Received: September 30, 2022. Revised: November 4, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Effective Groundwater Quality Classification Using Enhanced Whale Optimization Algorithm with Ensemble Classifier

N. D. S. S. Kiran Relangi[1,2]* **Aparna Chaparala[3]** **Radhika Sajja[4]**

[1]Department of Computer Science & Engineering,
ANU College of Sciences, Acharya Nagarjuna University, Guntur, AP, India
[2]Department of Computer Science Engineering,
Anil Neerukonda Institute of Technology and Sciences (Autonomous), Visakhaptnam, India
[3]Department of Computer Science Engineering, RVR & JC College of Engineering, Guntur, AP, India
[4]Department of Mechanical Engineering, RVR & JC College of Engineering,Guntur, AP, India
* Corresponding author's Email: kiran.cse@anits.edu.in

Abstract: In recent decades, the groundwater quality monitoring application gained more attention among the researcher community to assess the groundwater quality. The water quality index (WQI) is one of the effective models used for assessing the groundwater quality, which is not always superior in classifying the groundwater quality, especially in the large scale databases. Therefore, a new ensemble model is developed in this manuscript for classifying the ground-water quality. After collecting the data from the real-time and Indian water quality databases, the WQI calculation and the data denoising (Z-score and Min-Max normalization techniques) are accomplished. From the denoised data samples, the optimal features/attributes are chosen by implementing enhanced whale optimization algorithm (EWOA). Usually, the traditional WOA is computationally complex to explore the global solutions, therefore, a fitness function probability pro is included with the WOA for enhancing convergence speed and classification accuracy. The chosen optimal features/attributes are fed to the ensemble model: AlexNet and K-nearest neighbor (KNN) for classifying the types of groundwater quality. The introduced ensemble based EWOA model has achieved 99.88% and 99.98 (very near to 100%) of classification accuracy on the real time database and Indian water quality database.

Keywords: AlexNet, Groundwater quality classification, K-nearest neighbor, Min-Max normalization, Whale optimization algorithm, Z-score technique.

1. Introduction

In recent decades, the water pollution has become grimmer, due to increased urbanization and fast economic growth [1, 2]. Majority of the nations have started to implement effective environment water management systems for understanding the marine ecosystems quality [3, 4]. Groundwater is an important water supply source for people, where its quality is directly related to the people's health [5-7]. The intake of the contaminated groundwater leads to the severe health problems that significantly increases the mortality and morbidity rate [8, 9]. The WQI facilitates the water quality assessment, and it is one of the important tools to assess the quality of groundwater. The WQI is assessed by computing an extensive range of parameters such as organic matter, turbidity, pH, temperature, electrical conductivity etc. [10-12]. Hence, the WQI computing proved to be effective and time-intensive, but involved unintended errors [13, 14]. Therefore, numerous mathematical models are implemented based on both machine learning and deep learning methods [15-17]. With the advanced computing utilizing artificial intelligence, a novel model is developed in this paper for groundwater quality classification. The contributions of this work are stated as follows:

- After acquiring the real-time and Indian water quality databases, the WQI calculation and the data denoising is performed by utilizing Z-score and Min-Max normalization techniques. The data denoising process improves the acquired data quality by scaling its range.
- From the scaled data, the optimal attributes/features are chosen by developing EWOA technique where it utilizes probabilitypro fitness function for enhancing convergence speed and classification accuracy. In addition to this, the selection of optimal features /attributes decreases the complexity and computational time of the system.
- Finally, the chosen features/attributes are fed to the ensemble classifier for classifying the groundwater quality types like excellent, good, poor, and very poor. The efficacy of the ensemble based EWOA is analysed using the evaluation measures such as false discovery rate (FDR), Matthews correlation coefficient (MCC), sensitivity, accuracy and specificity.

The paper organization is depicted as follows: literature survey is done in section 2. The explanation about the ensemble based EWOA is specified in section 3. The validation results and the conclusion of this work is presented in sections 4 and 5.

2. Related works

Mallick [18] developed a new groundwater potentiality model (GPM) by integrating individual random forest classifier with an ensemble classifier that comprise artificial neural network (ANN), logistic regression (LR), and support vector machine (SVM). As denoted in the resulting section, the implemented model effectively improves the sustainability of the groundwater management plans, especially in the Bisha watersheds, Saudi Arabia. However, the integration of several machine learning classifiers increases the computational complexity of the system. Yang [19] used a random forest classifier to predict the interaction of the surface groundwater in the New-Zealand region by utilizing land use, geology, and hydrology data. Related to other machine learning techniques, the random forest classifier has achieved better simulation results with minimum misclassification error. However, the main drawback of random forest classifiers was that the larger number of trees makes this technique too slow, and in-effective in the real time predictions. Panahi [20] integrated both support vector regression (SVR) and convolutional neural network (CNN) methodologies for groundwater spatial prediction. The implemented SVR-CNN method has generated significant freshwater conservation and management strategies in the study area (South Korea). However, the CNN model was computationally expensive because it needs more data to achieve better prediction results.

