter, the analysis can be adopted in-situ instruments; however, BOD is recorded for at least five days. Accurate prediction of WQ parameters in a study area can save cost, energy, and time; this is why much effort is given to the modeling approaches when predicting these valuable parameters [24]. The modeling approaches are more important in developing countries where the budget for environmental quality assessment and monitoring is low compared to the developed countries. The research is conducted on the base to predict monthly scale BOD for Euphrates river located in Iraq region. Five different ensemble ML models were developed for this purpose. The selection of those models was owing to their massive implementation received and confirming their potential in hydrological, climatological and environmental researches [25]–[28]. The obtained modeling results were compared with several well-established literature on river WQ prediction of diverse region all around the world.

D. RESEARCH MOTIVATION AND OBJECTIVES

Several review research articles presented lately on the progress of ML development for river WQ [7], [29], [30]. The literature review emphasis on the exploration of new versions of ML models for modeling river WQ due the drawbacks of the associated limitations with the existed ML models. For instance, classical models such as artificial neural network (ANN), fuzzy logic (FL) and support vector machine (SVM) are associated with the drawbacks on tuning their internal parameters [31]–[34]. Another issue reported in the previous studies on the importance selecting the significant and related predictors for the targeted predicted parameters [16], [35]. As the prediction matrix is highly influenced by the input feature selection, integrating a prior approach for the better understanding the predictors effects is an essential step in ML models development. The previous studies have shown an admirable trend for this point of view. For instance, the integration of improved Grey relational analysis (IGRA) algorithm with Long-Short term Memory (LSTM) predictive model, to simulate the DO concentration at the Tai Lake and Victoria Bay [36]. In another study, water quality index (WQI) was predicted using the coupled Gaussian Naïve Bayes and several ML model at Rawal Lake [37]. Recently, some authors tested the capacity of the quantum teaching and learning based optimization as feature selection for WQI determination using weighted extreme learning machine model for groundwater samples collected at the Dharmapuri district in Tamil Nadu [38]. Several other scholars adopted similar methodologies for surface WQ simulation [39]–[41]. All those studies confirmed the significant of coupled ML models for modeling surface water quality for better understanding to the substantial correlation between the simulated WA parameters.

Hence, the current research was prompted on the base to explore more reliable and robust soft computing predictive models. In addition, the investigation of the highly influential parameters on the prediction of BOD in river located with semi-arid region. The objectives of the current research are (i) to explore the capacity of five ML models including Quantile regression forest (QRF), Random Forest

(RF), radial support vector machine (SVM), Stochastic Gradient Boosting (GBM) and Gradient Boosting Machines (GBM_H2O) for river BOD prediction, (ii) to identify the prediction matrix using the feasibility of the statistical correlation. The proposed ML models were further enhanced on their prediction capability by integrating two approaches of feature selections (GA and PCA). The ultimate goal of the current research was to develop a reliable and robust intelligence model for river water quality prediction.

II. CASE STUDY AND DATA DESCRIPTION

This study focused on the prediction of WQ parameters in the Euphrates River, Ramadi City, Anbar state, Iraq. The coordinates of the measured point are as follows: 33°26'15"N latitude and 43°16'52"E longitude (Figure 2). The laboratory measurement was conducted from a large drinking water plant treatment intake at Ramadi City. The climate of the region is semi-arid with extreme summer temperature “exceed 45 ^{o}C” and cold weather during winter [42]. Sampling was done monthly for 10 years (2004-2013). The quality of water in the Euphrates River basin is mainly affected by human activities, especially human domestic and agricultural activities. The salt level of the river has increased tremendously along the stream course. Furthermore, industrial discharge of untreated sewage water into the river also contributed to the sources of contaminants. Therefore, this study is relevant as it provides an intelligent system for WQ monitoring of the studied river. Until now, studies are yet to be reported in this perspective, hence, this is a novel contribution in consideration of the proposed methodology. The statistical properties of the WQ parameters presented in Table 1. WQ prediction models can aid in determining the trend of decline in WQ at any point. BOD and DO have been the commonly used parameters of WQ for decades, hence, this study focused on the prediction of both parameters as their accurate prediction is essential towards easing the protective initiatives.

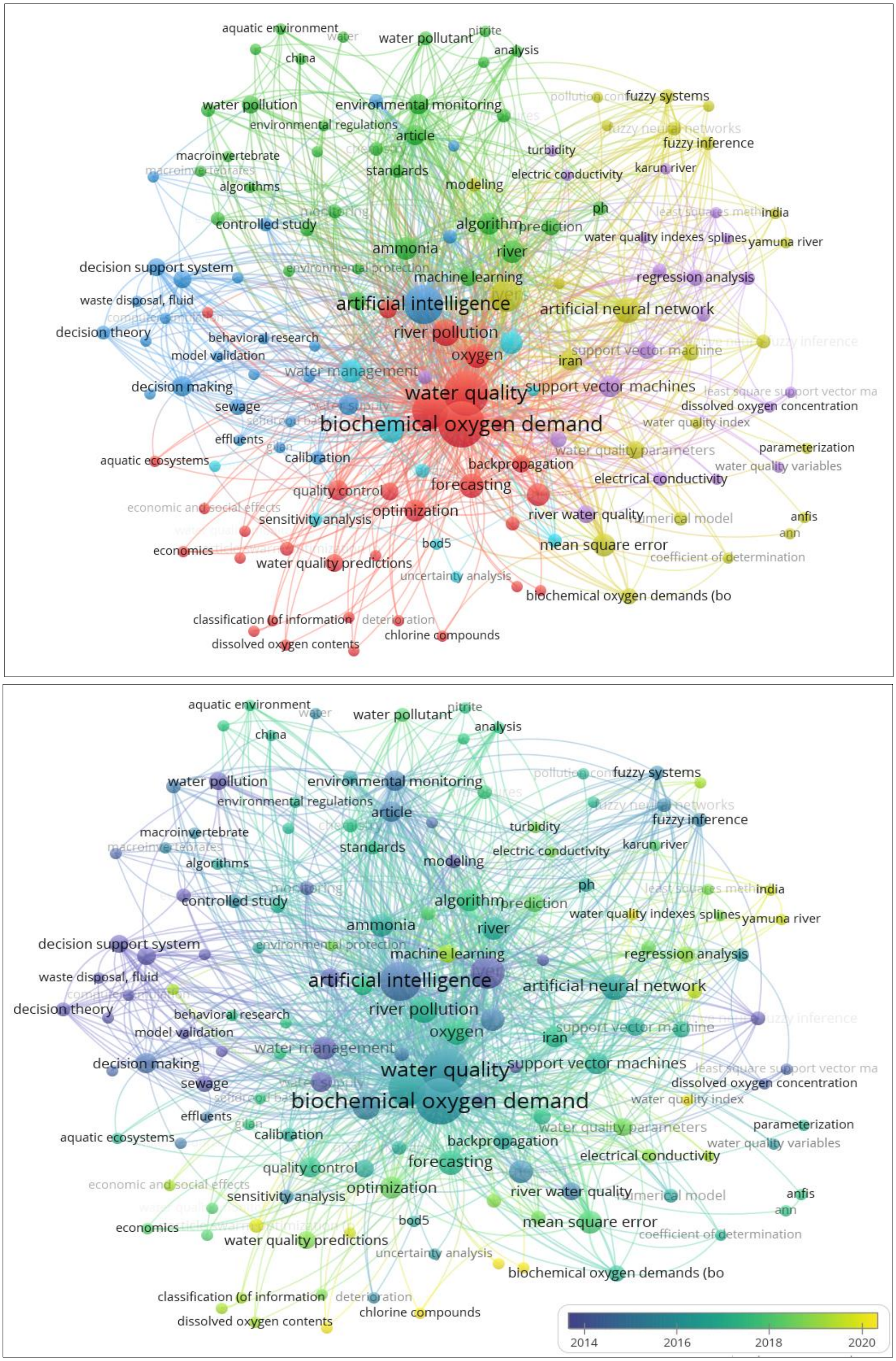


FIGURE 1: The VOSviewer algorithm results for the Scopus database research for the surveyed keywords "river biochemical oxygen demand modeling using artificial intelligence". 35 research articles were appeared over the time period 2014-2021.

TABLE 1:
The descriptive statistics of the water quality data at Ramadi City located on the Euphrates River.

