



OPEN

Computational assessment of groundwater salinity distribution within coastal multi-aquifers of Bangladesh

Mehdi Jamei¹, Masoud Karbasi², Anurag Malik³, Laith Abualigah⁴,
Abu Reza Md Towfiqul Islam⁵ & Zaher Mundher Yaseen^{6,7}✉

The rising salinity trend in the country's coastal groundwater has reached an alarming rate due to unplanned use of groundwater in agriculture and seawater seeping into the underground due to sea-level rise caused by global warming. Therefore, assessing salinity is crucial for the status of safe groundwater in coastal aquifers. In this research, a rigorous hybrid neurocomputing approach comprised of an Adaptive Neuro-Fuzzy Inference System (ANFIS) hybridized with a new meta-heuristic optimization algorithm, namely Aquila optimization (AO) and the Boruta-Random forest feature selection (FS) was developed for estimating the salinity of multi-aquifers in coastal regions of Bangladesh. In this regard, 539 data samples, including ten water quality indices, were collected to provide the predictive model. Moreover, the individual ANFIS, Slime Mould Algorithm (SMA), and Ant Colony Optimization for Continuous Domains (ACOR) coupled with ANFIS (i.e., ANFIS-SMA and ANFIS-ACOR) and LASSO regression (Lasso-Reg) schemes were examined to compare with the primary model. Several goodness-of-fit indices, such as correlation coefficient (R), the root mean squared error (RMSE), and Kling-Gupta efficiency (KGE) were used to validate the robustness of the predictive models. Here, the Boruta-Random Forest (B-RF), as a new robust tree-based FS, was adopted to identify the most significant candidate inputs and effective input combinations to reduce the computational cost and time of the modeling. The outcomes of four selected input combinations ascertained that the ANFIS-OA regarding the best accuracy in terms of (R = 0.9450, RMSE = 1.1253 ppm, and KGE = 0.9146) outperformed the ANFIS-SMA (R = 0.9406, RMSE = 1.1534 ppm, and KGE = 0.8793), ANFIS-ACOR (R = 0.9402, RMSE = 1.1388 ppm, and KGE = 0.8653), Lasso-Reg (R = 0.9358), and ANFIS (R = 0.9306) models. Besides, the first candidate input combination (C1) by three inputs, including Cl⁻ (mg/l), Mg²⁺ (mg/l), Na⁺ (mg/l), yielded the best accuracy among all alternatives, implying the role importance of (B-RF) feature selection. Finally, the spatial salinity distribution assessment in the study area ascertained the high predictability potential of the ANFIS-OA hybrid with B-RF feature selection compared to other paradigms. The most important novelty of this research is using a robust framework comprised of the non-linear data filtering technique and a new hybrid neuro-computing approach, which can be considered as a reliable tool to assess water salinity in coastal aquifers.

In many places of the world, groundwater is the most crucial water source for economic development and human survival¹; it is the typical source of drinking water in many parts of the world. Groundwater can be regarded as a renewable natural resource because it can be refilled continually in most circumstances². More than 2.5 billion

¹Faculty of Engineering, Shohadaye Hoveizeh Campus of Technology, Shahid Chamran University of Ahvaz, Dashte Azadegan, Iran. ²Water Engineering Department, Faculty of Agriculture, University of Zanjan, Zanjan, Iran. ³Punjab Agricultural University, Regional Research Station, Bathinda, Punjab, India. ⁴Faculty of Information Technology, Middle East University, Amman, Jordan. ⁵Department of Disaster Management, Begum Rokeya University, Rangpur 5400, Bangladesh. ⁶New Era and Development in Civil Engineering Research Group, Scientific Research Center, Al-Ayen University, Thi-Qar 64001, Iraq. ⁷Department of Earth Sciences and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia. ✉email: yaseen@ukm.edu.my

people globally depend on groundwater for water supply and to keep many key terrestrial ecosystems alive³. In many parts of the world, excess water exploration due to increasing water demand has threatened their long-term viability^{4,5}. The encroachment of seawater into coastal aquifers and the removal of water from coastal aquifers have caused erratic changes in the quality of groundwater and water flow patterns.

Recently, the growing demand for groundwater resources has subjected such natural resources to more strain than ever before^{6,7}. In unconfined coastal aquifers, the most typical problem is groundwater salinization, mainly when excessive groundwater pumping reduces the piezometric head⁸. Confining layers that partly separate seawater from groundwater and the vadose zone has complicated the hydrogeological processes in confined and semi-confined coastal aquifers⁹. As a result, seawater can easily intrude into semi-confined coastal aquifers. The problem of groundwater depression has been documented in various cases worldwide^{10–12}. The increasing level of salt accumulation in both plants and soil has significantly increased groundwater salinity, which has negatively impacted ecological health, the economic advancement of residents, and the productivity of coastal crops. Increased salinity also hurts drinking water quality, thereby jeopardizing human health^{13,14}. Furthermore, groundwater salinization results in an increased quantity of salt in the root zone, which creates an osmotic impact on plants, forcing them to expend more energy to absorb water from the soil, thereby limiting their ability to develop¹⁵.

Another issue is that excessive groundwater extraction and poor management have established local or regional groundwater depressions in several regions¹⁶. Prolonged over-extraction of groundwater can result in depression, as seen in the North China Plain, where several towns, including Cangzhou, Dezhou, and Tianjin, were in severe water depression¹⁷. Excessive groundwater extraction from aquifers in Dhaka, Bangladesh, has also been reported to cause an exponential fall in groundwater levels around the city and water quality-related issues¹⁸. Several wells were installed in Tripoli, Northwest Libya, to pump groundwater in the city; this act caused a sharp decline in groundwater level; a further construction of a groundwater depression caused the limited discharge of freshwater to the ocean¹⁹. Groundwater formation may also result in geological risks, such as ground cracks and land subsidence. In China, excessive groundwater extraction in some representative locations has been reported to cause geological problems due to prolonged over-extraction consistently distributed with groundwater depression²⁰.

Based on the reported literature, the salinity in groundwater is stochastic, and this is because several parameters affect its concentration and magnitude. Those parameters include upward intrusion from deep aquifers²¹, evaporation rate from soil²², irrigation saline water, and wastewater infiltration²³. Hence, understanding the actual mechanism of groundwater salinization and the affected sources is essential for water resources management and sustainability^{24,25}.

Worth mentioning, Electrical conductivity (EC) is used to explain the salinity of water; the concentration of dissolved salts is a metric to determine the EC of groundwater²⁶. In a thoroughly prepared groundwater sample, EC is generally tested by creating an electric current between the two electrodes of a salinometer. This approach is point-based and analyses the EC of the tested groundwater samples. Although this process is accurate, the preparation step makes it time-consuming to perform over a large area. In the local setting, direct current resistivity methods examine EC distribution. In this method, the potential field is determined using two additional electrodes; a current is created and delivered into the ground by point electrodes²⁷. However, this is a slow procedure that cannot be applied on a regional scale. Previous studies on mapping groundwater salinity based on a regional scale using the feasibility of remote sensing data have been conducted remarkably^{28–31}. However, the main concept of using geographical information system data for mapping the salinity is to estimate the salinity for unknown data points using the interpolation technique within a defined range of measured data points. The sensing technique has its merits. It is quick and straightforward to apply; however, it is connected with significant error calculation and is proportional to the square of the distance between data points³². Also, it does not consider the sample's distribution in high salinity areas. Hence, introducing new technology, such as computer aid models for solving this complex and significant natural issue, is one of the prioritized research topics in water and geo-science.

The ability to predict groundwater quality is crucial to comprehending its evolution trend^{33,34}. This is especially useful for determining groundwater quality in the groundwater depression cone zone. Numerical modeling and statistical prediction methods are available for predicting groundwater quality. However, machine learning (ML) models have been recently reported as a new method that substantially impacts groundwater quality modeling^{35,36}. As a known data analysis method, ML models can automate the framework of an analytical model. Artificial intelligence (AI) is based on the idea that machines can learn from data, recognize patterns, and make decisions with minimal human interaction^{37,38}. The ability of ML models to model groundwater salinity has been demonstrated via the establishment of a linear or non-linear relationship between water salinity and its control parameters (such as water table, evaporation, and distance to saltwater bodies) and using those relationships for the prediction of water salinity for regions with unavailable data points^{39,40}. Various versions of ML models have been reported in the literature, such as artificial neural network (ANN)^{41–45}, support vector machine (SVM)^{46–48}, adaptive neuro-fuzzy inference system (ANFIS)^{49,50}, ensemble ML models^{38,51,52}, group method of data handling (GMDH)⁵³, and Gaussian process scheme⁵⁴. The significant limitations associated with predictive ML models (1) the need for adequate input variables to explain the target data that may not be available everywhere^{55,56}, (2) the influence of well excessive pumping^{57,58}, (3) the reliability of the learning process of the predictive model where essential hyperparameters are optimized^{59,60}, (4) coupled ML models where a pre-processing technique was integrated for data time series decomposition^{61,62}. The ML model was adopted based on the inspiration of developing a new hybrid model for the ANFIS model. In groundwater quality modeling, hybrid ANFIS showed a promising result^{63,64}.

Due to the highly non-linear relationships between input predictors and water quality targets in coastal aquifers, using a scientific-based strategy to determine the optimal candidate input combinations for feeding the ML

methods is very important. It has received less attention in the previous literature. In most previous research, regardless of the behavior of the data, a certain number of possible input combinations have been examined using the ML methods, and superior results have been presented. However, selecting specific combinations without a scientific basis may increase the uncertainty and decrease the accuracy of the outcomes. This motivated this study focuses on three significant aspects. The aims of the current investigation, novel predictive models, were developed based on the hybridization of the ANFIS model with new nature-inspired optimization algorithms (i.e., Aquila optimization) for groundwater salinity prediction. The second aim is to inspect the highly influential predictors using the newly explored feature selection algorithm, Boruta-Random Forest. The outcomes of the primary model were examined with standalone ANFIS, ANFIS-SMA, ANFIS-ACOR, and Lasso-Reg approaches. Finally, the current research was adopted on a significant case study, “coastal areal of Bangladesh,” where the groundwater salinity is a vital issue for that region. The current research can provide an essential vision for introducing a reliable computer aid model.

Materials and methods

Study area description. With approximately 24,000 km², the coastal regions are primarily low-lying areas in the southern portion of Bangladesh along the Bay of Bengal (BoB). Diverse geomorphic characteristics of the coastal district include the deep funnel-shaped structures of the BoB’s northern landfill. Tidal fluctuation is prevalent in most river systems from the coastal area, causing fluctuations of 2–4 m. Groundwater pollution is caused by a rise in the relative sea level, rapid population increase, a poor drainage system, salinity intrusion, and other factors. Aquaculture and agricultural activities are the primary sources of income in the coastal regions. Due to rising soil salinity, aquaculture, particularly shrimp culture, has expanded while paddy crop production has decreased⁶⁵. The typical coastal climatic system is characterized by a warm and tropical environment dominated by the BoB’s southwest monsoonal flow. The average annual precipitation and temperature (June–September) are 2000–2500 mm and 25 °C, respectively.

The Bengal Basin’s coastal area started in the late Holocene to the Recent Age⁶⁶. The study area formed the basin’s deeper portion throughout the Holocene age. The lithology of this area is varied, with coarse-to fine-grained sandstone and peat soil combined with silty clay⁶⁵. Each sediment layer containing groundwater comprises coarse-grain sand, fine-grained silt, and clay⁶⁷. The coastal region’s hydrogeology comprises unconsolidated Quaternary alluvial sediments that are covered by a thick (3 to 7 m) silty-clay layer. The shallow aquifer depth ranged from 10 to 50 m, with salt concentrations ranging from 1500 mg/L to 2000 mg/L. Rainwater collecting, especially during the monsoon season, is another viable source of freshwater. The people rely on salinity-rich rivers, channels, and fishponds for their water supply⁶⁸. Furthermore, rainwater recharges the shallow aquifer during the dry winter season, which is invaded excessively⁶⁹.

There are three types of aquifers⁷⁰. The upper shallow aquifer is located northwest of the coastal area, with a thickness varying from a few meters to 60 m. Second, the shallow aquifer, which has a thickness ranging from 10 m to more than 100 m⁷¹, contains saltwater pockets, particularly abundant in coastal and estuarine flooded areas. Third, there is a deep aquifer with a thickness of more than 200 m and various features in the southern portion of the coastal zone.

The water in this region’s aquifer is frequently replenished by rainfall, rivers, and channels⁷². The groundwater is significantly depleted during the dry and monsoon seasons and subsequently refilled. Groundwater flow may have aided the saltwater infiltration into the water-bearing strata. The parent rock impacts the water chemistry, and numerous types of minerals found in the aquifer regulate the water quality. According to the available geological data, the aquifers in this area are either unconfined or semi-confined.

Data description and sampling technique. For using machine learning methods, a large dataset is required. The datasets used in this study came from⁶⁵ and⁷³, and the sampling design and analytical procedures are described below. During the wet season, 539 groundwater samples were collected from three campaigns between 2015 and 2017. Each sample was assigned an ID number, and coordinates were confirmed using a portable GPS device⁷⁴, as shown in Fig. 1. Before collecting the sample from the tube well, the groundwater was pumped for at least 10 min to remove any standing water.

The pumping of the sampling tube well was continued until the pH and electrical conductivity (EC) were both steady. The samples were collected in prewashed high-density polypropylene (HDPP) bottles^{66,75}. It is worth noting that each station collected two sets of replicated samples. The samples were collected and filtered using 0.45 m membranes from MF-Millipore™ in the United States. The samples’ HDPP bottles were kept at 4 °C in a more excellent box and subsequently sent to the laboratory for further analysis. While EC and pH were measured using portable pH/EC meters (Hanna HI 9811–5).

A field survey was used to measure groundwater depth and salinity during the wet seasons. Ion chromatography with Dionex ICS-90 was used to determine cations (Ca²⁺, Mg²⁺, Na⁺, K⁺) and anions (Cl⁻, HCO₃⁻, NO₃⁻, PO₄²⁻, SO₄²⁻, and F⁻). Five standard solutions (1, 5, 10, 15, and 20 mg/L) were employed during the calibration process. A conventional laboratory process and quality control checks were used to provide quality assurance. Three replicated samples were obtained simultaneously to ensure that the test results were accurate by cross-checking with a qualified laboratory. The ion charge balance error (ICBE), which was used to determine accuracy, ranged from 2.63 to 8.62 percent, with a mean of 8.24 percent, well within the permissible limit of 10%. Table 1 lists the descriptive statistics of datasets used in the salinity assessment of the multi-aquifers in coastal regions of Bangladesh. As can be seen, the maximum kurtosis values were owing to the PO₄⁻² (mg/l), K⁺ (mg/l), and Ca²⁺ (mg/l).

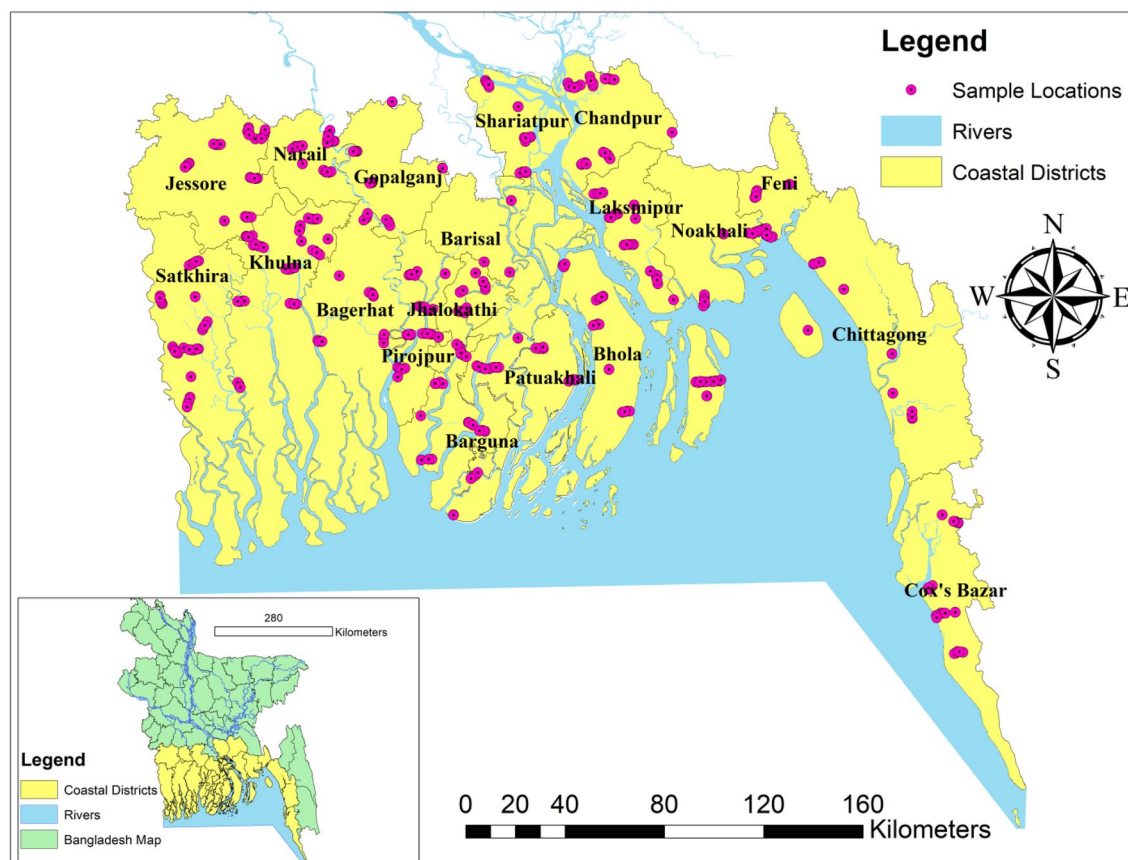


Figure 1. Location of sampling point in area study during the wet season; map of coastal regions of Bangladesh.