Mosavi [21] initially collected 339 groundwater resources, and then, the recursive feature elimination technique was applied for identifying the optimal features/attributes. Next the selected features were given to the ensemble models for groundwater potential prediction, where it includes random forest classifier, bagged classification and regression trees, boosted generalized assistive model and adaptive boosting classification tree. The experimental result states that the boosting model has attained better prediction performance and it out-performed other existing models in terms of recall, kappa, precision, and prediction accuracy. However, the computational time of the developed model was comparably higher related to the existing models. Ismael [22] used ANN model for classifying the groundwater quality in the Al-Zubayr and Safwan in Basra. In this literature, the developed ANN model helps in generating sustainable groundwater management strategies, but it was computationally expensive. Next, Singha [23] has initially collected 226 groundwater samples from Raipur district, India. Then, a novel deep learning model was developed in order to predict the groundwater quality. As stated in the resulting section, the developed model obtained high prediction performance compared to other machine learning models like ANN, random forest, and XGBoost. Hence, the implemented deep learning model was computationally expensive, where it requires high end processing systems.

El Bilali [24] collected 520 groundwater samples from Morocco, and further, the groundwater quality prediction was assessed by implementing four machine learning models: SVR, ANN, random forest and Adaboost. The extensive experimental result revealed that the random forest and Adaboost models attained higher prediction results related to the ANN and SVR. However, the ANN and SVR models were less sensitive and generalizable to the input variables than the random forest and Adaboost. Osman [25] implemented a significant groundwater level prediction model, here, the data were recorded from the highly populated regions of Malaysia. The collected data includes the attribute details like evaporation, rainfall and temperature for predicting groundwater levels. Related to the comparative models: ANN and SVR, the Xgboost model has achieved high prediction results, but the computational time was higher in this literature. Rizeei [26] has implemented an adaptive boosting logistic regression model for groundwater potential prediction. Hence, the presented model superiorly decreases the variance and bias in the database compared to other models. On the other hand, the presented model was sensitive to outliers and overfitting risks. Hmoud Al-Adhaileh, and Waselallah Alsaade, [27] has implemented adaptive neuro fuzzy inference system (ANFIS) for predicting WQI and then the KNN and ANN models were employed for classifying water quality. The developed model has achieved better classification results by means of accuracy, f-score and error rate, but computationally costly. T.H. Aldhyani [28] integrated long short term memory (LSTM) and Non-linear autoregressive neural network (NARNET) for WQI prediction. Further, the SVM classifier was implemented for classifying water quality, where it majorly supports binary classification. In order to address the afore-motioned issues, a novel ensemble based EWOA is proposed for effective groundwater quality classification, where it attained better performance related to other optimization techniques such as random selected leader based optimizer (RSLBO) [29] and squirrel search optimizer [30].

3. Materials and methods

The proposed groundwater quality classification model comprises four phases such as **data collection:** real time database and Indian water quality database, **data denoising**: WQI-calculation, Min-Max normalization and Z-score techniques, **feature optimization**: EWOA, and **groundwater quality classification:** ensemble model (KNN with AlexNet). Where, the work-flow of the proposed groundwater quality classification model is indicated in Fig. 1.

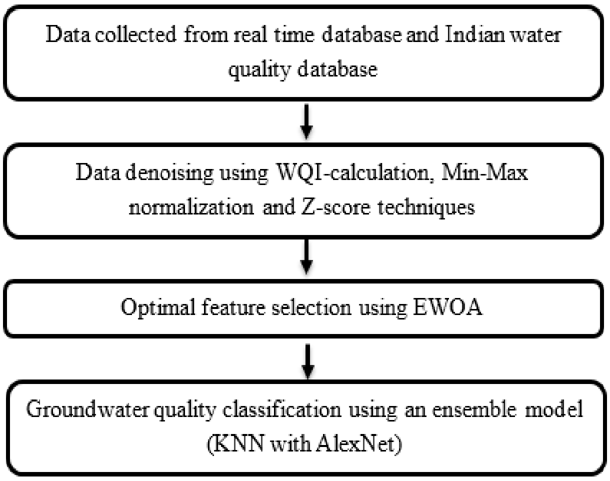
Data collected from real time database and Indian water quality database
Data denoising using WQI-calculation, Min-Max normalization and Z-score techniques
Optimal feature selection using EWOA
Groundwater quality classification using an ensemble model (KNN with AlexNet)

Figure. 1 Work-flow of the proposed groundwater quality classification model

3.1 Database description

The implemented groundwater quality classification model's (ensemble based EWOA) performance was analysed on a real time database and Indian water quality database. The real time database is recorded from a water quality monitoring lab, Narsapuram, Andhra Pradesh. The real time database comprises seven parameters such as total coliform, fecal coliform, pH, Conductivity, Temperature, Nitrate + Nitrite, and biological oxygen demand (BOD). In this real time database, the class labels of every groundwater sample is assessed by calculating the WQI.

In addition to this, the Indian water quality database is recorded from dissimilar Indian locations between the time periods of 2005 to 2014. This database contains 1679 samples and it is acquired from 666 dissimilar sources of lakes and rivers. This database consists of seven parameters like BOD, total coliform, temperature, fecal coliform, Nitrate, pH, and dissolved-oxygen. The Indian central government recorded the data for ensuring drinking water quality.