Statistics	**Temperature**	**Turbidity**	**PH**	**EC**	**Alk**	**Ca**	**COD**	**SO_4**	**TDS**	**TSS**	**BOD**
Mean	21.84	18.99	7.77	1484.43	125.46	86.37	11.48	383.52	1102.73	56.79	3.82
Standard Error	0.75	1.38	0.02	11.43	1.13	0.58	0.18	3.00	8.56	4.52	0.05
Median	22	12.45	7.8	1449.5	124	85	11.1	385	1113.5	35	3.7
Mode	21	13.2	7.8	1412	124	85	10.2	381	1095	14	3.3
Standard Deviation	8.26	15.10	0.17	125.23	12.41	6.32	1.93	32.90	93.81	49.51	0.58
Kurtosis	-0.98	2.71	0.20	1.33	0.15	0.68	-0.47	5.71	-0.37	0.98	-0.94
Skewness	0.10	1.87	-0.45	1.28	0.76	1.01	0.58	-1.45	-0.40	1.46	0.46
Range	29	66.1	0.8	572	58	28	8.3	237	426	184	2.1
Minimum	9	7.3	7.3	1324	104	77	8.3	212	863	12	2.9
Maximum	38	73.4	8.1	1896	162	105	16.6	449	1289	196	5

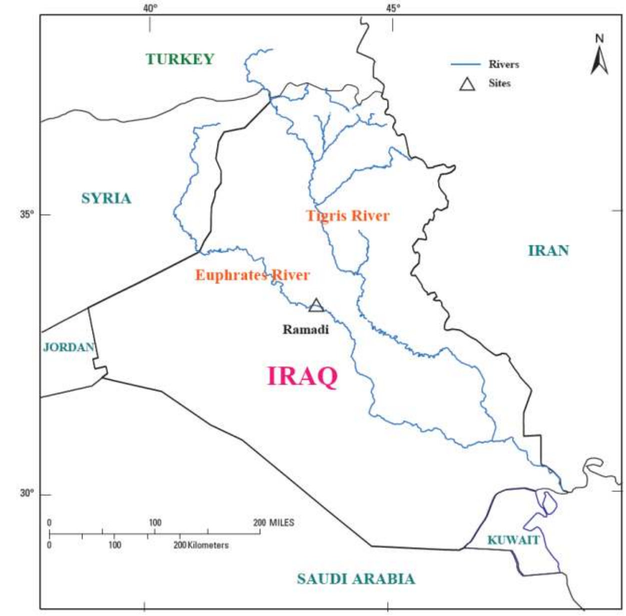


FIGURE 2. Ramadi water quality station location within Iraq region.

III. APPLIED ENSEMBLE MACHINE LEARNING MODELS

A. QUANTILE REGRESSION FOREST (QRF) MODEL

QRF model is one of the popular ML models in which was firstly developed in this study before applying quantile regressions at the last model prediction stage to achieve the quantile RF model. The regression algorithm was applied in this model since the model output (*i.e.*, BOD) is a continuous parameter. The concept of a RF model is based on the aggregation of several decision trees to establish the model output [43] as shown in Figure 3a. A decision tree (DT) refers to a decision support tool that relies on tree-like structures that consist of links and nodes to achieve potential model outputs. The starting point of each DT is a parent node that serves as a decision point; the parent node keeps creating branches until a decision is reached.

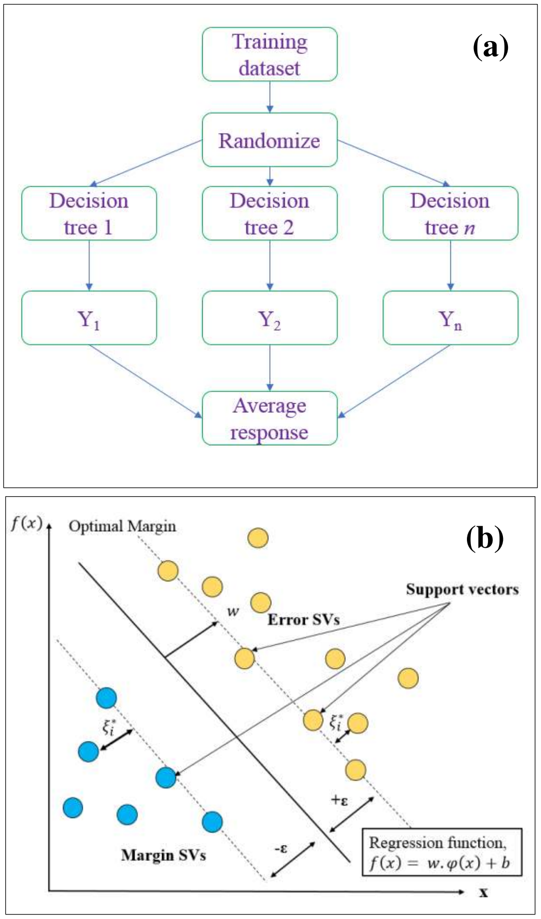


FIGURE 3. a) Quantile regression forest algorithm procedure flowchart. Y presents the average response and decision tree output. *N* denotes the sample sets in which randomized and produce n decision tree. b) the description of the regression support vector model.

During the modeling process, the training dataset is first randomly bootstrapped into sub-training sets (1, …, *n*) as in Figure 3a; each of the resulting bootstrapped sub-training sets are used to establish the DTs for different predictor parameters combinations starting with the topmost predictor. The response of each DT is an estimate of the response parameter; the response for 0.01 & 0.99 quintiles can be estimated from the number of DTs used to constructs an RF, while the mean response (*i.e.*, a quantile of 0.50) can be calculated as the final output of the model. The data within the range of 0.01 and 0.99 quintiles represents the prediction interval percentile. RF analysis demands a critical selection of the number of DTs; this ranges from some hundreds to thousands of trees. The optimum number of DTs in this study was determined using the out-of-bag (OOB) error technique [44]. A trained model is achieved when the prediction errors have been minimized and once this is achieved, the model is said to be the optimized/best-trained model. In this study, MSE was used to minimize the total error for each node of each DT. The error was estimated at each data splitting point – with the minimum MSE representing the best estimate. The study by [45] has earlier provided a detailed description of the formation of a DT and how RF works.

B. RANDOM FOREST (RF) MODEL

RF is a supervised learning approach that combined the bagging ensemble ML algorithm achieved from the classification and regression tree and the random subspace technique, introduced by Breiman [47]. Despite its simplicity, it is an effective tool that relies on the “divide and conquer” principle to solve multi-regression & prediction problems [48]–[50]. It has low sensitivity to multi-collinearity and achieves stable performances on unbalanced datasets. RF adopts the bootstrapping method in resampling the original dataset to generate subsets of similar sizes to the original set. Then, the tree construction is achieved by using the generated subsets, followed by the pooling of the results (prediction or regression) of the individual trees to arrive at the final outcome [50], [51]. RF has found successful application in environmental engineering [9] and other fields of study [52]. Detailed information on the mathematical formulation of RF models can be found in the studies presented by [46], [53], [54]. The *randomforest* and *caret* packages were used to train the predictive model. The RF model was initiated based on root mean square error-folds to control the model parameter. Grid search algorithm with randomly selected parameter.

C. SUPPORT VECTOR MACHINE (SVM) MODEL

This is an ML subcategory that was first proposed and developed by [55] for addressing both classification & regression problems. It is a robust approach that is based on the statistical learning theory [56]. The principle of the SVM model is hinged on first assessing the level of dependence of the target parameters ($\hat{y}$) on the predictive parameters (x) before obtaining a regression function using the relation [57]:

$$f(x) = \hat{y} = \omega \cdot \varphi(x) + b \tag{1}$$

where φ represents the functions for the replacement of complex nonlinear expressions with linear simpler ones. ω is the regression function weight while b is the regression function bias; both functions are generated via minimization of the deviation of $f(x)$ from the observed value ($\hat{y}$). SVM adopts the ε -insensitive loss function for the evaluation of this deviation (ϑ) [58], [59]:

$$|f(x) - y|_{\vartheta} = L(\vartheta) = \begin{cases} 0, |\vartheta| < \varepsilon \\ |\vartheta| - \varepsilon, |\vartheta| \geq \varepsilon \end{cases} \tag{2}$$

The following risk-structure function is also minimized to obtain the corresponding weight and bias:

$$S = \frac{1}{2}w^2 + C\sum_{i=1}^{n} |f(x_i) - y_i|_{\vartheta} \tag{3}$$

The hyperparameters are represented as C and ε. The Lagrange multiplier technique is used to minimize S to achieve the regression equation in Eq. 4, where the kernel function is represented as *K* [55]:

$$f(X) = \sum_{i} K(x, x_i) + b \tag{4}$$

Numerous kernel functions were tested in this study, and based on certain performance metrics and time efficiency, the linear function was selected for this study followed the reported literature [60], [61]. The regression function of the support vector machine model presented in Figure 3b.

