Intelligent Multi-level Analytics Approach to Predict Water Quality Index

Samaher Al-Janabi(✉) and Zahraa Al-Barmani

Faculty of Science for Women (SCIW), Department of Computer Science, University of Babylon, Babylon, Iraq
samaher@itnet.uobabylon.edu.iq

Abstract. In this paper will, building new miner called intelligent miner based on twelve concentrations to predict water quality called ($IM^{12}CP$-WQI). The main goal of that miner is to find water quality based on twelve types of concentrations that cause water pollution which is: Potential Hydrogen (PH), Total Dissolved Solids (TDS), Turbidity Unit NTU, Total Hardness (TH), Total Alkalinity, Calcium (Ca), Magnesium (Mg), Potassium (K), Sodium (Na), Chloride (Cl), Nitrogen Nitrate (NO3), and Sulfate (SO4). $IM^{12}CP$-WQI consists of four stages; the first stage related to data collection through two Seasons (i.e., summer & winter). The second stage, called pre-processing of data that include: (a) Normalization the dataset to make dataset in range (0, 1). (b) finding correlation between concentrations to know the direct or inverse correlation between those concentrations and their relationship with the water quality index WQI. The second stage involved building an optimization algorithm called DWM-Bat to find the optimum weight for each of the 12 compounds as well as the optimum number of M models for DMARS. The third phase involved building a mathematical model that combines these compounds, based on the development of MARS and drawing on the results of the previous stage, DWM-Bat. The last stage included the evaluation of the results obtained using three types of measures (R2, NSE, D) on the basis of which the value of WQI was determined based on that determined if the value of the WQI is less than 25, then it can be used for the purpose of drinking either between (26–50) it is used in fish lakes, as well as (51–75) it can be used in agriculture. Otherwise, it needs a refining process and reports are produced. Also, the results of the model ($IM^{12}CP$-WQI) were compared with the results of the models (MARS_Linear, MARS_poly, MARS sigmoid, MARS_RBF) under the same conditions and environment, finally; the results shown ($IM^{12}CP$-WQI) is pragmatic predictor of WQI.

Keywords: Deep learning · Multi-level analytics · $IM^{12}CP$-WQI · DWM-Bat · DMARS · Water Quality index

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 63–78, 2023.
https://doi.org/10.1007/978-3-031-27409-1_7

1 Introduction

Water is one of the most important resource for continuous life in the world. The source of water split into two types: "surface and groundwater water, in general, surface water is found in lakes, rivers, and reservoirs, while ground water lies under the surface of the land, it travels through and fills openings in the rocks". The water supply crisis is a harsh truth not only on a national level, but also on a global level. The recent Global Dangers report of the World Economic Forum lists the water supply crisis as one of the top five global risks to materialize over the next decade. On the basis of the current population trends and methods for water use, there is a strong indication that most African countries will exceed the limits of their usable water resources by 2025. The forecasted increases in temperature resulting from climate change will place additional demands on over-used water resources in the form of case dry's [1–6].

The major challenges of water are increasing water demand, water Scarcity, water pollution, inadequate access to safely, affordable water, sanitation, and climate change. That water pollution is the pollutant ion of water source such as oceans, rivers, seas, lakes, groundwater and aquifers by pollutant. Pollutants may end in the water by directly or indirectly application. This is the second most contamination type of the environmental after air pollution. The water quality depends on the eco-system and on human use, such as industrial pollution, wastewater and, more importantly, the overuse of water, which leads to reduce level of water. Water Quality is monitored by measurements taken at the original location and the assessment of water samples from the location achieving low costs and high efficiency in wastewater treatment is a popular challenge in developing states.

Prediction is one of the tasks achieve through data mining and artificial intelligent techniques; to find the discrete or continuous of facts based on the recent facts (i.e., the prediction techniques generated actual values if prediction build from real facet otherwise will generated the virtual values). Most prediction techniques based on the a statistical or probabilities tools for prediction of the future behaviors such as "Chi-squared Automatic Interaction Detection (CHAID), Exchange Chi-squared Automatic Interaction Detection (ECHAID), Random Forest Regression and Classification (RFRC), Multivariate Adaptive Regression Splines (MARS), and Boosted Tree Classifiers and Regression (BTCR)" [7].

Optimization is the process to finding of the best values dependent on the type of objective function for the problem identified. Generally speaking, the problem of maximizing or minimizing. There are many types of optimation namely *continuous optimization, bound constrained optimization, constrained optimization, derivative-free optimization, discrete optimization, global optimization, linear programming and non-differentiable optimization.* There are two types of objective function optimisation, a single objective function and a multiple objective function. In single-objective optimization, the decision to accept or decline solutions is based on the objective function value and there is only one search space. While one feature of multi-objective optimization involves potential conflicting objectives. There is therefore a trade-off between objectives, i.e. the improvement achieved for a single objective can only be achieved by

making concessions to a other objective. There is no optimal solution for all m objective functions at the same time. As a result, multiple-objective functions under a set of constrains specified [8].

