Environment and Water Engineering

Homepage: www.jewe.ir

ISSN: 2476-3683

Prediction of water quality parameters of Nahand River using random forest model optimized with genetic algorithm

Ali Ashraf Sadredini[1], Amin Sharifi[1], Saeed Samadianfard[1], and Milad Sharafi[2]

[1]Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[2]Department of Water Engineering, Faculty of Agriculture, Urmia University, Urmia, Iran

ARTICLE INFO

Paper Type: Research Paper

Received: 20 January 2025
Revised: 02 March 2025
Accepted: 01 April 2025
Published: 15 May 2025

Keywords
Biochemical Oxygen Demand
Dissolved Oxygen
Genetic Algorithm
Machine Learning
Random Forest
Water Quality Prediction

Corresponding author:
A.A. Sadredini
✉ sadraddini@tabrizu.ac.ir

ABSTRACT

This research aims to improve the accuracy of water quality predictions with machine learning models. The goal was to analyze the performance of the Random Forest (RF) model and the hybrid version with a Genetic Algorithm (GA-RF) for the prediction of biochemical oxygen demand (BOD) and dissolved oxygen (DO) for the Nahand River Basin, Iran. The hybrid model was performed using eleven years of daily water quality data from 2013-2023, with 11 input variables, including total nitrogen, total phosphate, nitrite, nitrate, phosphate, nephelometric turbidity unit, water temperature, air temperature, electrical conductivity, pH, and flow. A further six scenarios were created with the inputs from the water quality data set. The models were evaluated statistically with the coefficient of determination (R^2), root mean square error (RMSE), Nash-Sutcliffe efficiency (NS), and Willmott's index of agreement (WI). The results showed that the GA-RF consistently outperforms the standalone RF. In BOD prediction, also GA-RF-6 and RF-5 had R2 values of 0.563 and 0.548, respectively. Also, In DO prediction, GA-RF-5, RF-6 had R2 values of 0.81 and 0.792, respectively. The analysis showed that combining the Genetic Algorithm and Random Forest can improve predictive accuracy in water quality assessments to make better-informed sustainable water resource decisions.

Highlights

- Hybrid GA-RF model outperforms standalone Random Forest.
- Optimized model predicts BOD and DO with high accuracy.
- Genetic Algorithm enhances machine learning for water quality.

How to cite this paper:
Sadredini, A.A., Sharifi, A., Samadianfard, S., & Sharafi, M. (2025). Prediction of water quality parameters of Nahand River using random forest model optimized with genetic algorithm. *Environment and Water Engineering*, *11*(3), 346-355. https://doi.org/10.22034/ewe.2025.503129.2000

1. Introduction

Surface water resources represent a significant element of the Earth's hydrological cycle, supporting ecological and human systems. Rivers, lakes, and wetlands serve as dynamic reservoirs that store, move, and redistribute freshwater (Li et al., 2018). These ecosystems are habitat for biodiversity, upward to/making climate regulation (Gleeson et al., 2020). The surface water devices, including river systems, lakes, and wetlands, play a critical role in the global and regional water cycle, informing precipitation and regulating climate (Gleeson et al., 2020). Additionally, surface ecosystems maintain an amount of biodiversity and encapsulate a complex suite of food webs (Matveyeva and Chernov, 2019). Their hydrological flows and dynamics are important for gauging and anticipating climate impacts and adaptations, maintenance of, and foresight on sustainable management practices, and sustainability of the life web (Papa et al., 2023). Moreover, surface waters are critically important to agriculture. The global freshwater consumption devoted to irrigation is estimated to be 69%, and roughly 62% of this irrigation output emerges from surface water (Siebert et al., 2013).

Deteriorating water quality affects aquatic ecosystems, agriculture, and industry. Polluted water negatively affects crop quality and water industries (Deka et al., 2024). Monitor surface waters to ensure ecosystem health, human health and public health and sustainable water availability (Hidayat and

Kurniawan, 2024). Predictive modeling on key parameters such as biochemical oxygen demand (BOD) and dissolved oxygen (DO) are extremely critical to pollution management and determining appropriate water management.

With the increasing adoption of computational methods, intelligent models have been widely applied in studies predicting various water resource parameters. Researchers have highlighted the superior accuracy of these models compared to empirical relationships (Beiranvand & Rajaee, 2022; Moein et al., 2022). In recent decades, intelligent models have garnered significant attention across multiple disciplines, including water engineering. For instance, Raheli et al. (2017) conducted an uncertainty assessment of the MLP by developing a hybrid FFA-MLP (firefly algorithm-optimized MLP) model for BOD and DO prediction in Malaysia's Langat River. The analysis revealed that both MLP and FFA-MLP models achieved high predictive accuracy for BOD and DO simulations, demonstrating confidence levels of 72% and 91%, respectively. In a related study, Lorestani et al. (2020) examined longitudinal variations in water quality parameters within the Kalan Malayer Dam reservoir. The results indicated that BOD and chemical oxygen demand (COD) levels consistently exceeded WHO standards throughout the sampling period, rendering the water unsuitable for domestic and drinking purposes. This degradation was primarily attributed to upstream pollutant influx. In a separate methodological advancement, Cao et al. (2020) developed a novel DO prediction approach combining K-means clustering with GRU neural network modeling. The analysis demonstrated that the GRU model achieved superior accuracy and flexibility relative to the K-means approach, exhibiting an average absolute error of 0.26 and mean absolute percentage error (MAPE) of 3.5%.