Variables	Minimum	Maximum	Mean	Std. D	COV (%)	Skewness	Kurtosis
Depth (m)	18	336	113.1	62.84	55.56	1.957	3.141
pH	6.2	8.8	7.336	0.4449	6.065	0.0911	0.943
Ca ²⁺ (mg/l)	0.2511	902.4	89.82	111	123.6	3.479	17.15
Mg ²⁺ (mg/l)	0.0681	718.9	107.1	131.6	122.9	1.889	3.419
Na ⁺ (mg/l)	0.2	4211	722.5	836.8	115.8	1.295	0.789
K ⁺ (mg/l)	0.252	176.8	10.33	16.38	158.5	5.806	43.54
HCO ₃ ⁻ (mg/l)	24.4	823.5	221.7	119.3	53.81	1.123	1.91
SO ₄ ²⁻ (mg/l)	0	2926	22.87	141.1	617.2	16.9	336
PO ₄ ⁻² (mg/l)	0.03	57.7	1.359	3.498	257.4	10.34	142.6
Cl ⁻ (mg/l)	2.1	16,250	2054	2713	132.1	1.928	4.64
Salinity (ppt)	0	25.5	3.05	3.661	120.0	1.769	4.073

Table 1. descriptive statistics of all groundwater quality parameters for modeling the salinity using three hybrid ANFIS models and the Lasso-Reg approach.

Boruta-Random forest feature selection. The selection of features is critical in implementing machine learning algorithms⁷⁶. Boruta is an algorithm for feature selection. More precisely, it acts as a wrapper algorithm for Random Forest. This algorithm is named after a monster from Slavic folklore who inhabited pine trees. The stage of feature selection is critical in predictive modeling. This strategy is vital when a data set including several variables is provided for model construction. This is particularly true when the goal is to understand the mechanics behind the interest variable rather than merely build a high-prediction-accuracy black-box model. Boruta determines the Z-scores for each input predictor concerning the shadow property. The distribution of Z-score metrics reveals the essential characteristics of the predictors⁷⁷. A minimal-optimal feature selection technique was used by ranking and residuals based on the Boruta-determined relevance criteria, followed by stepwise model development⁷⁸.

1. To begin, it randomizes the input data set by making scrambled duplicates of all features (shadow features).

- Then, it trains a random forest classifier on the larger data set and evaluates the value of each feature using a feature importance measure (the default is Mean Decrease Accuracy), where greater equals more significant. The following Equation calculates the MDA:

$$MDA = \frac{1}{m_{tree}} \sum_{m=1}^{m_{tree}} \frac{\sum_{t \in OOB} I(y_t = f(x_t)) - \sum_{t \in OOB} I(y_t = f(x_t^n))}{|OOB|} \quad (1)$$

where OOB denotes out-of-bag (i.e., the prediction error for each of the training trials aggregated by bootstrap), whereas $(y_t = f(x_t))$ and $(y_t = f(x_t^n))$ denote the predicted values before and after permutation, respectively. Additionally, $I()$ denotes the indicator function.

- Each iteration determines if a genuine feature is more essential than the best of its shadow features (i.e., whether the feature has a higher Z score than the shadow features' maximum Z score) and continually eliminates features thought to be very irrelevant. The Z-score is computed as follows:

$$Z - score = \frac{MDA}{std} \quad (2)$$

where std is the standard deviation of accuracy losses, and then the maximum Z-score for duplicate attributes was computed (MZSA).

- If Z-scores are less than MZSA, the inputs are tagged "unimportant" and separated permanently until inputs with Z-scores more than MZSA are designated "Confirmed".
- Finally, the method terminates when all features have been validated or rejected or the required number of random forest iterations reached.

Lasso regression. Robert Tibshirani coined the term LASSO in 1996⁷⁹. It is a robust approach that accomplishes two primary tasks: regularization and feature selection. The Lasso approach constrains certain of the model parameters' absolute values. The total must be less than a preset value (upper bound). To do this, the approach employs a shrinkage (or regularization) procedure in which it penalizes the coefficients of regression variables, thereby shrinking them to zero⁸⁰. Incorporating a penalty item into linear regression may dramatically reduce the variance of a model by effectively shrinking the coefficient estimates, particularly in models with high-dimension predictors⁸¹. The optimized objective function of Lasso Regression (Lasso-Reg) is as follows:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \Gamma \sum_{j=1}^p |\beta_j| \quad (3)$$

where β_0 denotes the Lasso-Reg shift and β_j denotes the x_{ij} coefficients. In this relation, Γ is a regulation parameter, which means that if its value is equal to zero, the model becomes a normal regression, and all variables will be present, and if its value increases, the number of independent variables in the model will decrease. The determination of the value for this parameter is usually done by the cross-validation method⁸⁰.

Adaptive neuro-fuzzy inference system (ANFIS). The ANFIS technique is based on a knowledge-based mix of fuzzy inference systems (FIS) and artificial neural networks (ANN). The FIS can generate IF-THEN fuzzy rules from fuzzy sets with an adequate membership function (MF) to represent human thought, but its capabilities are restricted to adaptive learning⁸². While ANNs are capable of adaptive learning for decision-making, they cannot explain how the choice was formed. Thus, incorporating adaptive learning capabilities from ANNs into the IF-THEN fuzzy rules in FIS structures becomes more powerful and may be utilized to tackle complicated engineering or non-engineering issues in various applications⁸³.

The ANFIS model is used in this work due to its high capacity for learning and superior performance⁸⁴. The ANFIS model is structured in two parts: antecedent and consequent. To keep things simple, the ANFIS structure is configured with two inputs, x , and y , as seen in Fig. 2. The ANFIS model is composed of five levels structurally. Each level has a distinct role, detailed below⁸⁵.

- Layer 1 (Fuzzification layer): This layer accepts discrete input values and gives them membership functions.

$$O_{1,i} = \mu A_i(x) \quad i = 1, 2 \quad (4)$$

$$O_{1,i} = \mu B_{i-2}(y) \quad i = 3, 4 \quad (5)$$

The input nodes are represented by x and y . The linguistic variables are denoted by A and B . $A_i(x)$ and $B_{i-2}(y)$ are node membership functions.

- Layer 2 (Rule layer): Each rule's firing strength is created in this layer using the product operation.

$$O_{2,i} = w_i = \mu A_i(x) \mu B_i(y) \quad i = 1, 2 \quad (6)$$

where w_i denotes each node's output.

- Layer 3 (Normalization layer): This phase normalizes the firing strength of each rule to the total firing strength.

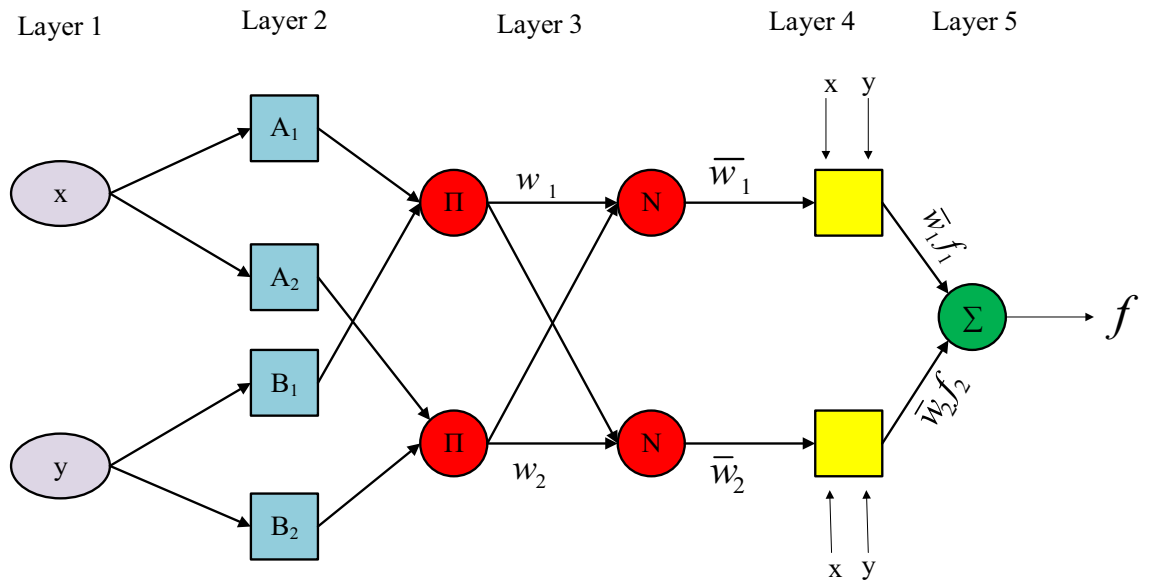


Figure 2. Structure of the ANFIS model.

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1 + w_2} \quad i = 1, 2 \tag{7}$$

4. Layer 4 (Defuzzification layer): This layer takes as normalized input values and their corresponding parameters ($p_i, q_i,$ and r_i). The defuzzified values are returned after combining these arguments.

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i (p_i x + q_i y + r_i) \tag{8}$$

5. Layer 5 (Output layer): The final output is a weighted average of the output from each rule.

$$O_{5,i} = \sum \bar{w}_i f_i = \frac{\sum \bar{w}_i f_i}{\sum w_i} \quad i = 1, 2 \tag{9}$$

The most significant notion in ANFIS is determining the number of membership functions. This may be regarded as a clustering problem; consequently, the FCM is employed to create a limited number of fuzzy rules. Bezdek invented the FCM in 1984⁸⁶. Each data point in the FCM method belongs to one of the clusters with a membership value that varies from zero to one. The FCM may be obtained by optimizing the objective function below⁸⁷.

$$J_{FCM} = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|x_j - v_i\|^2 \tag{10}$$

where u_{ij} is membership degree, c is the total number of clusters, m is a constant value, $\|x_j - v_i\|$ is Euclidean distance of x_j from i th cluster center v_i . In the FCM technique, the cluster center and membership degree may be computed using the following Eqs.⁸⁷:

$$v_i = \frac{\sum_{j=1}^n u_{ij} x_j}{\sum_{j=1}^n u_{ij}} \quad i = 1, 2, \dots, c \tag{11}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|} \right)^{\frac{2}{m-1}}} \tag{12}$$

Aquila optimizer (AO). Aquila Optimizer (AO) is a novel nature-inspired algorithm proposed by Abualgah et al. in⁸⁸. The following subsections explain how the AO models these processes.

AO simulates Aquila’s hunting behavior by demonstrating the actions taken at each hunt stage⁸⁹. The AO algorithm’s optimization processes are divided into four categories. The following is a mathematical model of the AO.

Step 1: Expanded exploration. The Aquila accepts the prey area and chooses the best hunting area by high soar with the vertical stoop in the first searching technique (X_1). Equation (3) represents this behavior mathematically⁹⁰.

$$X_1(t+1) = X_{best}(t) * \left(1 - \frac{t}{T}\right) + (X_m(t) - X_{best}(t) * rand) \quad (13)$$

where $X_1(t+1)$ is the solution created by the first search method for the next iteration of t . (X_1). The best-obtained solution at the t th iteration is $X_{best}(t)$, representing the prey's approximate location. $(1 - \frac{t}{T})$ is used to control the number of iterations in the expanded search (exploration). A random value between 0 and 1 is called $rand$. The current iteration and the maximum number of iterations are represented by t and T , respectively.

$$X_M(t) = \frac{1}{N} \sum_{i=1}^N X_i(t), \forall j = 1 \dots \dots Dim \quad (14)$$

where Dim is the problem's dimension size and N is the number of possible solutions in the population.

Step 2: Narrowed exploration. When the prey area is discovered from a high vantage point, the Aquila circles above the target prey prepares the land, and then attacks. Equation (15) represents this behavior mathematically⁹¹.

$$X_2(t+1) = X_{best}(t) * Levy(D) + X_R(t) + (y - x) * rand \quad (15)$$

where $X_2(t+1)$ is the solution created by the second search method (X_2) for the next iteration of t . $Levy(D)$ is the levy flight distribution function calculated using Eq. (16). At the i th iteration, $X_R(t)$ is a random solution taken in $[1 N]$.

$$Levy(D) = s * \frac{u * \sigma}{|v|^{\frac{1}{\beta}}} \quad (16)$$

where s is a constant value of 0.01, u , and v are random integers between 0 and 1, and σ is a value calculated by using Eq. (17).

$$\sigma = \left(\frac{\Gamma(1 + \beta) * \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) * \beta * 2^{\left(\frac{\beta-1}{2}\right)}} \right) \quad (17)$$

where β is a constant with a value of 1.5; in Eq. (15), the spiral shape in the search is represented by y and x , which are computed as follows.

Where

$$y = r * \cos(\theta), x = r * \sin(\theta), r = r_1 + U.D_1 \quad (18)$$

$$\theta = -\omega * D_1 * \theta_1, \theta_1 = \frac{3 * \pi}{2} \quad (19)$$

For a fixed number of search cycles, r_1 takes a value between 1 and 20, and U is a small value set to 0.00565. D_1 is an array of integer numbers ranging from 1 to the search space length (Dim), and ω is a small value set to 0.005. Figure 3 depicts the AO's behavior in a spiral shape.

Step 3: Expanded exploitation. When the prey area is precisely defined and the Aquila is ready to land and attack, the third technique (X_3) is used. Equation (20) represents this behavior mathematically⁹².

$$X_3(t+1) = (X_{best}(t) - X_M(t)) * \alpha - rand + ((UB_j - LB_j) * rand + LB_j) * \delta \quad (20)$$

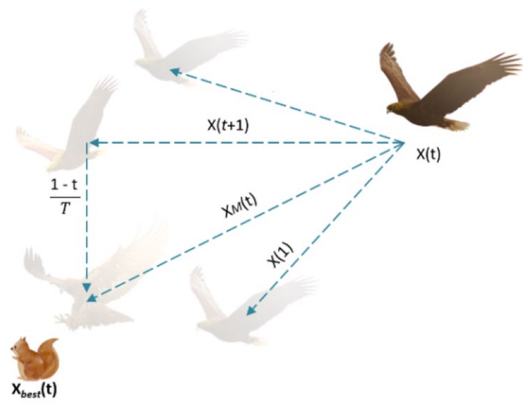
where $X_3(t+1)$ is the solution of the next iteration of t , which is generated by the third search method (X_3). $X_{best}(t)$ refers to the approximate location of the prey until i th iteration, and $X_M(t)$ denotes to the mean value of the current solution at t th iteration, which is calculated using Eq. (14). "rand" is a random value between 0 and 1. α and δ are the exploitation adjustment parameters fixed in this paper to a small value (please refer to Table 3).

Step 4: Narrowed exploitation. When the Aquila approaches the prey in the fourth technique (X_4), the Aquila attacks the prey over land based on their stochastic movements. Equation (21) represents this behavior mathematically⁸⁹.

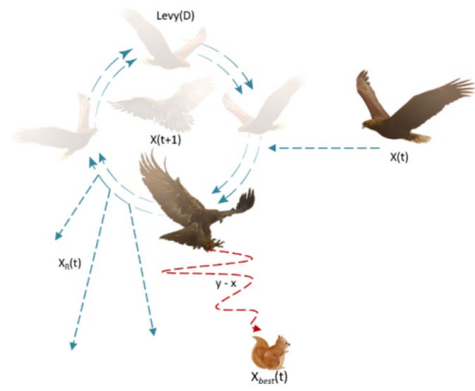
$$X_4(t) = QF * X_{best}(t) - (G_1 * X(t) * rand) - G_2 * Levy(D) + rand * G_1 \quad (21)$$

where $X_4(t+1)$ is the solution of the fourth search method's (X_4) for the next iteration of t , the quality function (Q_F) is used to balance the search strategies and is calculated using Eq. (21). G_1 refers to various AO motions that are used to track the prey during the elope and are generated using Eq. (21). G_2 shows decreasing values from 2 to 0, which represent the AO's flight slope as it follows the prey during the elope from the first (1) to the last (t) location, as calculated using Eq. (21). The current solution at the t th iteration is $X(t)$.