The detection of Water Quality Index (WQI) is one of the most important challenges; therefore, this paper suggests a method to build an intelligent miner to predict of WQI through combination between one of optimation algorithm after developing called (DWM-Bat) with one of the prediction algorithms that based on mathematical principle called (DMARS).

2 Building IM^{12}CP-WQI

The model presents in this paper consist of two phases, the first including build the station as electrical circuit to collect the data related to 12 concentrations in real time and saved it on the master computer to preparing and processing in next phase. The second phase focuses on processing dataset after split it based on season identifier, the processing phase pass on many levels of learning to product forecaster can deal with different size of dataset. All the actives of this researcher summarization in Fig. 1 while the algorithm of *IM^{12}CP-WQI* model described in main algorithm. The main hypothesis used

- The file of water have the following: pH, TDS (mg/l), Hardness (as CaCO3) (mg/l), Alkalinity (as CaCO3) (mg/l), Nitrate (mg/l), Sulfate (mg/l), Chloride (mg/l), Turbidity (NTU), Calcium (mg/l), Magnesium (mg/l), Sodium(mg/l), finally Potassium (mg/l).
- Limitation\range for each parameters from Permissible Limit to Maximum Limit: pH [6.5–8.5] to No relaxation, TDS (mg/l) [500 to 2000], Hardness (as CaCO3) (mg/l) [200 to 600], Alkalinity (as CaCO3) (mg/l) [200 to 600], Nitrate (mg/l) [45 to No relaxation], Sulfate (mg/l) [200 to 400], Chloride (mg/l) [250 to 1000], Turbidity (NTU) [5–10 to 12], Calcium (mg/l) [50 to No relaxation], Magnesium (mg/l) [50 to No relaxation], Sodium(mg/l) [200 to No relaxation], finally Potassium (mg/l) [12 to No relaxation] (see Table 1).

2.1 Data Preprocess Stage

Dataset collection through two seasons in region of Iraq. To building the predictor as follow.

- Split the dataset for each season and save it in separated file hold the name of this season.
- apply the normalization for each column in dataset related to each season. Normalize used to all the datasets (*PH, TDS, NTU, TH, TA, Ca, Mg, K, Na, Cl, NO3, and SO4)* to make the value of that concentration in the range [0, 1].
- Finally, apply the correlation for column in dataset related to each season. Correlation Pearson used to correlation all the datasets (*PH, TDS, NTU, TH, TA, Ca, Mg, K, Na,*

Table 1. Main Chemical Parameters related to determined WQI [9]

Parameters	Unit	Recommended water quality standards (Sn)
PH		6.5–8.5
Turbidityx (NTU)	NTU	5
Totalxdissolved solid (TDS)	(mg/L)	500
Calciumx (Ca)	(mg/L)	75
Magnesiumx (Mg)	(mg/L)	50
Chloridex (Cl)	(mg/L)	250
Sodiumx (Na)	(mg/L)	200
Potassiumx (K)	(mg/L)	12
Sulfatex (SO4)	(mg/L)	250
Nitratex (NO3)	(mg/L)	50
Totalxalkalinity (CaCO3)	(mg/L)	200
Totalxhardness (CaCO3)	(mg/L)	500

Cl, NO3, and SO4) to know the correlation between the concentrations. Algorithm 1 explains the main steps of that stage.

Algorithm#1: Pre-Processing
Input: *Water dataset include two season and twelve concentration of (PH, TDS, NTU, TH, TA, Ca, Mg, K, Na, Cl, NO3, and SO4)*
Output: *Split season [id_season] have data in range [0, 1]*
// Split water pollution dataset according [id_season]
For *all samples in water dataset*
  IF *season_name = id_season*
    create a file called season_name
    Put all the concentrations related to season_name in that file
  End IF
End For
// Apply normalization
For *each row in water dataset*
  For *each column in water dataset*
    Compute Normalization
    $\text{Normalization} = \frac{x}{maxx}$
  End For
End For
// Apply correlation
For *each row in water dataset*
  For *each column in water dataset*
    Compute Correlation_Pearson
    $P_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}$
  End For
End For
End pre-processing

Main Algorithm: IM[12]CP-WQI

Input: *water datasets of a two seasons include twelve concentrations* of water
(i.e., PH, TDS, NTU, TH, TA, Ca, Mg, K, Na, Cl, NO3, and SO4)

Output: *Predict the water quality index*

*// **Pre-Processing Stage***

```
1:    For each row in water dataset
2:        Call split Seasons
3:        For each column in water dataset
4:            Call Normalize
5:        End for
6:    End for
```