Rafiei et al. (2023) evaluated water quality indices and self-purification potential in the Balighlychay and Gharasu rivers using the QUAL2Kw Model. The evaluation considered the most important water quality characteristics, including dissolved oxygen (DO), biochemical oxygen demand (BOD), nitrate, and phosphate concentrations. The self-purification capacities, based on analyses, were extraordinary because during wet months of high-flow conditions, the systems had a self-purification capacity of 226.6% (DO) and 90.3% (BOD), and during dry months of low-flow conditions still had considerable potential with DO (281.7%), and BOD (89.1%) purification. Despite these natural remediation capabilities, the study identified substandard water quality conditions throughout most monitored river segments. Complementing these findings, Hassani and Ashafteh (2023) investigated DO dynamics in the Ekbatan reservoir using the CE-QUAL-W2 model, with a particular focus on thermal stratification effects. Their results established a clear inverse relationship between temperature and DO concentrations: a 68% temperature increase corresponded to a 37.5% DO reduction, while temperature decreases produced proportional DO increases. In a methodological advancement, Roshanghar and Davodi (2023) implemented an innovative deep learning approach for DO prediction, combining Long Short-Term Memory (LSTM) networks with dual pre-processing techniques. Their framework applied Discrete Wavelet Transform for spatial analysis and Complete Ensemble Empirical Mode Decomposition for temporal modeling across five Savannah River monitoring stations. This integrated approach achieved significant error reduction, with wavelet-based spatial modeling decreasing RMSE by 2% and empirical mode decomposition yielding a 15% RMSE improvement in temporal predictions. The optimal performance was observed in the one-day-ahead temporal modeling scenario, achieving an exceptional RMSE of 0.017 ppm.

The primary goal of the present research is to develop and evaluate a hybrid structure, which is referred to as a Random Forest- Genetic Algorithm (RF-GA) model, to predict the levels of Biochemical Oxygen Demand (BOD) and Dissolved Oxygen (DO) in the Nahand River. The study addresses two important shortcomings in the literature. First, the application of a combined RF-GA modeling framework to forecast BOD and DO is a new methodological contribution. Second, there are no previously published studies modelling the water quality of the Nahand River in Iran. The proposed RF-GA model is used in the Nahand watershed to improve the accuracy of prediction and to provide local water management strategies with a deeper tool.

2. Materials and Methods

2.1 Study area

This research was accomplished using BOD and DO data that was collected daily and represented the full 11-year time frame of January 1, 2013 to December 31, 2023 from sites in the Nahand River. The Nahand dam was constructed in the East Azerbaijan province of Iran, 43 km north of Tabriz (the largest city in the province). The dam was built to supply a portion of the water of the Nahand River (1 m3/s) for the city of Tabriz. The Namand River is one of the large branches of the Ajichay River. The Nahand Dam is an earthfill dam with a clay core in its center, height at a maximum of 35 m. The crest length of the dam is 730 m, and the width is 8 m. The foundation of the river bed where the dam is being constructed is at an elevation of 1570 m above sea level, and the thickness of the shell at that elevation is about 250 m. The mentioned area is situated at a geographical site of 38° 13′ north latitude and 28° 46′ east longitude. Figure 1 shows the geographical location of the river leading to the Nahand Dam.

Table 1 shows the statistical characteristics and parameters of the daily data used during the statistical period. The following parameters were used as input parameters, including total nitrate (TN), total phosphate (TP), nitrite (NIT), nitrate (NITR), phosphate (P), nephelometric turbidity unit (NTU), water temperature (C°), air temperature (AT), electrical conductivity (EC), potential of hydrogen (pH), and flow rate (Q) and the output parameters were Biochemical oxygen demand (BOD) and dissolved oxygen (DO). In addition, 70% of the data (257 data points) was utilized during the training stage of the models, and 30% of the data (121 data points) was used randomly for the testing stage.

Fig. 1 The geographical location of the Nahand Dam

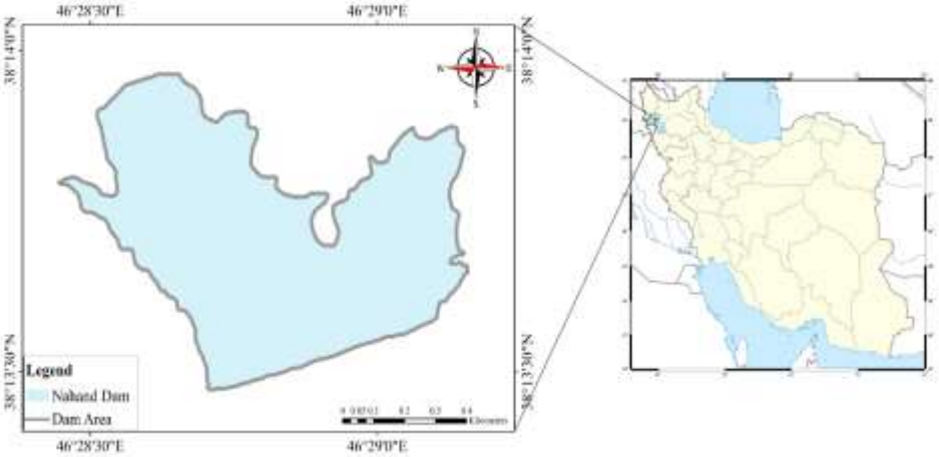


Table 1 Statistical parameters for the variables used

Parameter/ unit	TN ppm	TP ppm	NITR ppm	NIT ppm	P ppm	NTU -	WT °C	AT °C	Q m^3/s	pH -	EC µS/cm	DO ppm	BOD ppm
Minimum	0	0	0.1	0	0	0	0	0	0	5.5	376	0	0
Maximum	3.4	0.1	1.7	0.1	0.2	1000	25.3	36	15	8.9	3280	11.8	3.8
Mean	0.6	0	0.5	0	0	60.8	12.9	16	0.2	8.2	1019	7.1	2
Standard deviation	0.7	0	0.2	0	0	178.2	6.3	10.4	0.9	0.3	478.6	1.6	0.5

Table 2 shows the different input-output combinations used in the models, selected based on the Pearson correlation coefficient between the input parameters and the target variables (BOD and DO). The first scenario includes the parameters with the lowest correlation, while each subsequent scenario incorporates additional parameters with progressively higher correlations. Thus, TN and TP had the lowest correlation with the target parameters, and the discharge had the highest correlation with the target parameter.