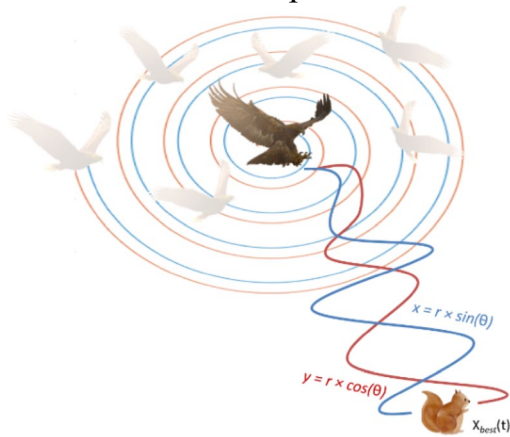
$$Q_F(t) = t^{\left(\frac{2*rand-1}{(1-T)^2}\right)}, G_1=2*rand-1, G_2 = 2 * \left(1 - \frac{t}{T}\right) \quad (22)$$



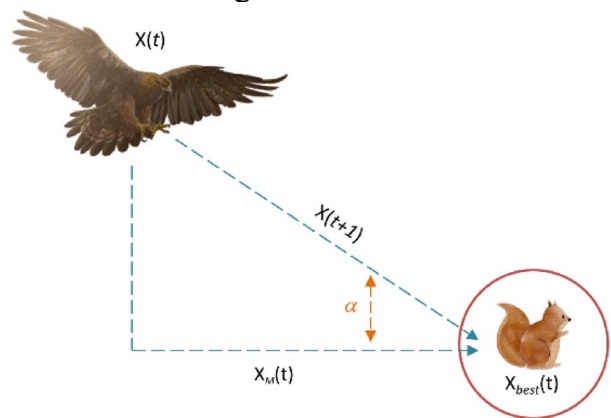
i. Aquila high soar with the vertical stoop



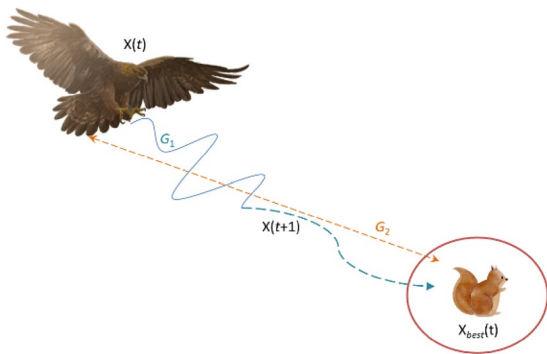
ii. The contour flight of the Aquila, with a short glide attack



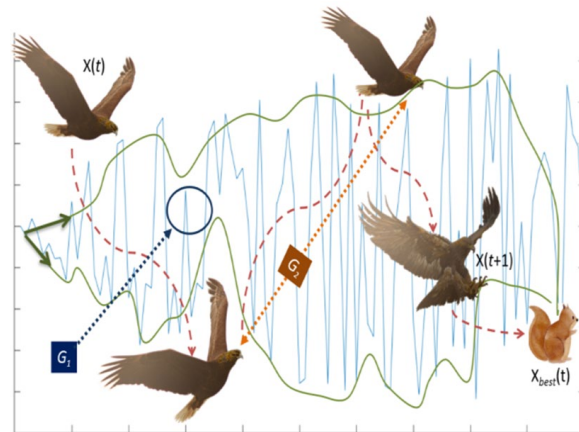
iii. The AO in a spiral shape



iv. The Aquila low-flying attacks with a slow descent



v. The Aquila walk and grab prey



vi. The effects of the Q_F , G_1 , and G_2

Figure 3. the schematic view of the OA algorithm.

The quality function value at the t th iteration is $Q_F(t)$, and the random value between 0 and 1 is $rand$. The current iteration and the maximum number of iterations are represented by t and T , respectively. The Levy flight distribution function is computed using Equation (6). The effects of the Q_F , G_1 , and G_2 on the AO's behavior are shown in Fig. 3. The Pseudo-code of the AO is detailed in Algorithm 1.

Algorithm 1

The Pseudo-code of the Aquila Optimizer (AO)

Initialization phase:

Initialize the population X of the AO.

Initialize the parameters of the AO.

WHILE ($t < T$)

Calculate the fitness function values.

Determine the best obtained solution according to the fitness values ($X_{best}(t)$).

FOR ($i=1,2,\dots,N$)

Update the mean value of the current solution $X_M(t)$.

Update the x , y , G_1 , G_2 , $Levy(D)$, etc.

IF ($t \leq (2/3) * T$)

IF ($rand \leq 0.5$)

Update the current solution using Equation (13).

>>Step 1: Expanded exploration

(X_1)

IF (Fitness $X_1(t+1) < \text{Fitness } X(t)$)

$X(t) = X_1(t+1)$

IF (Fitness $X_1(t+1) < \text{Fitness } (X_{best}(t))$)

$X_{best}(t) = X_1(t+1)$

ENDIF

ENDIF

ELSE

Update the current solution using Equation (15).

>>Step 2: Narrowed exploration

(X_2)

IF (Fitness $X_2(t+1) < \text{Fitness } X(t)$)

$X(t) = X_2(t+1)$

IF (Fitness $X_2(t+1) < \text{Fitness } (X_{best}(t))$)

$X_{best}(t) = X_2(t+1)$

ENDIF

ENDIF

ENDIF

ELSE

IF ($rand \leq 0.5$)

Update the current solution using Equation (20).

>>Step 3: Expanded exploitation

(X_3)

IF (Fitness $X_3(t+1) < \text{Fitness } X(t)$)

$X(t) = X_3(t+1)$

IF (Fitness $X_3(t+1) < \text{Fitness } (X_{best}(t))$)

$X_{best}(t) = X_3(t+1)$

ENDIF

ENDIF

ELSE

Update the current solution using Equation (21).

>>Step 4: Narrowed exploitation

(X_4)

IF (Fitness $X_4(t+1) < \text{Fitness } X(t)$)

$X(t) = X_4(t+1)$

IF (Fitness $X_4(t+1) < \text{Fitness } (X_{best}(t))$)

$X_{best}(t) = X_4(t+1)$

ENDIF

ENDIF

ENDIF

ENDIF

ENDFOR

ENDWHILE

Return the best solution (X_{best}).

Slime mould algorithm (SMA). The physarum polycephalum is frequently described in conjunction with the slime mould. Slime mould is so termed because it is classified as a fungus⁹³.

- *Approach food*

The following equations model depicts the SMA function. To replicate the constriction approach, the model equations are introduced:

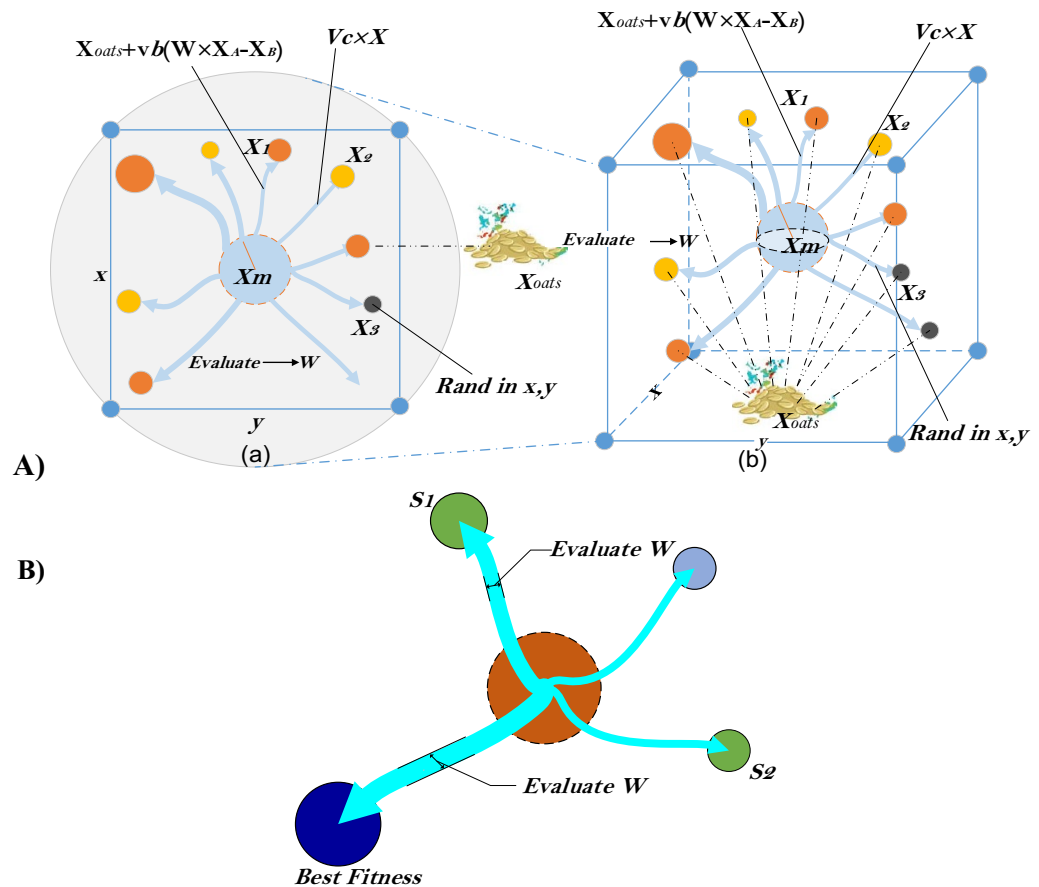


Figure 4. SMA algorithm stages; (A) Potential positions (B) Process of the fitness function.

$$\vec{X}(t+1) = \begin{cases} \vec{X}_b(t) + \vec{vb} \cdot (\vec{W} \cdot \vec{X}_A(t) - \vec{X}_B(t)), & r < p \\ \vec{vc} \cdot \vec{X}(t), & r \geq p \end{cases} \quad (23)$$

where \vec{vb} is a parameter utilized in $[-a, a]$, \vec{vc} a parameter values changes from 1 to 0. t is the t_{th} iteration, \vec{X}_b is the individual position of the current best, \vec{X} is the position of the current solution, \vec{X}_A and \vec{X}_B are two solutions selected randomly, \vec{W} are the weight of the current solution⁹⁴. The p value is determined as follows:

$$p = \tanh|S(i) - DF| \quad (24)$$

where $i \in \{1, 2, 3, \dots, n\}$, $S(i)$ is the fitness function of the current solution, DF is the best-obtained fitness value. The \vec{vb} is determined as follows:

$$\vec{vb} = [-a, a], a = \operatorname{arctanh}\left(-\left(\frac{t}{\max_iter}\right) + 1\right) \quad (25)$$

The \vec{W} is determined as follows:

$$\vec{W}(\operatorname{SmellIndex}(i)) = \begin{cases} 1 + r \cdot \log\left(\frac{bF - S(i)}{bF - wF} + 1\right), & \text{condition} \\ 1 - r \cdot \log\left(\frac{bF - S(i)}{bF - wF} + 1\right), & \text{others} \end{cases} \quad (26)$$

$$\operatorname{SmellIndex} = \operatorname{sort}(S)$$

where r is a random value in $[0, 1]$, bF is the best-obtained fitness value, wF is the worst obtained fitness value, $\operatorname{SmellIndex}$ is sorted fitness value. Figure 4a shows the impacts of potential positions⁹⁵.

- **Wrap food**

When the food product is pleased to stretch to a place where the food quantity is weak, the priority of that region decreases, causing researchers to shift their attention to other regions of food availability that are not as significant as the food product. Figure 4b depicts the rule for assessing slime mould fitness values.

The mathematical representation for updating positions is given as follows:

$$\vec{X}^* = \begin{cases} \text{rand} \cdot (UB - LB) + LB, \text{rand} < z \\ \vec{X}_b(t) + \vec{vb} \cdot (W \cdot \vec{X}_A(t) - \vec{X}_B(t)), r < p \\ \vec{vc} \cdot \vec{X}(t), r \geq p \end{cases} \quad (27)$$

where LB and UB are the lower bound and upper boundaries rand and r are random values in $[0,1]$, z is a parameter value in $[0, 0.1]$.

- *Grabble food*

\vec{vb} is an area of random numbers in $[-a, a]$. The \vec{vc} is given in $[-1,1]$. Although slime mould received a better feed supply, it would still spread organic material for seeking other locations for an upper-class food supply rather than investing all of it in a single area to discover a more reliable supply of nutrition. The SMA algorithm's mechanism is depicted in Algorithm 2.

Algorithm 2

Pseudo-code of SMA

Initialize the parameters *population size, dim, lb, ub, z, max_iter*;
 Initialize the set of random slime mould positions $X_i (i = 1, 2, \dots, n)$;
While ($t \leq \text{max_iter}$)
 Calculate and sort the fitness of all slime mould;
Update the bestFitness, the worst fitness
 Calculate the *weight of slime mould (W)*;
update the best fitness, the best position(X_b)
For each search agent
 Update p, vb, vc;
 Update the Position of search agents;
End For
 $t = t + 1$;
End While
Return *bestFitness, the best position(X_b)*;

Ant colony optimization (ACO). Ant colony optimization (ACO), invented by Dorigo⁹⁶, is a multi-agent approach used to tackle optimization issues. This method is based on observational data of real ants seeking food. Ants are small social insects in colonies and cooperate to ensure the colony's survival. While hunting for food, the ants inspect their surroundings and mark them with pheromones, which the colony's other members can follow. When ants locate a food supply, they attempt to nurture it by transporting it back to the colony via the nearest root⁹⁷. The ACO algorithm uses a discrete structure to determine the answer. The concept of discrete structure in ACO means that each decision variable in the defined interval is divided into a certain number of states. By discretizing the space of variables, there is a limit to the algorithm, reducing accuracy. In this regard, ACO generalization to continuous space was considered. If the decision variable space is assumed to be continuous, the algorithm will move in the R space of real numbers. The ACOR algorithm performs spatial integration in decision variables using a probability density function (PDF). Sossa and Dorigo proposed using a Gaussian function to create such a structure⁹⁸. A one-dimensional Gaussian function cannot produce a maximum of several points, whereas using a Gaussian kernel function, the sum of the weights of several single Gaussian functions, can perform such a task. The following Equation defines the weighted sum of 1-D Gaussian functions:

$$G^i(x) = \sum_{l=1}^k w_l g_l^i(x) = \sum_{l=1}^k w_l \frac{1}{\sigma_l^i \sqrt{2\pi}} e^{-\frac{(x-x_l^i)^2}{2\sigma_l^i{}^2}} \quad (28)$$

where i is the dimension of the problem, k denotes the total number of best solutions in the solution repository. w_l is the weight that each solution receives based on its rank and it can be calculated using following Eq. ⁹⁸:

$$w_l = \frac{1}{\sqrt{2\pi} qk} e^{-\frac{(r_l-1)^2}{q^2 k^2}} \quad (29)$$

r_l is rank of solutions. The q parameter (Intensification Factor) affects the minimum and maximum limits of w_l . When q is small, the solutions with the highest rankings are highly preferred. Equation 29 is used to compute the elements of the weight vector x . Following that, the sampling process is finished in two steps. The first step is

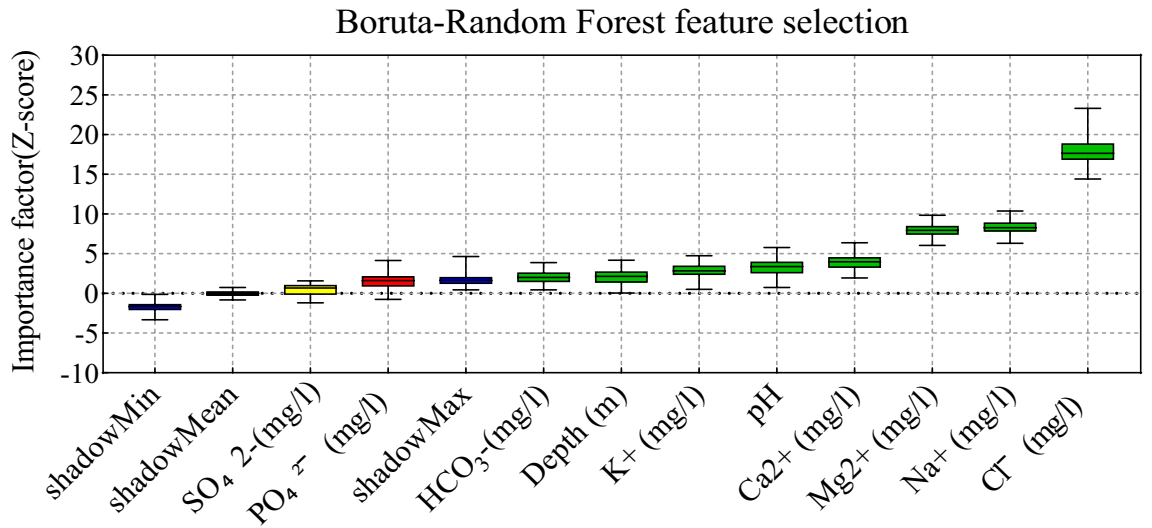


Figure 5. Feature selecting process using the Boruta-Random Forest method based on Z-score for all candidate input variables.

to select one of the Gaussian functions that comprise the Gaussian kernel PDF. The following formula expresses the probability of selecting the l th Gaussian function⁹⁸:

$$p_l = \frac{w_l^i}{\sum_{r=1}^k w_r^i} \tag{30}$$

The chosen Gaussian function is sampled in the second phase. This can be accomplished by employing a random number generator capable of producing random numbers based on a parameterized normal distribution. The standard deviation of a normal distribution PDF is σ_l^i and it is determined using the following Equation:

$$\sigma_l^i = \xi \sum_{e=1}^k \frac{|S_e^i - S_l^i|}{k - 1} \tag{31}$$

where e denotes the iteration and k denotes the solution's number in the solution archive. ξ (Deviation-Distance Ratio) is a parameter that controls the convergence speed. The algorithm's convergence speed decreases with increasing ξ . Solution S_l^i has rank l and S_e^i is the solution in the current iteration⁹⁸. The ACO_R method refines and regenerates the solution archive in each iteration by adding m new solutions ($k \rightarrow k + m$) and then removing the worst m solutions ($k + m \rightarrow k$) in order to maintain the solution archive's size constant (negative and positive update). As a result of the changes to the solutions recorded in the solution archive, the pheromone for each iteration is increased in optimized paths that do not improve the objective function. Thus, the ACO_R algorithm finds the optimal solution. For more simplicity, hereafter, the ACO_R is called ACO.