*// **Build DWM –Bat – DMARS Predictor***

```
7:    For each id_seasons
8:        For i in range (1: total number of samples [id_seasons])
9:            Split dataset according to 5- Cross-Validation into Training and
              Testing dataset
10:       End for
11:       For each Training part not used
12:           Call DWM –Bat     //determined the weights & number of the M
13:           Call DMARS        //predictive the value of twelve concentrations
14:       End for
15:       For each Testing part not used
16:           Test stopping conditions
17:               IF max error generation < Emax
18:                   Go to step 25
19:               Else
20:                   GO to step 8
21:               End IF
22:       End for
23:   End for
```

*// **Evaluation stage***

*24: **Call** Evaluation*

25: Generated report based on the result values of Water Quality Index (WQI)

End IM[12]CP-WQI

Fig. 1. Intelligent miner based on twelve concentrations to predict water quality

Where cov is the covariance between quantitative x and y, σ_x the standard deviation of x, σ_y the standard deviation of y, μ_x the average of x, μ_y the average of y, and E the expectation values.

2.2 Determine Weights of Concentrations and Number of Model (DWM-Bat)

In general, the BA is failing in satisfy the goal of it, when it arrives as max number of iterations without finding the goal, while it is a success in its' work when satisfy the following three steps (i.e., Evaluate the fitness of each Bat, Update individual and global bests, Update velocity and position of each Bat). These steps are repeated until some stopping condition is met. The goal of DWM-Bat is to determine the optimal (weight for each concentration, and number of base model of MARS "M"). Algorithm 2 shows the DWM-Bat step.

Algorithm#2: ***DWM –Bat***

Input: *Split seasons [id_seasons] have data in range [0, 1]*

Output: *Optimal of (weight for each concentration, number of base model "M")*

Initialization*: Pos: Position of population, NB :Population Size, W : weights of concentration, v_i :Velocity of population, Minimum (M),Maximum (M), Lower bound (Lb),Upper bound (Ub), f(pMBest), f(pwBest)=0.*

```
1:    For i=1 to max_iter          //max number of iteration
        // Determine number of base model(M)
2:      For number in base model (M) do
3:        Calculate fitness function
4:        IF fpM is better than f(pMBest)
5:          pMBest =fpM
6:        End IF
7:      End for
        // Determine weights
8:      For random weights w in W do
9:        Calculate fitness function
10:       IF fpw is better than f(pwBest)
11:         pwBest =fpw
12:       End IF
1         End for
3:
14:   End for
15:   gwBest = best w in W
16:   For random weights w in W do
17:     Calculate the velocity of weight      // as equation(1)
18:     Calculate the position of weight      // as equation(2)
19:   End for
20:     gmBest = best m in M
21:   For random weights m in M do
22:     Calculate the velocity of M       // as equation(3)
23:     Calculate the position of M       // as equation(4)
24:   End for
    End DWM –Bat
```

2.3 Develop MARS(DMARS)

Here we will train and predict concentrations movements for several epochs and see whether the predictions get better or worse over time. The Algorithm 3 shown how execution the DMARS.

Algorthim#3: DMARS

Input: *Training datasets, Split seasons [id_seasons] and the*
Optimal of (weight for each concentration, number of base model "M")

Output: *Prediction the value of WQI*

1: ***For*** *each id_seasons.*
2: *Specify Target variable Y.*
3: Building Model by following steps:
4: ***While*** complexity of Model<M, do following steps: //M determined through DWM-Bat
5: ***For*** each variable x, do following
6: ***For*** each Knot of variable x, Test each Knot according to equation:
7: $$(x-t)_+ = \begin{cases} x-t, & x > t \\ 0, & otherwise \end{cases}$$
8: Choose Knot for variable x, which decrease prediction error.
9: Add new basis function from variable x with knot to the Model.
10: ***IF*** complexity of Model>=M, then Stop building Model.
11: For each basis function in the Model make pruning as follow:
12: Calculate Generalized Cross Validation error according to equation:
13: $$GCV = \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{\left(1-\frac{C}{N}\right)^2}$$
14: Where N: number of samples, C=1+cd // d is equal # of independent BFs, c is the penalty for adding a BF.
15: Remove function with high Generalized Cross Validation error.
16: After Model is completed, predict value of Y in parallel as follow:
// linear equation:
17: $y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m * \boldsymbol{compute\ linear}$ // *as equation*$k(x_i, x_j) = x_i^T x_j$
// polynomial equation:
18: $y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m * \boldsymbol{compute\ polynoimal}$ // *as equation*$(k(x_i, x_j) = (x_i^T x_j + \gamma)^d \quad d, \gamma > 0$
// sigmoid equation:
19: $y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m * \boldsymbol{compute\ sigmoid}$ // *as equation* $\tanh(\gamma X_i^T X_j + r)$, γ>0
// RBF equation:
20: $y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m * \boldsymbol{compute\ RBF}$ // *as equation*$(\exp(-\gamma \|X_i^T - X_j\|)^2$, γ>0
*// **DWM –Bat***
21: $y = f(x) = 100 + \sum_{m=1}^{M}$(best weight for all $concentration *$ $best\ value\ for\ all\ concentration$)
22: ***End*** *For*
End DMARS

2.4 Evaluation Stage

In this section, we will explain the evaluation of the predictor based on the compute three measures called (R^2, NSE and D), for each season to all Concentrations as shown in Algorithm 4.