Table 2 Different combinations for predicting BOD and DO parameters

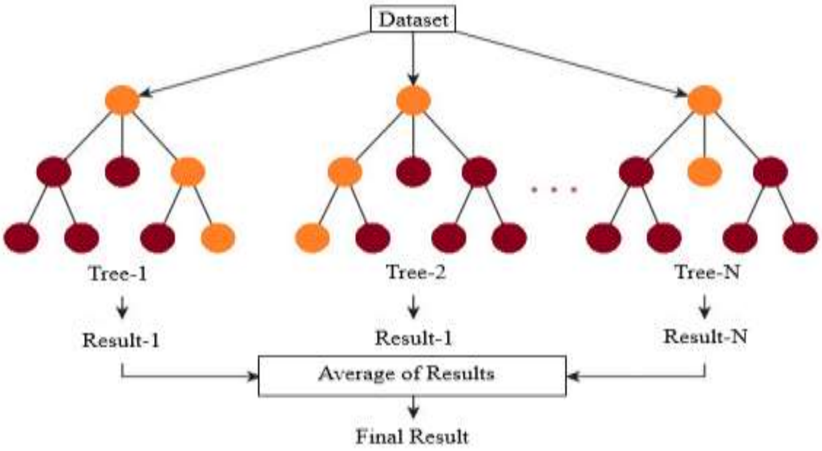
Scenario	Input											Output
1	TN	TP										BOD-DO
2	TN	TP	NIT	P								BOD-DO
3	TN	TP	NIT	P	AT	NTU						BOD-DO
4	TN	TP	NIT	P	AT	NTU	WT	EC				BOD-DO
5	TN	TP	NIT	P	AT	NTU	WT	EC	NITR	pH		BOD-DO
6	TN	TP	NIT	P	AT	NTU	WT	EC	NITR	pH	Q	BOD-DO

2.2 Random forest

The RF model was introduced first by Breiman (2001). The RF is a learning algorithm for a set of multiple regression trees. Compared with simple decision trees, RF runs more efficiently on high-dimensional datasets and is more accurate and robust against noise. Furthermore, RF has many advantages over older intelligent algorithms (Rodriguez-Galiano et al., 2015; Smith et al., 2013; Wang et al., 2015). RF algorithm exhibits several distinctive advantages that make it particularly suitable for water quality modeling. First, its computational efficiency enables rapid model training while maintaining robust performance with high-dimensional input data, including comprehensive variable importance analysis. Second, the algorithm incorporates intrinsic mechanisms for both generalization error estimation and missing data imputation, preserving predictive accuracy even under conditions of substantial data loss (Breiman, 2001).

In this study, the Random Forest (RF) model was implemented in Python using the scikit-learn library. The daily water quality dataset was divided into training and testing subsets. For each of the six scenarios, a separate RF model was trained, with hyperparameters such as the number of trees and maximum depth tuned through grid search and cross-validation. The model was executed multiple times to ensure stability and consistency of results. After training, the ensemble predictions were generated by averaging outputs from all decision trees. Variable importance was extracted to interpret model behavior and identify dominant factors influencing BOD and DO predictions.

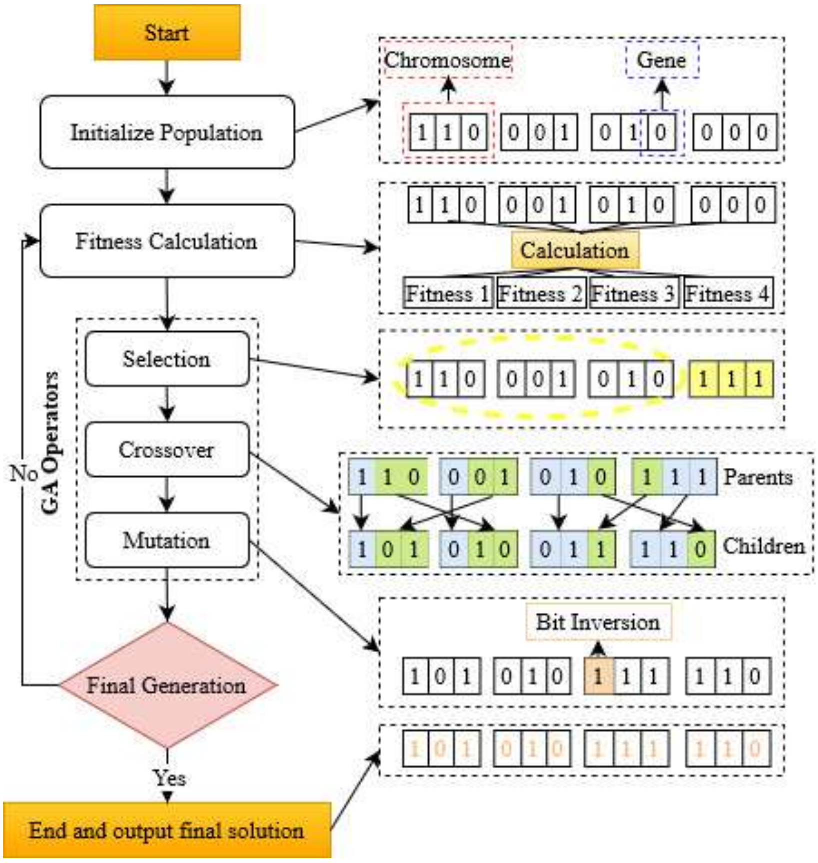
Fig. 2 Schematic structure of the RF



2.3 Genetic algorithm

GA was first introduced by Holland (1992). Although this algorithm is known as one of the oldest metaheuristic techniques based on Darwin's theory of evolution, it is one of the most efficient algorithms used to solve optimization problems. As shown in Fig. 3, the overall design of a GA consists of a population in which each component, called a chromosome, is considered the solution to each problem. The search in this algorithm begins with the random generation of the population. The subsequent generations of this population are expanded by operators (selection, crossover, and mutation). Based on the principle of survival, the population develops generation by generation to provide more suitable solutions. Like natural evolution, this method causes the next generation population to be better adapted to the environment than the previous generation, and the optimal individual among the final generation population can be considered an approximate optimal solution to the problem at hand (Goldberg and Holland, 1988).