Model performance evaluation. To assess the prediction performances of the various models, correlation coefficient (R), the root mean squared error (RMSE), Kling-Gupta efficiency (KGE) (Gupta et al., 2009), Willmott's agreement Index (I_A) (Willmott, 1982), Relative absolute error (RAE), Legate and McCabe's Index (E_{LM}) (Legates and McCabe, 2013) and coefficient of uncertainty with 95 confidence level ($U_{95\%}$) were utilized⁹⁹⁻¹⁰¹.

$$R = \frac{\sum_{i=1}^N (Salinity_{o,i} - \overline{Salinity_o}) (Salinity_{p,i} - \overline{Salinity_p})}{\sqrt{\sum_{i=1}^N (Salinity_{o,i} - \overline{Salinity_o})^2 \sum_{i=1}^N (Salinity_{p,i} - \overline{Salinity_p})^2}} \tag{32}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Salinity_{o,i} - Salinity_{p,i})^2} \tag{33}$$

$$KGE = 1 - \sqrt{(R - 1)^2 + (StD_p/StD_o - 1)^2 + (\overline{Salinity_p}/\overline{Salinity_o} - 1)^2} \tag{34}$$

$$I_A = 1 - \frac{\sum_{i=1}^N (Salinity_{o,i} - Salinity_{p,i})^2}{\sum_{i=1}^N (|Salinity_{o,i} - \overline{Salinity_o}| + |Salinity_{p,i} - \overline{Salinity_p}|)^2} \tag{35}$$

Scenario	Variables	No
C 1	Cl ⁻ (mg/l), Mg ²⁺ (mg/l), Na ⁺ (mg/l)	3
C 2	PH, Ca ²⁺ (mg/l), Cl ⁻ (mg/l), Mg ²⁺ (mg/l), Na ⁺ (mg/l), K ⁺ (mg/l)	6
C 3	Depth (m), PH, Ca ²⁺ (mg/l), Cl ⁻ (mg/l), Mg ²⁺ (mg/l), Na ⁺ (mg/l), K ⁺ (mg/l)	7
C 4	Depth (m), PH, Ca ²⁺ (mg/l), Cl ⁻ (mg/l), Mg ²⁺ (mg/l), Na ⁺ (mg/l), K ⁺ (mg/l), HCO ₃ ⁻ (mg/l)	8

Table 2. Selected scenarios of candidate input components for modeling the salinity of groundwater obtained via Burota-random forest feature selection.

Model/algorithm	Setting-tuning	Value
ANFIS	Epoch number	220
	Step Size Decrease	0.9
	Initial Step Size	0.01
	Step Size Increase	1.1
	Cluster number	3–5
SMA	Iteration	500
	Population	20
	z	0.03
ACOR	Iteration	500
	Population	20
	Intensification Factor	0.5
	Deviation-Distance Ratio	1
OA	Iteration	500
	Population	20
	α	0.5
	δ	0.5
	ω	0.005
	β	1.5
Lasso-Reg	Γ	0–1

Table 3. Optimal setting parameters owing to the standalone ANFIS model, optimization algorithm, and Lasso-Reg model.

$$RAE = \frac{\sum_{i=1}^N |Salinity_{o,i} - Salinity_{p,i}|}{\sum_{i=1}^N |Salinity_{o,i} - Salinity_{p,i}|} \quad (36)$$

$$U_{95\%} = 1.96 \sqrt{SD_e^2 + RMSE^2} \quad (37)$$

$$E_{LM} = 1 - \frac{\sum_{i=1}^N (Salinity_{o,i} - Salinity_{p,i})^2}{\sum_{i=1}^N (Salinity_{o,i} - Salinity_o)^2} \quad (38)$$

where the predicted and actual values of salinity are $Salinity_{p,i}$ and $Salinity_{o,i}$, respectively. $\overline{Salinity_p}$ is the average of the predicted outcomes. $Salinity_o$ is the average value of observed salinity values. The number of samples in the training or testing stage is denoted by N . SD_e is the standard deviation of estimation error. StD_p and StD_o are the standard deviation of predicted and observed values, respectively. It should be emphasized that a model with $R = 1$, $E_{LM} = 1$, $RMSE = 0$, $RAE = 0$, $KE = 1$, $I_A = 1$ and $U_{95\%} = 0$ is a great model.

Model development

Feature selection procedure. Here, the salinity of multi-aquifers in coastal regions of Bangladesh was modeled based on ten parameters, as reported in Table 1. As stated in literature, optimal input feature selection is one of the most important steps in developing an efficient predictive model with numerous features. On the other hand, linear regression-based methods such as correlation analysis and the best subset approaches may not correctly capture non-linear interactions between input and target parameters. Therefore, adopting an efficient strategy is inevitable. Recently, various strategies have been proposed that are appropriately able to detect non-linear aspects between data well. In the current research, the Boruta-Random forest feature selection¹⁰², was

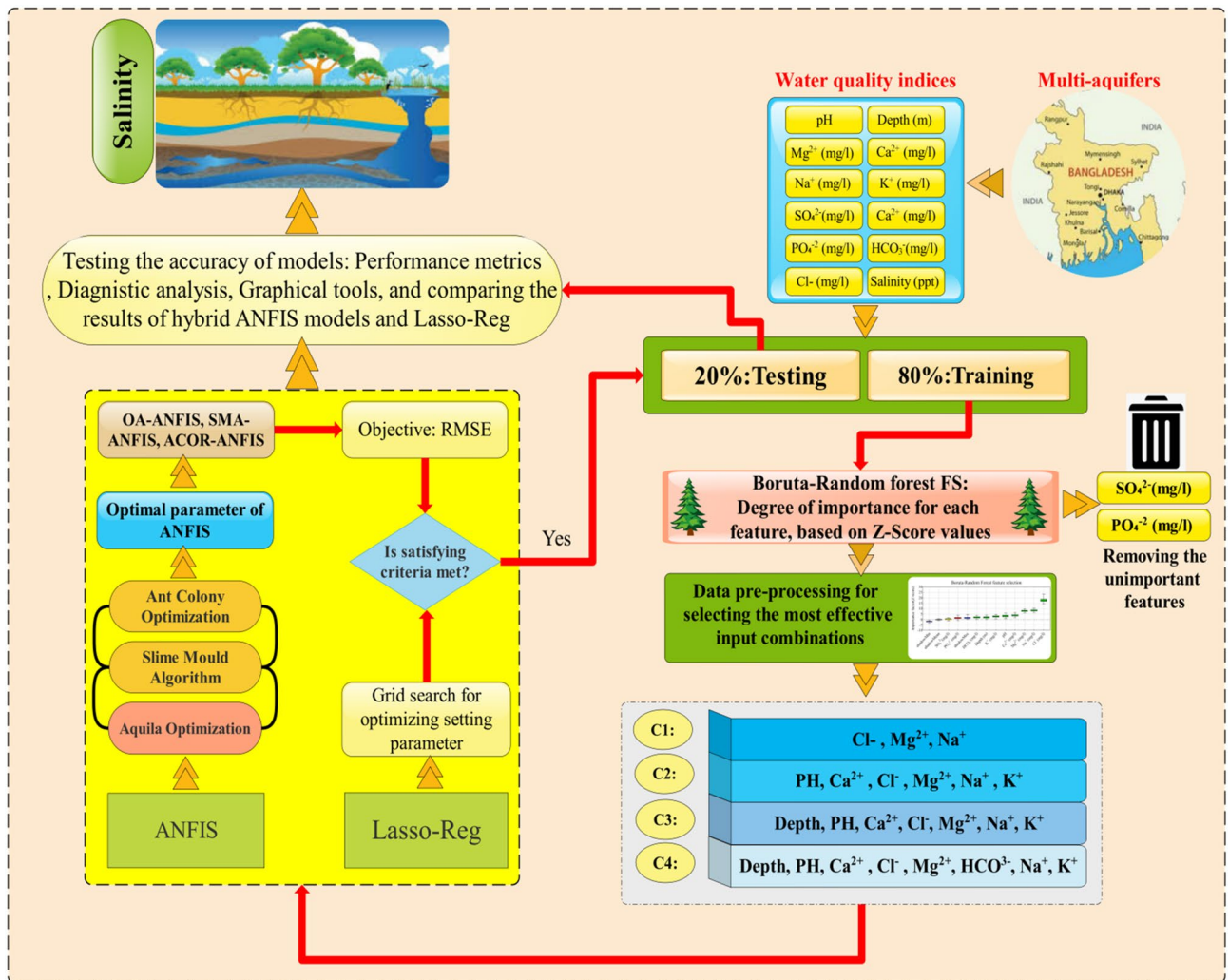


Figure 6. Workflow of salinity modeling based on a new ML-based hybrid strategy for multi-aquifers in coastal regions of Bangladesh.

employed as a tree-based powerful feature selection to optimize the input combination and assess the critical degree of each input feature. The outcomes of the Boruta-Random forest feature selection based on the Z-score criterion are illustrated in Fig. 5, which implies the importance of each feature versus salinity. It can be included that Cl⁻, Na⁺, and Mg²⁺ features have a greater impact on modeling the salinity than the other parameter. Thus, those features were employed in all candidate input combinations. Also, the PH, K⁺, Depth, and HCO₃⁻ were stood in the next rank and sequentially added to the significant features as the other candidate input combinations. In addition, the PO₄²⁻ and SO₄²⁻ regarding fewer Z-scores than the shadow max criterion were ignored to simulate the salinity of the multi-aquifers. The optimal candidate input combinations for more assessment via the predictive models were reported in Table 2.

model providing an optimal setting-parameter. Here, five efficient data-intelligent systems, including standalone ANFIS, OA-ANFIS, SMA-ANFIS, ACO-ANFIS, and Lasso-Reg, were examined to estimate the salinity of multi-aquifer in coastal regions of Bangladesh. The ANFIS and hybrid-ANFIS model were developed via Matlab 2019a, and the Lasso-Reg approach was adopted in Python platform 3.8 through the open-source sci-kit-learn library¹⁰³, on a system with Intel Core (TM)i7-6700 CPU with 3.40 GHz. In order to optimize the most significant setting parameter of Lasso-Reg (i.e., Γ), the great search strategy was adopted in the range of $\Gamma \in (0, 1)$. Besides, the standalone ANFIS approach was optimized based on a trial and error procedure by examining the cluster number in the range of (3–5). In the hybrid ANFIS model, membership functions of ANFIS were optimized using three algorithms (e.i., OA, SMA, and ACO) in popuse of the computational cost reduction and accuracy enhancement. Table 3 lists the optimum ANFIS setting parameters, critical values of the algorithm, and the Lasso-Reg tuning parameter. In this research, first, whole datasets (539 data points) are randomly divided into two subsets of training (80%) and testing (20%). Also, to avoid overfitting, the Monte Carlo approach has been employed with 500 runs, and the average results obtained from the Monte Carlo method have been considered. The workflow of salinity estimating the multi-aquifer in coastal regions of Bangladesh is demonstrated in Fig. 6.

Models	Phase	R	RMSE	RAE	E_{LM}	KGE	I_A	$U_{95\%}$
C1								
ANFIS	Training	0.9596	1.0535	0.1646	0.9208	0.9428	0.9791	2.9218
	Testing	0.9306	1.2139	0.2131	0.8639	0.9222	0.9643	3.3699
ANFIS-OA	Training	0.9559	1.1183	0.1835	0.9108	0.9032	0.9760	3.0775
	Testing	0.9450	1.1253	0.1878	0.8871	0.9146	0.9696	3.0632
ANFIS-SMA	Training	0.9283	1.4083	0.2105	0.8585	0.9260	0.9632	3.9054
	Testing	0.9406	1.1534	0.1896	0.8830	0.8793	0.9690	3.1632
ANFIS-ACO	Training	0.9121	1.5501	0.2484	0.8286	0.8317	0.9488	4.2946
	Testing	0.9402	1.1688	0.2205	0.8802	0.8653	0.9656	3.1822
Lasso-Reg	Training	0.9013	1.6215	0.2348	0.8124	0.8602	0.9461	4.4971
	Testing	0.9358	1.1863	0.2091	0.8700	0.8552	0.9628	3.2766
C2								
ANFIS	Training	0.9761	0.8140	0.1529	0.9527	0.9662	0.9878	2.2577
	Testing	0.9090	1.3857	0.2238	0.8226	0.8965	0.9525	3.8451
ANFIS-OA	Training	0.9300	1.3851	0.2000	0.8631	0.9208	0.9639	3.8397
	Testing	0.9374	1.1520	0.2091	0.8774	0.9100	0.9671	3.1931
ANFIS-SMA	Training	0.9292	1.4446	0.2531	0.8511	0.9119	0.9632	4.0001
	Testing	0.9307	1.2801	0.2436	0.8486	0.8790	0.9626	3.5099
ANFIS-ACO	Training	0.9153	1.5079	0.2538	0.8378	0.8802	0.9544	4.1822
	Testing	0.9322	1.1930	0.2276	0.8685	0.9022	0.9642	3.3110
Lasso-Reg	Training	0.9149	1.5110	0.2477	0.8371	0.8776	0.9541	4.1907
	Testing	0.9331	1.1874	0.2232	0.8697	0.8982	0.9644	3.2936
C3								
ANFIS	Training	0.9740	0.8479	0.1604	0.9487	0.9633	0.9867	2.3516
	Testing	0.9205	1.3513	0.2412	0.8313	0.9073	0.9586	3.7540
ANFIS-OA	Training	0.9491	1.1871	0.2035	0.8994	0.9144	0.9730	3.2817
	Testing	0.9378	1.1685	0.2083	0.8738	0.8849	0.9662	3.2104
ANFIS-SMA	Training	0.9002	1.6635	0.2433	0.8025	0.8861	0.9476	4.6038
	Testing	0.9314	1.2634	0.2226	0.8525	0.8616	0.9619	3.4353
ANFIS-ACO	Training	0.9161	1.5012	0.2543	0.8392	0.8797	0.9548	4.1635
	Testing	0.9312	1.2025	0.2317	0.8664	0.8995	0.9635	3.3366
Lasso-Reg	Training	0.9154	1.5070	0.2493	0.8380	0.8782	0.9543	4.1796
	Testing	0.9327	1.1907	0.2245	0.8690	0.8977	0.9641	3.3031
C4								
ANFIS	Training	0.9482	1.1888	0.2145	0.8992	0.9268	0.9729	3.2971
	Testing	0.9172	1.3274	0.2513	0.8372	0.9068	0.9571	3.6828
ANFIS-OA	Training	0.9392	1.2878	0.2227	0.8817	0.9202	0.9683	3.5697
	Testing	0.9402	1.1489	0.2073	0.8780	0.9121	0.9687	3.1689
ANFIS-SMA	Training	0.9168	1.4958	0.2454	0.8404	0.8735	0.9547	4.1479
	Testing	0.9335	1.1882	0.2179	0.8696	0.8887	0.9641	3.2897
ANFIS-ACO	Training	0.9226	1.4447	0.2557	0.8511	0.8901	0.9586	4.0064
	Testing	0.9367	1.1602	0.2326	0.8756	0.8851	0.9653	3.2158
Lasso-Reg	Training	0.9165	1.4977	0.2403	0.8399	0.8796	0.9550	4.1538
	Testing	0.9344	1.1781	0.2140	0.8718	0.8966	0.9650	3.2649

Table 4. The goodness of fit of the predictive model for estimation of the salinity of groundwater. Significant values are in bold.

Application results and analysis

The selection of the optimal input variables in modeling complex hydrological processes is very crucial. The present study utilized the Boruta-Random forest (B-RF) feature selection technique to nominate the most significant variables which affect the output and cast to develop the efficient data-intelligent models (standalone & hybrid). After the application of B-RF, four scenarios, i.e., C1 with 3 inputs, C2 with 6 inputs, C3 with 7 inputs, and C4 with 8 inputs (Table 2), were built to estimate groundwater's salinity in Bangladesh through standalone & hybrid machine learning (ML) models.

The results of the standalone and hybrid ML/data-intelligent models (i.e., ANFIS, ANFIS-OA, ANFIS-SMA, ANFIS-ACO, and Lasso-Reg) under C1 to C4 scenarios were evaluated based on seven statistical indicators including R, RMSE, RAE, E_{LM} , KGE, I_A , and $U_{95\%}$. Table 4 summaries the values of R, RMSE, RAE, E_{LM} , KGE,

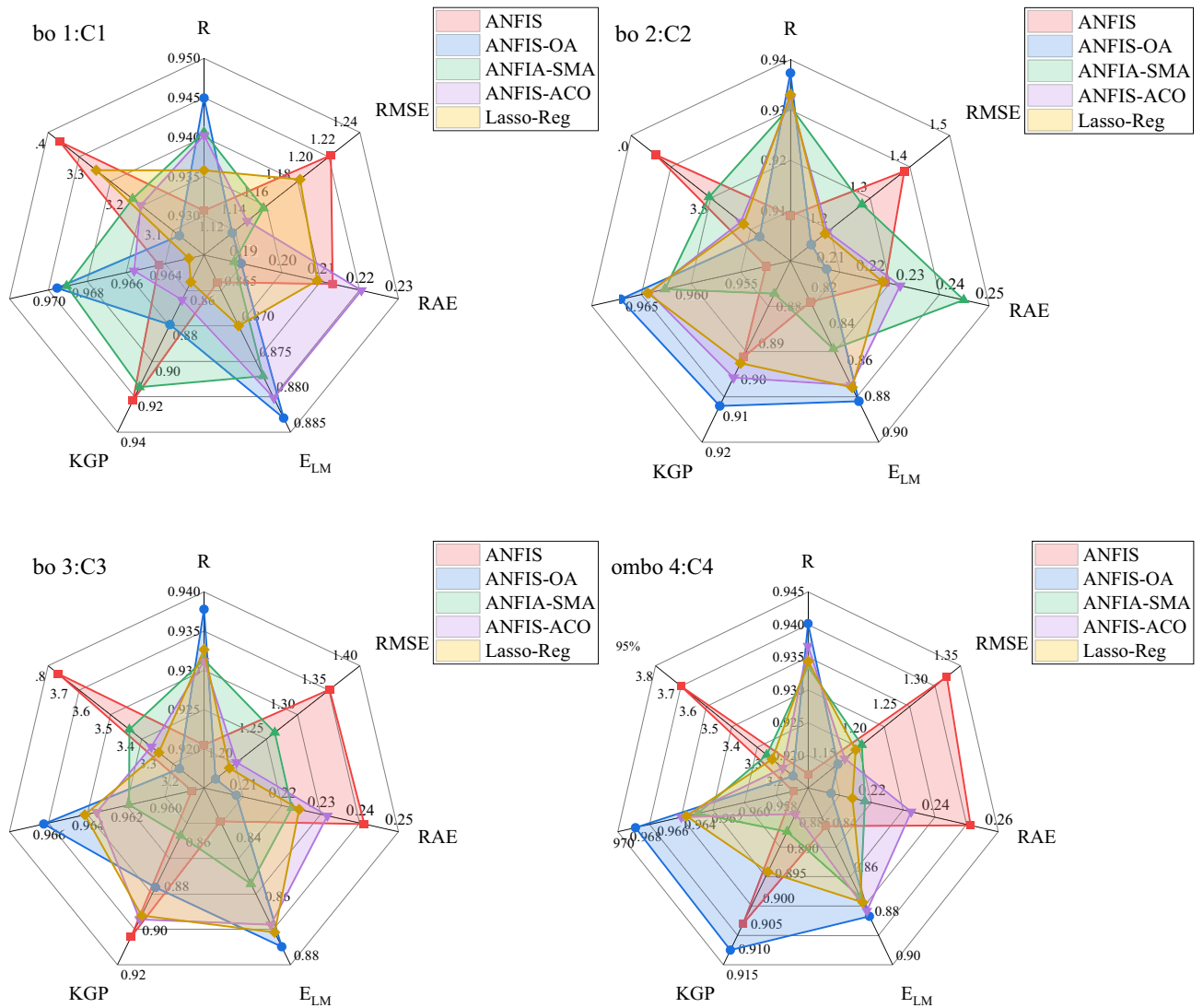


Figure 7. the presentation of spider plot for the developed models and input combinations.