D. STOCHASTIC GRADIENT BOOSTING (GBM) MODEL

Friedman [62] first developed the GBM algorithm as a combination of the gradient descent with the boosting algorithm. Hence, GBM was developed as an ensemble learning algorithm that merged boosting and DTs; the new model was built following the gradient descent path of the loss function of the earlier model. GBM algorithm was developed for the training of the classification function *F* * (*X*) which will minimize the loss function between the real function and the classification function. The loss function distribution is important in the implementation of the GBM model [62] even though the model can be applied to all loss functions. Friedman [63] suggested the surrogate loss function (multi-class log-loss) for the *K*-class problem. The mathematical expression of the loss function is as follows:

$$\psi(y_k, F_k(X)_1^K) = -\sum_{k=1}^{K} y_k \log\, p_k(X) \tag{5}$$

$$= -\sum_{k=1}^{K} y_k \log \left[\exp\,(F_k(X)) \middle/ \sum_{l=1}^{K} \exp\,(F_l(X)) \right]$$

where $X = \{x_1, x_2, \ldots, x_n\}$ represent the input parameter, *y* is the output parameter, *k* represents the number of classes, and the probability is represented as $p_k(X)$. This gives rise to the following equation:

$$\tilde{y}_{im} = -\left[\frac{\partial \psi\left(y_i, F_j(x_i)\right)_{j=1}^{K}}{\partial F(x_i)}\right]_{\left(F_j(X)=F_{j,m-1}(X)\right)_1^K} = y_i^k - p_k(x_i) \tag{6}$$

where $y_i^k - p_k(x_i)$ = the existing residuals; hence, K-trees are induced, leading to the production of K trees each with L-terminal nodes at iteration m, R_{klm}. For each tree, a separate line search can be used to resolve the terminal node as shown.

$$\gamma_{lm} = \arg \min_{\gamma} \sum_{x_i \in R_{lm}} \psi(y_i, F_{m-1}(x_i) + \gamma) \tag{7}$$

The updating of each of the functions leads to the formation of the GBM. The GBM algorithm has earlier been detailed in the study by [64]. The GBM algorithm depends on three key parameters which are (i) the number of trees (boosting interactions, M), (ii) the depth of the interaction (the max tree depth, J), and (iii) the shrinkage (the learning rate, v). Better performance and generalization of the GBM model depends on a proper tunning of these hyper-parameters.

E. GRADIENT BOOSTING MACHINE (GBM_H2O)

Another popular supervised ML model is the GBM_H2O model which was developed by [62], [63]; it is an efficient tool in solving both classification and regression problems [65], [66]. Boosting learns multiple classifiers via manipulation of the sample weights during the training phase and later linearly merges these classifiers to improve the classification performance. Friedman presented an extension of Boosting to regression tasks in 2011 via the introduction of the GBM to come up with an additive model that can ensure minimization of the loss function. The GBM model is first initialized to a constant value that minimizes the loss function, followed by the estimation of the negative gradient of the loss function in each iterative training process as the current models' residual value. Then, a new RT is trained to fit the current residual, followed by the addition of the current RT to the previous model and the updating of the residual. The algorithmic process is continued until the maximum iteration number set by the user is attained. The GBM model has been improved on the aspect of its poor performance (when using data) by ensuring that the RT is continuously used to fit the residuals.

F. PRICIPAL COMPONENT ANALYSIS

The principal component analysis (PCA) is a well-recognized feature selection approach that works based on un-supervised pattern recognition. It abstracts the frequent pattern that scores the highest in the simulated matrix [67]. The mathematical procedure of the PCA approach is working on the base to allocate the minimum error between the observed and predicted values due to the variance of the principal component [68]. The variance component $(a_{k,i})$ is calculated using:

$$\sigma^2(e_k) = E\left[a_{k,i}^2\right] = e_k^T S e_k \tag{8}$$

where $e_k = [e_1, k^{e2}, k \ldots e_d, k]^T$, e_k is the d-by-1 vector, S is the matrix scatter eigen values. the magnitude of the e_k vectors are eigen vector (k value is ranged from 1 to d'). For the case where $d' < d$, where a reduction in the dimension is attained, the $d'$is calculated through:

$$E_{d'} = \frac{1}{2} \sum_{i=d'+1}^{d} \lambda_k \tag{9}$$

where λ_k presents the scatter matrix of eigen values against the e_k. The variance direction is followed the direction of the eigen values [69].

G. GENETIC ALGORITHM

GA is a realistic method that is based Darwin's principle [70]. In the current study, GA was adopted to sort the best fit predictors using its potential on the base of evolutionary process [71]. The successive iterations were calculated then after the filtered values and the optimal solution configured. The mathematic aspect of the best value for the optimal feature is computed as follows [72]:

$$f(\chi^{th}) = k. R(\chi^{th}) \tag{10}$$

χ^{th}presents the individual feature "predictors water quality parameters", k indicates the constant variable for the selective pressure between 1 and 2. The last term $R(\chi^{th})$ defines the ranking of the individual features.

IV. MODELING RESULTS AND ANALYSIS

The modeling procedure adopted in this research was exhibited in a form of flowchart presented in Figure 4.

A. Predictors selection

In this study, the development of five different ensemble data-intelligence models (i.e., QRF, RF, SVM, GBM and GBM_H2O) were established for surface water BOD prediction. In addition, the integration of the PCA and GA feature selection approaches was investigated as the second modeling scenario. The models' performances were compared based on multiple statistical criteria including determination coefficient (R^2), root mean square error (RMSE), mean absolute error (MAE), Nash-Sutcliffe model efficiency coefficient (NSE), Willmott index (d), and percent bias (PBIAS) [73], [74], and graphical presentation. Owing to the fact that the wise selection of which predictor "water quality parameters" to be included in the prediction formula, it has more advantageous effects on overall performance than the choice of the modeling algorithm itself and thus the feature selection approaches were employed to identify the minimal subset of features for optimal learning.

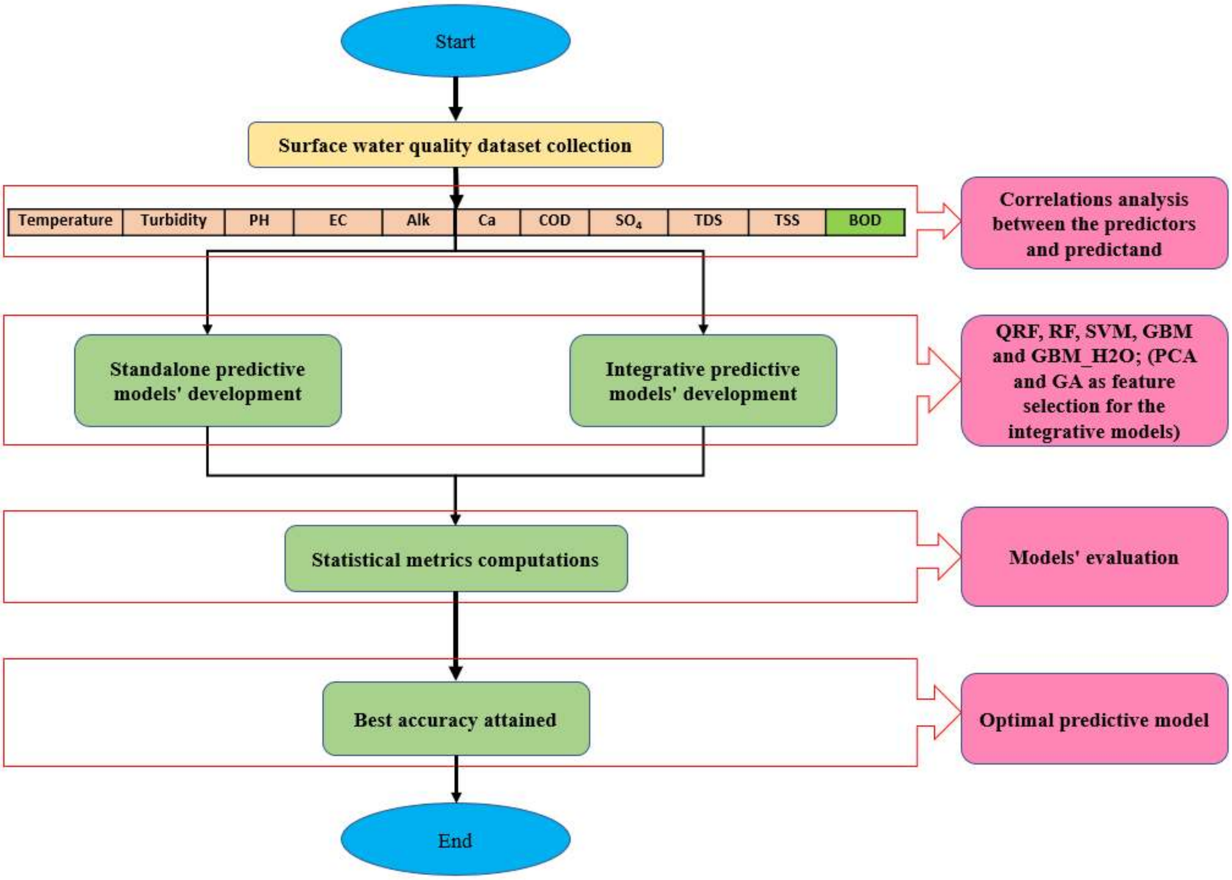

Start
Surface water quality dataset collection
Temperature
Turbidity
PH
EC
Alk
Ca
COD
SO$_4$
TDS
TSS
BOD
Correlations analysis between the predictors and predictand
Standalone predictive models' development
Integrative predictive models' development
QRF, RF, SVM, GBM and GBM_H2O; (PCA and GA as feature selection for the integrative models)
Statistical metrics computations
Models' evaluation
Best accuracy attained
Optimal predictive model
End

FIGURE 4. The modeling flowchart of the adopted methodology of the current research.