Fig. 3 The general process of the GA model (Achite et al., 2023)



In the present study, the RapidMiner Studio version 2.10 model was used to implement the RF and GA-RF models. Table 3 shows the hyperparameters used to achieve the best performance in the GA-RF model to predict the target parameters. The hyperparameters were obtained by the hybrid model in the form of achieving the minimum error value. So that the hybrid model finds the best value in each parameter that reduces the prediction error value, to improve the accuracy of the model.

Table 3 Hyperparameters used in the hybrid model

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
Number of Trees	21	10	60	33	90	46
Maximum Depth	4	5	29	100	92	64
Confidence Level	0.05	0.34	0.29	0.23	0.27	0.3
Minimum Leaf Size	93	3	16	63	63	82
Minimum Size for Splitting	50	23	36	59	47	14
Number of Pre-Pruning Iterations	28	17	25	26	37	52
Subset Ratio	0.22	0.44	0.79	0.17	0.89	0.14
Local Random Function	54	28	47	73	68	69

2.4 Evaluation criteria

The performance of the used models was evaluated using R, RMSE, N-S coefficient, and WI using Eqs. 1 through (Amini et al., 2014).

$$R = \frac{\left(\sum_{i=1}^{N}(Oi-\bar{O})-\frac{1}{N}(Pi-\bar{P}\right)}{\sqrt{\sum_{i=1}^{N}(O_i-\bar{O})^2\sum_{i=1}^{N}(P_i-\bar{P})^2}} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Pi - Oi)^2} \quad (2)$$

$$NS = 1 - \left[\frac{\sum_{i=1}^{N}(Oi-Pi)^2}{\sum_{i=1}^{N}\left(Oi-\overline{O}i\right)^2}\right] \quad (3)$$

$$WI = 1 - \left[\frac{\sum_{i=1}^{N}(Oi-Pi)^2}{\sum_{i=1}^{N}\left(\left|Pi-\overline{O}i\right| + \left|Oi-\overline{O}i\right|\right)^2}\right] \quad (4)$$

where O_i and P_i are the observed and predicted values, respectively, and N is the number of observations.

3. Results and Discussion

This study applied Random Forest (RF) and Genetic Algorithm-optimized Random Forest (GA-RF) models to predict biochemical oxygen demand and dissolved oxygen concentrations in the Nahand River using daily water quality data from 2013 to 2023. The performance of the models was evaluated under six scenarios. Table 4 presents the evaluation metrics for each scenario during the testing period.

3.1 BOD Prediction results

The performance of the RF and GA-RF models in predicting biochemical oxygen demand across six input scenarios is presented in Table 4. The RF model showed a steady improvement from scenario 1 to scenario 5. In RF-1, the prediction capability was notably poor with a coefficient determination of 0.01, a root mean square error of 0.37 ppm, a negative Nash–Sutcliffe of -0.17, and a low Willmott index of 0.39. These values indicate that the model was not able to capture the patterns in the BOD data with the limited set of inputs.

However, with the addition of key water quality parameters in subsequent scenarios, the model's accuracy improved significantly. The RF-2 model showed a clear jump in results (R^2 = 0.449, RMSE = 0.25 ppm, NS = 0.45, WI = 0.79). It showed continual gains in scenarios 3 to 5, and RF-5 was show to be the best RF model compared to the standard RF (R^2 = 0.548, RMSE = 0.23 ppm, NS = 0.54, WI = 0.83) and was equal to the results of RF-6 (R^2 = 0.533, RMSE = 0.23 ppm, NS = 0.53, WI = 0.83). The application of a genetic algorithm to optimize input selection and model parameters greatly improved prediction accuracy from the RF base models. The genetic algorithm-RF model (GA-RF), in every scenario, was superior in prediction accuracy to the base RF. In GA-RF-1, it's seen that the GA-RF model was still not great (R^2 = 0.044, RMSE = 0.34 ppm), but outperformed RF-1. The optimal values occurred in GA-RF-6 (R^2 = 0.563, RMSE = 0.22 ppm, NS = 0.57, WI = 0.84), providing a more stable and accurate estimation of BOD values. GA-RF-4 and GA-RF-5 had competitive results with only slightly lower and comparable accuracy levels. The GA-RF not only reduced prediction error but also improved the reliability of the model through optimizing feature combinations and parameter settings.

3.2 DO Prediction results

The results of the dissolved oxygen prediction using RF and GA-RF models across various scenarios also reveal significant insights. The RF-1 model had a weak performance with R^2 = 0.176, RMSE = 1.32 ppm, NS = 0.16, and WI = 0.54, reflecting poor model capability when limited inputs were used.

The models improved significantly and were performing well from scenario 2 onward. The models' improved accuracy to a great extent, starting from scenarios 3 to 6. The RF-4, RF-5, and RF-6 models all had high coefficients of determination at approximately 0.792, low RMSE at approximately 0.66 to 0.67 ppm, and high NS and WI values (NS > 0.78; WI > 0.93) were observed in these scenarios, as well. In addition, results showed that variables related to air temperature and turbidity are therefore significant predictors of dissolved oxygen concentrations or factors for oxygen solubility and microbial activity.

GA-RF was the best-performing of all the models. The GA-RF-5 and GA-RF-6 models had the same results (R^2 = 0.81, RMSE = 0.63 ppm, NS = 0.81, and WI = 0.95) and slightly improved performance when compared with the RF models. The GA-RF-5 and GA-RF-6 scenarios also exemplified the model's ability to consistently predict DO concentrations in the existence of different environmental background conditions. GA-RF-4 followed closely with similar accuracy (R^2 = 0.81, RMSE = 0.65 ppm, NS = 0.80, WI = 0.94), making it a competitive alternative.