I_A , and $U_{95\%}$ during training and testing stages of various ML models under different scenarios. It was observed from Table 4 the standalone ANFIS model had poor performance among the other ML models based on statistical indicators in all four scenarios. The hybrids of ANFIS and Lasso-Reg models had improved performance over the ANFIS model in C1 to C4 scenarios. However, the superior performance of hybrid ANFIS-OA model was noted in terms of statistical indicators. The hybrid ANFIS-OA model had values of $R = 0.9450$, $RMSE = 1.1253$ ppt, $RAE = 0.1878$, $E_{LM} = 0.8871$, $KGE = 0.9146$, $I_A = 0.9696$ and $U_{95\%} = 3.0632$ in scenario C1, $R = 0.9374$, $RMSE = 1.1520$ ppt, $RAE = 0.2091$, $E_{LM} = 0.8774$, $KGE = 0.9100$, $I_A = 0.9671$, and $U_{95\%} = 3.1931$ in scenario C2, $R = 0.9378$, $RMSE = 1.1685$ ppt, $RAE = 0.2083$, $E_{LM} = 0.8738$, $KGE = 0.8849$, $I_A = 0.9662$, and $U_{95\%} = 3.2104$ in scenario C3, and $R = 0.9402$, $RMSE = 1.1489$ ppt, $RAE = 0.2073$, $E_{LM} = 0.8780$, $KGE = 0.9121$, $I_A = 0.9687$, and $U_{95\%} = 3.1689$ in scenario C4 during testing stage. The results indicate the significant improvement in the performance of ANFIS model while optimizing with OA algorithm than other algorithms. The comparison of employed ML models' outcomes among the different scenarios are marked as $C1 > C4 > C2, C3$. Additionally, Habibi et al.¹⁰⁴ proved the potential of hybrid ML model i.e., ANN-GA (artificial neural network-genetic algorithm) against the ANN, PLSR (partial least square regression), and DT (decision tree) models for soil salinity prediction in central Iran. Pouladi et al.¹⁰⁵ predicted the soil salinity in Miandoab city of Iran by employing the MLP-FFA (multilayer perceptron-firefly algorithm) model using remote sensing and topography data. The outcomes of MLP-FFA model were compared with MLP model based on several statistical indices. Evaluation of results show that the MLP-FFA model had highest value of determination coefficient ($R^2 = 0.66$), and lowest values of mean absolute error ($MAE = 0.45$ dS m^{-1}), and $RMSE = 0.54$ dS m^{-1} than standalone MLP ($R^2 = 0.34$, $MAE = 0.54$ dS m^{-1} , $RMSE = 0.67$ dS m^{-1}) model.

The results were also appraised using the spider chart in Fig. 7 to compare the performance of ML models in terms of R , $RMSE$, RAE , E_{LM} , KGE , I_A , and $U_{95\%}$. These figures also clearly demonstrate the superiority of the hybrid ANFIS-OA model over the other ML models. Thus, the outcomes of the applied ML models in C1, C2, C3, and C4 scenarios are ranked in the following order, i.e., ANFIS-OA > ANFIS-SMA > ANFIS-ACO > Lasso-Reg > ANFIS.

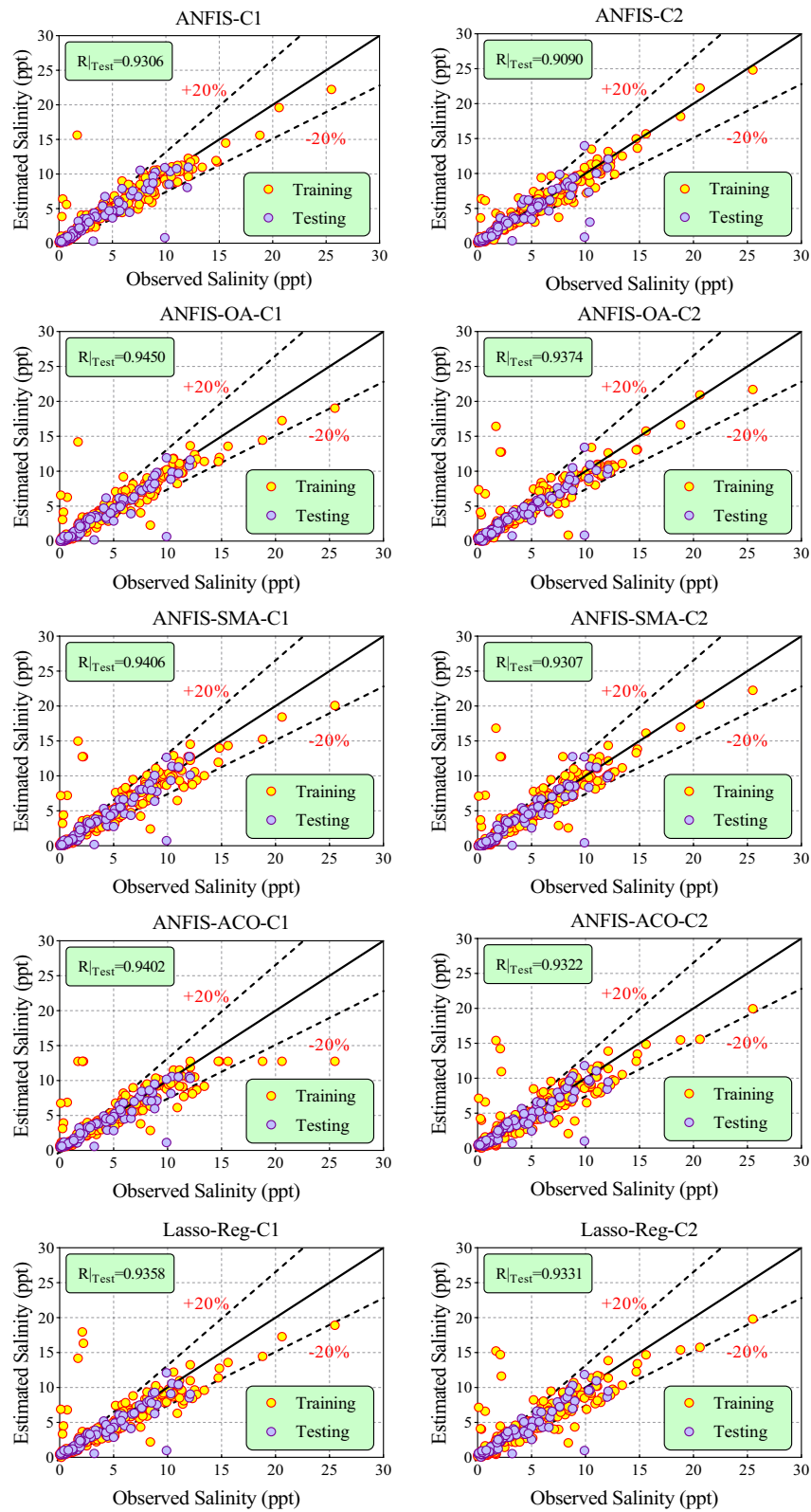


Figure 8. Scatter plot of observational and predicted values of salinity of groundwater using the provided models in all the scenarios.

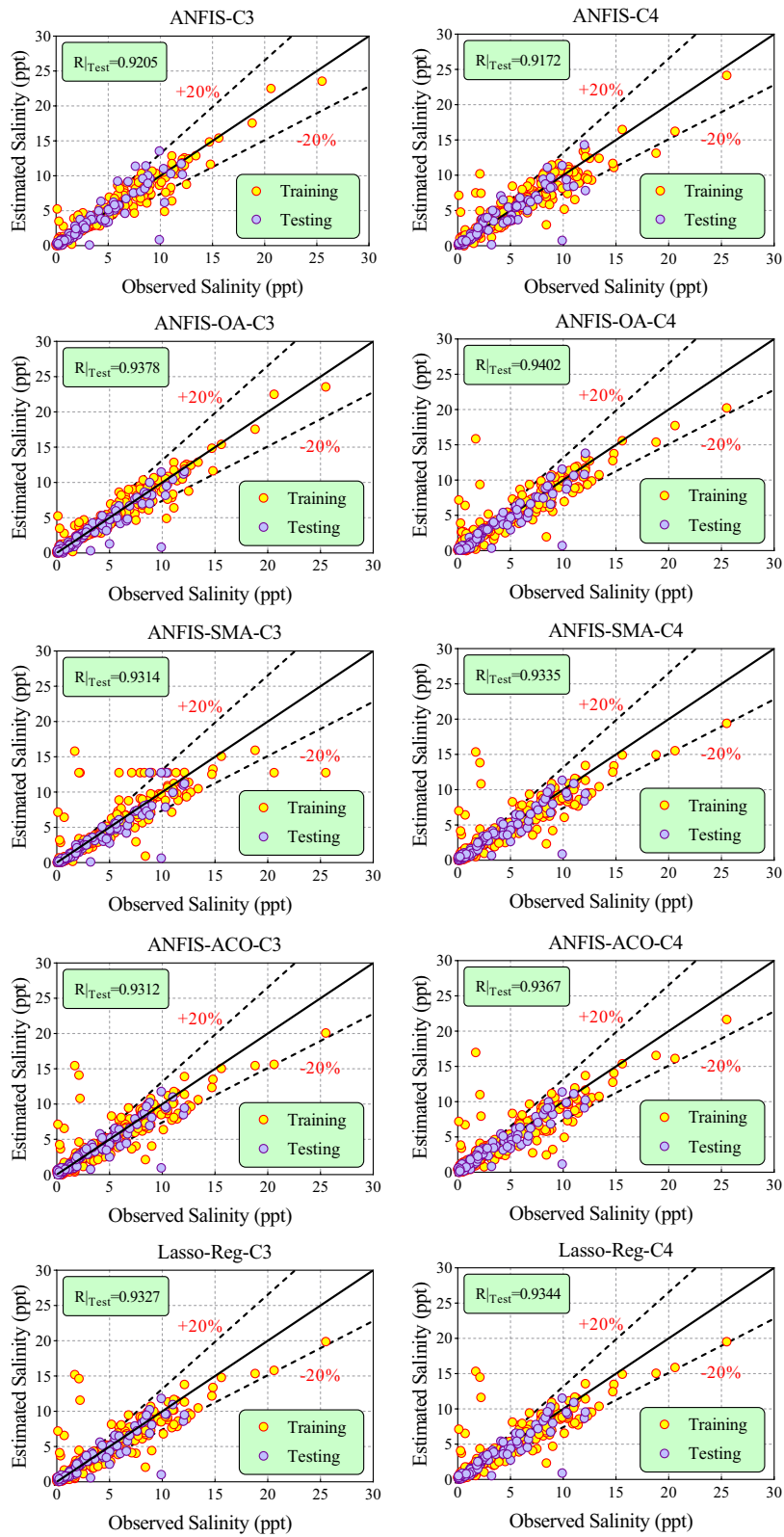


Figure 9. Continue; scatter plot of observational and predicted values of salinity of groundwater using the provided models in all the scenarios.

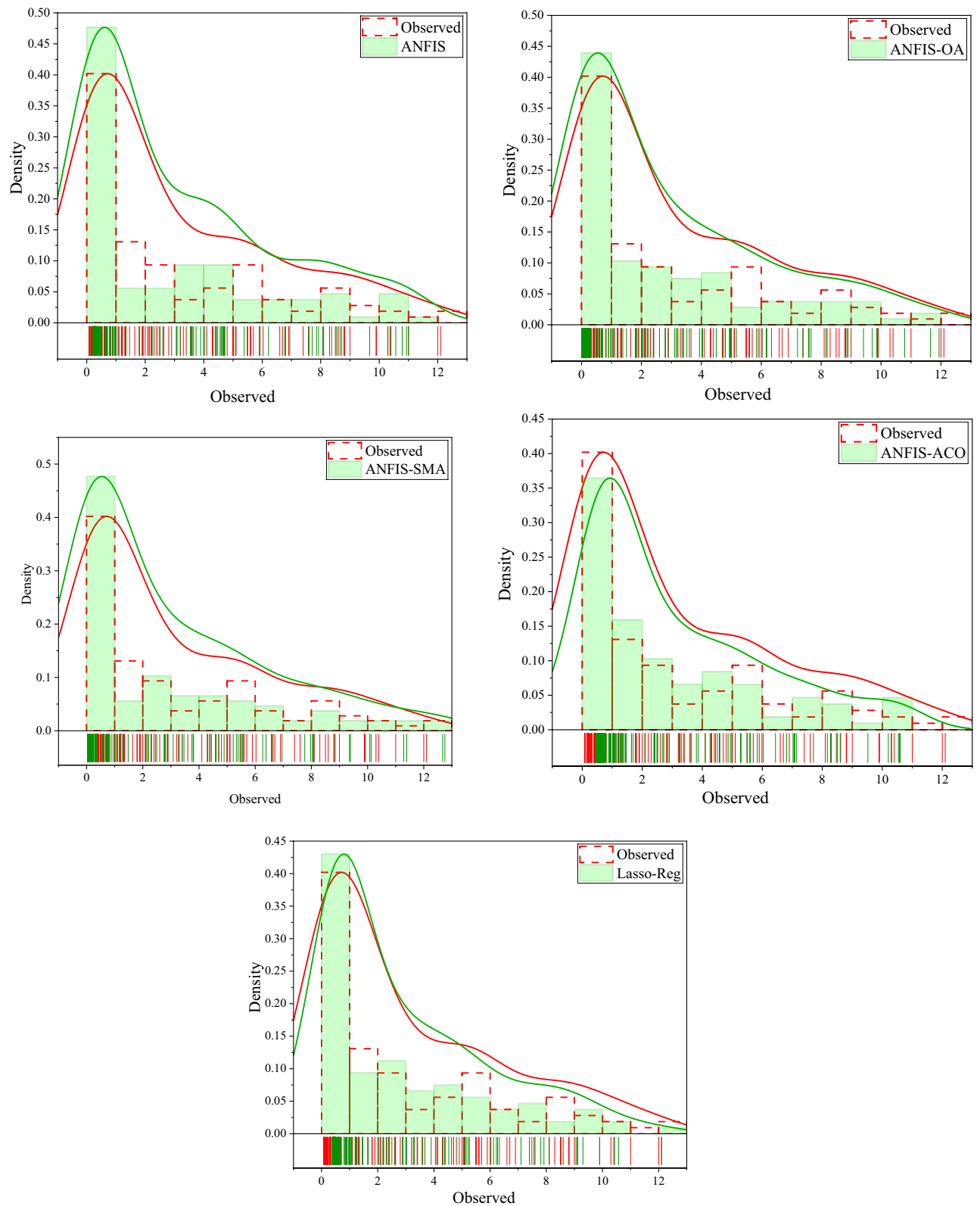


Figure 10. Control Rug and density distribution of Histogram Plot for all the provided models in the optimal scenario in the testing phase for predicting the salinity of groundwater.