The performance of the feature selection techniques was compared to the benchmark models comprising a full set of data covering potential casual parameters including temperature (T), Turbidity, pH, EC, Alkalinity (Alk), Ca, COD, SO_4, TDS, and TSS data. Initially, the collected data was analyzed in terms of their correlations as shown in Figures 5 and 6. Figure 5 shows the Pearson correlation coefficient at significant level 0.05 between the BOD (predictand) and the predictors. It is noticeable from Figure 5 that all these parameters are more or less correlated with each other and hence introduce a lot of multi-collinearity when all predictors are used in a model to predict BOD. The correlation coefficient values of T-BOD, Turbidity-BOD, pH-BOD, EC-BOD, Alk-BOD, Ca-BOD, COD-BOD, SO_4-BOD, TDS-BOD, and TSS-BOD were 0.67, -0.25, -0.36, -0.33, -0.28, -0.12, 0.44, -0.04, -0.20, and -0.27, respectively. According to the above Pearson correlation coefficients, all the parameters were significantly correlated at 0.05 level with BOD except the SO_4 and Ca parameters. The T parameter exhibits the strong positive correlation with the BOD values while the remaining parameter show inverse significant relationships. It seems that the biological reactor of this particular case study is mainly influencing due to the water temperature as is could be due to the climate characteristics of this region. Thus, it can be concluded that T, Turbidity, pH, EC, Alk, COD, TDS, and TSS are significantly associated at 0.05 level with the BOD.

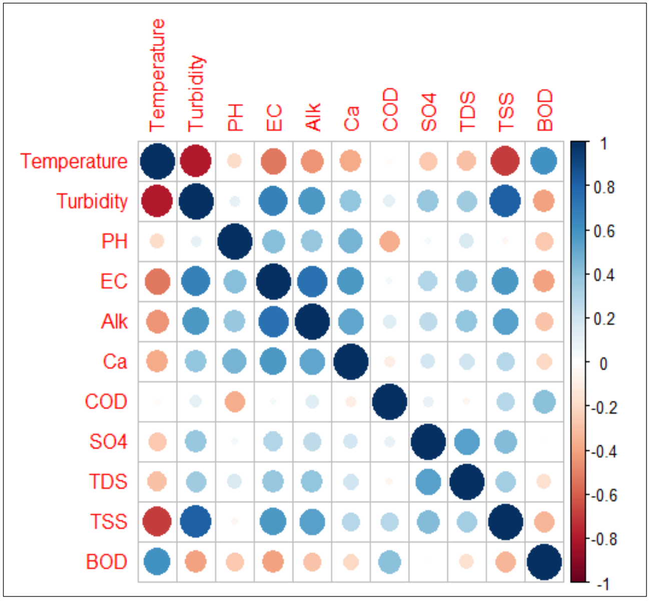
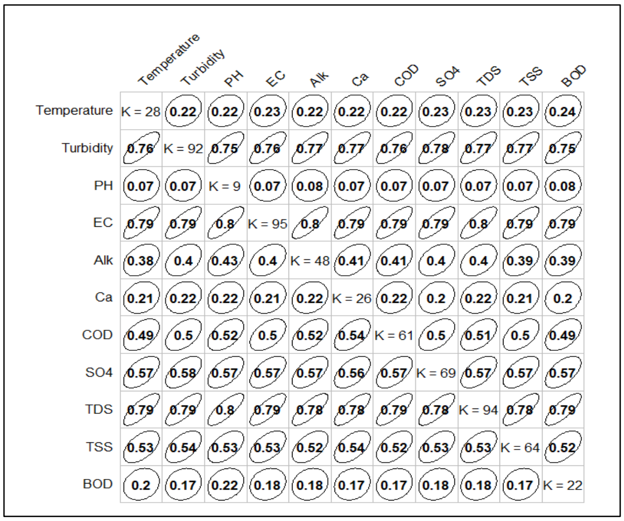


FIGURE 5. The correlation matrix of water quality data toward the BOD parameter.

The statistical results of Goodman and Kruskal tau measurement were presented in Figure 6 in which presenting the correlation between the input parameters and the target parameter. The distinguished values of each input parameters were presented in diagonal elements. The forward and backward tau measures were reported in the form of off-diagonal elements. The associations from T, Turbidity, pH, EC, Alk, Ca, COD, SO_4, TDS, and TSS to BOD were 0.24, 0.75, 0.08, 0.79, 0.39, 0.20, 0.49, 0.57, 0.79, and 0.52, respectively. Apparently, all predictors were associated with the predictand values and thence suggesting of potential predictability from the selected parameters to BOD. From the above analysis, it is judgeable that each test presented different concept on the association between the water quality parameters. Therefore, the entire data set was used to build the benchmark models.



	Temperature	Turbidity	PH	EC	Alk	Ca	COD	SO4	TDS	TSS	BOD
Temperature	K = 28	0.22	0.22	0.23	0.22	0.22	0.22	0.23	0.23	0.23	0.24
Turbidity	0.76	K = 92	0.75	0.76	0.77	0.77	0.76	0.78	0.77	0.77	0.75
PH	0.07	0.07	K = 9	0.07	0.08	0.07	0.07	0.07	0.07	0.07	0.08
EC	0.79	0.79	0.8	K = 95	0.8	0.79	0.79	0.79	0.8	0.79	0.79
Alk	0.38	0.4	0.43	0.4	K = 48	0.41	0.41	0.4	0.4	0.39	0.39
Ca	0.21	0.22	0.22	0.21	0.22	K = 26	0.22	0.2	0.22	0.21	0.2
COD	0.49	0.5	0.52	0.5	0.52	0.54	K = 61	0.5	0.51	0.5	0.49
SO4	0.57	0.58	0.57	0.57	0.57	0.56	0.57	K = 69	0.57	0.57	0.57
TDS	0.79	0.79	0.8	0.79	0.78	0.78	0.79	0.78	K = 94	0.78	0.79
TSS	0.53	0.54	0.53	0.53	0.52	0.54	0.52	0.53	0.53	K = 64	0.52
BOD	0.2	0.17	0.22	0.18	0.18	0.17	0.17	0.18	0.18	0.17	K = 22

FIGURE 6. **Goodman and Kruskal test between the predictors and predictands.**

Using the GA approach, the search of the feature space was conducted repeatedly within resampling iteration. Hence, the training data were split according to the 5 fold-cross validation resampling method specified in the control function. Hence, the entire GA approach was performed in 5 separate times. For the first fold, one fifth of the data were employed in the search while the remaining fifth was employed to estimate the external performance since the data points were not used in the search. The internal and external average accuracy estimates computed from the 5-out samples prediction were exhibited in Figure 7. Using the R software package "gafs() function using 100 generation and 50 individuals", was implemented to perform the assessment for the chromosomes of each generation. This was conducted by random forest model and 5-fold cross validation. Therefore, in the final search using the entire training set; only 4 features (among the ten) were selected at iteration 49 which included T, pH, COD and TSS with RMSE, R^2, and MAE values of 0.3463, 0.7138, 0.2742, respectively based on the external performance. Accordingly, the optimal four predictors were used to build the integrative GA-ML models.

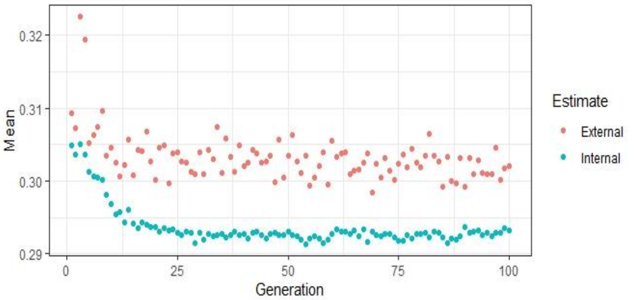


FIGURE 7. **The internal and external performances of genetic algorithm.**

In the same manner, the collected data was analyzed in terms of their associations using the PCA approach. Figure 8 depicts biplot for the first two most variance components. As it can noticed that the first component is the most dominant by the T parameter. While the second component was dominated by pH and Ca. Moreover, the scree plot that explains the most of variability in the data was plotted as shown in Figure 9. Where the x-axis and the y-axis represents the component and the importance of that component, respectively. As it can be seen from the figure that after the second component there is a significant drop-off to the incremental impact of each additional component. The eigen value per component was calculated as given in Table 2. The only parameter which has an eigen value close to 1 were included. The idea behind this is that if the eigen value is much less than 1, then the component accounts for less variance than a single parameter contributed. Upon that, the only first four components were used to build the models.

