



A Comparative Study of Artificial Intelligence Models and A Statistical Method for Groundwater Level Prediction

Mojtaba Poursaeid^{1,2} · Amir Houssain Poursaeid³ · Saeid Shabanlou⁴

Received: 6 September 2021 / Accepted: 12 January 2022 / Published online: 22 January 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022, corrected publication 2022

Abstract

Today, various methods have been developed to extract drinking water resources, which scientists use to simulate the quantitative and qualitative water resources parameters. Due to Iran's geographical and climatic characteristics, this region is located on the drought belt in Asia. In this research, some Artificial Intelligence (AI) and mathematical models have been used for groundwater level prediction. The AI models used for this research are Extreme Learning Machine (ELM), Least Square Support Vector Machine (LSSVM), Adaptive Neuro-Fuzzy Inference System (ANFIS), and Multiple Linear Regression (MLR) model. In this study, simultaneously, these models were used to simulate and estimate groundwater level (GWL). The database used in the simulation is the data related to the Total Dissolved Solids (TDS), Electrical Conductivity (EC), Salinity (S), and Time (t) parameters. The results showed that ELM was more accurate than other methods. In Uncertainty Wilson Score Method (UWSM) analysis, ELM had an Underestimation performance and was determined as the more precise model.

Keywords Extreme Learning Machine · Least square Support Vector Machine · Multiple Linear Regression · Adaptive Neuro-Fuzzy Inference System · MLR · LSSVR · ELM · ANFIS · Water Parameters · Quantity Parameters

Highlights

- Extreme Learning Machine (ELM) method is the best performance in GWL simulation.
- Simulation results showed better performance of ELM and LSSVM in modeling groundwater. The MLR and ANFIS methods ranked next respectively.
- Evaluation of the models was done by seven approaches.
- The ELM model was the superior model in all seven approaches.

✉ Mojtaba Poursaeid
mojtaba.poursaeid.ac@gmail.com

¹ MPO-Plan and Budget Organization, Khorramabad, Iran

² Department of Civil Engineering, Faculty of Technical and Engineering, Payam Noor University, Khorramabad, Iran

³ Department of Electrical Engineering, Lorestan University, Khorramabad, Iran

⁴ Department of Water Engineering, Islamic Azad University, Kermanshah Branch, Kermanshah, Iran

1 Introduction

1.1 Aim & Scope

Today, water scarcity is one of the most critical problems for humans. Also, water quality modeling is one of the main challenges in water resources management (Kheradpisheh et al. 2015; Qu et al. 2020). Use for agricultural, industrial, and drinking purposes is one of the reasons for the high importance of our water quality management. One of the management challenges is to predict the future state of water resources. Also, in today's research, water resources have been studied and modeled by different scenarios (Chang et al. 2021).

However, water scarcity has occurred worldwide due to population growth, industrial development, and increasing water use, especially in arid and semi-arid regions (Sharafati et al. 2020). Recently, multiple and continuous droughts in different parts of Iran have occurred. It should be noted that available water resources are unstable, so that there is no guarantee that they will be usable. At this particular time, with the advancement of science and various models for studying climate change, water quality, air pollution, etc. Meanwhile, One of these advances has occurred in the case of AI, which contributes to many types of research, and AI models have been more successful than the other approaches (Cao et al. 2020; Lyu and Liu 2021).

1.2 Literature Review

Recently, many types of research have been done in water resources management. These studies have been conducted to qualitative and quantitative modeling, optimization of the system, and estimating the parametric changes of water resources. Scientific advances in engineering applications led to the development of AI that makes it easy to analyze non-linear and complex problems. These methods are divided into several categories such as Artificial Neural Network (ANN), Machine Learning (ML), Metaheuristic Optimization Algorithms (MOA), Fuzzy Inference System (FIS), and Combination of ANNs and MOAs.

On the one hand, some researchers studied water resources management using ANNs, which not only increase the speed of evaluation but also improve the accuracy of the groundwater level and other water parameters estimation (Yang et al. 2014; Shahid and Ehteshami 2015; Heddami et al. 2016). On the other hand, In many articles, authors modeled the water variables using ML models (Elkiran et al. 2019; Majumder and Eldho 2020; Qu et al. 2020; Kakhodazadeh and Farzin 2021). Moreover, in some of the papers, parameters simulated by ML models and ANNs conjunctions (Zhu and Heddami 2019), while in other works, ANNs and ML models have been used separately for quantitative and qualitative assessment of water resources parameters, water resource monitoring, estimation (Azad et al. 2015; Shi et al. 2018; Ye et al. 2019; Patki et al. 2021; Zhang et al. 2021). Other studies used optimization algorithms to improve the modeling and simulation of water resources. MOAs such as genetic algorithm, differential evolution, and particle swarm optimization was implemented to increase the accuracy and precision of water engineering (Heddami et al. 2016; Guneshwor et al. 2018; Jekhouni et al. 2020). In the rest of the articles, a combination of MOAs, wavelet transform, and ANNs and MLs is considered a solution to improve the modeling optimization, optimal design, accurate estimation, and prediction of water resources. (Jaddi and Abdullah 2017; Alizadeh et al. 2018; Zhang et al. 2019; Poursaeid et al. 2020; Noori et al. 2020). Also, fuzzy logic, fuzzy neural networks,

and FISs were used for parametric modeling of water, water pollution. (Tokachichu and Gaddam 2021; Sada and Ikpeseni 2021; Niu et al. 2021; Asgari et al. 2021).

1.3 Contribution

In this paper, the groundwater level was predicted using ML techniques and mathematical methods. Although various studies have been done in this field, for the first time in the study area, the GWL was estimated with ELM, LSSVM, ANFIS models, and MLR model as a comparative study to simultaneously estimate groundwater parameters. In quantitative and qualitative water resource management, various practical factors include Cl^- , EC, TDS, SO_4^{2+} , Ca^{2+} , Mg^{2+} , etc. Among these factors, the most widely used water quality parameters are TDS, EC, salinity, and time considered the input vector to the abovementioned models.

The rest of the paper is organized as follows: The water quality and its parameters are explained in Sect. 2. In Sect. 3, materials and methods such as various AI models and MLR formulation are described. In Sect. 4, the study area and its steps are expressed. The results are presented and discussed in Sect. 5. Finally, the conclusion is summarized in Sect. 6.

2 Problem Description

In this section, water quality and its parameters are explained. Also, some international water quality standards are presented.

2.1 Water Quality (WQ)

WQ management plays a critical role in the quality management of water resources and sustainable use of these water resources (Ahmadianfar et al. 2020). Several factors affect water quality. These factors are divided into two categories of quantitative and qualitative factors. Among the quantitative factors, such as rainfall, temperature, etc., can be mentioned. Qualitative factors in terms of number are much more than quantitative factors. The following are some of the most handful of water quality parameters (Lukawska-Matuszewska and Urbański 2014).

2.1.1 Total Dissolved Solids (TDS)

Total Dissolved Solids (TDS) is one of the reliable and valid parameters for WQ, which means the numerical sum of all types of soluble solids in water (Jamei et al. 2020). On the one hand, TDS measures the number of solutes remaining after evaporation of the measured volume of purified water (Mokhatab et al. 2019). On the other hand, this parameter is widely known as a measure of water suitability for drinking and irrigation purposes, which is measured in milligram per liter (mg/l). TDS includes various types of mineral salts such as sodium (Na^+), magnesium (Mg^{+2}), calcium (Ca^{+2}), potassium (K^+), chloride (Cl^-), sulfate (SO_4^{-2}), nitrates (NO^{-3}), soluble bicarbonates (HCO^{-3}), and organic matter (Ahmadianfar et al. 2020).