Table 1 Statistical Analyses of the M_5 Tree and Kstar ModelTable 4 Results of the evaluation of the studied models for the BOD and DO parameters

Model	BOD				DO			
	R^2	RMSE (ppm)	NS	WI	R^2	RMSE (ppm)	NS	WI
RF-1	0.01	0.37	-0.17	0.39	0.18	1.32	0.16	0.54
RF-2	0.45	0.25	0.45	0.79	0.42	1.10	0.42	0.77
RF-3	0.49	0.24	0.49	0.82	0.74	0.77	0.71	0.90
RF-4	0.50	0.24	0.50	0.82	0.79	0.67	0.78	0.94
RF-5	0.55	0.23	0.54	0.83	0.79	0.67	0.79	0.93
RF-6	0.53	0.23	0.53	0.83	0.79	0.66	0.79	0.93
GA-RF-1	0.04	0.34	0.04	0.35	0.31	1.21	0.29	0.61
GA-RF-2	0.52	0.24	0.52	0.81	0.59	0.93	0.59	0.84
GA-RF-3	0.53	0.24	0.52	0.82	0.77	0.70	0.77	0.93
GA-RF-4	0.56	0.23	0.55	0.84	0.81	0.65	0.80	0.94
GA-RF-5	0.55	0.23	0.55	0.83	0.81	0.63	0.81	0.95
GA-RF-6	0.56	0.22	0.57	0.84	0.81	0.63	0.81	0.95

3.3 Influence of input scenarios on BOD and DO prediction

The performance of the RF and GA-RF models under different input scenarios revealed the important role of specific water quality variables in improving the prediction accuracy for both BOD and DO parameters. Evaluating scenarios 1 through 6 demonstrated that the inclusion of additional inputs, especially nutrient concentrations and physical water characteristics, had a significant impact on the results.

For the prediction of biochemical oxygen demand, scenario 2 showed noticeable improvement compared to scenario 1. This improvement was due to the addition of nitrite and phosphate parameters. nitrite is a key source of nitrogen for aquatic plants and supports their growth in freshwater ecosystems. As the plant biomass increases, the demand for oxygen required for microbial decomposition of organic matter also increases. This leads to higher levels of biochemical oxygen demand. This relationship has been noted in previous studies, including Antia et al. (1963). Phosphate also aids this process because it promotes algal and aquatic vegetation growth. When this organic matter is decomposed, it leads to increased oxygen use in the water, and, therefore, affects water quality and reduces oxygen available to aquatic organisms. This relationship was confirmed in the findings of Bhateria and Jain (2016).

For dissolved oxygen, the prediction in scenario three, which included air temperature and turbidity measured in NTU, had the best model performance. An increase in air temperature corresponds with increases in water temperature, which amounts to decreases in the solubility of oxygen. The relationship between warmer temperatures and lower dissolved oxygen concentrations has been discussed by Kadlec and Reddy (2001). Turbidity affects light availability in the water and has positive or negative effects on biological activity that impacts dissolved oxygen. Increased turbidity may limit photosynthesis and also may indicate suspended particles that affect oxygen levels via microbial activity and biochemical processes. These effects are described in the work of Schenk and Bragg (2021).

The results of all six scenarios demonstrate that the models' ability to relate changes in oxygen-related parameters is increased with the inclusion of substantive physical and chemical input variables. Nutrients such as nitrite and phosphate appear to be significant factors in adding organic load and therefore oxygen demand, while air temperature and turbidity quantity affect the availability of oxygen, oxygen saturation, and biological activity in the water. The overall improvement of the model's performance with the addition of these inputs emphasizes the need to select meaningful variables that adequately note the real-world dynamics of water quality processes. This outcome is also consistent with several simple ecological principles and adds credibility to the use of intelligent models for the management of complex river systems. Careful selection of input parameters not only increases the accuracy of predictions when the model is evaluated, but can also assist with assessing the key pollutants and primary regulating mechanisms used to assess afterpollutant and oxygen variations in the water.

3.4 Comparative analysis with existing literature

Comparison with previous studies further highlights the effectiveness of the proposed models. For instance, Li et al. (2017) applied Multivariate Adaptive Regression Splines, Artificial Neural Networks, and SVM (optimized by Particle Swarm Optimization (PSO)) to predict DO in a river system. The SVM-PSO predicted the DO with 1.27 mg/l RMSE, which is considerably higher than the result of 0.63 mg/l obtained in this research using GA-RF.

Similarly, Fathima et al. (2014) predicted BOD using data mining algorithms and achieved an RMSE of 0.47 mg/l, which is again higher than the 0.22 mg/l reported here using GA-RF. These comparisons emphasize the superior accuracy and robustness of the hybrid GA-RF model proposed in this research for both BOD and DO prediction.

Fig. 4 shows the scatter plots of BOD and DO parameters for the best scenario in both the single and hybrid models. The criterion for selecting the best scenario was based on having the highest correlation and the lowest error. However, in a scenario that had the same correlation and error as the other scenario, the scenario with the lowest inputs was preferred. In all scatter plots, the 1:1 line, a 45-degree line, indicates that the closer the points are to it, the higher the model accuracy and the closer the observed data will be to the predicted data. Comparing the coefficients of determination in the four plots shows that the GA-RF performed better than the individual RF model and increased the correlation between the observed and predicted data. However, the fifth scenario also performed well in predicting the BOD and DO parameters, and showed that the model could achieve adequate accuracy even in the absence of the river flow parameter.

Fig. 4 Scatter plots of BOD and DO parameters for the best scenario in each model: a) RF-5, b) RF-6, c) GA-RF-6, (d) GA-RF-5.