3 Experiment and Results

Select the suitable parameters of any learning algorithm is considered one of the main challenges in the science, in general, MARS take a very long time in implementation

to give the result, therefore this section shows how DWM –Bat solves this problem and exceed this challenge.

In other words, the determination of weights and the number of model (M) are essential parameters that fundamentally affect DMARS performance. In general, the MARS based on the dynamic principle in selecting the parameters of it, the main parameters of DWM –Bat shown in Table 2.

Table 2. The Parameters Utilize in DWM –Bat

Parameter	Value
Number of bats (swarm size) (NB)	720
Minimum (M)	2
Maximum (M)	12
Determine frequency (pulse_frequency)	Pulse_frequency = 0*ones*(row_num,col_num)*
Loudness of pulse	1
Loudness decreasing factor(alpha)	0.995
Initial emission rate (init_emission_rate)	0.9
Emission rate increasing factor (gamma)	0.02
Bats initial velocity (init_vel)	0
Determine vector of initial velocity(velocity)	velocity = init_vel*ones*(row_num,col_num)*
Population Size (row_num)	*60*
(col_num)	*12*
Minimum value of observed matrix (min_val)	*0.0200*
Miaximum value of observed matrix (max_val)	*538*
Maximum number of iteration (max_iter)	*250*
Number of cells (n_var)	$n\_var = row\_num*col\_num$
Lower bound (lb)	$lb = min\_val*ones(row\_num,col\_num)$
Upper bound (ub)	$ub = max\_val*ones(row\_num,col\_num)$
Position of bat (Pos)	$Pos = lb + rand(row\_num,col\_num)*(ub\text{-}lb)$
rand1, rand2	Random numbers that are in the range [0, 1]
Calculate velocity and position of weight of each concentrations	$vw = vw + rand1 * (pwBest - w) + rand2 * (gwBest - w)$ (1) $w = w + vw$ (2)
Calculate velocity and position of the # of M	$vm = vm + rand1 * (pmBest - m) + rand2 * (gmBest - m)$ (3) $m = m + vm$ (4)

By apply the DWM–Bat get the best weight of each the 12 contractions as follow: *PH = 0.247, NTU = 0.420, TDS = 0.004, Ca = 0.028, Mg = 0.042, Cl = 0.008, Na = 0.011, K =, 0.175, SO4 = 0.008, NO3 = 0.042, CaCO3(TA) = 0.011, and CaCO3(TH) = 0.004, while the optimal number of M related to winter and summer dataset is 9.*

DMARS is mainly based on the MARS algorithm, which is capable of handling the dynamic principle in selecting the parameters of it.

In this stage, forward the parameters result from DWM–Bat to DMARS that represents the weight of each material, number of model (M) with the dataset of that seasons generated from the best split of five cross-validations to represent training of DMARS the main parameters of that algorithm represent in Table 3. Then compute the prediction values based on the best split result from five cross-validations.

With respect to Eq. (5), the proposed approach found that TH, TDS, K, NO3, Na, PH, TA, Cl and Ca had a very important contribution in the prediction of the WQI in winter season from any of the remaining concentrations.

Example #1: Proof the accuracy of the proposed model through some of samples related to winter season, taking into account that the data is limited between 0 and 1 due to the normalization of it.

The use of the ideal (M) model number and the ideal weights that were determined from ***DWM-BA***, which are as follows: *M = 9; Weights = [PH = 0.247, NTU = 0.420, TDS = 0.004, Ca = 0.028, Mg = 0.042, Cl = 0.008, Na = 0.011, K = 0.175, SO4 = 0.008, NO3 = 0.042, CaCO3(TA) = 0.011, and CaCO3(TH) = 0.004]. In general, the* ranges of WQI based on the stander measures and possible use shown below (see Table 4).

Proof:

1-IF PH = 0.991; TDS = 0.675; Cl = 0.667; TA = 0.7939; Ca = 0.8634; TH = 0.825; NO3 = 0.194; Na = 0.300; K = 0.0012.

WQI (1) = 100*[0.991**0.247* + 0.675**0.004* + 0.667**0.008* + 0.794* *0.011* + 0.864* 0.*028* + 0.825* 0.*004* + 0.194**0.042* + 0.300**0.011* + 0.002**0.175]*

WQI (1) = 100* 0.300837 = 30.0837

Obviously, ***the WQI score is dependent on Case #2***

2-IF PH = 1.000; TDS = 0.729; Cl = 0.750; TA = 0.786;Ca = 0.0773; TH = 0.850; NO3 = 0.186;Na = 0.300;K = 0.002.