2.1.2 Electrical Conductivity (EC)

Electrical conductivity (EC) is one of the essential WQ parameters used in several articles to model the WQ resources. EC is one of the most important factors in WQ analysis which is equivalent to the concept of Salinity (Serrano-Finetti et al. 2019). Therefore, EC is a parameter that indicates the degree of electrical transmission in water, which is closely related to the number of water-soluble salts. The salts concentration of water is a crucial factor in determining WQ's suitability for drinking and irrigation purposes. It should be noted that EC is measured in micro Siemens per centimeter ($\mu\text{S}/\text{cm}$). Same with TDS, EC is dependent on the amount of dissolved ionic solutes such as sodium (Na^+), chloride (Cl^-), magnesium (Mg^{+2}), sulfate (SO_4^{-2}), And calcium (Ca^{+2}) in water. However, the amount of ionic salts in the water reduces its drinking quality (Ahmadianfar et al. 2020).

2.1.3 Salinity

The water salinity is a qualitative parameter that is one of the criteria for WQ assessment. This parameter is known as the concentration of salt in the water. However, salinity is defined as natural salinity in water resources, but some factors such as high evaporation rates or increased human consumption cause it to increase (Harris 2009). In other words, salinity can be defined as the concentration of soluble mineral salts in water and soil based on volume or weight per unit area (Sparks 2003).

2.1.4 Hardness

This parameter included the presence of some of the water-soluble salts. It can also mean the amount of calcium and magnesium in water. Moreover, its ordinary meaning is the concentration of calcium carbonate in water. These salts have different forms which the most common form are fluorides, carbonates, sulfates of calcium, and magnesium (Mtaita 2003). Since water hardness is a criterion of WQ measurement, it is considered slightly different from Salinity (Ansell 2005).

2.1.5 Dissolved Oxygen (DO)

Dissolved Oxygen (DO) is one of the WQ parameters, which means the amount of soluted oxygen present in water. This parameter plays a significant role in the science of water resource management (Yang et al. 2021). Also, it represents the health of the water and is a criterion for understanding the health status of the river (Tiyasha et al. 2021).

2.2 WQ Standards

There are several standards for WQ parameters, shown in Table 1. These standards include WHO, BIS, and SSMQO (Ahuja et al. 2019).

Table 1 Standards of WQ

Parameter	WHO(2017)	BIS(10,500:2012)	SSMQO(2003)
pH	–	6.5- 8.5	6.5–9.2
Electrical Conductivity	< 2500 $\mu\text{S/cm}$	–	1500 $\mu\text{S/cm}$
Total Dissolved Solids	300–600 ppm ¹	500–2000 ppm	–
Total hardness	300–600 ppm ²	200–600 ppm	500 ppm
Ca	–	75–200 ppm	200 ppm
Mg	–	ppm	150 ppm

¹TDS Value for WHO (2011)

²Total Hardness for WHO (1997)

3 Material and Methods

3.1 AI Models

In this section, AI models and the MLR method are expressed.

3.1.1 Least Square-Support Vector Machine (LSSVM)

The support vector machine (SVM) is based on Vapnik theory (Sapankevych and Sankar 2009). This type of ML uses the method of minimizing structural risk, while some other methods of AI use the experimental method of minimizing the risks (Cristianini and Shawe-Taylor 2000; Dibike et al. 2001). The SVM can be used for classification and regression problems. In this theory, in a quadratic programming problem, an equation is obtained that determines the constant parameters of the model. Then, the optimal values for the constants of SVM can be obtained using MOAs. SVMs were initially used for classification, but they can be used for time-series prediction (Cristianini and Shawe-Taylor 2000; Campbell 2002; Schölkopf and Smola 2002; Suykens et al. 2002).

By mathematical definition, the least squares support vector machine (LSSVM) is considered as if x_i and y_i are the input and output data for the model, respectively, then the non-linear regression function is also defined as follows (Valyon and Horvath 2007):

$$f(x) = \sum_{i=1}^k w_i \varphi_i(x) + b = w^T \cdot \varphi(x_i) + b \quad (1)$$

where w is the weight vector, b is the bias, and φ are nonlinear functions for mapping data into large feature spaces:

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix}, \quad \varphi = \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_k \end{bmatrix} \quad (2)$$

or

$$w = [w_1, \dots, w_k]^T, \quad \varphi = [\varphi_1, \dots, \varphi_k]^T$$

The nonlinear regression problem can be solved by minimizing the following quadratic programming problem:

$$\text{Min}_{w,e,b} \left\{ G(w, e) = \frac{1}{2} w^T w + \frac{1}{2} C \sum_{i=1}^n e_i^2 \right\} \tag{3}$$

where C has the role of tradeoff variable between two terms of the equation. The result is defined as follows:

$$Y_i = w^T \cdot \varphi(x_i) + b + e_i, \quad Y_i = y_i + \lambda_i \tag{4}$$

λ_i is the system noise. Also, for each x_i in LSSVM, the result is a weighted sum of n kernel functions, in which the central variable of the kernel functions is obtained using trained inputs. The Lagrangian form of the equation with these explanations is shown in Eq. (5).

$$\Psi(w, b, e, \alpha) = G(w, e) - \sum_{i=1}^n \alpha_i (w^T \cdot \varphi(x_i) + b + e_i - Y_i) \tag{5}$$

In Eq. (5), α_i 's are Lagrangian multipliers. Then, a constrained optimization problem can be solved. Optimization constraints will be defined as Eq. (6).

$$\begin{aligned} \frac{\partial \Psi}{\partial w} = 0 &\quad \rightarrow w = \sum_{i=1}^n \alpha_i \varphi(x_i) \\ \frac{\partial \Psi}{\partial b} = 0 &\quad \rightarrow \sum_{i=1}^n \alpha_i = 0 \\ \frac{\partial \Psi}{\partial e_i} = 0 &\quad \rightarrow \alpha_i = C e_i, \quad i = 1, 2, \dots, n \\ \frac{\partial \Psi}{\partial \alpha_i} = 0 &\quad \rightarrow w^T \varphi(x_i) + b + e_i - Y_i = 0, \quad i = 1, 2, \dots, n \end{aligned} \tag{6}$$

At the end of the above steps, the final solution of the problem is as follows:

$$\begin{bmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & \Phi_{11} + \frac{1}{C} & \Phi_{12} & \dots & \Phi_{1n} \\ 1 & \Phi_{21} & \Phi_{12} + \frac{1}{C} & \dots & \Phi_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \Phi_{(n-1)1} & \Phi_{(n-1)2} & \dots & \Phi_{(n-1)n} \\ 1 & \Phi_{n1} & \Phi_{n2} & \dots & \Phi_{nn} + \frac{1}{C} \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} 0 \\ Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \tag{7}$$

$$Y = [Y_1, Y_2, \dots, Y_n], \alpha = [\alpha_1, \alpha_2, \dots, \alpha_n] \tag{8}$$

$, 1 = [1, 1, \dots, 1]$

Furthermore, in Eq. (7), the Φ_{ij} is the kernel matrix, and $H(x_i, x_j)$ is the kernel functions, which will be written as follows:

$$\Phi_{ij} = H(x_i, x_j), \quad \Phi_{ij} = \varphi^T(x_i) \cdot \varphi(x_j), \quad i, j = 1, \dots, n \tag{9}$$