TABLE 2
Eigen values of the principal components of water quality parameters.

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
4.28	1.97	1.31	0.99	0.63	0.47	0.43	0.36	0.25	0.15	0.09

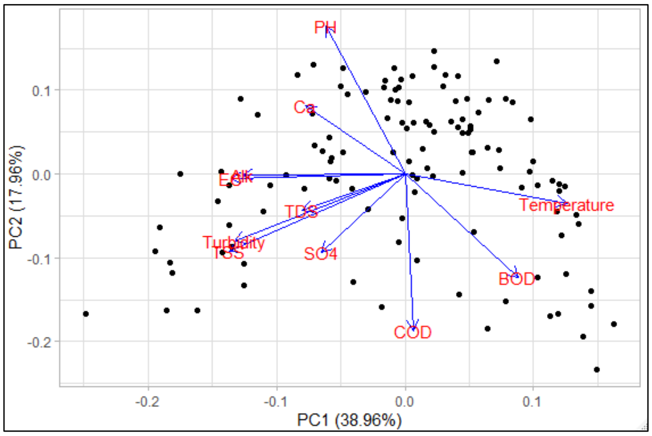


FIGURE 8. Principal component analysis (PCA) plot for the input parameter used for BOD (mg/l) prediction.

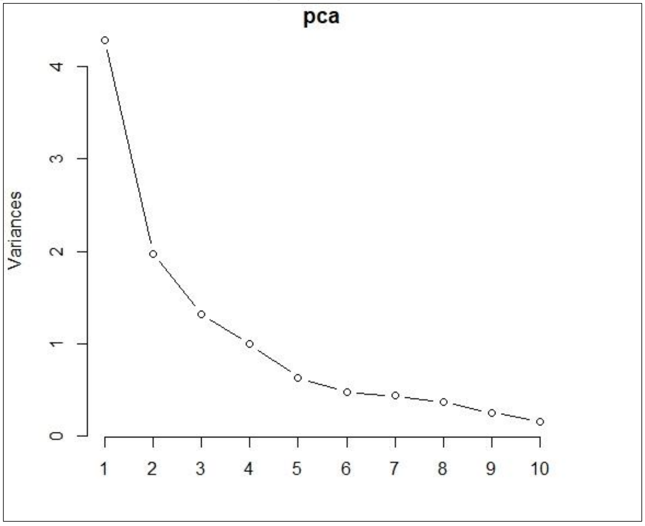


FIGURE 9. Scree plot of principal component analysis for the input parameters used for BOD (mg/l) prediction.

B. Models performances

The performances of the five-ensemble ML models developed in the study were evaluated based on learning accuracy using WQ data collected form the Euphrates River. Before applying the data into the models, it was randomly partitioned into 75% for training the models and the remaining 25% for validation [14], [16], [75]. The statistical performance of the developed five ML models was reported in Table 3. In general, the results indicated a highly competitive among the five models; however, an acceptable level of predictability performance was observed. The prediction power of the adopted models was ranked based on their average performance across the six statistical measures. The graphical presentation for the attained results were selected using Taylor diagram and boxplots over the validation phase. The statistical metrics records of R^2, RMSE, MAE, NSE, d, and PBIAS form QRF- GBM_H2O were 0.9, 0.16, 0.07, 0.87, 0.97, and 0- 0.84, 0.19, 0.1, 0.81, 0.96, and 0, respectively (Table 3). In other words, the lowest RMSE, MAE, PBIAS and the highest R^2, NSE, and d were from these two models. However, the remaining models had performances less accurate than the QRF and GBM_H2O. Figure 10a presented the boxplot results of the established five ML models in comparison with the benchmark observed dataset over the validation modeling phase. The middle line indicates the magnitude of the BOD and the whiskers are presented by the minimum and maximum magnitudes of the samples. The 25th and 75th percentiles are the referred to the lower and the upper edges. It is clearly can be observed that the QRF model could achieve the identical prediction accuracy as it is the nearest shape to the observed dataset. Whereas, the GBM model reported the worst prediction accuracy in comparison with the other models. Taylor diagram presented the results in the form of 2-dimensions graph where the observed dataset was indicated as a circle along the abscissa and other models were exhibited their performance based on the distance from the observed data based on the RMSE, standard deviation and the correlation statistic (Figure 10b). In harmony with the boxplot, the QRF model was presented the nearest coordination to the observed dataset and the GBM model was the furthest. The correlation coefficient of the QRF model was within the range of 0.95 and the centered pattern RMS difference between the two pattern was 0.16.

TABLE 3
The statistical performance criteria of the developed ML models over the calibration and validation phases.

	calibration							validation					
	R^2	RMSE	MAE	NSE	d	PBIAS		R^2	RMSE	MAE	NSE	d	PBIAS
GBM	0.78	0.26	0.21	0.69	0.93	-0.2		0.73	0.26	0.2	0.61	0.91	1
GBM_H2O	0.97	0.1	0.07	0.96	0.99	0		0.84	0.19	0.1	0.81	0.96	0
SVM	0.81	0.24	0.16	0.75	0.94	-0.3		0.83	0.21	0.15	0.71	0.94	0.3
RF	0.96	0.13	0.1	0.94	0.99	0.1		0.79	0.23	0.17	0.67	0.93	1.1
QRF	0.99	0.07	0.02	0.98	0.99	0		0.9	0.16	0.07	0.87	0.97	0
GA-GBM_GA	0.75	0.28	0.22	0.61	0.92	0.2		0.73	0.27	0.21	0.6	0.91	2
GA-GBM_H2O	0.85	0.22	0.17	0.8	0.96	0		0.72	0.26	0.21	0.63	0.91	0.7
GA-SVM	0.79	0.26	0.19	0.69	0.93	0.1		0.7	0.27	0.21	0.53	0.9	1.7
GA-RF	0.93	0.17	0.13	0.87	0.97	0.2		0.79	0.24	0.18	0.59	0.92	0.9
GA-QRF	0.97	0.11	0.05	0.96	0.99	-0.2		0.85	0.19	0.09	0.82	0.96	0
PCA-GBM	0.96	0.11	0.09	0.95	0.99	-0.2		0.88	0.17	0.13	0.84	0.96	0.3
PCA-GBM_H2O	0.99	0.06	0.05	0.99	0.99	0		0.89	0.16	0.09	0.87	0.97	0.3
PCA-SVM	0.94	0.14	0.1	0.93	0.98	-0.3		0.89	0.17	0.11	0.84	0.97	0.2
PCA-RF	0.97	0.11	0.09	0.95	0.99	0.1		0.92	0.17	0.13	0.79	0.96	0.8
PCA-QRF	0.99	0.04	0.01	0.99	0.99	0		0.94	0.12	0.05	0.93	0.98	0.3

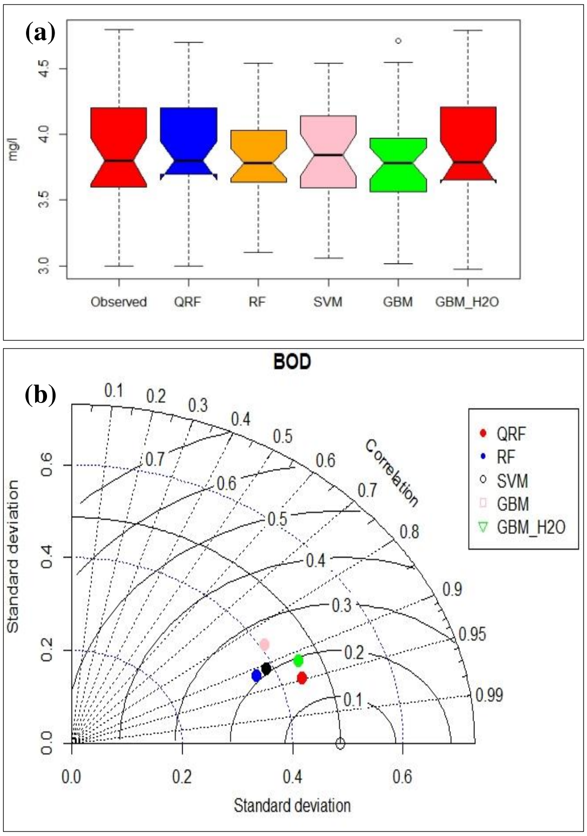


FIGURE 10. a) Boxplot and b) Taylor diagram for the benchmark predictive models.