Figures 8 and 9 illustrate the scatter plots of observed versus predicted values of groundwater salinity by the ANFIS, ANFIS-OA, ANFIS-SMA, ANFIS-ACO, and Lasso-Reg models to C1, C2, C3, and C4 scenarios

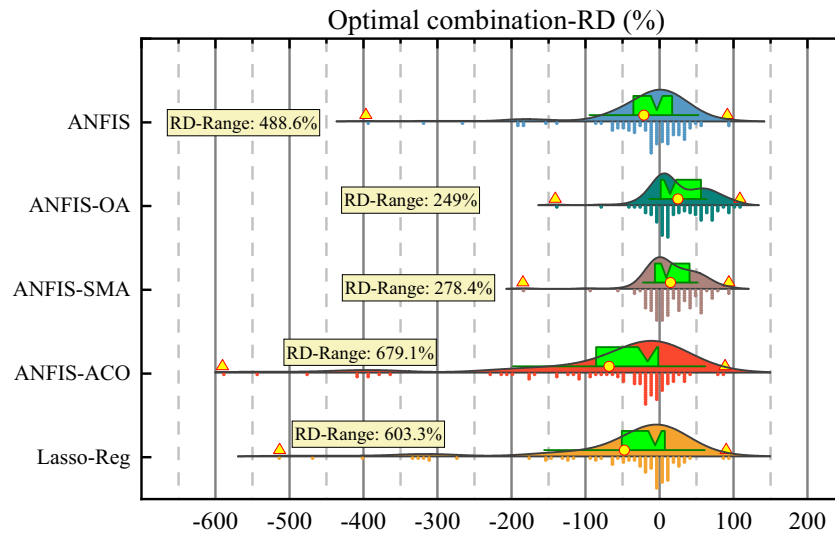


Figure 11. Relative devotion to the predictive model in the optimal scenario for diagnostic analysis for the testing stage.

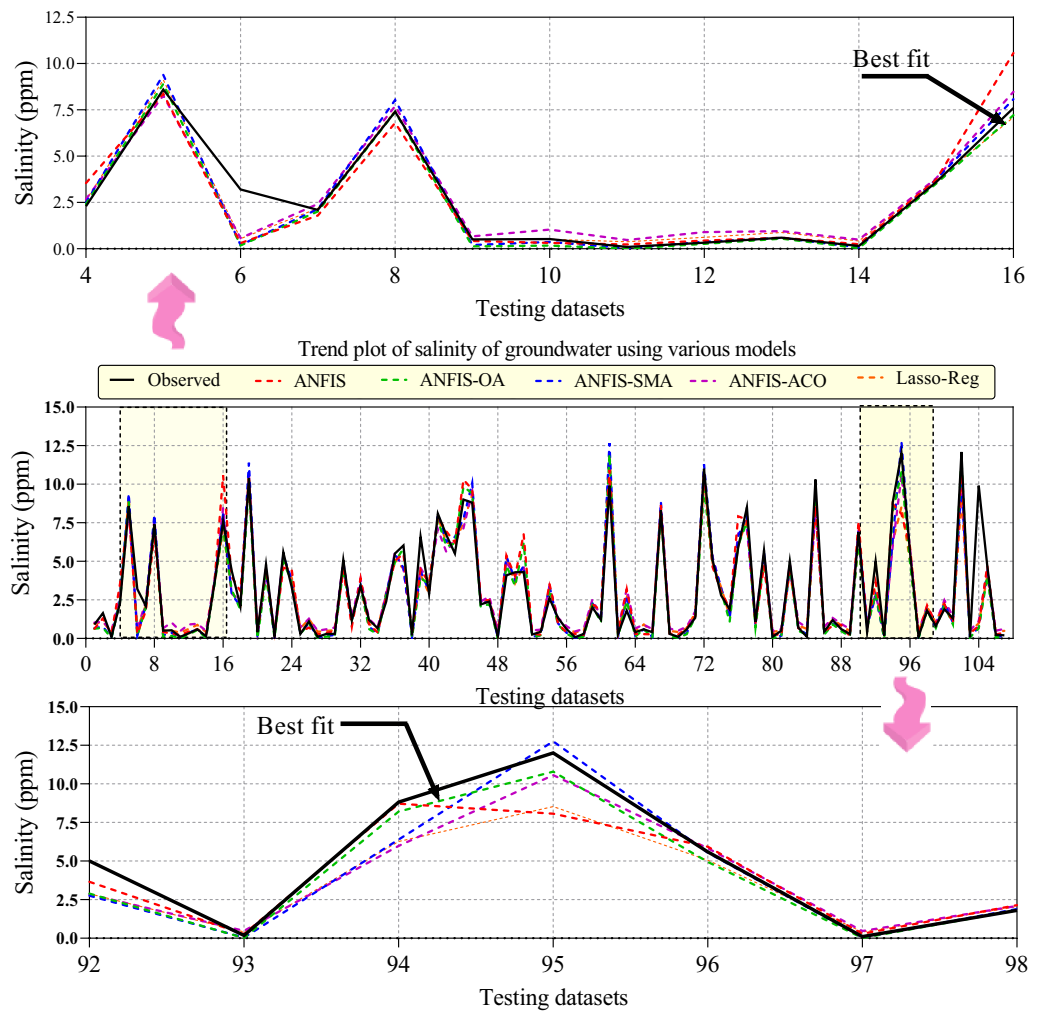


Figure 12. comparison between the observational and values of salinity of groundwater via the provided models in the optimal scenario.

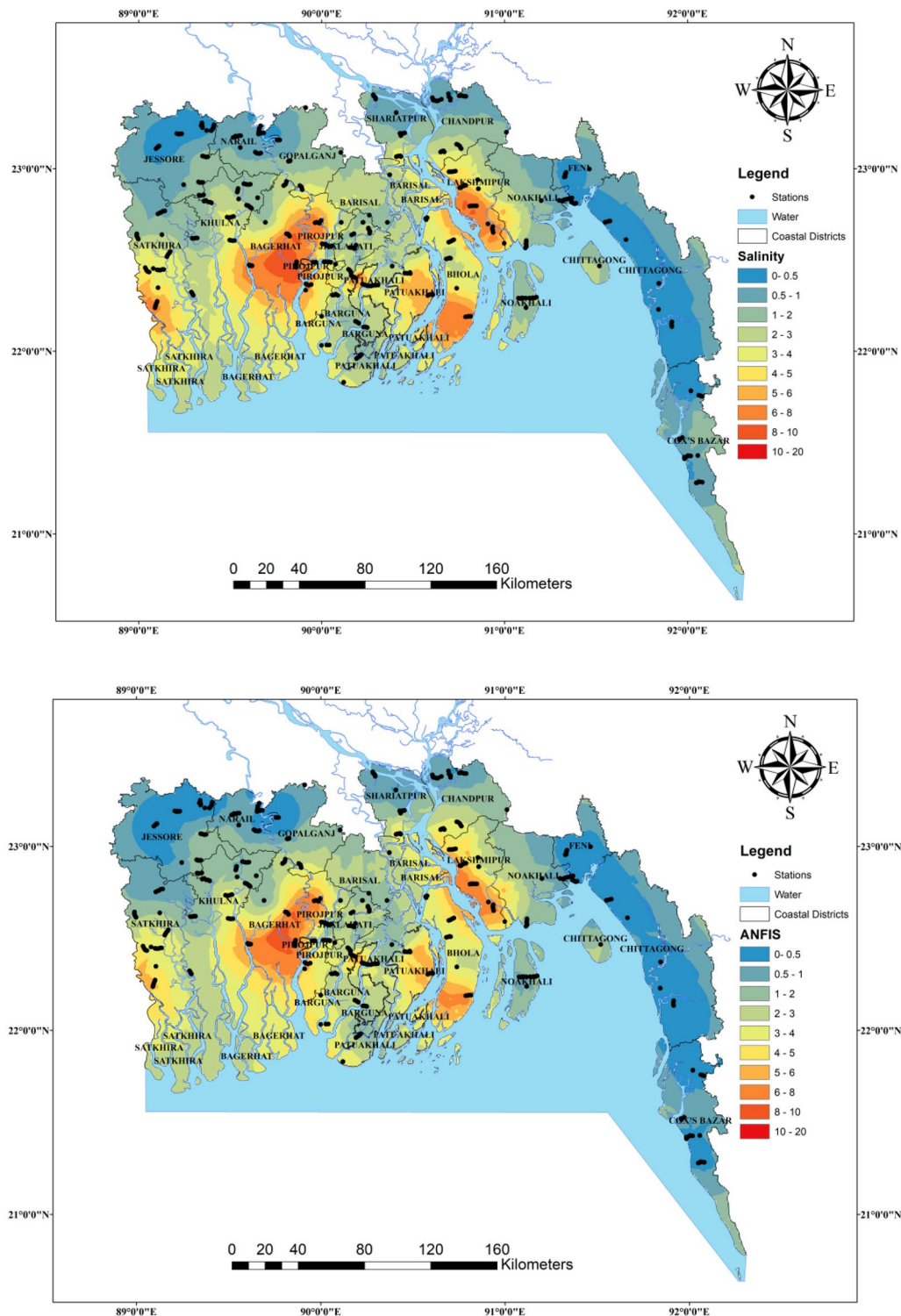


Figure 13. Salinity susceptibility maps for the groundwater samples collected from the coastal regions of Bangladesh.

during training and testing phases. The outputs yielded by the ML models were plotted about the 45° (1:1 line or best fit line)—solid black line) along with relative error bands of ± 20% (dash-dot lines). Another metric i.e., R (coefficient of correlation) between observed versus predicted values of groundwater salinity in testing, was displayed to assess the effectiveness of the ML models. These figures show that the ANFIS model optimized with the OA algorithm has predicted groundwater salinity values close to the observed values or most of the predicted data centered towards the 1:1 line within ± 20% error bands in all scenarios. Furthermore, the highest value

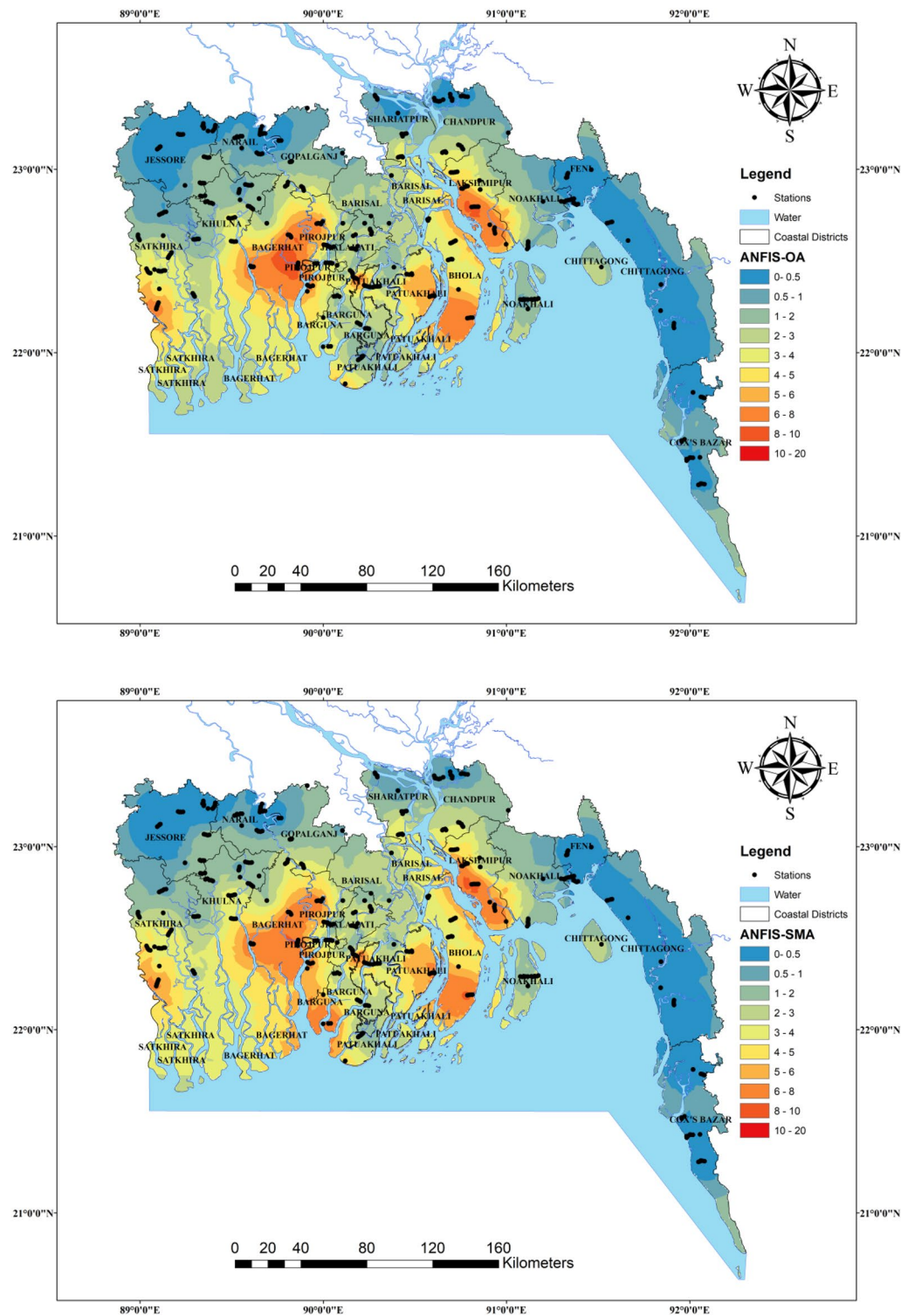


Figure 13. (continued)

of $R = 0.9450, 0.9374, 0.9378,$ and 0.9402 was gained by the hybrid ANFIS-OA model than other ML models in C1, C2, C3, & C4 scenarios during the testing stage. Comprehensively, the ANFIS model tuned with an AO nature-inspired algorithm can be considered a robust and reliable model for predicting groundwater salinity in the study region.

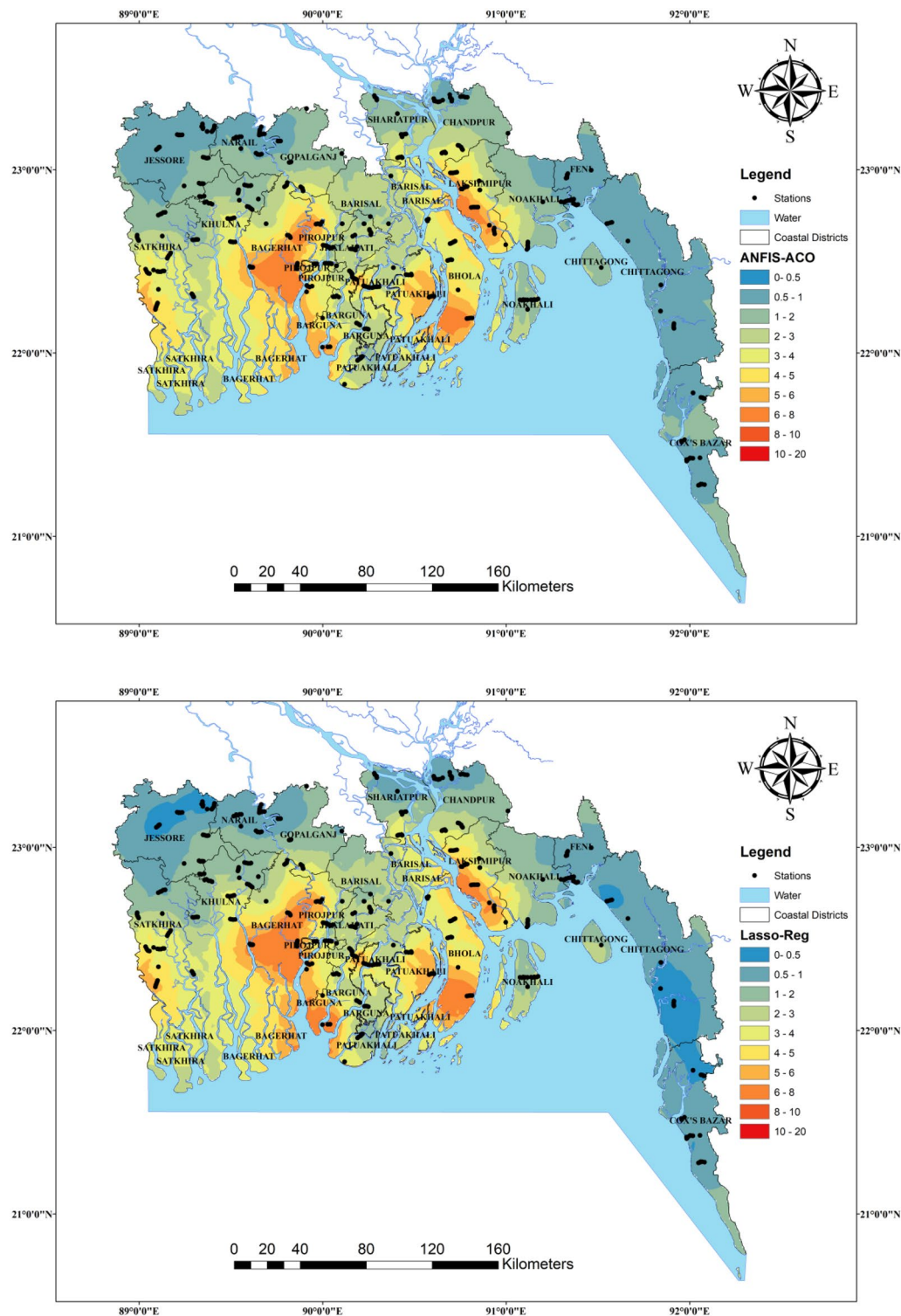


Figure 13. (continued)

Discussion

Accurate monitoring and prediction of soil salinity are essential for sustainable development, land management, water quality, and agricultural activities, especially in arid and semi-arid regions¹⁰⁶. Therefore, other criteria to examine the performance of applied hybrid and standalone ML models under different scenarios for predicting groundwater salinity are the control rug and density distribution. As mentioned in Table 4, scenario C1 was considered optimal for groundwater salinity prediction in the study area according to the comparison results. So, Fig. 10 demonstrates the control rug and density distribution histogram of ANFIS, ANFIS-OA, ANFIS-SMA,

ANFIS-ACO, and Lasso-Reg models corresponding to C1 in the testing stage. These figures also confirm the supremacy of the OA algorithm over the others in groundwater salinity prediction.