3.1.2 Extreme Learning Machine (ELM)

The theory of this algorithm was proposed by a Singaporean scientist named Bin in 2004 (Bin et al. 2004). This model of AI is one of the learning machines, and in various researches, its superiority over other methods of AI has been proved due to its single-layer feed-forward neural network (Bin et al. 2006, 2012). If we have n neurons in the hidden layer, we can define the single-layer feed-forward network as Eq. (10) (Liang et al. 2006).

$$y_j = \sum_{i=1}^n \beta_i g(x_k; c_i, a_i) \quad , j = 1, 2, 3, \dots, k \tag{10}$$

where g , c_i and β_i are the transfer function between input and output layers, respectively. The weights that connect the output nodes to the hidden layer nodes and the biases are initialized randomly. The equation, as mentioned earlier, can be rewritten in the form of the following equations.

$$H \beta = Y \tag{11}$$

$$H = \begin{bmatrix} g(x_1; c_1, w_1) & \dots & g(x_1; c_k, w_k) \\ \vdots & \ddots & \vdots \\ g(x_n; c_1, w_1) & \dots & g(x_n; c_k, w_k) \end{bmatrix}_{n \times k} \tag{12}$$

$$\beta = (\beta_1^T, \beta_2^T, \dots, \beta_h^T)_{h \times m}^T \tag{13}$$

Finally, the output weights of the learning machine can be calculated in the hidden layer using the Moore–Penrose generalized inverse matrix method:

$$\beta = H^\dagger Y \tag{14}$$

3.1.3 Adaptive Neuro-Fuzzy Inference System (ANFIS)

Adaptive Neuro-Fuzzy Inference System (ANFIS) is a feed-forward neural network that simulates based on fuzzy logic. In this network, two types of Fuzzy Inference Systems (FIS) based on fuzzy logic (Tokachichu and Gaddam 2021; Arora and Keshari 2021):

- Fuzzy inference system-based network, called Mamdani, known as M-FIS for short.
- Takagi–Sugeno fuzzy inference system-based network, known as TS-FIS for short.

In these networks, at least there are two inputs, D_1 and D_2 , which will be the two *if-then* conditional principles for each output as O_i for a network based on the TS-FIS fuzzy inference system. The conditional rules of these fuzzy networks are as follows:

- 1) If x is input D_1 and output O_1 , then we have:

$$f_1 = a_1 x_1 + b_1 y_1 + c_1$$

2) If x is input D_2 and output O_2 , then we have:

$$f_2 = a_2x_2 + b_2y_2 + c_2$$

Neuro-fuzzy networks are organized with an input layer and five other layers, which can be a multi-layered neural network.

- Layer 0: Input layer with n Input Nodes
- Layer 1: This layer provides a membership function for points using Gaussian principles by fuzzifying each node.

$$\mu_{Di}(x) = \exp \left\{ - \left[\left(\frac{x - h_i}{z_i} \right)^2 \right]^{t_i} \right\} \tag{15}$$

where z_i , t_i , and h_i are the parameter of adaptive functions in the network.

- Layer 2: all fuzzified data is passed into operators. D_i , O_i The membership parameters $\mu_{Di}(x)$ and $\mu_{Oi}(x)$, Are the antecedent parameters of rule (1).

$$w_i = \mu_{Di}(x) \times \mu_{Oi}(x) \tag{16}$$

- Layer 3: All of the nodes is normalized as:

$$\bar{w}_i = \frac{w_i}{\sum_{i=1}^T w_i} \tag{17}$$

where the \bar{w}_i second layer is the sum of the operator in the i^{th} order.

- Layer 4: In each node, the corresponding linear function is calculated, and the coefficient of the functions is calculated using the backpropagation neural network error.

$$\bar{w}_i f_i = \bar{w}_i (a_0x_0 + a_1x_1 + a_2) \tag{18}$$

where a_i is the input i and \bar{w}_i as the output of layer 3. This model is trained using the least-squares approximation method.

- Layer 5: This layer is the sum of the output of each node from the fourth layer, which is calculated as below:

$$\sum \bar{w}_i f_i = \frac{\sum w_i f_i}{\sum w_i} \tag{19}$$

3.1.4 Multiple Linear Regression (MLR)

Multiple linear regression (MLR) methods are a statistical method used to examine and infer the relationship between the dependent variable and multivariate primary variables. These methods are written as the following equation based on the mathematical relationships between the primary and secondary variables (predictors and responses):

$$f(x_i) = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n + e \tag{20}$$

where $f(x_i)$ is a secondary variable, x_i 's are multiple primary variables, a_i are regression multipliers, and e is a random error (Mustapha and Abdu 2012).

4 Case Study and Data Collection

4.1 Study Area

In this paper, the study area is Mighan plain in Arak, located in Markazi province in Iran. According to the statistical results provided by Synoptic stations in the region, the maximum and minimum rainfall varies from 461 mm in the northeast to 208 mm in the center of Arak plain. Figure 1 shows the geographical location of the study area and Vismeh well. In this work, TDS, EC, water salinity, and time parameters were used as models dataset for GWL simulation.

4.2 Study Steps and Data Analysis

In this study, the time-series database was first collected through the database of the Regional Water Company of Markazi province, and then the dataset was categorized. The K-Fold cross-validation method was used to increase the simulation's reliability and

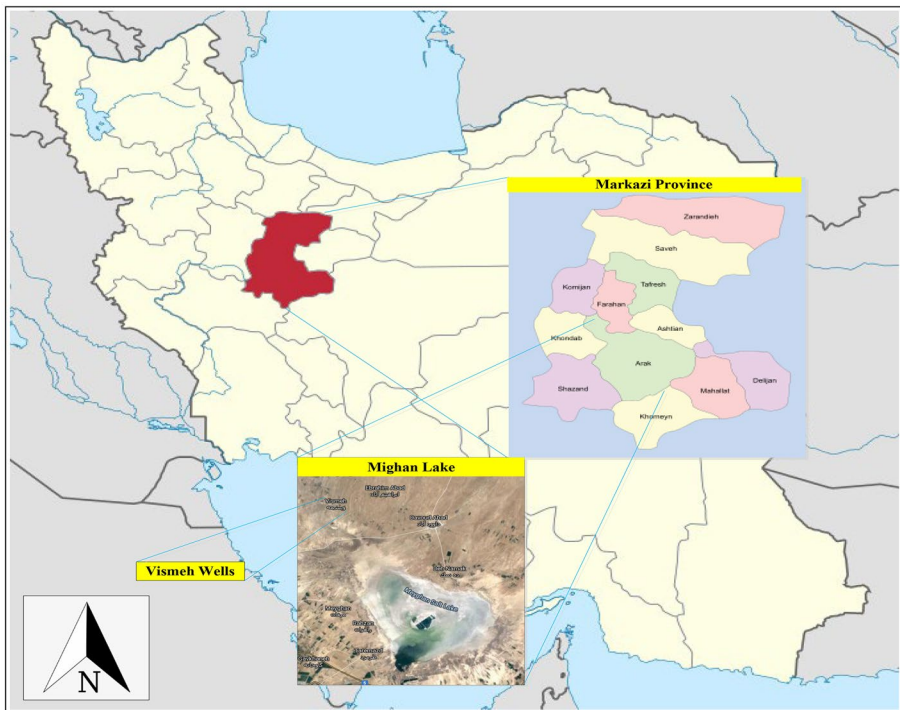


Fig. 1 Mighan Lake (Source: Wikimedia & Google Map)

accuracy by removing the data trend (detrending data) and data randomization (Poursaeid et al. 2021).