Link:
https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data

3.2 Data denoising

The data denoising is an important section in the groundwater quality classification, which helps in improving the quality of collected data. In this section, the WQI is calculated from 7 parameters in the databases, and then, the water samples are classified based on WQI values. In addition to this, the Min-Max normalization and Z-score techniques are employed for data normalization in order to achieve superior classification accuracy. Firstly, the WQI decreases the acquired data into a specific value, which helps in easy understanding of the water quality information. In WQI, a weight function W_i is assigned to every parameter on the basis of its importance. The effectiveness of the ensemble based EWOA is evaluated on both databases with 7 quality parameters, and the WQI is computed utilizing Eq. (1).

$$WQI = \frac{\sum_{i=1}^{N} q_i \times W_i}{\sum_{i=1}^{N} W_i} \tag{1}$$

Where, q_i indicates quality estimation Scale (QES) of every parameter$i$, W_i represents unit weight of every parameter and N states total parameters. The term QES q_i is mathematically stated in Eq. (2).

Table 1. Permissible limits of the 7 parameters and its unit weights

Parameters	Permissible limits	Unit weight W_i
Total coliform/100 mL	1000	0.0022
Fecal coliform/100 mL	100	0.0221
Nitrate, mg/L	45	0.0492
BOD, mg/L	5	0.4426
Conductivity, $\mu S/cm$	1000	0.0022
Ph	8.5	0.2604
Dissolved-oxygen, mg/L	10	0.2213

Table 2. Water quality classification

Range of water quality index	Classification
0 to 25	Excellent
26 to 50	Good
51 to 75	Poor
76 to 100	Very poor

$$q_i = 100 \times \left(\frac{V_i - V_{Ideal}}{S_i - V_{Ideal}}\right) \quad (2)$$

Where,V_{Ideal} states ideal value (pH=7, dissolved-oxygen=14.60 mg/L, and other parameters are zero), S_i denotes standard value, and V_i indicates measured value. The weight function W_i is computed using Eq. (3).

$$W_i = \frac{K}{S_i} \quad (3)$$

Where, K indicates proportionality constant, and it is computed utilizing Eq. (4). Hence, the permissible limits of the 7 parameters and its unit weights are depicted in Table 1. Further, the water quality classification is indicated in Table 2.

$$K = \frac{1}{\sum_{i=1}^{N} S_i} \quad (4)$$

After performing WQI, the Min-Max normalization technique rescales the collected data into lower and upper bounds, which usually ranges between 0 to 1 and -1 to 1. Additionally, the Z-score technique is employed for normalizing the collected data by calculating standard-deviation and mean values. The Z-score technique scales the parametric values between 0 to 1. The mathematical expressions of Min-Max normalization and z-score techniques are stated in Eq. (5) and (6).

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (5)$$

$$Z - score = \frac{(x - \mu)}{\sigma} \quad (6)$$

Where, x indicates tested samples in the databases,μ represents mean value, σ states standard deviation value, $x_{max}andx_{min}$ states maximum and minimum attribute values. In addition, the rescaled data are dimensionally decreased by implementing EWOA, where this procedure helps in reducing the computational time and complexity of the system.

3.3 Feature optimization

The metaheuristic optimization algorithm: WOA mimics hump back whales behaviour for resolving the feature optimization issues. Firstly, the populations are randomly generated, and further the optimal prey location is searched by employing either bubble-net or encircling approaches. In the encircling approach, the best location of the hump-back whales is found by utilizing Eq. (7) and (8).

$$D = |B \odot P^*(t) - P(t)| \quad (7)$$

$$P(t+1) = |P^*(t) - A \odot D| \quad (8)$$

Whereas, $A$and B represents coefficient variables, t denotes iteration number, $P(t)$ indicates hump-back whales position, and D represents distance between two preys$P^*(t)$. Though, the coefficient variables are computed utilizing Eq. (9) and (10).

$$A = 2l \odot r_v - l \quad (9)$$

$$B = 2r_v \quad (10)$$

Where, l states a linear function that ranges between zeros to two and r_v represents a random vector $\in [0,1]$. On the other hand, a bubble-net approach is utilized for identifying the prey's optimal location by encircling a shrinkage and updating the spiral position. The mathematical expressions of the bubble-net approach are given in Eq. (11) and (12).

$$P(t+1) = \acute{D} \odot e^{AB} \odot cos(2\pi A) + P^*(t) \quad (11)$$

$$P(t+1) = \begin{cases} P^*(t) - A \odot Difp \geq 0.5 \\ \acute{D} \odot e^{AB} \odot cos(2\pi A) + P^*(t)ifp < 0.5 \end{cases} \quad (12)$$

Where, $\acute{D} = |P^*(t) - P(t)|$ represents the distance between prey and hump-back whale, and $\odot$ states element multiplication process. Hence, the hump-back whale position is updated by replacing the random search agent with the best search agent in the exploration phase and it is specified in Eq. (13) and (14).