Similarly, Fig. 11 shows the relative deviation (RD) of predicted groundwater salinity values by ANFIS, ANFIS-OA, ANFIS-SMA, ANFIS-ACO, and Lasso-Reg models for the C1 scenario during the testing period regarding the observed values. The RD was minimum in for the ANFIS-OA (249%) model than ANFIS (488.6%), ANFIS-SMA (278.4%), ANFIS-ACO (679.1%), and Lasso-Reg (603.3%) models. Because of RD%, the ML models attain ANFIS-OA > ANFIS-SMA > ANFIS > Lasso-Reg > ANFIS-ACO pattern. This RD analysis also supports the OA algorithm's effectiveness in optimizing the ANFIS model's performance in groundwater salinity prediction in the study region.

Figure 12 displays the temporal variation of groundwater salinity predicted by the ANFIS, ANFIS-OA, ANFIS-SMA, ANFIS-ACO, and Lasso-Reg models corresponding to the C1 optimal scenario in the testing stage. The predicted values of the salinity of groundwater are distributed or plotted concerning the observed values of groundwater salinity (solid black line). The outcomes of the hybrid ANFIS-OA model (dash-dot green line) have much better matching with experimental groundwater salinity values and designate the dominance of the ANFIS-OA model.

Figure 13 shows the spatial pattern of groundwater salinity yielded by ANFIS, ANFIS-OA, ANFIS-SMA, ANFIS-ACO, and Lasso-Reg models under the C1 scenario in the testing phase concerning the observed values. The spatial interpolation was done by combining the training and testing datasets using IDW (inverse distance weighted) method. According to these spatial maps, the difference between the observed (left top figure) and the ANFIS, ANFIS-SMA, ANFIS-ACO, and Lasso-Reg models is high but much similar smooth pattern ANFIS-OA model estimates. Thus, the ANFIS with OA algorithm can distinguish the different ranges of groundwater salinity from the other ML models. Wu et al.¹⁰⁷ predicted and mapped the spatial distribution of soil salinity in central Mesopotamia of Iraq by employing the support vector machine (SVM) and random forest regression (RFR) algorithms. They found that the RFR model provided better estimates than the SVM model and stated that the spatial map of soil salinity prepared by the RFR algorithm outcomes helps maintain the agricultural activities and sustainable development in the study area. In addition, the spatial distribution maps of the present study will help understand the vulnerability of salinity in groundwater on a regional scale and adopt preventive measures according to the level of salinity of groundwater.

Furthermore, to make the current research more concur and impactful, results were compared with the existed studies on soil salinity prediction using ML models in a different part of the world^{108–110}. Wang et al.¹¹¹ predicted soil salinity in three regions, i.e., Qitai, Kuqa, and Yuta oases of China, by using five ML algorithms, including MARS (multiple adaptive regression splines), CART (classification and regression trees), RF (random forest tree ensembles), SGT (stochastic gradient tree boost), and LASSO (least absolute shrinkage and selection operator). According to the results of the comparison, the SGT algorithm was found most suitable for predicting the soil salinity in three distinct regions. Ma et al.¹⁰⁶ applied XGBoost (extreme gradient boosting), CART, and RF models to predict the soil salinity in the Ogan-Kuqa river oasis in China using remote sensing and topographical observations. They found that the XGBoost model achieved better prediction ($R^2 = 0.68$, $RMSE = 10.56 \text{ dS m}^{-1}$) than CART ($R^2 = 0.57$, $RMSE = 12.20 \text{ dS m}^{-1}$), and RF ($R^2 = 0.63$, $RMSE = 11.41 \text{ dS m}^{-1}$) models. Wang et al.¹¹² employed three ML algorithms, i.e., SVM, ANN, and RF, to predict the soil salinity in China's KPNR (Kongterik Pasture Nature Reserve).

Results show the better performance of SVM ($R^2 = 0.88$, $RMSE = 4.89 \text{ dS m}^{-1}$) model than RF ($R^2 = 0.27$, $RMSE = 10.61 \text{ dS m}^{-1}$) and ANN ($R^2 = 0.57$, $RMSE = 8.15 \text{ dS m}^{-1}$) models. The literature above also established the supremacy of ML algorithms for predicting the soil salinity in different environmental conditions and strongly supports the outcomes of the present study, and endorses the application of hybrid ML models to predict the salinity of groundwater in the study region. The occurrence of major seawater intrusion in the multi-aquifers in the southwest coastal zone was indicated by extremely high salinity levels. Climate change effects such as sea-level rise, storm surges, and water logging, to the best of our knowledge, also increase salt intrusion (Na-Cl type water) along the coast. As a result, salinity-induced water is projected to move further inland, increasing contamination intensity¹¹³. The lack of surface freshwater supplies, such as downstream river flow, long dry periods, shrimp farming, and rainfall uncertainty, may cause changes in the coastal hydrogeologic environment, causing instability in groundwater recharge, storage, and flow. Our study presents a novel technique to aid water managers and decision-makers in safeguarding groundwater resources against saline water intrusion. Constructing accurate salinity susceptibility maps can lead to increased groundwater resource management strategies and environmental sustainability. Comparing the current study's results with the previous similar regional investigation demonstrates that the ANFIS-AO regarding higher accuracy ($R = 0.945$) resulted in the promising outcomes than the Catboost model ($R = 0.916$) for predicting the groundwater salinity of multi-layer aquifers of Mekong Delta, Vietnam¹¹⁴ and extreme gradient boosting (EGB) model ($R = 0.943$) for estimating the groundwater salinity in the southern coastal aquifer of the Caspian Sea, Iran³⁸.

Conclusion and future direction

In this research, a novel Nature-inspired ANFIS model (i.e., ANFIS-AO) along with ANFIS-SMA, ANFIS-ACOR, individual ANFIS, and Lasso-Reg coastal multi-aquifers in some regions of Bangladesh based on ten water quality indices promised of depth, pH, Ca^{2+} (mg/l), Mg^{2+} (mg/l), Na^+ (mg/l), K^+ (mg/l), HCO_3^- (mg/l), SO_4^{2-} (mg/l), PO_4^{2-} (mg/l), Cl^- (mg/l). In the pre-processing stage, the training dataset was explored using the B-RF feature selection, indicating each feature's importance degree in modeling salinity. The outcomes of pre-processing ascertained that SO_4^{2-} (mg/l) and were neglected as the candidate inputs and Cl^- (mg/l), Na^+ , and Mg^{2+} (mg/l) were indicated as the most significant features. Based on the mentioned feature selecting process, four input combinations were examined to assess the compatibility of the predictive ML-based models. A careful

review of the statistical criteria and graphical analysis of the employed ML models shows that the best results are related to the C1, C4, C2, and C3, respectively ($C1 > C4 > C2, C3$). The hybrid ANFIS-OA approach regarding the most promising metrics ($R = 0.9450$, $RMSE = 1.1253$ ppt, $KGE = 0.9146$, and $U_{95\%} = 3.0632$) in the testing phase was superior to the ANFIS-SMA ($R = 0.9406$, $RMSE = 1.1534$ ppm, $KGE = 0.8793$, and $U_{95\%} = 3.1632$), ANFIS-ACOR ($R = 0.9402$, $RMSE = 1.1388$ ppm, $KGE = 0.8653$, and $U_{95\%} = 3.1822$), Lasso-Reg ($R = 0.9358$, $RMSE = 1.1863$ ppm, and $KGE = 0.8552$), and ANFIS ($R = 0.9306$, $RMSE = 1.2139$ ppm, and $KGE = 0.9222$) models. Besides, the diagnostic assessment of the superior candidate input combination (C1) ascertained that the ANFIS-OA concerning the least RD value (249%) attained the most reliable predicted salinity in coastal multi-aquifer followed by the ANFIS-SMA (278.4%), ANFIS (488.6%), ANFIS-ACO (679.1%), and Lasso-Reg (603.3%), respectively. Finally, a comparison of the spatial distribution of salinity in the aquifers of the coastal areas of Bangladesh shows well that the ANFIS-OA method has the best agreement with the actual salinity distribution, while the ANFIS-SMA method is in second place. It is worth noting that the IDW method was employed to interpolate the data points of each predictive model to provide the spatial pattern contours. Since the frameworks presented in this research are based on a robust pre-processing aim to identify the most influential input combinations, the degree of uncertainty of the models will be shallow. Besides, the hybrid neuro-fuzzy models with minor limitations can be used for other water quality indices even in other areas of study. We added this augmentation into the discussion section.

The current research was adopted to develop a hybrid version of ML models, and the research findings were successfully approached. Future research direction can be established on different aspects. For instance, studying the data, models, and input parameters uncertainties, investigating the seawater intrusion and its effects on the salinity concentration, and identifying the essential connection between groundwater salinity and crops/plantations' health and contamination. As another alternative future study, a robust classification method can assess the salinity risks in the entire coastal multi-aquifer system. To lessen and control the deterioration of groundwater quality in the coastal zone, it is critical to avoid leaching salinity intrusion and toxic soil contents into groundwater. The study outcomes can assist the policy-makers and respective agencies in managing and protecting the water resources in the coastal region of Bangladesh. Regionalization of salinity estimation has potential implications for rationally utilizing and developing water resource strategies and plans for reducing vulnerability. The classification methods developed as a future direction can be a possible alternative for salinity assessment in any coastal plain with similar aquifer features and hydrogeologic settings.

Data availability

The data generated or analyzed during this study are available from the corresponding author on reasonable request.

Received: 7 April 2022; Accepted: 17 June 2022

Published online: 01 July 2022

References

- Wada, Y. *et al.* Global depletion of groundwater resources. *Geophys. Res. Lett.* **37**, n/a–n/a (2010).
- Dohare, D., Deshpande, S. & Kotiya, A. Analysis of ground water quality parameters: A review. *Res. J. Eng. Sci.* **2278**, 9472 (2014).
- Gleeson, T., Wada, Y., Bierkens, M. F. P. & Van Beek, L. P. H. Water balance of global aquifers revealed by groundwater footprint. *Nature* <https://doi.org/10.1038/nature11295> (2012).
- Motevalli, A. *et al.* Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. *J. Clean. Prod.* <https://doi.org/10.1016/j.jclepro.2019.04.293> (2019).
- Awadh, S. M., Al-Mimar, H. & Yaseen, Z. M. Groundwater availability and water demand sustainability over the upper mega aquifers of Arabian Peninsula and west region of Iraq. *Environ. Dev. Sustain.* <https://doi.org/10.1007/s10668-019-00578-z> (2020).
- Vijayanand, S., Akkara, P. J., Gretel, A., John, N. & Josephine, M. Water quality analysis of ground water from locations in North Bangalore. *Curr. Trends Biotechnol. Pharm.* **15**, 437–443 (2021).
- Khatri, N., Tyagi, S., Rawtani, D., Tharmavaram, M. & Kamboj, R. D. Analysis and assessment of ground water quality in Satlasana Taluka, Mehsana district, Gujarat, India through application of water quality indices. *Groundw. Sustain. Dev.* **10**, 100321 (2020).
- Buvaneshwari, S. *et al.* Potash fertilizer promotes incipient salinization in groundwater irrigated semi-arid agriculture. *Sci. Rep.* **10**, 1–14 (2020).
- Khairy, H. & Janardhana, M. R. Hydrogeochemical features of groundwater of semi-confined coastal aquifer in Amol-Ghaemshahr plain, Mazandaran Province, Northern Iran. *Environ. Monit. Assess.* **185**, 9237–9264 (2013).
- Ferguson, G. & Gleeson, T. Vulnerability of coastal aquifers to groundwater use and climate change. *Nat. Clim. Change* **2**, 342–345 (2012).
- Ranjan, S. P., Kazama, S. & Sawamoto, M. Effects of climate and land use changes on groundwater resources in coastal aquifers. *J. Environ. Manag.* **80**, 25–35 (2006).
- Post, V. E. A., Vandenbohede, A., Werner, A. D. & Teubner, M. D. Groundwater ages in coastal aquifers. *Adv. Water Resour.* <https://doi.org/10.1016/j.advwatres.2013.03.011> (2013).
- Gholami, V., Khaleghi, M. R. & Sebghati, M. A method of groundwater quality assessment based on fuzzy network-CANFIS and geographic information system (GIS). *Appl. Water Sci.* **7**, 3633–3647 (2017).
- Khaleefa, O. & Kamel, A. H. On the evaluation of water quality index: case study of Euphrates River, Iraq. *Knowl. Based Eng. Sci.* **2**, 35–43 (2021).
- Shah, S. H. H. *et al.* Stochastic modeling of salt accumulation in the root zone due to capillary flux from brackish groundwater. *Water Resour. Res.* **47**, (2011).
- Lv, C., Ling, M., Wu, Z., Guo, X. & Cao, Q. Quantitative assessment of ecological compensation for groundwater overexploitation based on emergy theory. *Environ. Geochem. Health* **42**, 733–744 (2020).
- Li, Y. *et al.* Evolution characteristics and influence factors of deep groundwater depression cone in North China Plain, China—A case study in Cangzhou region. *J. Earth Sci.* <https://doi.org/10.1007/s12583-014-0488-5> (2014).
- Hoque, M. A., Hoque, M. M. & Ahmed, K. M. Declining groundwater level and aquifer dewatering in Dhaka metropolitan area, Bangladesh: Causes and quantification. *Hydrogeol. J.* <https://doi.org/10.1007/s10040-007-0226-5> (2007).