The closest distribution around the line 1:1 was observed from the QRF model with values of R^2, RMSE, MAE, NSE, d, and PBIAS equal to 0.85, 0.19, 0.09, 0.82, 0.96, and 0, respectively. While the remaining models performed with less accuracy than QRF. The boxplot of the results obtained from the evaluated the integrative GA-ML modelling methods during the validation stage were given as shown in Figure 11a. With the same manner of the benchmark models, the distribution of the QRF model was the most similar to the observed followed by GBM_H2O>GBM> SVM>RF. Taylor diagram (Figure 11b) confirmed that the optimal performance was from the QRF model while the RF and SVM are the furthest, and the other evaluated methods in between. The correlation coefficient between the QRF and the observed data is less than 0.95, and the centered pattern RMS difference between the two patterns is ~0.19. The performances of the integrative GA-ML models were not as good as to those from the benchmark models. Indicating that the selected features by the GA was not representative to the entire data of BOD.

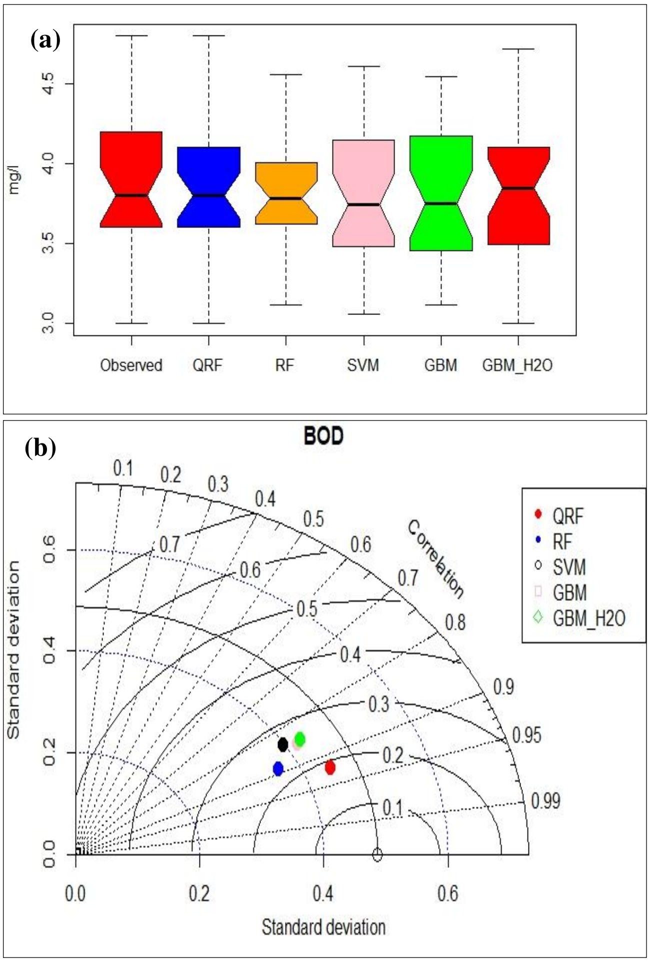


FIGURE 11. a) boxplot and b) Taylor diagram for each of the integrative GA-ML models

Overall, the applied ML models using the PCA approach performed better than GA approach by producing the lowest prediction error. However, QRF outperformed the GBM_H2O in terms of the statistical performance metrics. R^2, RMSE, MAE, NSE, d, and PBIAS of QRF were 0.94, 0.12, 0.05, 0.93, 0.98, and 0.3, respectively. While those of GBM_H2O were 0.89, 0.16, 0.09, 0.87, 0.97, and 0.3, respectively. The performance from QRF and GBM_H2O was followed by SVM> GBM> RF. The boxplot of the results obtained from the evaluated the integrative PCA-ML modelling methods during the validation stage were analyzed as shown in Figure 12a. It can be confirmed that the distribution from QRF was the most similar to that from the observed. The interquartile of the QRF model was almost the closest one to the observed values. Then followed by GBM_H2O>GBM> SVM>RF. This fact was further confirmed by Taylor diagram (Figure 12b) which prove that the optimal performance was from the QRF model while the RF and SVM were the worst, and the other evaluated methods in between. The correlation coefficient between the QRF and the observed data is greater than 0.95, and the centered pattern RMS difference between the two patterns is ~0.12. The subset selection using the PCA approach outperformed that of the benchmark and GA-ML models.

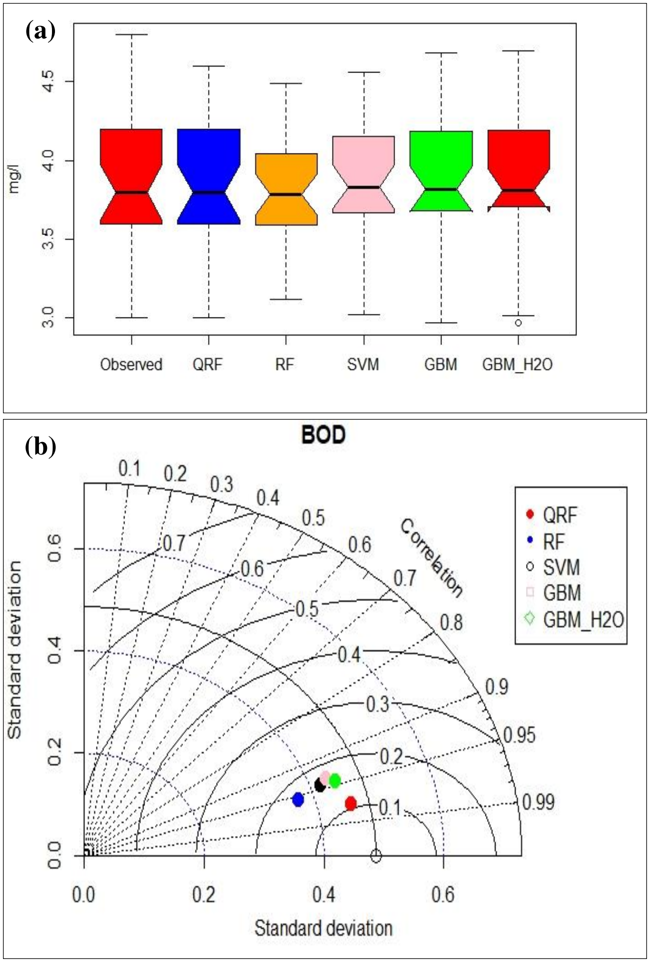


FIGURE 12. a) Boxplot and b) Taylor diagram for the developed PCA-ML models.

V. DISCUSSION

The redundant and irrelevant predictors significantly deteriorate the performances of regression models and causes overfitting problem in the prediction models. Therefore, extracting a smaller subset of predictors with most relevant predictors might be useful since it saves time in data collection and computation [76], [77].

In this study, two-feature selection were integrative with five different ensemble learning artificial intelligence models (i.e., QRF, RF, GBM_H2O, GBM, SVM) in order to improve the surface BOD water quality prediction accuracy at the Euphrates River. These two-feature selections can be broadly categorized into filter methods (PCA) and wrapper methods (genetic algorithm) [78]. It was concluded that the performance from PCA outperforms the predictability performances of GA approach and the benchmark models. The GA works by searching the space of possible feature subsets and then evaluating a subset of features using a ML algorithm. This method is known as greedy algorithms owing to the fact that they aim to find the best possible combination of features, which result in the best performant algorithm model [79]. This in turn would be computationally expensive, and impractical in the case of exhaustive search. While in PCA, each predictor is evaluated with a statistical performance metric and then ranked according to its performance indicator. Then after, the top-performing features is selected through the truncation selection before applying a ML models. Hence, the method is considered as a pre-processing step as it doesn't consider the complex interactions between predictors and are independent of learning algorithms [80]. As mentioned earlier, it is well identified that the PCA method is computationally efficient [81]. However, one shortcoming was pointed when applying this method is being stuck in local optimum when the complex interactions among predictors are ignored [81], [82]. Many researchers argued that wrapper methods (the GA) take into consideration the interaction among predictors but they are not as computationally efficient as filter methods (the PCA) because of the larger space to search [83]–[85]. It is well pointed out that the main drawback of applying GA is the necessity to be applied with a higher population size and larger number of generation, which are mostly time consuming [78]. It is prevailed that the optimal features selection returns by GA and the better the network perform in prediction can be attained when there are a large population size and number of generations. Small data set for feature selection may cause the problem of overfitting which is why the performance of GA in this study was not superior in comparison to the baseline models.

The combination of PCA with quantile regression forest model outperforms all the applies models in terms of the statistical performances criteria. In QRF model, the conditional quantiles can be inferred which was introduced by Meinshausen [86] as a generalization form of random forests [46]. The robustness of QRF method attributed to its non-parametric accurate way of estimating conditional quantiles for high-dimensional predictor parameters. The method is proved to be consistent when applied with multiple different scenarios, suggesting that the algorithm is competitive in terms of predictive power.