$$D = |B \odot P_{rand} - P(t)| \quad (13)$$

$$P(t+1) = |P_{rand} - A \odot D| \quad (14)$$

Where, P_{rand} represents a random position, where it is computed from the present population size. The conventional WOA is computationally complex in exploring the global solutions, so EWOA is implemented for improving the classification accuracy, reliability of prey searching and convergence speed. After every iteration, a number between the ranges of zeros to one is extracted for each hump-back whale. If the obtained random number is <0.5, Eq. (17) is selected for updating the hump-back whale's position. Otherwise, Eq. (11) is chosen for updating the hump-back whale's position. In addition, the hump-back whale's component is changed with a fitness function (probabilitypro) in a search space, and it is stated in Eq. (15).

$$pro = 0.3(1 - \frac{iter}{iter_{max}}) \quad (15)$$

Where, $iter_{max}$ denotes maximum iteration and $iter$ states present iteration number. In order to select a design value x_j, a random number is selected between the ranges of one topro. At last, a number n is extracted between the intervals of zero to one based onpro value. Hence, the selected value x_j is altered utilizing Eq. (16) and (17).

$$x_j = x_{jmin} + random \times (x_{jmax} - x_{jmin}) \quad (16)$$

$$P(t+1) = |P^*(t) - A \odot D^{x_j}| \quad (17)$$

The proposed EWOA significantly maintains a better balance between diversification inclinations and intensifications. From 4928 attributes, the optimal 3940 attributes are chosen for groundwater quality classification. The assumed parameters of the EWOA are: shrinking-encircling=0.5, random search ability=0.1, iteration numbers=150, population size=100, and spiral updating probability=0.5. Flow-chart of the EWOA is depicted in Fig. 2.

3.4 Groundwater quality classification

The dimensionally decreased 3940 attributes are given to the ensemble model: KNN with AlexNet for groundwater quality classification. The KNN classifier utilizes k-neighbourhood values for finding the closest points between the data objects. Additionally, k-value is utilized for identifying the closest points in the selected attributes, where it should be unique. Here, 3 k-values are appropriate in obtaining superior classification results and the Euclidean distance measure Eu_i is employed in determining the nearest neighbors in the selected attributes and it is mathematically defined in Eq. (18).

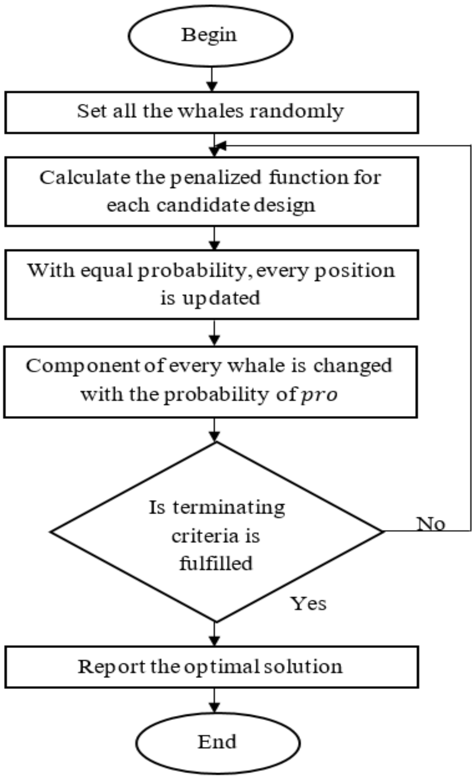
Begin
Set all the whales randomly
Calculate the penalized function for each candidate design
With equal probability, every position is updated
Component of every whale is changed with the probability of pro
Is terminating criteria is fulfilled
No
Yes
Report the optimal solution
End

Figure. 2 Flow-chart of the EWOA

$$Eu_i = \sqrt{(x_1 - x_2) + (y_1 - y_2)^2} \quad (18)$$

Where, $x_1, x_2, y_1, and y_2$ indicates input data variable. In addition to this, the AlexNet model includes 3 fully connected layers and 5 convolutional layers along with Rectifier Linear Unit (ReLU) activation function and Max-pooling operation for groundwater quality classification. The assumed hyper-parameter of the AlexNet is: training model=stochastic gradient descent, validation frequency=30, momentum=0.6, learning rate=0.15, maximum epoch-10 and the L2 regularization=1.000e-04. The AlexNet configuration is denoted in Table 3, and the developed ensemble model superiorly classifies the groundwater quality into four classes like excellent, good, poor and very poor. The experimental outcomes of the ensemble based EWOA is stated in section 4.

Table 3. AlexNet configuration

Layers		Functions	Configurations
Convolutional layers	1	Max-pooling	850 filters with 7 × 7 size
	2		850 filters with 5 × 5 size
	3		680 filters with 5 × 5 size
	4		680 filters with 5 × 5 size
	5		450 filters with 2 × 2 size
Fully-connected layers	1	ReLU	2096 nodes
	2		2096 nodes
	3		400 nodes

4. Experimental results

In the groundwater quality classification, the proposed ensemble based EWOA is simulated utilizing a python software environment on the computer with windows 10 operating system, i7 Intel core processor and 16GB random access memory. The performance measures such as sensitivity, FDR, specificity, accuracy, and MCC are utilized to evaluate the efficiency of the ensemble based EWOA in predicting the WQI and classifying the groundwater quality. Then, the mathematical representations of the undertaken performance measures such as sensitivity, FDR, specificity, accuracy, and MCC are depicted in Eq. (19-23).

$$Sensitivity = \frac{TP}{TP+FN} \times 100 \tag{19}$$

$$FDR = \frac{FP}{FP+TP} \times 100 \tag{20}$$

$$Specificity = \frac{TN}{TN+FP} \times 100 \tag{21}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{22}$$

$$MCC = \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \times 100 \tag{23}$$

Where, FP, FN, TP and TN state false positive, false negative, true positive and true negative values.