It should be noted that 173 months of sampling data were used in the training of models. In most articles on AI, the combination of test and train data percent is 80 to 20 or 70 to 30. (Reynolds et al. 2019; Jang et al. 2019; Sada and Ikpeseni 2021; Hameed et al. 2021; Hou et al. 2021). Therefore, due to better validation of the model, it was decided that 70% of the data would be used for training and the remaining 30% for the testing phase in modeling. Then, the same training dataset is entered for all of the LSSVM, ELM, ANFIS, and MLR models. Moreover, the observed data were *TDS*, *Salinity*, *t*, and *EC*, as Primary variables, and *GWL*, as a response parameter. Finally, the performance and accuracy of the models were compared. Statistical indices made this comparison according to Eqs. (21) to (24).

$$R^2 = \frac{(n \sum_{i=1}^n O_i I_i - \sum_{i=1}^n O_i \sum_{i=1}^n I_i)^2}{(n \sum_{i=1}^n (O_i)^2 - \sum_{i=1}^n (O_i))^2 (n \sum_{i=1}^n (I_i)^2 - \sum_{i=1}^n (I_i))^2} \quad (21)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - I_i)^2} \quad (22)$$

$$MAPE = \frac{\sum_{i=1}^n |O_i - I_i|}{n} \times 100 \quad (23)$$

$$SI = \frac{RMSE}{\bar{I}} \quad (24)$$

where I_i and O_i are the input values and output values, respectively. Also, for all AI and MRL models, \bar{I} it is considered the mean of observational values and equal to n the number of observational values. In the following, the accuracy of different models for estimating *GWL* parameter values is investigated.

5 Results and Discussions

First, the input vectors were applied to all models, and *GWL* is considered the output vector. Then, according to the evaluation indices, the performance of the models was evaluated. In this research, seven approaches were used to assess the performance of models in simulation (Figs. 2 and 3).

5.1 Response and Correlation Plots

The response plot shows the actual and the predicted values for every sample that mapped on each other. Also, the correlation plot is a scatter diagram used to demonstrate the linear correlation between the actual values and their corresponding predicted ones. The response and correlation plot of all models are drawn in Figs. 4 and 5, respectively.

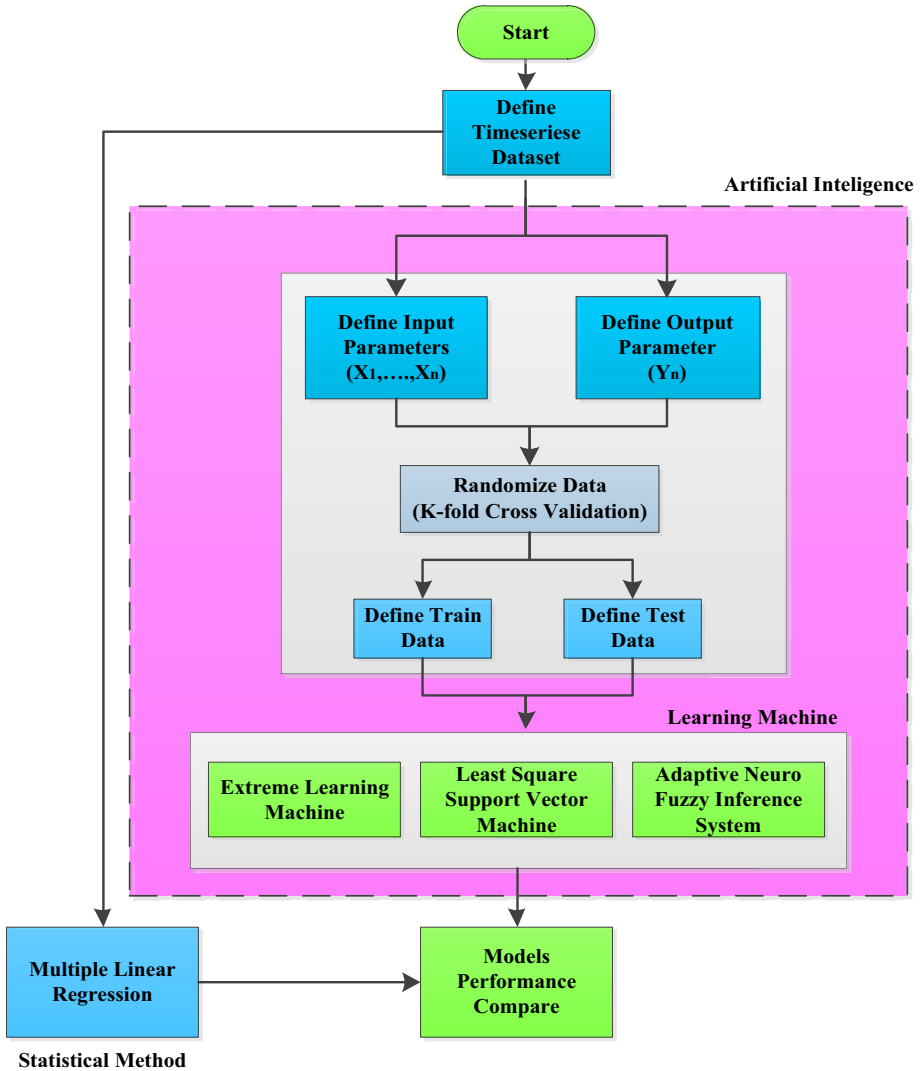


Fig. 2 Research Flowchart

To determine the superior model, the simulation results are drawn by response plots. According to Fig. 4, the ELM model was the most accurate GWL prediction. In Fig. 5, the ELM model had the best correlation between observed and predicted data and was determined as a superior model. However, The least accuracy and performance were assigned to the ANFIS model. As shown in Fig. 5, it had the lowest correlation between responses and observed GWL data.

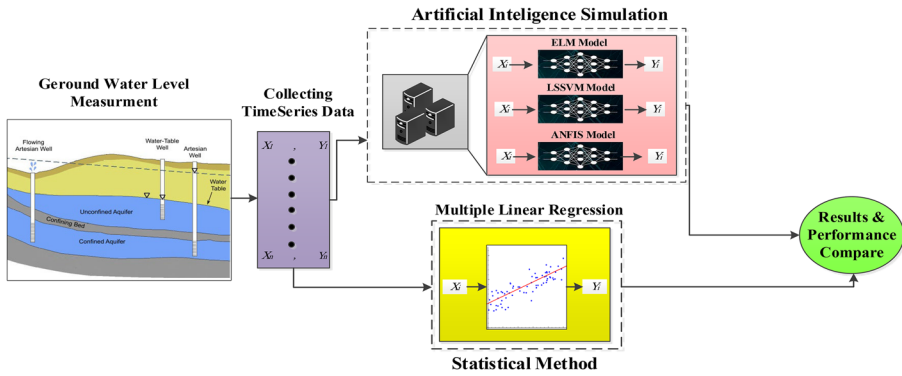


Fig. 3 Graphical Abstract

5.2 Statistical Indicators

The results that compare the accuracy of all models based on different performance indices are shown in Table 2. As can be seen in Table 2, the statistical indices for the ELM model were the most accurate, which has the lowest RMSE, MAPE, and SI value compared to the other methods while it has the closest value to 1 for R^2 . The RMSE, MAPE, SI, and R^2 values for ELM equal to 0.1562, 0.0067, 0.000094, and 0.988, respectively. Besides, the LSSVM model with indices equal to 0.3952, 0.0165, 0.000238, and 0.927 is known as the second accurate model. The MLR and ANFIS models placed in third and fourth, respectively.