WQI (2) = 100*[1.000**0.247* + 0.729**0.004* **+** 0.750**0.008* **+** 0.786* *0.011* + 0.773* **0**.*028* + 0.850* 0.*004* + 0.186 **0.042* **+** 0.300**0.011* **+** 0.002**0.175]*

WQI (2) = 100*[0.301068] **=** 30.1068

Obviously, ***the WQI score is dependent on Case #2***

As for prediction values to WQI for two seasons winter and summer based on the best result of a split of five cross validations for $IM^{12}CP$-*WQI* model, where data for each season were divided into two parts, 80% samples training and 20% samples testing, and Ranging for all material from 0 to 1. We notice that the prediction values are very close to the real values and this indicates that the $IM^{12}CP$-*WQI* predictor is a good

Table 3. The Parameters Utilize in DMARS

Parameter	Description
Number of input variable(d)	d = 12
Datasets (x)	x = samples of winter season or samples of summer season
Number of columns (m)	m = 13
Number of row (n)	n = 60
Training data cases (Xtr, Ytr)	Xtr(i,:), Ytr(i), i = 1, …, n
Vector of maximums for input variables (x_max)	x_max(winter) = [0.06, 7.55, 538, 42.60, 381.66, 417.424, 88, 397.984, 15.32, 9.28, 457.20, 135.69, 94.27] x_max(sumer) = [0.060, 7.470, 539, 24.850, 325, 417.760, 92, 447.424, 6.700, 3.800, 427.760, 137.945, 87.707]
Vector of minimums for input variables (x_min)	x_min(winter) = [0.02, 7.240, 363, 21.300, 300, 28.800, 36, 2.35, 1.859, 1.780, 0.89, 20.146, 12.233] x_min(summer) = [0.0200, 6.900, 390, 14.200, 235, 24, 33.600, 2.355, 1, 0.920, 0.630, 64.857, 11.449]
Size of dataset (x_size)	*x_size(n,m) = x_size(60, 12)*

BF	Equation
BF_Z1	0.175*K // k = 0.985
BF_Z2	0.011*TH // TH = 0.86
BF_Z3	0.042*NO3 // NO3 = 0.761
BF_Z4	0.004*TDS // TDS = 0.55
BF_Z5	0.011*Na // Na = 0.415
BF_Z6	0.247*PH // PH = 0.371
BF_Z7	0.011*CaCo3(TA) // TA = 0.37
BF_Z8	0.008*Cl // Cl = 0.362
BF_Z9	0.028*Ca //Ca = 0.317

$$WQI = 100 * \sum_{(K=0)}^{M} (BF\_ZK) = 100 * (\text{BF\_Z1} + \text{BF\_Z2} + \text{BF\_Z3} + \text{BF\_Z4} + \text{BF\_Z5} + \text{BF\_Z6} + \text{BF\_Z7} + \text{BF\_Z8} + \text{BF\_Z9}) \quad (5)$$

predictor as it was able to predict the real values well, so it is a better predictor compare with MARS linear, MARS_Sig, MARS_RBF and MARS_Poly. As shown in Figs. 2, 3, 4, and 5.

Table 4. Generated report of WQI based on four cases

Case	WQI	Possible use
Case#1	Value in rang (0–25)	Drinkable
Case#2	Value in rang (26, 50)	Fit for aquarium and animal drinking
Case#3	Value in rang (51, 75)	Not suitable for drinking, but suitable for watering crops
Case#4	Value in rang (76, 100)	Unusable pollutant must go to recurrence