4.1 Quantitative investigation on a real time database

In this scenario, the proposed ensemble based EWOA's performance is validated on a real time database in light of sensitivity, FDR, specificity, accuracy and MCC. In this manuscript, the proposed ensemble based EWOA's effectiveness is validated by performing k-fold cross-validation such as 3-fold, 5-fold, and 7-fold. In that, the 5-fold cross validation (80:20% data training and testing) attained high classification results related to other cross fold validations, and the experimental results are depicted in Table 4. In this manuscript, the inclusion of cross-fold validation reduces the computational time, bias, and variance of the developed ensemble based EWOA.

As depicted in Table 4, the experimental results are validated with dissimilar classifiers such as SVM, multi-SVM (MSVM), KNN, AlexNet and ensemble classifier with and without performing EWOA. By inspecting Table 4, the combination: ensemble classifier with EWOA attained maximum classification results with accuracy of 99.88%, sensitivity of 99.34%, FDR of 99.74%, specificity of 99.78% and MCC of 99.90%. The obtained experimental results are high compared to other individual classifiers like SVM, MSVM, KNN and AlexNet on a real time database. Graphical comparison of the ensemble based EWOA on a real time database is stated in Fig. 3.

4.2 Quantitative investigation on an Indian water quality database

In this segment, the proposed ensemble based EWOA's efficacy is investigated on an Indian water quality database by means of sensitivity, FDR, specificity, accuracy, and MCC. The Indian water quality database has 1679 samples in which 80:20% of the data are used for model training and testing. As specified in Table 5, the ensemble based EWOA has obtained a maximum classification result with sensitivity of 100%, FDR of 99.82%, specificity of 100%, accuracy of 99.98%, and MCC of 99.98% on an Indian water quality database. In addition, the achieved experimental result is better related to the individual classifiers like SVM, MSVM, KNN, and AlexNet. Graphical comparison of the ensemble based EWOA on an Indian water quality database is represented in Fig. 4. On the other hand, the optimal feature selection of EWOA significantly decreases the computational time and complexity of the system.

4.3 Comparative investigation

The proposed ensemble based EWOA's effectiveness is compared with an existing model in light of accuracy and regression coefficient. Hmoud Al-Adhaileh, and Waselallah Alsaade, [27] implemented ANFIS model for predicting the WQI, and then

Received: September 30, 2022. Revised: November 4, 2022.

integrated KNN and ANN for classifying the

Table 4. Experimental results of the ensemble based EWOA on a real time database

Without performing EWOA					
Classifiers	**Sensitivity (%)**	**FDR (%)**	**Specificity (%)**	**Accuracy (%)**	**MCC (%)**
SVM	90.60	91.65	91.60	91.58	90.43
MSVM	94.40	94.96	92.54	94.58	92.14
KNN	96.95	97.54	94.78	97.36	96.06
AlexNet	98.58	98.35	95.18	98.52	98.68
Ensemble	99.08	99.56	99.72	99.66	99.77
With performing EWOA					
Classifiers	**Sensitivity (%)**	**FDR (%)**	**Specificity (%)**	**Accuracy (%)**	**MCC (%)**
SVM	92.44	95.09	95.65	94	96.22
MSVM	93.47	96.90	96.53	95.50	96.66
KNN	97.64	98.55	98.28	98.77	98.02
AlexNet	99.02	99.36	99.10	99.06	99.34
Ensemble	99.34	99.74	99.78	99.88	99.90