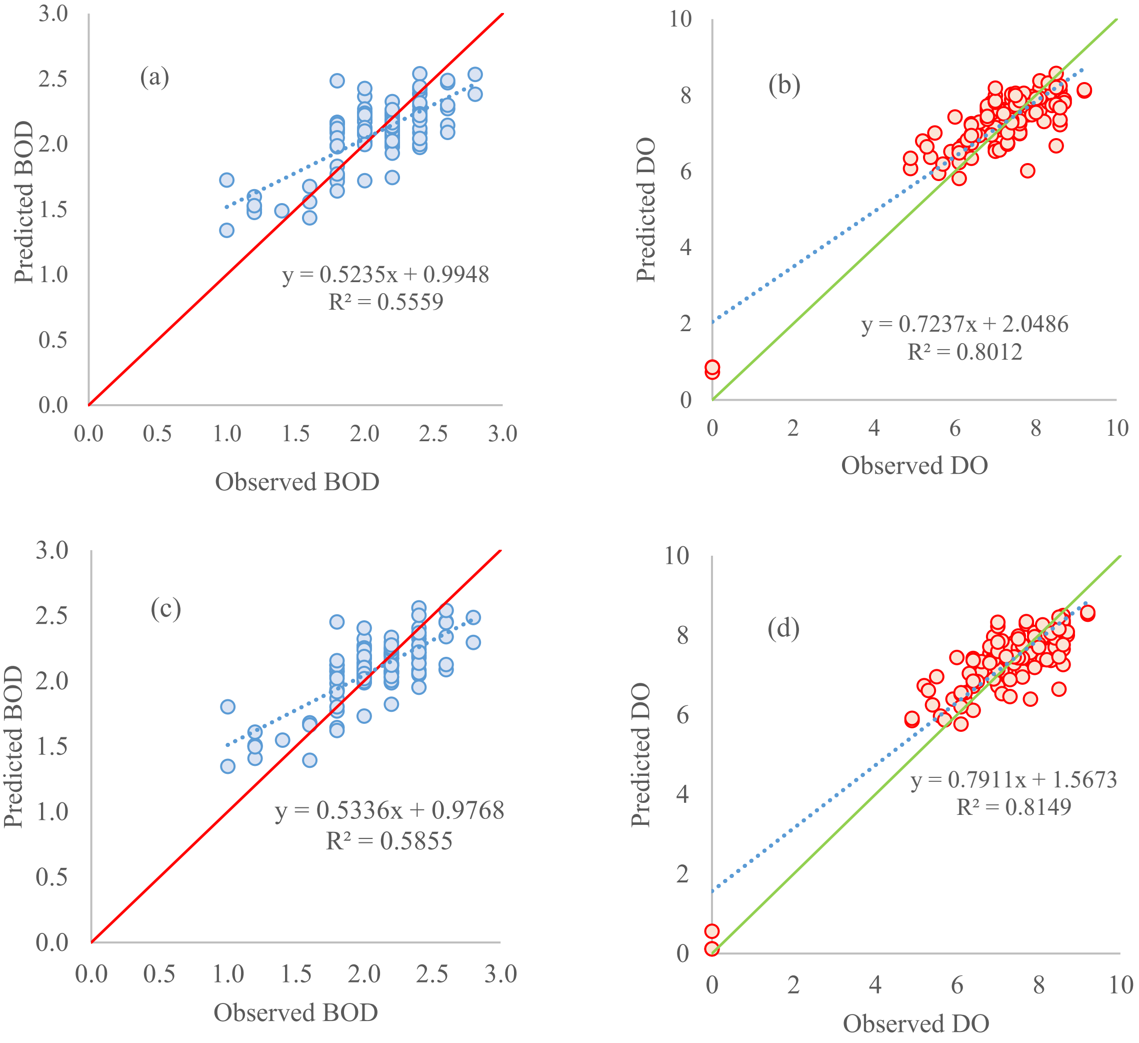
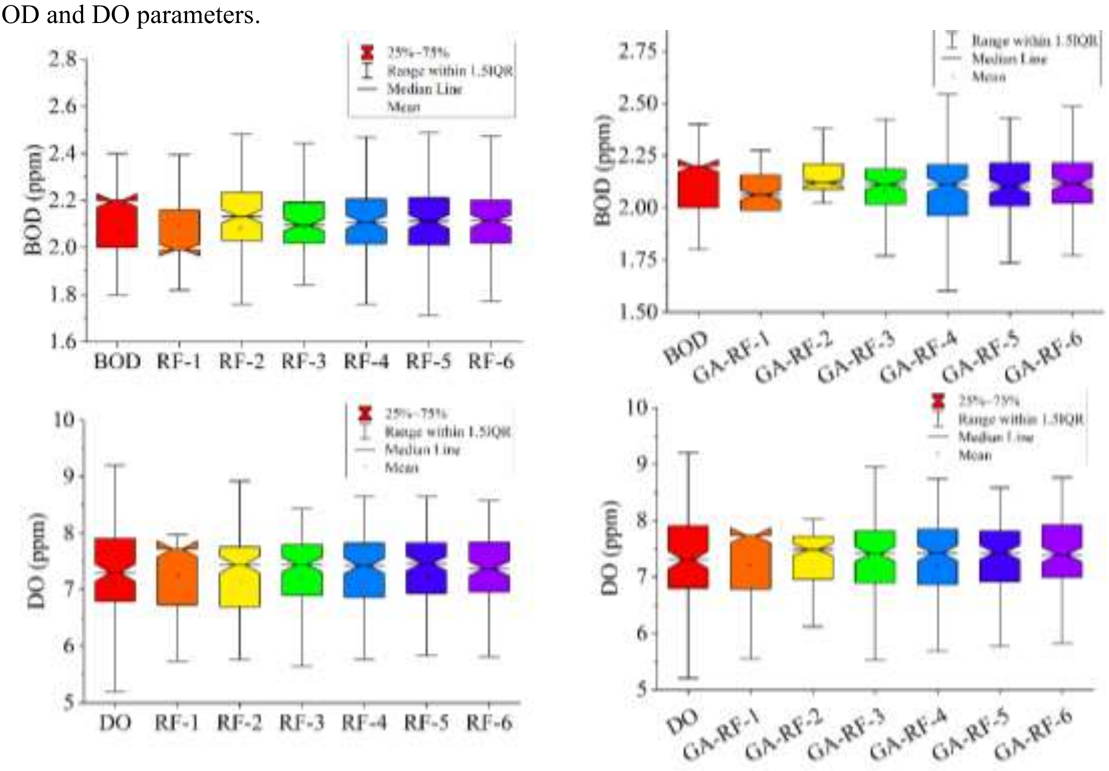