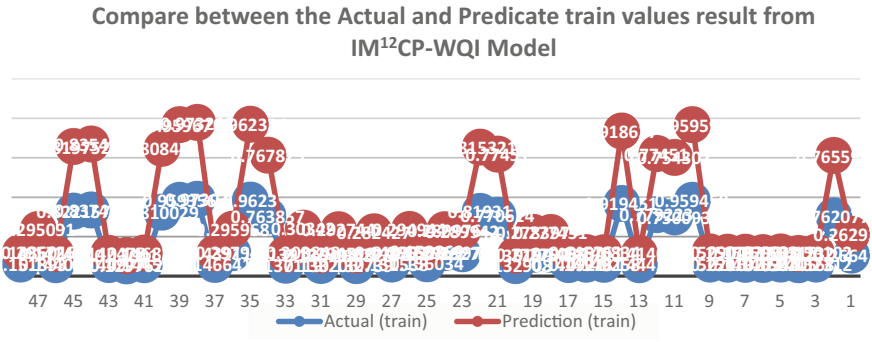

Fig. 2. Predictive Model *IM*^{12}*CP-WQI* for Training Dataset of Winter Season

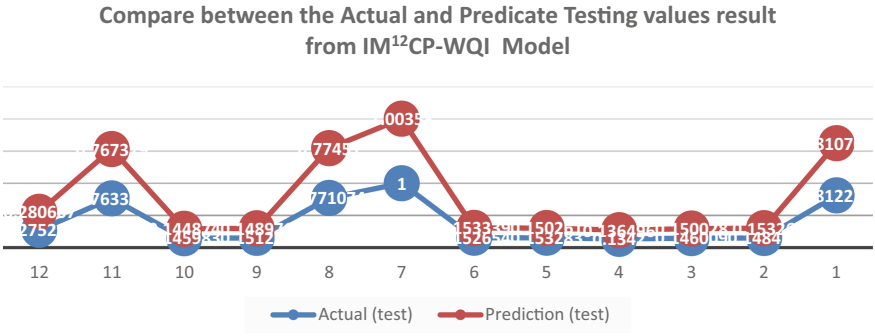

Fig. 3. Predictive Model *IM*^{12}*CP-WQI* for Testing dataset for Winter Season

The results shown ***IM*^{12}*CP-WQI*** model, were located closer to the reference point, indicating better performance compared to the other models. A comparison showed that the ***IM*^{12}*CP-WQI*** model generally converged faster and to a lower error value than the eithers model under same input combinations. The novel hybrid ***IM*^{12}*CP-WQI*** model showed more accurate WQI estimates with faster convergence rate than the other models.

The performances of the all models test in this study (i.e., MARS Linear, MARS_Poly, MARS_Sig, MARS_RBF, and MARS_DWM-BA) to predict the WQI

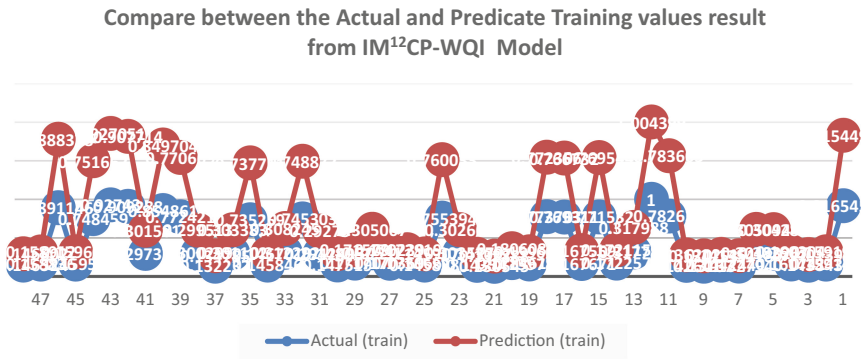

Fig. 4. Predictive Models $IM^{12}CP\text{-}WQI$ for Training Dataset to Summer Season

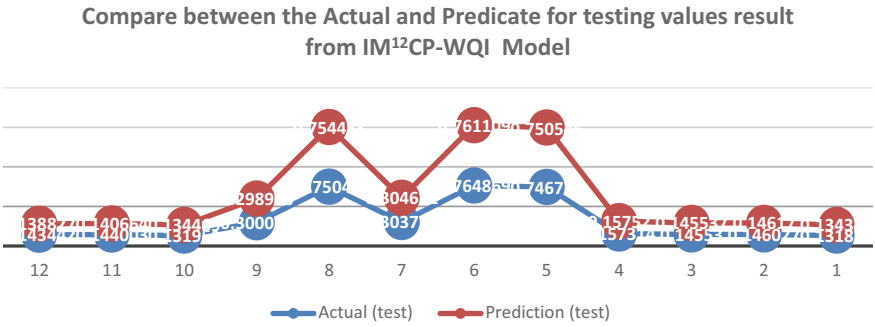

Fig. 5. Predictive Model $IM^{12}CP\text{-}WQI$ for Testing Dataset to Summer Season

were investigated for both training and testing stages for both seasons (winter and summer).

- In the training phase at winter season, for the prediction of WQI, $IM^{12}CP\text{-}WQI$ provided more accurate performance (R2 = 0.2202, NSE=0.9999, and D = 1) compare with other models, and MARS_RBF provided less accurate performance (R2 = −0.1148, NSE = −2.3411, and D = −16.6417) compare with other the models.
- While, in the testing phase at winter season for the prediction of WQI, IM12CP-WQI provided more accurate performance (R2 = 0.7919, NSE = 0.9999, and D = 1) compare with other models, in other side; MARS_RBF provided less accurate performance (R2 = −0.2034, NSE = −1.4032, and D = −2.5096).
- While, the evaluation of the summer season proves the training dataset of IM12CP-WQI give the best performance based on the three evaluation measures (R2 = 0.2331, NSE = 0.9999, and D = 1) compare with other models, and MARS_RBF provided less accurate performance (R2=0.751, NSE = −2.2284, and D = −12.0533) compare with the other models.
- Also, IM12CP-WQI provided more accurate performance for the three measures of testing dataset (R2 = 1.2688, NSE = 0.9999, and D = 1) compare with other models,

while; MARS_RBF provided less accurate performance (R2 = 2.7051, NSE = −2.185, and D = −2.6243).