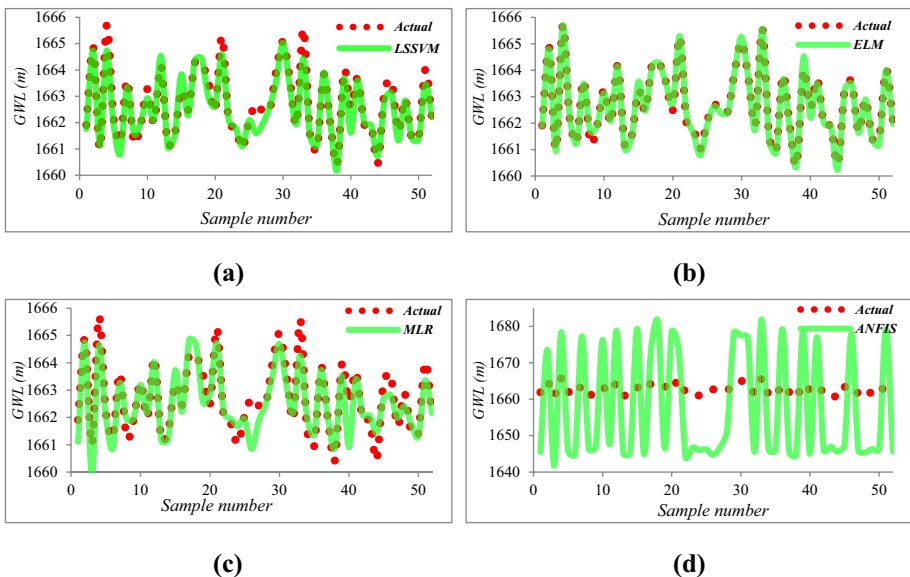


Fig. 4 Models Prediction Performance: (a) LSSVM, (b) ELM, (c) MLR, (d) ANFIS

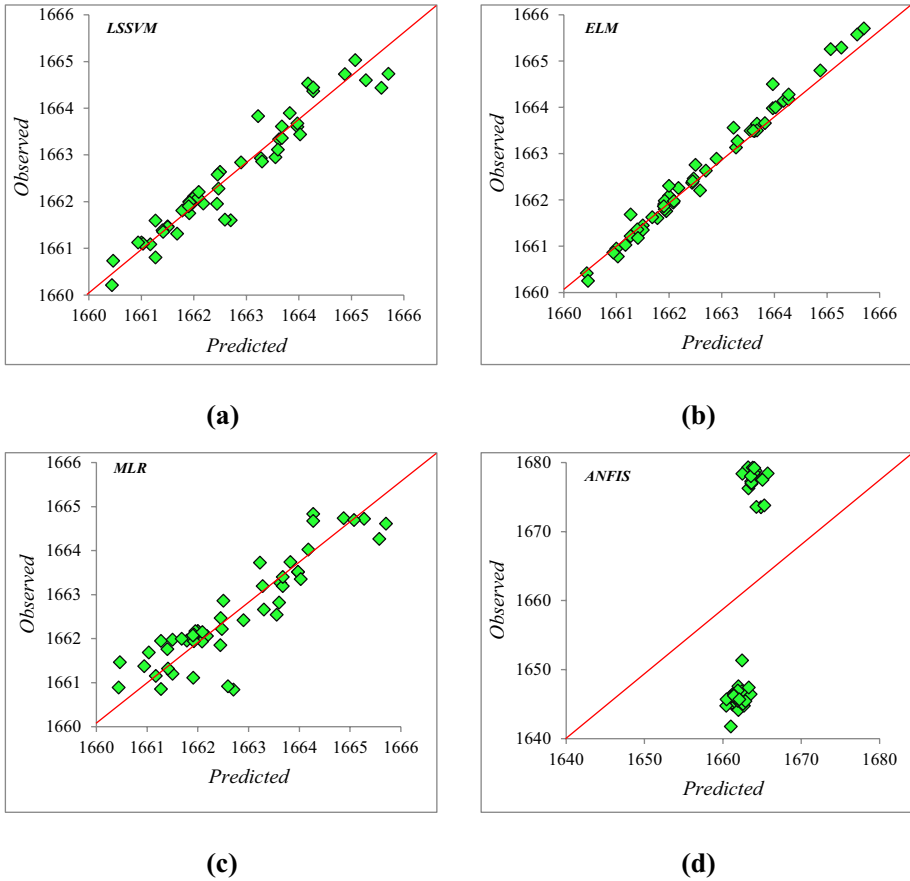


Fig. 5 Observed data and Results Correlation: (a) LSSVM, (b) ELM, (c) MLR, (d) ANFIS

To better visualize Table 2, the various performance indices are compared to each other, shown in Fig. 6. ELM model was determined as the superior model.

5.3 Uncertainty Analysis by Wilson Score Method (WSM)

Each of the four methods, as mentioned above, has errors between the actual values and the predicted ones, which are evaluated using uncertainty analysis by WSM analysis

Table 2 Statistical Indices

Model	MAPE	RMSE	SI	R ²
ELM	0.0067	0.1562	9.39444e-05	0.98804
LSSVM	0.0165	0.3952	0.000237688	0.92756
MLR	0.028	0.6161	0.000370545	0.80461
ANFIS	8.56667	15.2365	0.009163786	0.68459

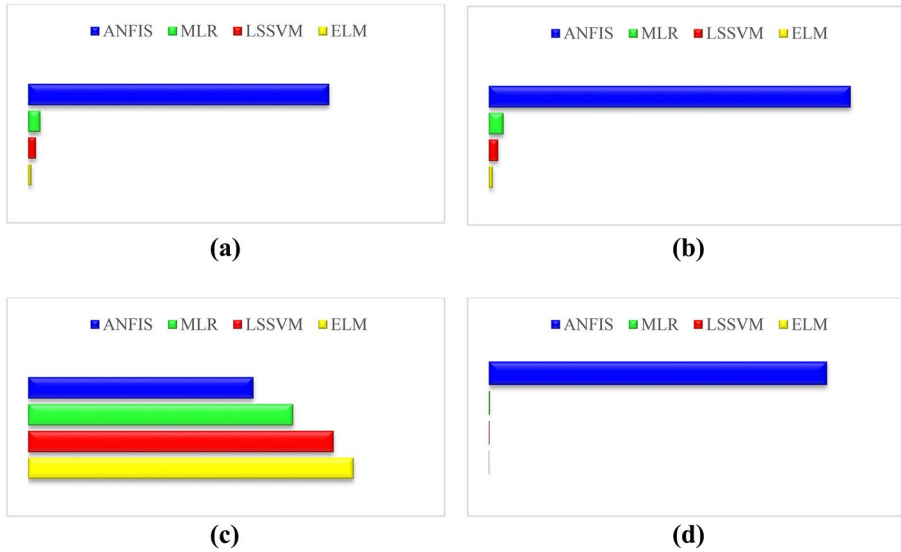


Fig. 6 Accuracy Compare: (a) RMSE, (b) SI, (c) R², (d) MAPE

(Bonakdari et al. 2020). This analysis can be calculated independently according to the computational error in the simulation of each model. However, some of the uncertainties are related to data sampling errors, which is impossible to investigate this type of uncertainty due to the limited number of data or the accessibility to the monthly gathering of the datasets. Computational parameters in WSM analysis are forecast error Er_i , average prediction error *Average (Eri)*, and standard deviation of error values S_e , which are calculated according to as follows:

$$Er_i = O_i - I_i \tag{25}$$

$$Avg(Er) = \frac{1}{n} \sum_{i=1}^n Er_i \tag{26}$$

$$S_e = \sqrt{\sum_{i=1}^n (Er_i - Avg(Er))^2 / (n - 1)} \tag{27}$$

Table 3 Models Uncertainty Analysis

Model	WUB	Average (Er _i)	S _e	95% CI	
				Upper bound	Lower bound
ELM	0.048	-0.02744	0.1633	0.02056	-0.07544
LSSVM	0.1263	-0.1562	0.3666	0.11068	-0.14192
MLR	0.2598	0.1818	0.5945	0.4416	-0.078
ANFIS	1.6589	5.1414	14.4828	6.8003	3.4825

where I_i and O_i are the input and output values, respectively, while n is the number of observation samples. The results of the WSM analysis are shown in Table 3 by considering the Width of Uncertainty Band (WUB) of 95% and applying $\pm 1.64 S_e$, which causes the formation of confidence interval equal to 95% (5% error) approximately and denotes by 95% CI.