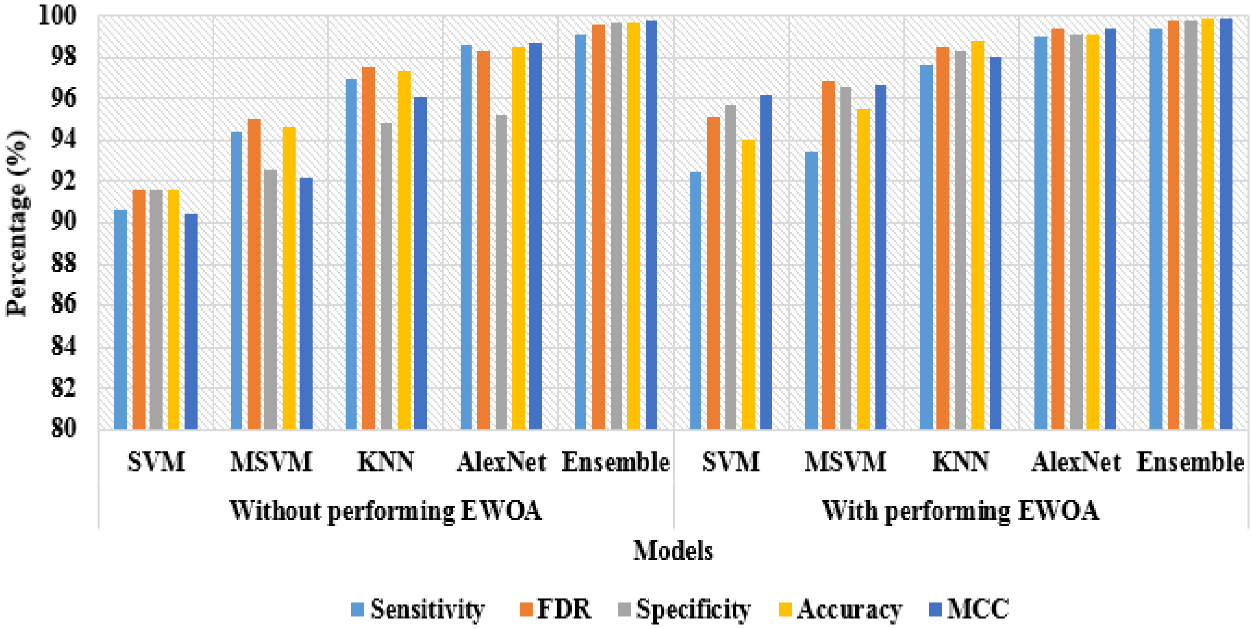

Figure. 3 Graphical comparison of the ensemble based EWOA on a real time database

Table 5. Experimental results of the ensemble based EWOA on an Indian water quality database

Without performing EWOA					
Classifiers	**Sensitivity (%)**	**FDR (%)**	**Specificity (%)**	**Accuracy (%)**	**MCC (%)**
SVM	92.65	94.90	93.68	92.20	94.46
MSVM	95.76	95.54	95.70	93.56	95.74
KNN	96.06	96.58	96.64	93.30	96.66
AlexNet	97.76	97.30	98.58	94.56	97.96
Ensemble	99.09	98.95	99.35	98.85	98.95
With performing EWOA					
Classifiers	**Sensitivity (%)**	**FDR (%)**	**Specificity (%)**	**Accuracy (%)**	**MCC (%)**
SVM	98.90	98.90	98.68	98.72	98.88
MSVM	98.96	98.94	98.98	98.50	98.96
KNN	99.14	99.12	99.18	99.72	99.02
AlexNet	99.82	99.70	99.80	99.82	99.67
Ensemble	100	99.82	100	99.98	99.98

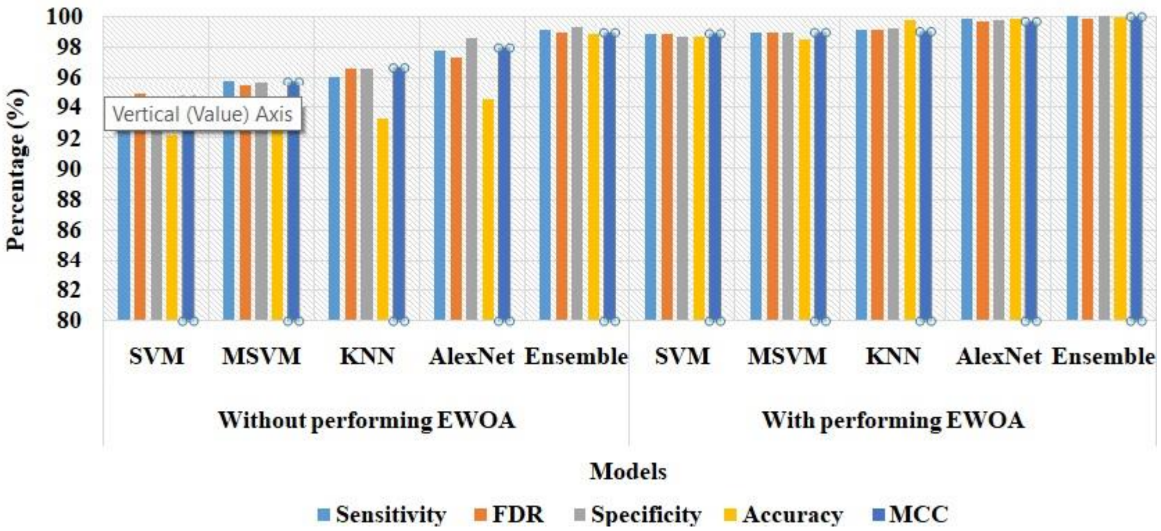

Figure. 4 Graphical comparison of the ensemble based EWOA on an Indian water quality database

Table 5. Comparative results of the existing and the proposed model

Models	Accuracy (%)	Specificity (%)	Sensitivity (%)
KNN with ANN [27]	100	99.61	99.61
LSTM and NARNET with SVM [28]	97.01	97.78	99.23
Ensemble based EWOA	99.98	100	100

water quality. In this literature study, the ANFIS model was implemented based on 7 statistical parameters such as dissolved oxygen, temperature, fecal coliform, BOD, total coliform, Nitrate, and pH. The experimental investigations showed that the implemented model has achieved 100% of classification accuracy and 99.61% of sensitivity and specificity on the Indian water quality database. T.H. Aldhyani [28] has combined both LSTM and NARNET for WQI prediction. Next, the SVM classifier was developed for classifying water quality. The extensive experiment indicates that the implemented model has achieved 97.01% of classification accuracy, 99.23% of sensitivity and 97.78% of specificity on the Indian water quality database. Related to these existing works, the developed ensemble based EWOA has attained high classification results in the groundwater quality classification with the accuracy very near to 100%, sensitivity and specificity of 100% on the Indian water quality database.