Fig. 5 shows the box plots of BOD for the models used in six scenarios. The first box corresponds to the observed value, and the other boxes correspond to the scenario used in each model. The hollow cube also indicates the mean value of each series, so that the closer the boxes are to the observed mean, the greater the accuracy of the model in that scenario. A comparison of different scenarios for the RF model reveals that none of the scenarios had a value similar to the observed value, and therefore none of the scenarios achieved very high accuracy. However, the last two scenarios (fifth and sixth) had a mean value closer to the observed value. Additionally, for the RF model, comparing the observed value with the scenarios reveals that the DO has a wide range of approximately 2.5 to 2.9, whereas all scenarios have a shorter range than the observed value; therefore, the model has underestimated the DO in all scenarios. However, the sixth scenario had the best performance in the single model due to having the closest mean and median to the observed value. Comparison of scenarios for the GA-RF model also shows that the fifth and sixth scenarios had a box closer to the observed value. However, the sixth scenario had a slight overestimation in the 75-25% range of the top of the box, and therefore, the fifth scenario had a more appropriate performance.

Fig. 5 Box plots of BOD and DO parameters.



4. Conclusion

The objective of this research was to evaluate the predictive ability of the Random Forest (RF) model and its hybrid with the Genetic Algorithm (GA-RF) model for two important water quality variables for rivers, biochemical oxygen demand and dissolved oxygen concentration. These variables must be adequately predicted to monitor, manage, and protect freshwater ecosystems efficiently. The main goal of the study was to identify the best combinations of inputs and model structures to create accurate and reliable predictions of BOD and DO. To do this, the study examined six different input scenarios of different water quality indicators to determine their predictive capability and how the models behaved.

1. The RF model, with five inputs of temperature, pH, nitrate, nitrite, and phosphate in the fifth scenario, performed with reasonable accuracy with a Nash–Sutcliffe coefficient of 0.54 and Willmott's index of 0.83. The GA-RF model performed better with the sixth scenario (using all input variables), achieving an NS of 0.57 and WI of 0.84.

2. For dissolved oxygen estimation, the sixth scenario provided the best performance in the RF model (NS = 0.79, WI = 0.93). In contrast, the GA-RF model achieved the highest accuracy in the fifth scenario (NS = 0.81, WI = 0.95), indicating that careful selection and optimization of inputs significantly influence prediction performance for DO.

3. Across nearly all scenarios, the GA-RF hybrid model outperformed the standalone RF model in both BOD and DO prediction tasks. This increase in predictive ability was likely a result of the GA's ability to optimize feature selection, as well as model hyperparameters, which led to models that were more robust and generalized.

The present research can be extended in multiple directions, including additional environmental and hydrological variables such as land cover characteristics, precipitation, and streamflow, which may improve predictive accuracy and applicability of future models. Using different metaheuristic optimization algorithms may also provide additional insights into the improvement in model performance. There may be similar potential for using deep learning approaches like Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) when analyzing temporal water quality data. This will support more effectively made and data-driven designs for water resource management.

Statements and Declarations

Data availability

The data used in this research will be available on reasonable request.

Conflicts of interest

The author of this paper declared no conflict of interest regarding the authorship or publication of this paper.

Author contribution

M. Sharafi: Methodology, Investigation, Conceptualization, Writing Original Draft. A.A. Sadredini: Supervision. S. Samadianfard: Review-Editing. A. Sharifi: Review-Editing.

AI Use Declaration

Portions of the language editing and writing refinement in this manuscript were assisted by the use of artificial intelligence tools, under the full supervision and final approval of the authors.

References

Achite, M., Samadianfard, S., Elshaboury, N., & Sharafi, M. (2023). Modeling and optimization of coagulant dosage in water treatment plants using hybridized random forest model with genetic algorithm optimization. *Environment, Development and Sustainability*, 25(10), 11189-11207. DOI: https://doi.org/10.1007/s10668-022-02523-z.

Antia, N., McAllister, C., Parsons, T., Stephens, K., & Strickland, J. (1963). Further measurements of primary production using a large-volume plastic sphere. *Limnology and Oceanography*, 8(2),166–183. DOI: https://doi.org/10.4319/lo.1963.8.2.0166.

Amini, A., Ghazvinei, P. T., Javan, M., & Saghafian, B. (2014). Evaluating the impacts of watershed management on runoff storage and peak flow in Gav-Darreh watershed, Kurdistan, Iran. *Arabian Journal of Geosciences*. DOI: https://doi.org/10.1007/s12517-013-0950-1

Beiranvand, B., & Rajaee, T. (2022). Application of artificial intelligence-based single and hybrid models in predicting seepage and pore water pressure of dams: A state-of-the-art review. *Advances in Engineering Software*, 173(1),103-117.DOI: https://doi.org/10.1016/j.advengsoft.2022.103121.

Bhateria, R., & Jain, D. (2016). Water quality assessment of lake water: a review. *Water Resources Management*, 2(2),161–173. DOI: https://doi.org/10.1007/s40899-015-0014-7.

Breiman, L., (2001). Random forests. Mach. Learn. 45(1):5–32. DOI: https://doi.org/10.1023/A:1010933404324.

Cao, X., Liu, Y., Wang, J., Liu, C., & Duan, Q. (2020). Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network. *Aquacultural Engineering*, 91(1),102-121. DI: https://doi.org/10.1016/j.aquaeng.2020.102121.