4 Discussion

In this section, a quite few statistical measures are presented to evaluate the performance of the proposed models. Moreover, the results of the IM^{12}CP-WQI and MARS technologies compared with more than one core. The results proved that the IM^{12}CP-WQI model gives the best results according to the evaluation measures in two seasons related to the training and testing dataset, in general, this study answers the following questions [10–16]

- How Bat optimization algorithm can be useful in building an intelligent Miner?
- BOA works to modify the behavior of each in a particular environment gradually, depending on the behavior of their neighbors until they obtained the optimal solution.
- On the other hand, the MARS use the principle of the try and error in the selection of the basic parameters of their own and modified gradually to reach the values accepted for those parameters.
- Depending on the BOA and MARS of the above subject, we used the BOA principle to find the optimal weights for each concentration and the number of based models of the MARS.
- How to build a multi-level model with a combination of two technologies)MARS with BOA)?

Through, building new miner called IM^{12}CP-WQI that combining between the DWM –Bat and the DMARS. Where DWM –Bat used to find the best values of wights to each concentration with best number of M to DMARS while DMARS used to predict the water quality index (WQI).

- Is three evaluation measures enough to evaluate the results of suggested Miner?
- Yes, that measures are sufficient to evaluate the results of the miner to the both seasons.
- What is the benefit result from building miner by combination between DWM_Bat and DMARS?

By combining DWM_Bat and DMARS, reduce the execution time by defining MARS parameters but at the same time will increase the computational complexity.

5 Conclusions

We can summarize the main point performance in that paper as the follows: Water quality index dataset is a sensitive data need to accuracy techniques to extract a useful knowledge from it. Therefore; IM^{12}CP-WQI was able to solve this problem by giving results of high predictive accuracy, but on the other hand, it increased the mathematical complexities to obtain of that results. The main purpose of the normalization process is to convert data within a specified range of values to be handled more precisely at subsequent processing

stages. Especially since the concentrations are within different ranges and are measured in different units, so a normalization has been made to make them within a specific range to work on. Where the concentrations were placed between range (0, 1). This study proves the correlation between WQI and the important concentrations are k = 0.985, TH = 0.86, NO3 = 0.761, TDS = 0.55, Na = 0.415, PH = 0.371, TA = 0.37, Cl = 0.362, Ca = 0.317. This step focus on determined the important concentrations are Total Hardness (TH) that have negative relation with WQI and TDS. By apply the DWM-Bat get the best weight of each concentration as follow: W-PH = 0.247, W-NTU = 0.420, W-TDS = 0.004, W-Ca = 0.028, W-Mg = 0.042, W-Cl = 0.008, W-Na = 0.011, W-K = 0.175, W-SO4 = 0.008, W-NO3 = 0.042, W-CaCO3(TA) = 0.01l, and W-CaCO3(TH) = 0.004. While the optimal number of M related to both datasets are 9. This stage increases the accuracy of results and reduces the time required to training the MARS algorithm. Selection the best activation function to build the predictor based on mathematical concept, through build DMARS that replace the core of MARS by four types of functions (i.e., polynomial, sigmoid, RFB and linear). Results indicated that the MARS technique with linear and sigmoid kernel functions have stood at higher level of accuracy rather than the MARS approaches developed by other types of kernel functions. As the results of both training and testing indicated that MARS-linear and MARS-sig methods have provided relatively precise prediction for WQI, compared to the MARS_RBF and MARS_Poly. IM^{12}CP-WQI give pragmatic model of water quality index for different seasons indicates the water become high quality when the value of WQI is small value not exceed twenty-five will used to drink while other values highest than twenty-five to fifty. It is possible use to other uses, such as watering crops, fish lakes, and factories, except that requires a refining process to the water.

The following point give good idea for features works; Using other optimization algorithms based on search agent algorithm such as Whale Optimization Algorithm (WOA) or Particle Swarm Optimization (PSO) or Ant Lion Optimization (ALO). Investigation other prediction algorithm that adopts the mining principle such as Gradient Boosting Machine (GBM) or extreme gradient boosting (XGBoost). Verification from the prediction results based on other evaluation measures such as (Accuracy, Recall, Precision, F, and FB). Test the model on the new dataset that contain other concentrations rather than these used in this study.

Author Contributions. All authors contributed to the study conception and design. Data collection and analysis were performed by [Samaher Al-Janabi] and Zahra A. The first draft of the manuscript was written by [Samaher Al-Janabi] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Declarations.

Conflict of Interest:. The authors declare that they have no conflict of interest.