As mentioned earlier, the feature optimization is a main integral part of this research. Hence, the selection of optimal features significantly decreases the system complexity to linear $O(N)$ where, order of magnitude is indicated as O and input size is represented as N. Additionally, the computational time of the developed model is 42.1 and 33.2 seconds on the real time and Indian water quality databases and it is superior to the conventional models. These are major concerns depicted in the literature section, and the comparative results are stated in Table 5.

5. Conclusion

In this research article, a new ensemble based EWOA model is implemented for effective groundwater quality classification. Initially, the WQI calculation and data denoising (Z-score technique and Min-Max normalization technique) are performed for enhancing acquired data quality. Next, the optimal features/attributes are selected that are relevant to the groundwater quality by proposing EWOA technique. Lastly, the dimensionally reduced features/attributes are fed to the ensemble classification model for classifying water quality types like excellent, good, poor and very poor. The ensemble classification model integrates KNN and AlexNet for groundwater quality classification. The evaluation measures like sensitivity, FDR, specificity, accuracy and MCC are utilized for analyzing the effectiveness of the proposed model. Hence, the ensemble based EWOA model has achieved 99.88% and 99.98% of classification accuracy on the real time database and Indian water quality database. The achieved experimental result is maximum related to the traditional machine learning classifiers like SVM, MSVM, KNN, and AlexNet. In addition, the selection of optimal features/attributes significantly reduces the computational time and complexity of the system. As a future extension, a new deep learning based ensemble classification model is implemented, and validated with a multimodal data to further improve groundwater quality classification.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2ndand 3rdauthor.

References

[1] R. Barzegar, A. A. Moghaddam, R. Deo, E. Fijani, and E. Tziritis, "Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms", *Science of the Total Environment*, Vol. 621, pp. 697-712, 2018.

[2] Y. Chen, W. Chen, S. C. Pal, A. Saha, I. Chowdhuri, B. Adeli, S. Janizadeh, A. A. Dineva, X. Wang, and A. Mosavi, "Evaluation efficiency of hybrid deep learning algorithms with neural network decision tree and boosting methods for predicting groundwater potential", *Geocarto International*, Vol. 37, No. 19, pp. 5564-5584, 2021.

[3] S. Elmahdy, T. Ali, and M. Mohamed, "Regional mapping of groundwater potential in ar rub al khali, arabian peninsula using the classification and regression trees model", *Remote Sensing*, Vol. 13, No. 12, p. 2300, 2021.

[4] P. Prasad, V. J. Loveson, M. Kotha, and R. Yadav, "Application of machine learning techniques in groundwater potential mapping along the west coast of India", *GIScience & Remote Sensing*, Vol. 57, No. 6, pp. 735-752, 2020.

[5] P. Yariyan, M. Avand, E. Omidvar, Q. B. Pham, N. T. T. Linh, and J. P. Tiefenbacher, "Optimization of statistical and machine learning hybrid models for groundwater potential mapping", *Geocarto International*, pp. 1-35, 2020.

[6] A. Mosavi, F. S. Hosseini, B. Choubin, M. Abdolshahnejad, H. Gharechaee, A. Lahijanzadeh, and A. A. Dineva, "Susceptibility prediction of groundwater hardness using ensemble machine learning models", *Water*, Vol. 12, No. 10, p. 2770, 2020.

[7] M. S. Jaafarzadeh, N. Tahmasebipour, A. Haghizadeh, H. R. Pourghasemi, and H. Rouhani, "Groundwater recharge potential zonation using an ensemble of machine learning and bivariate statistical models", *Scientific Reports*, Vol. 11, No. 1, pp. 1-18, 2021.

[8] S. M. Guzman, J. O. Paz, M. L. M. Tagert, and A. E. Mercer, "Evaluation of seasonally classified inputs for the prediction of daily groundwater levels: NARX networks vs support vector machines", *Environmental Modeling & Assessment*, Vol. 24, No. 2, pp. 223-234, 2019.

[9] P. M. Santos, and P. Renard, "Mapping groundwater potential through an ensemble of big data methods", *Groundwater*, Vol. 58, No. 4, pp. 583-597, 2020.

[10] A. Arabameri, A. Arora, S. C. Pal, S. Mitra, A. Saha, O. A. Nalivan, S. Panahi, and H. Moayedi, "K-fold and state-of-the-art metaheuristic machine learning approaches for groundwater potential modelling", *Water Resources Management*, Vol. 35, No. 6, pp. 1837-1869, 2021.