It is worth to mention that span of the dataset used for the current study provided a satisfactory information for the ML models development and the learning process. It is true that several data span were adopted over the literature; however, in this study, the monthly scale of ten years observations were adequately construct the ML models.

The current research modeling is associated with some limitations such as tuning the internal parameters of the SVM model with other advanced non-linear function [87]. In addition, using metaheuristic optimization algorithms can be another option to enhance the performance of the ML models learning process [88].

VI. CONCLUSION

This study was proposed five relatively new explored ML models for BOD of surface WQ prediction. These models were considered in this work as a robust approach towards the prediction of WQ parameters rather than relying on laboratory analysis. Further enhancement, two feature selection approaches (GA and PCA) were integrated with the developed ML models to enhance their predictability performance. Various categories of water parameters, including physical, chemical, and biological parameters were used for the development of the proposed models as the input attributes. The data for the model construction was 10 years period laboratory information covering 2004-2013. The outcome of the research showed that PCA-QRF model provided a reliable performance of the BOD prediction compared to the other established models. Furthermore, the proposed model exhibited less approximation of the input parameters that are extremely for the catchments with less environmental or ecological information. Generally, the proposed ML models performed an accurate prediction of the WQ parameters of the Euphrates River. Future studies are aimed at the prediction of other WQ parameters, as well as the inclusion of more input attributes, such as climatological or hydrological factors.

Conflict of interest : The authors declare no conflict of interest to any party.

REFERENCES

[1] Z. Luo, Q. Shao, Q. Zuo, and Y. Cui, "Impact of land use and urbanization on river water quality and ecology in a dam dominated basin," *Journal of Hydrology*, vol. 584, p. 124655, May 2020.

[2] S. Loos, C. M. Shin, J. Sumihar, K. Kim, J. Cho, and A. H. Weerts, "Ensemble data assimilation methods for improving river water quality forecasting accuracy," *Water research*, vol. 171, p. 115343, 2020.

[3] M. Okumah, A. S. Yeboah, and S. K. Bonyah, "What matters most? Stakeholders' perceptions of river water quality," *Land Use Policy*, vol. 99, p. 104824, 2020.

[4] S. B. H. S. Asadollah, A. Sharafati, D. Motta, and Z. M. Yaseen, "River water quality index prediction and uncertainty analysis: A comparative study of machine learning models," *Journal of Environmental Chemical Engineering*, 2020.

[5] Z. Z. Al-Janabi, A.-R. Al-Kubaisi, and A.-H. M. J. Al-Obaidy, "Assessment of water quality of Tigris River by using water quality index (CCME WQI)," *Al-Nahrain Journal of Science*, vol. 15, no. 1, pp. 119–126, 2012.

[6] Y. R. Ding, Y. J. Cai, P. D. Sun, and B. Chen, "The use of combined neural networks and genetic algorithms for prediction of river water quality," *Journal of applied research and technology*, vol. 12, no. 3, pp. 493–499, 2014.

[7] Tiyasha, T. M. Tung, and Z. M. Yaseen, "A survey on river water quality modelling using artificial intelligence models: 2000–2020," *Journal of Hydrology*. 2020.

[8] T. R. Zolnikov, "The maladies of water and war: Addressing poor water quality in Iraq," *American Journal of Public Health*, vol. 103, no. 6, pp. 980–987, 2013.

[9] A. J. Kadhem, "Assessment of Water Quality in Tigris River-Iraq by Using GIS Mapping," *Natural Resources*, vol. 04, no. 06, pp. 441–448, 2013.

[10] A. S. Alsaqqar, B. H. Khudair, and S. K. Ali, "Evaluating water stability indices from water treatment plants in Baghdad City," *Journal of Water Resource and Protection*, vol. 6, no. 14, p. 1344, 2014.

[11] K. A. Rahi and T. Halihan, "Changes in the salinity of the Euphrates River system in Iraq," *Regional Environmental Change*, vol. 10, no. 1, pp. 27–35, 2010.

[12] S. H. Abbas, B. H. Khudair, and M. S. Jaafar, "River Water Salinity Impact on Drinking Water Treatment Plant Performance

Using Artificial neural network," *Journal of Engineering*, vol. 25, no. 8, pp. 149–159, 2019.

[13] Z. F. Makki, A. A. Zuhaira, S. M. Al-Jubouri, R. K. S. Al-Hamd, and L. S. Cunningham, "GIS-based assessment of groundwater quality for drinking and irrigation purposes in central Iraq," *Environmental monitoring and assessment*, vol. 193, no. 2, pp. 1–27, 2021.

[14] S. I. Abba *et al.*, "Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination," *Journal of Hydrology*, vol. 587, p. 124974, Aug. 2020.

[15] A. Azad, H. Karami, S. Farzin, A. Saeedian, H. Kashi, and F. Sayyahi, "Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (Case study: Gorganrood River)," *KSCE Journal of Civil Engineering*, vol. 00, no. 0000, pp. 1–8, 2017.

[16] Y. Chen, L. Song, Y. Liu, L. Yang, and D. Li, "A review of the artificial neural network models for water quality prediction," *Applied Sciences*, vol. 10, no. 17, p. 5776, 2020.

[17] T. Rajaee, S. Khani, and M. Ravansalar, "Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review," *Chemometrics and Intelligent Laboratory Systems*, vol. 200, no. February, p. 103978, 2020.

[18] R. Barzegar, J. Adamowski, and A. A. Moghaddam, "Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran," *Stochastic Environmental Research and Risk Assessment*, vol. 30, no. 7, pp. 1797–1819, 2016.

[19] M. Ehteram, S. Q. Salih, and Z. M. Yaseen, "Efficiency evaluation of reverse osmosis desalination plant using hybridized multilayer perceptron with particle swarm optimization," *Environmental Science and Pollution Research*, 2020.

[20] Z. M. Yaseen, M. Ehteram, A. Sharafati, S. Shahid, N. Al-Ansari, and A. El-Shafie, "The Integration of Nature-Inspired Algorithms with Least Square Support Vector Regression Models: Application to Modeling River Dissolved Oxygen Concentration," *Water*, vol. 10, no. 9, p. 1124, Aug. 2018.

[21] M. Bayatvarkeshi, M. A. Imteaz, O. Kisi, M. Zarei, and Z. M. Yaseen, "Application of M5 model tree optimized with Excel Solver Platform for water quality parameter estimation," *Environmental Science and Pollution Research*, pp. 1–18, 2020.

[22] S. Park, S. Jung, H. Lee, J. Kim, and J.-H. Kim, "Large-Scale Water Quality Prediction Using Federated Sensing and Learning: A Case Study with Real-World Sensing Big-Data," *Sensors*, vol. 21, no. 4, p. 1462, 2021.

[23] C. S. Akratos, J. N. E. Papaspyros, and V. A. Tsihrintzis, "An artificial neural network model and design equations for BOD and COD removal prediction in horizontal subsurface flow constructed wetlands," *Chemical Engineering Journal*, vol. 143, no. 1–3, pp. 96–110, 2008.

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

[25] I. A. & M. B. Ali Naseri, Mehdi Jamei, "Nanofluids Thermal Conductivity prediction applying a Novel Hybrid Data-Driven Model Validated using Monte Carlo based Sensitivity Analysis," *Engineering with Computers*, 2020.

[26] I. Ahmadianfar, M. Jamei, M. Karbasi, A. Sharafati, and B. Gharabaghi, "A novel boosting ensemble committee-based model for local scour depth around non-uniformly spaced pile groups," *Engineering with Computers*, pp. 1–23, 2021.

[27] M. Gholizadeh, M. Jamei, I. Ahmadianfar, and R. Pourrajab, "Prediction of nanofluids viscosity using random forest (RF) approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 201, p. 104010, Jun. 2020.

[28] G. Papacharalampous *et al.*, "Probabilistic hydrological post-processing at scale: Why and how to apply machine-learning quantile regression algorithms," *Water*, vol. 11, no. 10, p. 2126, 2019.

[29] S. Giri, "Water Quality Prospective in Twenty First Century: Status of Water Quality in Major River Basins, Contemporary Strategies and Impediments: A Review," *Environmental Pollution*, p. 116332, 2020.

[30] V. Sagan *et al.*, "Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing," *Earth-Science Reviews*. 2020.

[31] M. I. Shah, M. F. Javed, and T. Abunama, "Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques," *Environmental Science and Pollution Research*, vol. 28, no. 11, pp. 13202–13220, 2021.

[32] S. He, W. Chen, X. Mu, and W. Cui, "Constrained optimization model of the volume of initial rainwater storage tank based on ANN and PSO," *Environmental Science and Pollution Research*, vol. 27, no. 17, pp. 21057–21070, 2020.