Ethical Approval:. This article does not contain any studies with human participants or animals performed by any of the author.

References

1. Hudson, Z.: The applicability of advanced treatment processes in the management of deteriorating water quality in the Mid-Vaal river system. Environmental Sciences at the Potchefstroom Campus of the North-West University or Natural and Agricultural Sciences [1709] (2015). http://hdl.handle.net/10394/16075
2. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R., García-Nieto, J.: Efficient Water Quality Prediction Using Supervised Machine Learning, vol. 11, p. 2210 (2019). https://doi.org/10.3390/w11112210
3. Aghalari, Z., Dahms, H.U., Sillanpää, M., et al.: Effectiveness of wastewater treatment systems in removing microbial agents: a systematic review. Global Health **16**, 13 (2020). https://doi.org/10.1186/s12992-020-0546-y
4. Singh, P., Kaur, P.D.: Review on data mining techniques for prediction of water quality. Int. J. Adv. Res. Comput. Sci. **8**(5), 396–401 (2017)
5. Qiu, Y., Li, J., Huang, X., Shi, H.: A feasible data-driven mining system to optimize wastewater treatment process design and operation. **10**, 1342 (2018). https://doi.org/10.3390/w10101342
6. Al-Janabi, S.: Smart system to create an optimal higher education environment using IDA and IOTs. Int. J. Comput. Appl. **42**(3), 244–259 (2020). https://doi.org/10.1080/1206212X.2018.1512460
7. Al-Janabi, S.: A novel agent-DKGBM predictor for business intelligence and analytics toward enterprise data discovery. J. Babylon Univ./Pure Appl. Sci. **23**(2) (2015)
8. Alkaim, A.F., Al-Janabi, S.: Multi objectives optimization to gas flaring reduction from oil production. In: Farhaoui, Y. (eds.) Big Data and Networks Technologies. BDNT 2019. Lecture Notes in Networks and Systems, vol 81. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-23672-4_10
9. Ameen, H.A.: Spring water quality assessment using water quality index in villages of Barwari Bala, Duhok, Kurdistan Region, Iraq. Appl. Water Sci. **9**(8), 1–12 (2019). https://doi.org/10.1007/s13201-019-1080-z
10. Al-Janabi, S., Mahdi, M.A.: Evaluation prediction techniques to achievement an optimal biomedical analysis. Int. J. Grid and Utility Comput. **10**(5), 512–527 (2019).https://doi.org/10.1504/IJGUC.2019.102021.7
11. Al-Janabi, S., Patel, A., Fatlawi, H., Al-Shourbaji, I., Kalajdzic, K.: Empirical rapid and accurate prediction model for data mining tasks in cloud computing environments. In: 2014 International Congress on Technology, Communication and Knowledge (ICTCK), pp. 1–8 (2014). https://doi.org/10.1109/ICTCK.2014.7033495
12. Al_Janabi, S., Yaqoob, A., Mohammad, M.: Pragmatic method based on intelligent big data analytics to prediction air pollution. In: Big Data and Networks Technologies, BDNT 2019. Lecture Notes in Networks and Systems, pp. 84–109, Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23672-4_8
13. Al-Janabi, S., Alkaim, A.F., Adel, Z.: An Innovative synthesis of deep learning techniques (DCapsNet & DCOM) for generation electrical renewable energy from wind energy. Soft. Comput. **24**, 10943–10962 (2020). https://doi.org/10.1007/s00500-020-04905-9
14. Al-Janabi, S., Alkaim, A.F.: A comparative analysis of DNA protein synthesis for solving optimization problems: a novel nature-inspired algorithm. In: Abraham, A., Sasaki, H., Rios, R., Gandhi, N., Singh, U., Ma, K. (eds.) Proceedings of the 11th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2020) held during December 16–18. IBICA 2020. Advances in Intelligent Systems and Computing, vol. 1372, pp. 1–22. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73603-3_1
15. Al-Janabi, S., Kad, G.: Synthesis biometric materials based on cooperative among (DSA, WOA and gSpan-FBR) to water treatment. In: Abraham, A., et al. (eds.) Proceedings of

the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020). SoCPaR 2020. Advances in Intelligent Systems and Computing, vol. 1383, pp. 20–33. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73689-7_3

16. Al-Janabi, S ., Mohammad, M., Al-Sultan, A.: A new method for prediction of air pollution based on intelligent computation. Soft. Comput. **24**(1), 661–680 (2019).https://doi.org/10.1007/s00500-019-04495-1
17. Sharma, T.: Bat Algorithm: an Optimization Technique. Electrical & Instrumentation Engineering Department Thapar University, Patiala Declared as Deemed-to-be-University u/s 3 of the UGC Act., 1956 Post Bag No. 32, PATIALA–147004 Punjab (India) (2016). https://doi.org/10.13140/RG.2.2.13216.58884