[11] Y. N. Kontos, T. Kassandros, K. Perifanos, M. Karampasis, K. L. Katsifarakis, and K. Karatzas, "Machine learning for groundwater pollution source identification and monitoring network optimization", *Neural Computing and Applications*, pp. 1-31, 2022.

[12] A. M. Al. Abadi, and J. J. Alsamaani, "Spatial analysis of groundwater flowing artesian condition using machine learning techniques", *Groundwater for Sustainable Development*, Vol. 11, p. 100418, 2020.

[13] S. K. Sarkar, S. Talukdar, A. Rahman, and S. K. Roy, "Groundwater potentiality mapping using ensemble machine learning algorithms for sustainable groundwater management", *Frontiers in Engineering and Built Environment*, 2021.

[14] S. Sachdeva, and B. Kumar, "Comparison of gradient boosted decision trees and random forest for groundwater potential mapping in Dholpur (Rajasthan), India", *Stochastic Environmental Research and Risk Assessment*, Vol. 35, No. 2, pp. 287-306, 2021.

[15] S. Miraki, S. H. Zanganeh, K. Chapi, V. P. Singh, A. Shirzadi, H. Shahabi, and B. T. Pham, "Mapping groundwater potential using a novel hybrid intelligence approach", *Water Resources Management*, Vol. 33, No. 1, pp. 281-302, 2019.

[16] P. T. Nguyen, D. H. Ha, M. Avand, A. Jaafari, H. D. Nguyen, N. A. Ansari, T. V. Phong, R. Sharma, R. Kumar, H. V. Le, and L. S. Ho, "Soft computing ensemble models based on logistic

regression for groundwater potential mapping", *Applied Sciences*, Vol. 10, No. 7, p. 2469, 2020.

[17] M. Kulisz, J. Kujawska, B. Przysucha, and W. Cel, "Forecasting water quality index in groundwater using artificial neural network", *Energies*, Vol. 14, No. 18, p. 5875, 2021.

[18] J. Mallick, S. Talukdar, and M. Ahmed, "Combining high resolution input and stacking ensemble machine learning algorithms for developing robust groundwater potentiality models in Bisha watershed, Saudi Arabia", *Applied Water Science*, Vol. 12, No. 4, pp. 1-19, 2022.

[19] J. Yang, J. Griffiths, and C. Zammit, "National classification of surface-groundwater interaction using random forest machine learning technique", *River Research and Applications*, Vol. 35, No. 7, pp. 932-943, 2019.

[20] M. Panahi, N. Sadhasivam, H. R. Pourghasemi, F. Rezaie, and S. Lee, "Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR)", *Journal of Hydrology*, Vol. 588, pp. 125033, 2020.

[21] A. Mosavi, F. S. Hosseini, B. Choubin, M. Goodarzi, A. A. Dineva, and R. E. Sardooi, "Ensemble boosting and bagging based machine learning models for groundwater potential prediction", *Water Resources Management*, Vol. 35, No. 1, pp. 23-37, 2021.

[22] A. N. Ismael, H. A. Abed, and M. A. Abed, "Classification of Groundwater Quality using Artificial Neural Networks in Safwan and Al-Zubayr in Basra", *Advances in Computer, Signals and Systems*, Vol. 4, No. 1, pp. 25-35, 2020.

[23] S. Singha, S. Pasupuleti, S. S. Singha, R. Singh, and S. Kumar, "Prediction of groundwater quality using efficient machine learning technique", *Chemosphere*, Vol. 276, p. 130265, 2021.

[24] A. E. Bilali, A. Taleb, and Y. Brouziyne, "Groundwater quality forecasting using machine learning algorithms for irrigation purposes", *Agricultural WaterManagement*, Vol. 245, p. 106625, 2021.

[25] A. I. A. Osman, A. N. Ahmed, M. F. Chow, Y. F. Huang, and A. E. Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia", *Ain Shams Engineering Journal*, Vol. 12, No. 2, pp. 1545-1556, 2021.

[26] H. M. Rizeei, B. Pradhan, M. A. Saharkhiz, and S. Lee, "Groundwater aquifer potential modeling using an ensemble multi-adoptive boosting logistic regression technique", *Journal of Hydrology*, Vol. 579, p. 124172, 2019.

[27] M. H. A. Adhaileh, and F. W. Alsaade, "Modelling and prediction of water quality by using artificial intelligence", *Sustainability*, Vol. 13, No. 8, p. 4259, 2021.

[28] T. H. Aldhyani, M. A. Yaari, H. Alkahtani, and M. Maashi, "Water quality prediction using artificial intelligence algorithms", *Applied Bionics and Biomechanics*, 2020.

[29] F. A. Zeidabadi, M. Dehghani, and O. P. Malik, "RSLBO: Random Selected Leader Based Optimizer", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 5, pp. 529-538, 2021, doi: 10.22266/ijies2021.1031.46.

[30] M. Sumanl, V. P. Sakthivel, and P. D. Sathya, "Squirrel search optimizer: nature inspired metaheuristic strategy for solving disparate economic dispatch problems", *International Journal of Intelligent Engineering and Systems*, Vol. 13, pp. 111-121, 2020, doi: 10.22266/ijies2020.1031.11.