[33] A. Azad, S. Farzin, H. Sanikhani, H. Karami, O. Kisi, and V. P. Singh, "Approaches for Optimizing the Performance of Adaptive Neuro-Fuzzy Inference System and Least-Squares Support Vector Machine in Precipitation Modeling," *Journal of Hydrologic Engineering*, vol. 26, no. 4, p. 4021010, 2021.

[34] O. Kisi, A. Azad, H. Kashi, A. Saeedian, S. A. A. Hashemi, and S. Ghorbani, "Modeling groundwater quality parameters using hybrid neuro-fuzzy methods," *Water resources management*, vol. 33, no. 2, pp. 847–861, 2019.

[35] A. Mosavi, P. Ozturk, and K. W. Chau, "Flood prediction using machine learning models: Literature review," *Water (Switzerland)*. 2018.

[36] J. Zhou, Y. Wang, F. Xiao, Y. Wang, and L. Sun, "Water quality prediction method based on IGRA and LSTM," *Water*, vol. 10, no. 9, p. 1148, 2018.

[37] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient water quality prediction using supervised machine learning," *Water*, vol. 11, no. 11, p. 2210, 2019.

[38] J. Charles, G. Vinodhini, and R. Nagarajan, "An Efficient Feature Selection with Weighted Extreme Learning Machine for Water Quality Prediction and Classification Model," *Annals of the Romanian Society for Cell Biology*, pp. 1969–1994, 2021.

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

[40] D. Wu, H. Wang, and R. Seidu, "Smart data driven quality prediction for urban water source management," *Future Generation Computer Systems*, vol. 107, pp. 418–432, 2020.

[41] Q. Zou, Q. Xiong, Q. Li, H. Yi, Y. Yu, and C. Wu, "A water quality prediction method based on the multi-time scale bidirectional long short-term memory network," *Environmental Science and Pollution Research*, pp. 1–12, 2020.

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

[43] L. Breiman, "Random Forrests," *Machine learning*, 2001.

[44] S. Han and H. Kim, "On the optimal size of candidate feature set in random forest," *Applied Sciences*, vol. 9, no. 5, p. 898, 2019.

[45] A. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.

[46] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[47] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[48] D. Tsagkrasoulis and G. Montana, "Random forest regression for manifold-valued responses," *Pattern Recognition Letters*, vol. 101, pp. 6–13, 2018.

[49] F. B. de Santana, A. M. de Souza, and R. J. Poppi, "Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 2018.

[50] W. Chen *et al.*, "Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and

random forest methods," *Science of The Total Environment*, vol. 701, p. 134979, 2020.

[51] D. R. Cutler *et al.*, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[52] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[53] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016.

[54] E. Goel, E. Abhilasha, E. Goel, and E. Abhilasha, "Random forest: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 1, 2017.

[55] V. N. Vapnik, *The Nature of Statistical Learning Theory*, vol. 8, no. 6. 2000.

[56] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[57] B. Keshtegar, M. Bagheri, and Z. M. Yaseen, "Shear strength of steel fiber-unconfined reinforced concrete beam simulation: Application of novel intelligent model," *Composite Structures*, vol. 212, pp. 230–242, Mar. 2019.

[58] R. G. Brereton and G. R. Lloyd, "Support Vector Machines for classification and regression," *The Analyst*, vol. 135, no. 2, pp. 230–267, 2010.

[59] H. Azimi, H. Bonakdari, and I. Ebtehaj, "Design of radial basis function-based support vector regression in predicting the discharge coefficient of a side weir in a trapezoidal channel," *Applied Water Science*, vol. 9, no. 4, pp. 1–12, 2019.

[60] K. P. Singh, N. Basant, and S. Gupta, "Support vector machines in water quality management," *Analytica Chimica Acta*, vol. 703, no. 2, pp. 152–162, 2011.

[61] Y. Xiang and L. Jiang, "Water quality prediction using LS-SVM and particle swarm optimization," in *2009 Second International Workshop on Knowledge Discovery and Data Mining*, 2009, pp. 900–904.

[62] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[63] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 2001.

[64] A. Jasra and C. C. Holmes, "Stochastic boosting algorithms," *Statistics and Computing*, vol. 21, no. 3, pp. 335–347, 2011.

[65] J. Zhou, E. Li, M. Wang, X. Chen, X. Shi, and L. Jiang, "Feasibility of stochastic gradient boosting approach for evaluating seismic liquefaction potential based on SPT and CPT case histories," *Journal of Performance of Constructed Facilities*, vol. 33, no. 3, p. 4019024, 2019.

[66] S. Touzani, J. Granderson, and S. Fernandes, "Gradient boosting machine for modeling the energy consumption of commercial buildings," *Energy and Buildings*, vol. 158, pp. 1533–1543, 2018.

[67] S. K. Bhagat, T. Tiyasha, S. M. Awadh, T. M. Tung, A. H. Jawad, and Z. M. Yaseen, "Prediction of sediment heavy metal at the Australian Bays using newly developed hybrid artificial intelligence models," *Environmental Pollution*, 2020.

[68] S. K. Bhagat, T. Tiyasha, T. M. Tung, R. R. Mostafa, and Z. M. Yaseen, "Manganese (Mn) removal prediction using extreme gradient model," *Ecotoxicology and Environmental Safety*, vol. 204, no. August, p. 111059, 2020.

[69] A. Malhi and R. X. Gao, "PCA-based feature selection scheme for machine defect classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 6, pp. 1517–1525, 2004.

[70] S. Mirjalili, "Genetic Algorithm," in *Evolutionary Algorithms and Neural Networks*, Springer, 2019, pp. 43–55.

[71] S. Forrest, "Genetic algorithms: principles of natural selection applied to computation," *Science*, vol. 261, no. 5123, pp. 872–878, 1993.

[72] A. Al Imran, M. R. I. Rifat, and R. Mohammad, "Enhancing the Classification Performance of Lower Back Pain Symptoms Using Genetic Algorithm-Based Feature Selection," in *Proceedings of International Joint Conference on Computational Intelligence*, 2020, pp. 455–469.

[73] M. Jamei, I. Ahmadianfar, X. Chu, and Z. M. Yaseen, "Prediction of surface water total dissolved solids using hybridized wavelet-multigene genetic programming: New approach," *Journal of Hydrology*, 2020.

[74] Z. M. Yaseen, "An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions," *Chemosphere*, p. 130126, 2021.

[75] M. Al-Mukhtar, "Modeling the monthly pan evaporation rates using artificial intelligence methods: a case study in Iraq," *Environmental Earth Sciences*, vol. 80, no. 1, 2021.

[76] S. K. Bhagat *et al.*, "Prediction of copper ions adsorption by attapulgite adsorbent using tuned-artificial intelligence model," *Chemosphere*, vol. 276, p. 130162, Aug. 2021.

[77] S. K. Bhagat *et al.*, "Prediction of lead (Pb) adsorption on attapulgite clay using the feasibility of data intelligence models," *Environmental Science and Pollution Research*, Feb. 2021.

[78] F. Amini and G. Hu, "A two-layer feature selection method using Genetic Algorithm and Elastic Net," *Expert Systems with Applications*, vol. 166, no. December 2019, p. 114072, 2021.

[79] A. Oztekin, L. Al-Ebbini, Z. Sevkli, and D. Delen, "A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology," *European Journal of Operational Research*, vol. 266, no. 2, pp. 639–651, 2018.

[80] Z. Hu, Y. Bao, T. Xiong, and R. Chiong, "Hybrid filter-wrapper feature selection for short-term load forecasting," *Engineering Applications of Artificial Intelligence*, vol. 40, pp. 17–27, 2015.

[81] J. H. Cheng, D. W. Sun, and H. Pu, "Combining the genetic algorithm and successive projection algorithm for the selection of feature wavelengths to evaluate exudative characteristics in frozen-thawed fish muscle," *Food Chemistry*, vol. 197, pp. 855–863, 2016.

[82] R. A. Welikala *et al.*, "Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 64–77, 2015.

[83] M. Monirul Kabir, M. Monirul Islam, and K. Murase, "A new wrapper feature selection approach using neural network," *Neurocomputing*, vol. 73, no. 16–18, pp. 3273–3283, 2010.

[84] H. Aytug, "Feature selection for support vector machines using Generalized Benders Decomposition," *European Journal of Operational Research*, vol. 244, no. 1, pp. 210–218, 2015.

[85] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2016.

[86] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, no. Jun, pp. 983–999, 2006.

[87] H. Sharafi, I. Ebtehaj, H. Bonakdari, and A. H. Zaji, "Design of a support vector machine with different kernel functions to predict scour depth around bridge piers," *Natural Hazards*, vol. 84, no. 3, pp. 2145–2162, 2016.

[88] I. Ebtehaj and H. Bonakdari, "A support vector regression-firefly algorithm-based model for limiting velocity prediction in sewer pipes," *Water Science and Technology*, vol. 73, no. 9, pp. 2244–2250, 2016.