According to Table 3, ELM and LSSVM models have an underestimation performance, while MLR and ANFIS have an overestimation performance. The ELM model with an average prediction error equal to 0.02744 is considered the most accurate model.

5.4 Regression Receiver Operating Characteristic (RROC) Curve and Area over the RROC Curve (AOC)

The Receiver Operating Characteristic (ROC) curve is a two-dimensional (2D) curve used in classification problems. This criterion is used for the effectiveness of the factor in various issues (Fluss et al. 2012). Analysis of classification issues using this curve is known as ROC analysis. The ROC curve shows the classification problem performance and brought in regression issues, known as the RROC curve. The RROC curve is equivalent to the concept of the ROC curve but in the case of regression problems.

Moreover, the RROC curve shows estimation accuracy and proposed that shows the Over-estimation against Under-estimation (Hernández-Orallo 2013). Also, for comparing

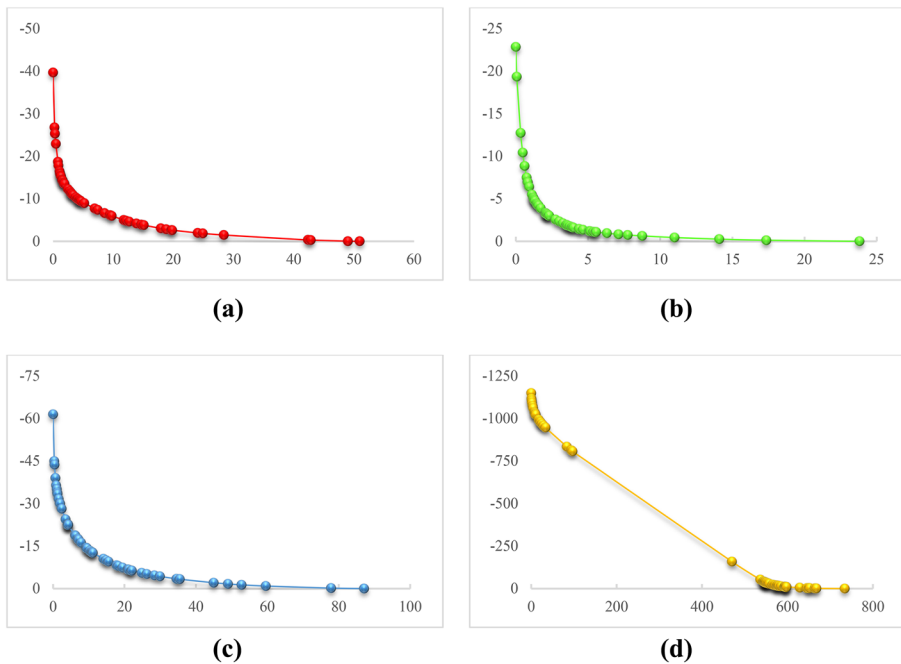


Fig. 7 Models RROC Curve: (a) LSSVM, (b) ELM, (c) MLR, (d) ANFIS

Table 4 AOC values of the Models

Model	AOC
LSSVM	178.2307
ELM	28.7048
MLR	468.59
ANFIS	2.7e+05

the regression modeling and prediction, the Area Over the RROC Curve (AOC) can be implemented (Poursaeid and Namdari 2022). The smaller the value of AOC shows, the higher accuracy of the modeling.

Based on the values range in RROC curves which are shown in Fig. 7, it is concluded that the model has the most accuracy in predicting GWL. The ELM range and domain on 2D Axis conducted that the ELM simulation was better than other models. Then LSSVM model earned second place in the ranking. Meanwhile, the MLR and ANFIS models were ranked next. Also, based on the values of AOC in Table 4, considering that the ELM model has a minimum value of AOC equal to 28.7048, so is known as the superior model. The LSSVM model also has an AOC equal to 178.2307, in second place while MLR and ANFIS models with AOC values equal 468.59 and 270,000 were ranked third and fourth, respectively.

5.5 Discrepancy Ratio (DR)

According to the mathematical concept of the DR as Eq. (28), the closeness of its value to a horizontal line (DR = 1) shows that the predicted values are close to the actual ones

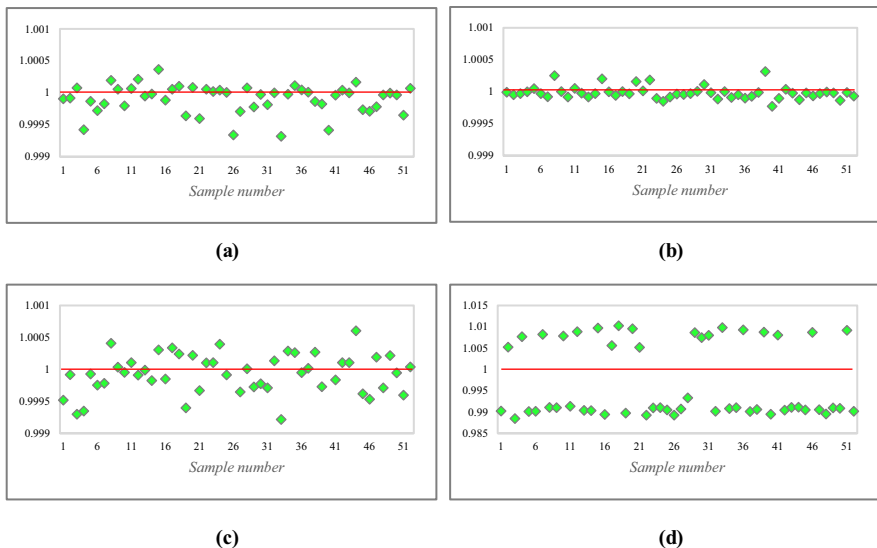


Fig. 8 Discrepancy Ratio: (a) LSSVM, (b) ELM, (c) MLR, (d) ANFIS

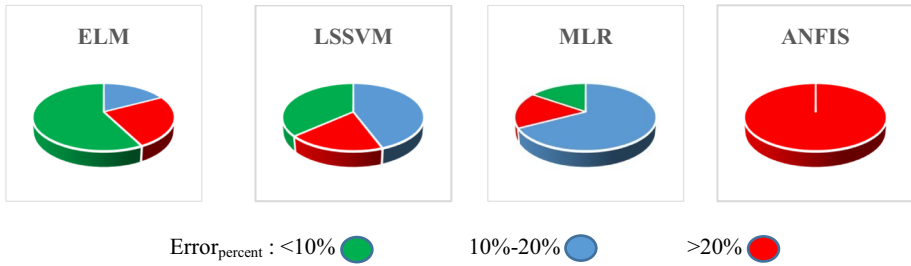


Fig. 9 Models Error Distribution

(Poursaeid et al. 2020). Moreover, this diagram shows the superiority of the ELM model to other approaches. The ANFIS model had the worst result.

$$DR = \left[\frac{GWL_{predicted}}{GWL_{observed}} \right] \tag{28}$$

Based on the diagrams of Fig. 8, the ELM model had the most accurate prediction that can be detected according to the closeness of points on the line DR = 1. The LSSVM, MLR, and ANFIS were ranked next after the ELM.