Deka, P., Dutta, P. K., Kalita, S., Nath, R. K., & Dutta, P., (2024). Water Management in Organic Farming. In Advances in Organic Farming. *Apple Academic Press*, (pp.151–161). https://doi.org/10.1201/9781003338681-11.

Fathima, A., Mangai, J. A., & Gulyani, B. B. (2014). An ensemble method for predicting biochemical oxygen demand in river water using data mining techniques. *International Journal of River Basin Management*, 12(4):357–366. DOI: https://doi.org/10.1080/15715124.2014.917318.

Gleeson, T., Wang-Erlandsson, L., Porkka, M., Zipper, S. C., Jaramillo, F., Gerten, D., Fetzer, I. Cornell, S.E., Piemontese, L., & Gordon, L.J. (2020). Illuminating water cycle modifications and Earth system resilience in the Anthropocene. *Water Resources Research*, 56(4): 1-18. DOI: https://doi.org/10.1029/2019WR024957.

Goldberg, D., & Holland, J. (1988). Genetic algorithms and machine learning. I*n Proceedings of the sixth annual conference on Computational learning theory*, 3(2),95–99. DOI: https://doi.org/10.1023/A:1022602019183.

Hassani, S.Z., & Ashofteh, P.-S. (2023). Modeling of Dissolved Oxygen in Ekbatan Reservoir Using CE-QUAL-W2 Model. *Water Irrigation Management*, 13(4), 983-1000 (In Persian). DOI: https://doi.org/10.22059/jwim.2023.359526.1077

Hidayat, R. D. X., & Kurniawan, A. (2024). Sustainable Water Management in Urban Areas through Smart Water Circulation Systems. *Environmental Earth Sciences*, 1, 1416-1429. DOI: https://doi.org/10.1088/1755-1315/1416/1/012019.

Holland, J. H. (1992). Genetic algorithms. Scientific American, 267(1), 66–73. DOI: https://doi.org/10.1038/scientificamerican0792-66.

Kadlec. RH., & Reddy, K. (2001). Temperature effects in treatment wetlands. *Water Environment Research*, 73(5): 543–557. DOI: https://doi.org/10.2175/106143001X139614.

Li, X., Cheng, G., Ge, Y., Li, H., Han, F., Hu, X., Tian, W., Tian, Y., Pan, X., & Nian, Y. (2018). Hydrological cycle in the Heihe River Basin and its implication for water resource management in endorheic basins. *Journal of Geophysical Research: Atmospheres*, 123(2) :890–914. DOI: https://doi.org/10.1002/2017JD027889.

Li, X., Sha, J., & Wang, Z.-l., (2017). A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. *Hydrology Research*, 48(5): 1214–1225. https://doi.org/10.2166/nh.2016.258.

Lorestani, B., Merrikhpour, H., & Cheraghi, M. (2020). Cross-Sectional Study of Water Quality Changes in Lake of Kalan Malayer Dam(Case Study: 2017-2018). *Journal of Environmental Health Science & Engineering*, 8(1), 99-116 (In Persian). DOI: https://dx.doi.org/10.52547/jehe.8.1.99.

Matveyeva, N., & Chernov, Y., (2019). Biodiversity of terrestrial ecosystems, The Arctic. Routledge England. 40 pp.

Moein, M.M ,.Saradar, A., Rahmati, K., Mousavinejad, S. H. G., Bristow, J., Aramali, V., & Karakouzian, M. (2022). Predictive models for concrete properties using machine learning and deep learning approaches: A review. *Journal of Building Engineering*. 19: 105017. DOI: https://doi.org/10.1016/j.jobe.2022.105017.

Papa, F., Crétaux, J.-F., Grippa, M., Robert, E., Trigg, M., Tshimanga, R.M., Kitambo, B., Paris, A., Carr, A., & Fleischmann, A.S. (2023). Water resources in Africa under global change: monitoring surface waters from space. *Surveys in Geophysics*, 44(1), 43–93. DOI: https://doi.org/10.1007/s10712-022-09721-4.

Rafiei, G., Moezzi, F., Poorbagher, H., Rezaei Tavabe, K., & Nematollahi, M.A. (2023). Assessing Water Quality Indices and Autopurification Capacity of Balighli-Chai and Ghare-Sou Rivers using QUAL2Kw Model. *Environment and Water Engineering*, 9(3), 335-351 (In Persian). DOI: https://doi.org/10.22034/jewe.2022.336023.1756.

Raheli, B., Aalami, M.T., El-Shafie, A., Ghorbani, M.A., & Deo, R.C. (2017). Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River. *Environmental Earth Sciences*, 76(51), 1–16. DOI: https://doi.org/10.1007/s12665-017-6393-1.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71(12), 804–818. DOI: https://doi.org/10.1016/j.oregeorev.2015.06.009.

Roushangar, k., & Davoudi, S. (2023). Dissolved Oxygen Modeling Using Deep Learning and Pre-Processor Methods. *Water Irrigation Management*, 12(4), 983-890 (In Persian). DOI: https://doi.org/10.22059/jwim.2022.345864.1005.

Schenk, L., & Bragg, H. (2021). Sediment transport, turbidity, and dissolved oxygen responses to annual streambed drawdowns for downstream fish passage in a flood control reservoir. *Journal of Environmental Management*, 295(1), 113–125. DOI: https://doi.org/10.1016/j.jenvman.2021.113021.

Siebert, S., Henrich, V., Frenken, K., & Burke, J. (2013). Update of the digital global map of irrigation areas to version 5. *Agricultural Water Management*, 10(2), 2660-2728. DOI: https://doi.org/10.1016/j.agwat.2013.05.007

Smith, P.F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods*, 220(1), 85–91. DOI: https://doi.org/10.1016/j.jneumeth.2013.08.004.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527(241), 1130–1141. DOI: https://doi.org/10.1016/j.jhydrol.2015.05.057.