19. Alfarrah, N. & Walraevens, K. Groundwater overexploitation and seawater intrusion in coastal areas of arid and semi-arid regions. *Water (Switzerland)* <https://doi.org/10.3390/w10020143> (2018).
20. Huang, F., Wang, G. H., Yang, Y. Y. & Wang, C. B. Overexploitation status of groundwater and induced geological hazards in China. *Nat. Hazards* <https://doi.org/10.1007/s11069-014-1102-y> (2014).
21. Vengosh, A., Spivack, A. J., Artzi, Y. & Ayalon, A. Geochemical and boron, strontium, and oxygen isotopic constraints on the origin of the salinity in groundwater from the Mediterranean coast of Israel. *Water Resour. Res.* <https://doi.org/10.1029/1999WR900024> (1999).
22. Rajmohan, N., Masoud, M. H. Z. & Niyazi, B. A. M. Impact of evaporation on groundwater salinity in the arid coastal aquifer, Western Saudi Arabia. *CATENA* **196**, 104864 (2021).
23. Han, D. *et al.* A survey of groundwater levels and hydrogeochemistry in irrigated fields in the Karamay Agricultural Development Area, northwest China: Implications for soil and groundwater salinity resulting from surface water transfer for irrigation. *J. Hydrol.* **405**, 217–234 (2011).
24. Masciopinto, C., Liso, I. S., Caputo, M. C. & De Carlo, L. An integrated approach based on numerical modelling and geophysical survey to map groundwater salinity in fractured coastal aquifers. *Water (Switzerland)* <https://doi.org/10.3390/w9110875> (2017).
25. Samsudin, A. R., Haryono, A., Hamzah, U. & Rafek, A. G. Salinity mapping of coastal groundwater aquifers using hydrogeochemical and geophysical methods: A case study from north Kelantan, Malaysia. *Environ. Geol.* **55**, 1737–1743 (2008).
26. Brown, N. L. & Hamon, B. V. An inductive salinometer. *Deep Sea Res.* **8**, 65-IN8 (1961).
27. El Bastawesy, M., Gebremichael, E., Sultan, M., Attwa, M. & Sahour, H. Tracing Holocene channels and landforms of the Nile Delta through integration of early elevation, geophysical, and sediment core data. *Holocene* <https://doi.org/10.1177/0959683620913928> (2020).
28. Ferchichi, H., Ben Hamouda, M. F., Farhat, B. & Ben Mammou, A. Assessment of groundwater salinity using GIS and multivariate statistics in a coastal Mediterranean aquifer. *Int. J. Environ. Sci. Technol.* **15**, 2473–2492 (2018).
29. Elmahdy, S. I. & Mohamed, M. M. Relationship between geological structures and groundwater flow and groundwater salinity in Al Jaaw Plain, United Arab Emirates; Mapping and analysis by means of remote sensing and GIS. *Arab. J. Geosci.* <https://doi.org/10.1007/s12517-013-0895-4> (2014).
30. Haselbeck, V., Kordilla, J., Krause, F. & Sauter, M. Self-organizing maps for the identification of groundwater salinity sources based on hydrochemical data. *J. Hydrol.* <https://doi.org/10.1016/j.jhydrol.2019.06.053> (2019).
31. Rina, K., Singh, C. K., Datta, P. S., Singh, N. & Mukherjee, S. Geochemical modelling, ionic ratio and GIS based mapping of groundwater salinity and assessment of governing processes in Northern Gujarat, India. *Environ. Earth Sci.* **69**, 2377–2391 (2013).
32. Diaconu, D. C., Bretcan, P., Peptenatu, D., Tanislav, D. & Mailat, E. The importance of the number of points, transect location and interpolation techniques in the analysis of bathymetric measurements. *J. Hydrol.* <https://doi.org/10.1016/j.jhydrol.2018.12.070> (2019).
33. Naganna, S. R., Beyaztas, B. H., Bokde, N. & Armanuos, A. M. On the evaluation of the gradient tree boosting model for groundwater level forecasting. *Knowl. Based Eng. Sci.* **1**, 48–57 (2020).
34. Cui, F. *et al.* Boosted artificial intelligence model using improved alpha-guided grey wolf optimizer for groundwater level prediction: Comparative study and insight for federated learning technology. *J. Hydrol.* **606**, 127384 (2022).
35. Yaseen, Z. M., Sulaiman, S. O., Deo, R. C. & Chau, K.-W. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *J. Hydrol.* **569**, 387–408 (2018).
36. Yaseen, Z. M., El-shafie, A., Jaafar, O., Afan, H. A. & Sayl, K. N. Artificial intelligence based models for stream-flow forecasting: 2000–2015. *J. Hydrol.* **530**, 829–844 (2015).
37. Adnan, R. M. *et al.* Predictability performance enhancement for suspended sediment in rivers: Inspection of newly developed hybrid adaptive neuro-fuzzy system model. *Int. J. Sediment Res.* <https://doi.org/10.1016/j.ijsrc.2021.10.001> (2021).
38. Sahour, H., Gholami, V. & Vazifedan, M. A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. *J. Hydrol.* <https://doi.org/10.1016/j.jhydrol.2020.125321> (2020).
39. Nordin, N. F. C. *et al.* Groundwater quality forecasting modelling using artificial intelligence: A review. *Groundw. Sustain. Dev.* **14**, 100643 (2021).
40. Danandeh Mehr, A. *et al.* Genetic programming in water resources engineering: A state-of-the-art review. *J. Hydrol.* <https://doi.org/10.1016/j.jhydrol.2018.09.043> (2018).
41. Banerjee, P., Singh, V. S., Chattopadhyay, K., Chandra, P. C. & Singh, B. Artificial neural network model as a potential alternative for groundwater salinity forecasting. *J. Hydrol.* <https://doi.org/10.1016/j.jhydrol.2010.12.016> (2011).
42. Barzegar, R. & Moghaddam, A. A. Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction. *Model. Earth Syst. Environ.* **2**, 26 (2016).
43. Nasr, M. & Zahran, H. F. Using of pH as a tool to predict salinity of groundwater for irrigation purpose using artificial neural network, Egypt. *J. Aquat. Res.* <https://doi.org/10.1016/j.ejar.2014.06.005> (2014).
44. Nozari, H. & Azadi, S. Experimental evaluation of artificial neural network for predicting drainage water and groundwater salinity at various drain depths and spacing. *Neural Comput. Appl.* **31**, 1227–1236 (2019).
45. Seyam, M. & Mogheir, Y. Application of artificial neural networks model as analytical tool for groundwater salinity. *J. Environ. Prot.* **2**, 56 (2011).
46. Lal, A. & Datta, B. Development and implementation of support vector machine regression surrogate models for predicting groundwater pumping-induced saltwater intrusion into coastal aquifers. *Water Resour. Manag.* **32**, 2405–2419 (2018).
47. Mosavi, A. *et al.* Susceptibility mapping of groundwater salinity using machine learning models. *Environ. Sci. Pollut. Res.* <https://doi.org/10.1007/s11356-020-11319-5> (2021).
48. Alagha, J. S., Seyam, M., Md Said, M. A. & Mogheir, Y. Integrating an artificial intelligence approach with k-means clustering to model groundwater salinity: The case of Gaza coastal aquifer (Palestine). *Hydrogeol. J.* <https://doi.org/10.1007/s10040-017-1658-1> (2017).
49. Jeihouni, M., Delirhasannia, R., Alavipanah, S. K., Shahabi, M. & Samadianfard, S. Spatial analysis of groundwater electrical conductivity using ordinary kriging and artificial intelligence methods (Case study: Tabriz plain, Iran). *Geofizika* <https://doi.org/10.15233/gfz.2015.32.9> (2015).
50. Nazari, H., Taghavi, B. & Hajizadeh, F. Groundwater salinity prediction using adaptive neuro-fuzzy inference system methods: A case study in Azarshahr, Ajabshir and Maragheh plains, Iran. *Environ. Earth Sci.* **80**, 1–10 (2021).
51. Lal, A. & Datta, B. Performance evaluation of homogeneous and heterogeneous ensemble models for groundwater salinity predictions: A regional-scale comparison study. *Water Air Soil Pollut.* **231**, 1–21 (2020).
52. Mosavi, A., Hosseini, F. S., Choubin, B., Goodarzi, M. & Dineva, A. A. Groundwater salinity susceptibility mapping using classifier ensemble and Bayesian machine learning models. *IEEE Access* **8**, 145564–145576 (2020).
53. Lal, A. & Datta, B. Application of the group method of data handling and variable importance analysis for prediction and modeling of saltwater intrusion processes in coastal aquifers. *Neural Comput. Appl.* **33**, 4179–4190 (2021).
54. Cui, T., Pagendam, D. & Gilfedder, M. Gaussian process machine learning and Kriging for groundwater salinity interpolation. *Environ. Model. Softw.* **144**, 105170 (2021).

55. Kassem, Y., Gökçekuş, H. & Maliha, M. R. M. Identifying most influencing input parameters for predicting chloride concentration in groundwater using an ANN approach. *Environ. Earth Sci.* <https://doi.org/10.1007/s12665-021-09541-6> (2021).
56. Afan, H. A. *et al.* Input attributes optimization using the feasibility of genetic nature inspired algorithm: Application of river flow forecasting. *Sci. Rep.* **10**, 1–15 (2020).
57. Seyam, M. *et al.* Investigation of the influence of excess pumping on groundwater salinity in the Gaza coastal aquifer (Palestine) using three predicted future scenarios. *Water (Switzerland)* <https://doi.org/10.3390/w12082218> (2020).
58. Armanuos, A., Ahmed, K., Shiru, M. S. & Jamei, M. Impact of increasing pumping discharge on groundwater level in the Nile Delta Aquifer, Egypt. *Knowl. Based Eng. Sci.* **2**, 13–23 (2021).
59. Shiri, N. *et al.* Development of artificial intelligence models for well groundwater quality simulation: Different modeling scenarios. *PLoS ONE* <https://doi.org/10.1371/journal.pone.0251510> (2021).
60. Diop, L. *et al.* Annual rainfall forecasting using hybrid artificial intelligence model: Integration of multilayer perceptron with whale optimization algorithm. *Water Resour. Manag.* <https://doi.org/10.1007/s11269-019-02473-8> (2020).
61. Poursaeid, M., Mastouri, R., Shabanlou, S. & Najarchi, M. Estimation of total dissolved solids, electrical conductivity, salinity and groundwater levels using novel learning machines. *Environ. Earth Sci.* <https://doi.org/10.1007/s12665-020-09190-1> (2020).
62. Jamei, M., Ahmadianfar, I., Chu, X. & Yaseen, Z. M. Prediction of surface water total dissolved solids using hybridized wavelet-multigene genetic programming: New approach. *J. Hydrol.* <https://doi.org/10.1016/j.jhydrol.2020.125335> (2020).
63. Kisi, O. *et al.* Modeling groundwater quality parameters using hybrid neuro-fuzzy methods. *Water Resour. Manag.* **33**, 847–861 (2019).
64. Jeihouni, E., Mohammadi, M., Eslamian, S. & Zareian, M. J. Potential impacts of climate change on groundwater level through hybrid soft-computing methods: a case study—Shabestar Plain. *Iran. Environ. Monit. Assess.* **191**, 1–16 (2019).
65. Islam, A. R. M. T., Al Mamun, A., Rahman, M. M. & Zahid, A. Simultaneous comparison of modified-integrated water quality and entropy weighted indices: Implication for safe drinking water in the coastal region of Bangladesh. *Ecol. Indic.* **113**, 106229 (2020).
66. Md Towfiqul Islam, A. R., Siddiqua, M. T., Zahid, A., Tasnim, S. S. & Rahman, M. M. Drinking appraisal of coastal groundwater in Bangladesh: An approach of multi-hazards towards water security and health safety. *Chemosphere* **255**, 126933 (2020).
67. Adhikari, D. K. *et al.* Urban geology: A case study of Khulna City Corporation, Bangladesh. *J. Life Earth Sci.* **1**, 17–29 (2006).
68. Kabir, M. M. *et al.* Salinity-induced fluorescent dissolved organic matter influence co-contamination, quality and risk to human health of tube well water, southeast coastal Bangladesh. *Chemosphere* **275**, 130053 (2021).
69. Islam, M. S., Idris, A. M., Islam, A. R. M. T., Ali, M. M. & Rakib, M. R. J. Hydrological distribution of physicochemical parameters and heavy metals in surface water and their ecotoxicological implications in the Bay of Bengal coast of Bangladesh. *Environ. Sci. Pollut. Res.* <https://doi.org/10.1007/s11356-021-15353-9> (2021).
70. UNDP. The Hydrogeological Condition of Bangladesh. United Nations Development Programme (Technical Report, DP/UN/BGD-74-009/1). *Groundw. Surv.* 1–13 (1982).
71. DPHE. DPHE, 1999. Main report and volumes S1–S5, report on phase I. Groundwater Studies for Arsenic Contamination in Bangladesh, Dhaka, Bangladesh. *Main Rep. Vol. S1–S5* (1999).
72. Islam, A. R. M. T. *et al.* Co-distribution, possible origins, status and potential health risk of trace elements in surface water sources from six major river basins, Bangladesh. *Chemosphere* **249**, 126180 (2020).
73. Rahman, M. M., Bodrud-Doza, M., Siddiqua, M. T., Zahid, A. & Islam, A. R. M. T. Spatiotemporal distribution of fluoride in drinking water and associated probabilistic human health risk appraisal in the coastal region, Bangladesh. *Sci. Total Environ.* **724**, 138316 (2020).
74. Pinder, S. Long Island Pine Barren Ponds: Water Quality. (2007).
75. Federation, W. E. & Association, A. P. H. *Standard Methods for the Examination of Water and Wastewater* (Am. Public Heal. Assoc., Washington, 2005).
76. Gholami, V., Khaleghi, M. R. & Sebghati, M. A method of groundwater quality assessment based on fuzzy network-CANFIS and geographic information system (GIS). *Appl. Water Sci.* <https://doi.org/10.1007/s13201-016-0508-y> (2017).
77. Kursu, M. B., Jankowski, A. & Rudnicki, W. R. Boruta: A system for feature selection. *Fundam. Informaticae* **101**, 271–285 (2010).
78. Kursu, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
79. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
80. Shafiee, S. *et al.* Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. *Comput. Electron. Agric.* <https://doi.org/10.1016/j.compag.2021.106036> (2021).
81. Zhang, S. *et al.* A temporal LASSO regression model for the emergency forecasting of the suspended sediment concentrations in coastal oceans: Accuracy and interpretability. *Eng. Appl. Artif. Intell.* <https://doi.org/10.1016/j.engappai.2021.104206> (2021).
82. Omeje, O. E., Maccido, H. S., Badamasi, Y. A. & Abba, S. I. Performance of hybrid neuro-fuzzy model for solar radiation simulation at Abuja, Nigeria: A correlation based input selection technique. *Knowl. Based Eng. Sci.* **2**, 54–66 (2021).
83. Ali, N. S. M. *et al.* Power peaking factor prediction using ANFIS method. *Nucl. Eng. Technol.* **54**, 608–616 (2022).
84. Tur, R. & Yontem, S. A comparison of soft computing methods for the prediction of wave height parameters. *Knowl. Based Eng. Sci.* **2**, 31–46 (2021).
85. Penghui, L. *et al.* Metaheuristic optimization algorithms hybridized with artificial intelligence model for soil temperature prediction: Novel model. *IEEE Access* **8**, 51884–51904 (2020).
86. Bezdek, J. C., Ehrlich, R. & Full, W. FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **10**, 191–203 (1984).
87. Pei, Z. & Wei, Y. Prediction of the bond strength of FRP-to-concrete under direct tension by ACO-based ANFIS approach. *Compos. Struct.* **282**, 115070 (2022).
88. Abualigah, L. *et al.* Computers & industrial engineering Aquila optimizer: A novel meta-heuristic optimization algorithm. *Comput. Ind. Eng.* **157**, 107250 (2021).
89. Wang, S., Jia, H., Abualigah, L., Liu, Q. & Zheng, R. An improved hybrid Aquila optimizer and Harris Hawks algorithm for solving industrial engineering optimization problems. *Processes* **9**, 1551 (2021).
90. El Shinawi, A., Ibrahim, R. A., Abualigah, L., Zelenakova, M. & Abd Elaziz, M. Enhanced adaptive neuro-fuzzy inference system using reptile search algorithm for relating swelling potentiality using index geotechnical properties: A case study at El Sherouk City, Egypt. *Mathematics* **9**, 3295 (2021).
91. Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M. & Gandomi, A. H. The arithmetic optimization algorithm. *Comput. Methods Appl. Mech. Eng.* **376**, 113609 (2021).
92. Abualigah, L. *et al.* Aquila optimizer: A novel meta-heuristic optimization algorithm. *Comput. Ind. Eng.* **157**, 107250 (2021).
93. Lin, S., Jia, H., Abualigah, L. & Altalhi, M. Enhanced slime mould algorithm for multilevel thresholding image segmentation using entropy measures. *Entropy* **23**, 1700 (2021).
94. Abualigah, L., Diabat, A. & Elaziz, M. A. Improved slime mould algorithm by opposition-based learning and Levy flight distribution for global optimization and advances in real-world engineering problems. *J. Ambient Intell. Humaniz. Comput.* <https://doi.org/10.1007/s12652-021-03372-w> (2021).
95. Hassan, M. H., Kamel, S., Abualigah, L. & Eid, A. Development and application of slime mould algorithm for optimal economic emission dispatch. *Expert Syst. Appl.* **182**, 115205 (2021).
96. Dorigo, M. & Di Caro, G. Ant colony optimization: A new meta-heuristic. In *Proceedings of the 1999 Congress on Evolutionary Computation, CEC 1999* (1999). <https://doi.org/10.1109/CEC.1999.782657>.

97. Mehdizadeh, S., Mohammadi, B., Bao Pham, Q., Nguyen Khoi, D. & Thi Thuy Linh, N. Implementing novel hybrid models to improve indirect measurement of the daily soil temperature: Elman neural network coupled with gravitational search algorithm and ant colony optimization. *Measurement* **165**, 108127 (2020).
98. Dorigo, M. & Socha, K. Ant colony optimization. In *Handbook of Approximation Algorithms and Metaheuristics* (2007). <https://doi.org/10.1201/9781420010749>
99. Jamei, M. *et al.* The assessment of emerging data-intelligence technologies for modeling Mg^{2+} and SO_4^{2-} surface water quality. *J. Environ. Manag.* **300**, 113774 (2021).
100. Jamei, M., Ahmadianfar, I., Chu, X. & Yaseen, Z. M. Estimation of triangular side orifice discharge coefficient under a free flow condition using data-driven models. *Flow Meas. Instrum.* **77**, 101878 (2021).
101. Yaseen, Z. M. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. *Chemosphere* **277**, 130126 (2021).
102. Ahmed, A. A. M. *et al.* Deep learning hybrid model with Boruta-Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. *J. Hydrol.* **599**, 126350 (2021).
103. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
104. Habibi, V., Ahmadi, H., Jafari, M. & Moeini, A. Machine learning and multispectral data-based detection of soil salinity in an arid region, Central Iran. *Environ. Monit. Assess.* **192**, 759 (2020).
105. Pouladi, N., Jafarzadeh, A. A., Shahbazi, F. & Ghorbani, M. A. Design and implementation of a hybrid MLP-FFA model for soil salinity prediction. *Environ. Earth Sci.* **78**, 159 (2019).
106. Ma, G., Ding, J., Han, L., Zhang, Z. & Ran, S. Digital mapping of soil salinization based on Sentinel-1 and Sentinel-2 data combined with machine learning algorithms. *Reg. Sustain.* **2**, 177–188 (2021).
107. Wu, W. *et al.* Soil salinity prediction and mapping by machine learning regression in Central Mesopotamia, Iraq. *L. Degrad. Dev.* **29**, 4005–4014 (2018).
108. Ghorbani, M. A., Deo, R. C., Kashani, M. H., Shahabi, M. & Ghorbani, S. Artificial intelligence-based fast and efficient hybrid approach for spatial modelling of soil electrical conductivity. *Soil Tillage Res.* <https://doi.org/10.1016/j.still.2018.09.012> (2019).
109. Habibi, V., Ahmadi, H., Jafari, M. & Moeini, A. Mapping soil salinity using a combined spectral and topographical indices with artificial neural network. *PLoS ONE* <https://doi.org/10.1371/journal.pone.0228494> (2021).
110. Melesse, A. M. *et al.* River water salinity prediction using hybrid machine learning models. *Water* **12**, 2951 (2020).
111. Wang, F., Yang, S., Yang, W., Yang, X. & Jianli, D. Comparison of machine learning algorithms for soil salinity predictions in three dryland oases located in Xinjiang Uyghur Autonomous Region (XJUAR) of China. *Eur. J. Remote Sens.* **52**, 256–276 (2019).
112. Wang, J. *et al.* Soil salinity mapping using machine learning algorithms with the Sentinel-2 MSI in Arid Areas, China. *Remote Sens.* **13**, 305 (2021).
113. Islam, A. R. T. *et al.* A coupled novel framework for assessing vulnerability of water resources using hydrochemical analysis and data-driven models. *J. Clean. Prod.* **336**, 130407 (2022).
114. Tran, D. A. *et al.* Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam. *Ecol. Indic.* **127**, 107790 (2021).

Author contributions

M.J.: Project leader, Software, Writing up, Analysis, Validation, Methodology, Writing up. M.K.: Methodology, Assessment and Analysis, Writing up. Anurag Malik: Writing up, Validation. L.A.: Supervision, Writing up, Revision, Investigation. A.R.M.T.I.: Data collection, Writing up, Revision, Investigation. Z.M.Y.: Supervision, Writing up, Analysis, Validation, Discussion.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.M.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022