5.6 Error Distribution Plots

According to Eqs. (29) and (30), the concept of error is based on absolute Error definition, which is introduced as the "difference between the actual value and the predicted value." The differences obtained for each model are calculated. Then, it is written in percent.

$$Error = GWL_{predicted} - GWL_{observed} \tag{29}$$

and

$$Error_{Percent} = \left| \frac{GWL_{predicted} - GWL_{observed}}{GWL_{observed}} \right| \times 100 \tag{30}$$

Based on the results of prediction error distribution in Fig. 9, The ELM model has the most Error-percent in the range of less than 10%. Moreover, the ELM model has the least Error-percent in the range greater than 20%, so it is determined as superior to other models.

Table 5 Models Simulation Time

Model	Simulation Time (sec)
LSSVM	1.276492
ELM	1.119735
MLR	4.972179
ANFIS	6.083326

5.7 Comparison of Testing Times

The computer with specification Intel® Core™ i5-4510U CPU @ 2 GHz calculates the testing phase of all four models. In this part, the consumed time for testing the models is compared, shown in Table 5. As depicted in mentioned Table, the ELM model was the fastest model, which the testing time equal to 1.119735. Next, the LSSVM model was more rapid than other models. The MLR and ANFIS were ranked next position.

6 Conclusions

This study used the qualitative and quantitative parameters of groundwater in Arak plain, Markazi province, in Mighan wetland, simultaneously to predict the GWL. The parameters used as input are sampling time, TDS, EC, and salinity, which are used to estimate GWL by implementing four models consisting of three AI models and one statistical model. The AI models are ELM, LSSVM, ANFIS, while the mentioned statistical model is MLR. After analyzing the results, based on statistical indices, the best results were recorded for ELM and LSSVR models with the less amount of RMSE, MAPE, and SI and closer value to 1 for R^2 . Moreover, based on the response plot, the best performance was assigned to the ELM model by better mapping the predicted values on the target ones.

The uncertainty analysis by WSM, the ELM model with changes in the confidence bound from 0.02056 to 0.07544, and average errors equal to 0.02744 was the best and the most accurate in the simulation of GWL, which has underestimation performance. In the case of DR, the ELM model had the most concentration of output points in the closeness of the $DR=1$ line. Also, Based on the error distribution method, the best accuracy in the prediction was assigned to the ELM, according to the least simulation errors in range ($>20\%$) and the most error, in range ($<10\%$).

Regarding the RROC analysis, ELM Model was considered the superior model due to having the smallest range of changes in the coordinate axes of Fig. 7. Also, based on AOC values, the ELM model had the lowest value of AOC and was known as the most accurate model. Finally, the ELM was the fastest model in the aspect of consumed time in the testing phase.

Author Contributions Conceptualization and methodology, Simulations and results, Review and editing: Mojtaba Poursaeid; Simulations, Review, and editing: Amirhussain Poursaeid, Review and editing: Saeid Shabanlou.

Funding Not applicable; no funding was received.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors declare that they have no competing interests.

References

- Ahmadianfar I, Jamei M, Chu X (2020) A novel Hybrid Wavelet-Locally Weighted Linear Regression (W-LWLR) Model for Electrical Conductivity (EC) Prediction in Surface Water. *J Contam Hydrol* 232:103641. <https://doi.org/10.1016/j.jconhyd.2020.103641>
- Ahuja AK., Singh P, Singh V (2019) Physico-chemical Characterization of Ground Water with Reference to Water Quality Index and Their Seasonal Variation in Vicinity of Thermal Power Plant at Yamuna Nagar, Haryana. *Int J Adv Sci Res Manag* 4
- Alizadeh Z, Yazdi J, Moridi A (2018) Development of an Entropy Method for Groundwater Quality Monitoring Network Design. *Environ Process* 5:769–788. <https://doi.org/10.1007/s40710-018-0335-2>
- Ansell RO (2005) ION-SELECTIVE ELECTRODES | Water Applications. In: Worsfold P, Townshend A, Poole C (eds) *Encyclopedia of Analytical Science*, 2nd edn. Elsevier, Amsterdam, pp 540–545. <https://doi.org/10.1016/B0-12-369397-7/00298-3>
- Arora S, Keshari AK (2021) ANFIS-ARIMA modelling for scheming re-aeration of hydrologically altered rivers. *J Hydrol* 601:126635. <https://doi.org/10.1016/j.jhydrol.2021.126635>
- Asgari G, Komijani E, Seid-Mohammadi A, Khazaei M (2021) Assessment the Quality of Bottled Drinking Water Through Mamdani Fuzzy Water Quality Index. *Water Resour Manag* 35:5431–5452. <https://doi.org/10.1007/S11269-021-03013-Z>
- Azad S, Debnath S, Rajeevan M (2015) Analysing predictability in Indian monsoon rainfall: A data analytic approach. *Environ Process* 2:717–727. <https://doi.org/10.1007/S40710-015-0108-0/TABLES/6>
- Bin HG, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multiclass classification. *IEEE Trans Syst Man Cybern Part B Cybern* 42:513–529. <https://doi.org/10.1109/TSMCB.2011.2168604>
- Bin HG, Zhu QY, Siew CK (2004) Extreme learning machine: A new learning scheme of feed-forward neural networks. *IEEE Int Conf Neural Networks Conf Proc* 2:985–990. <https://doi.org/10.1109/IJCNN.2004.1380068>
- Bin HG, Zhu QY, Siew CK (2006) Extreme learning machine: Theory and applications. *Neurocomputing* 70:489–501. <https://doi.org/10.1016/J.NEUCOM.2005.12.126>
- Bonakdari H, Gholami A, Mosavi A et al (2020) A novel comprehensive evaluation method for estimating the bank profile shape and dimensions of stable channels using the maximum entropy principle. *Entropy* 22:1–23. <https://doi.org/10.3390/e22111218>
- Campbell C (2002) Kernel methods: A survey of current techniques. *Neurocomputing* 48:63–84. [https://doi.org/10.1016/S0925-2312\(01\)00643-9](https://doi.org/10.1016/S0925-2312(01)00643-9)
- Cao X, Liu Y, Wang J et al (2020) Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network. *Aquac Eng* 91:102122. <https://doi.org/10.1016/J.AQUAENG.2020.102122>
- Chang CL, Chung SC, Fu WL, Huang CC (2021) Artificial intelligence approaches to predict growth, harvest day, and quality of lettuce (*Lactuca sativa* L.) in a IoT-enabled greenhouse system. *Biosyst Eng* 212:77–105. <https://doi.org/10.1016/J.BIOSYSTEMSENG.2021.09.015>
- Cristianini N, Shawe-Taylor J (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, England. <https://doi.org/10.1017/CBO9780511801389>
- Dibike YB, Velickov S, Solomatine D, Abbott MB (2001) Model Induction with Support Vector Machines: Introduction and Applications. *J Comput Civ Eng* 15:208–216. [https://doi.org/10.1061/\(ASCE\)0887-3801\(2001\)15:3\(208\)](https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208))
- Elkiran G, Nourani V, Abba SI (2019) Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. *J Hydrol* 577:123962. <https://doi.org/10.1016/J.JHYDROL.2019.123962>
- Fluss R, Reiser B, Faraggi D (2012) Adjusting ROC curves for covariates in the presence of verification bias. *J Stat Plan Inference* 142:1–11. <https://doi.org/10.1016/J.JSPI.2011.03.016>
- Guneshwor L, Eldho TI, Vinod Kumar A (2018) Identification of Groundwater Contamination Sources Using Meshfree RPCM Simulation and Particle Swarm Optimization. *Water Resour Manag* 32:1517–1538. <https://doi.org/10.1007/S11269-017-1885-1>
- Hameed K, Chai D, Rassau A (2021) Texture-based latent space disentanglement for enhancement of a training dataset for ANN-based classification of fruit and vegetables. *Inf Process Agric*. <https://doi.org/10.1016/J.INPA.2021.09.003>
- Harris G (2009) Salinity *Enycl Inl Waters* 1:79–84. <https://doi.org/10.1016/B978-012370626-3.00103-4>
- Heddam S, Lamda H, Filali S (2016) Predicting Effluent Biochemical Oxygen Demand in a Wastewater Treatment Plant Using Generalized Regression Neural Network Based Approach: A Comparative Study. *Environ Process* 3:153–165. <https://doi.org/10.1007/S40710-016-0129-3>

- Hernández-Orallo J (2013) ROC curves for regression. *Pattern Recognit* 46:3395–3411. <https://doi.org/10.1016/J.PATCOG.2013.06.014>
- Hou Z, Guertler CA, Okamoto RJ et al (2021) Estimation of the mechanical properties of a transversely isotropic material from shear wave fields via artificial neural networks. *J Mech Behav Biomed Mater* 126:105046. <https://doi.org/10.1016/J.JMBBM.2021.105046>
- Jaddi NS, Abdullah S (2017) A cooperative-competitive master-slave global-best harmony search for ANN optimization and water-quality prediction. *Appl Soft Comput* 51:209–224. <https://doi.org/10.1016/J.ASOC.2016.12.011>
- Jamei M, Ahmadianfar I, Chu X, Yaseen ZM (2020) Prediction of surface water total dissolved solids using hybridized wavelet-multigene genetic programming: New approach. *J Hydrol* 589:125335. <https://doi.org/10.1016/J.JHYDROL.2020.125335>
- Jang J, Baek J, Leigh SB (2019) Prediction of optimum heating timing based on artificial neural network by utilizing BEMS data. *J Build Eng* 22:66–74. <https://doi.org/10.1016/J.JOBE.2018.11.012>
- Jeihouni M, Toomanian A, Mansourian A (2020) Decision Tree-Based Data Mining and Rule Induction for Identifying High Quality Groundwater Zones to Water Supply Management: a Novel Hybrid Use of Data Mining and GIS. *Water Resour Manag* 34:139–154. <https://doi.org/10.1007/S11269-019-02447-W/FIGURES/11>
- Kadkhodazadeh M, Farzin S (2021) A Novel LSSVM Model Integrated with GBO Algorithm to Assessment of Water Quality Parameters. <https://doi.org/10.21203/RS.3.RS-465707/V1>
- Kheradpisheh Z, Talebi A, Rafati L et al (2015) Groundwater quality assessment using artificial neural network A case study of Bahabad plain, Yazd, Iran. *Desert* 20:65–71. <https://doi.org/10.22059/JDESERT.2015.54084>
- Liang NY, Bin HG, Saratchandran P, Sundararajan N (2006) A fast and accurate online sequential learning algorithm for feed-forward networks. *IEEE Trans Neural Networks* 17:1411–1423. <https://doi.org/10.1109/TNN.2006.880583>
- Lukawska-Matuszewska K, Urbański JA (2014) Prediction of near-bottom water salinity in the Baltic Sea using Ordinary Least Squares and Geographically Weighted Regression models. *Estuar Coast Shelf Sci* 149:255–263. <https://doi.org/10.1016/J.ECSS.2014.09.003>
- Lyu W, Liu J (2021) Artificial Intelligence and emerging digital technologies in the energy sector. *Appl Energy* 303:117615. <https://doi.org/10.1016/J.APENERGY.2021.117615>
- Majumder P, Eldho TI (2020) Artificial Neural Network and Grey Wolf Optimizer Based Surrogate Simulation-Optimization Model for Groundwater Remediation. *Water Resour Manag* 34:763–783. <https://doi.org/10.1007/S11269-019-02472-9>
- Mokhatab S, Poe WA, Mak JY (2019) Utility and Offsite Systems in Gas Processing Plants. In: Mokhatab S, Poe WA, Mak JY (eds) *Handbook of Natural Gas Transmission and Processing*. Elsevier, Amsterdam, pp 537–578. <https://doi.org/10.1016/B978-0-12-815817-3.00018-6>
- Mtata TA (2003) Food. In: Hazeltine B, Bull C (eds) *Field Guide to Appropriate Technology*. Elsevier, Amsterdam, pp 277–480. <https://doi.org/10.1016/B978-012335185-2/50047-4>
- Mustapha A, Abdu A (2012) Application of Principal Component Analysis & Multiple Regression Models in Surface Water Quality Assessment. *J Environ Earth Sci* 2:16–23
- Niu C, Tan K, Jia X, Wang X (2021) Deep learning based regression for optically inactive inland water quality parameter estimation using airborne hyperspectral imagery. *Environ Pollut* 286:117534. <https://doi.org/10.1016/J.ENVPOL.2021.117534>
- Noori N, Kalin L, Isik S (2020) Water quality prediction using SWAT-ANN coupled approach. *J Hydrol* 590:125220. <https://doi.org/10.1016/J.JHYDROL.2020.125220>
- Patki VK, Jahagirdar S, Patil YM et al (2021) Prediction of water quality in municipal distribution system. *Mater Today Proc*. <https://doi.org/10.1016/J.MATPR.2021.02.826>
- Poursaeid AH, Namdari F (2022) Real-time voltage stability monitoring using weighted least square support vector machine considering overcurrent protection. *Int J Electr Power Energy Syst* 136:107690. <https://doi.org/10.1016/J.IJEPES.2021.107690>
- Poursaeid M, Mastouri R, Shabanlou S (2020) Najarchi M (2020) Estimation of total dissolved solids, electrical conductivity, Salinity and groundwater levels using novel learning machines. *Environ Earth Sci* 79:1–25. <https://doi.org/10.1007/S12665-020-09190-1>
- Poursaeid M, Mastouri R, Shabanlou S, Najarchi M (2021) Modelling qualitative and quantitative parameters of groundwater using a new wavelet conjunction heuristic method: wavelet extreme learning machine versus wavelet neural networks. *Water Environ J* 35:67–83. <https://doi.org/10.1111/WEJ.12595>
- Qu X, Chen Y, Liu H et al (2020) A holistic assessment of water quality condition and spatiotemporal patterns in impounded lakes along the eastern route of China's South-to-North water diversion project. *Water Res* 185:116275. <https://doi.org/10.1016/J.WATRES.2020.116275>

- Reynolds J, Ahmad MW, Rezgui Y, Hippolyte JL (2019) Operational supply and demand optimisation of a multi-vector district energy system using artificial neural networks and a genetic algorithm. *Appl Energy* 235:699–713. <https://doi.org/10.1016/J.APENERGY.2018.11.001>
- Sada SO, Ikpeseni SC (2021) Evaluation of ANN and ANFIS modeling ability in the prediction of AISI 1050 steel machining performance. *Heliyon* 7:e06136. <https://doi.org/10.1016/J.HELIYON.2021.E06136>
- Sapankevych N, Sankar R (2009) Time series prediction using support vector machines: A survey. *IEEE Comput Intell Mag* 4:24–38. <https://doi.org/10.1109/MCI.2009.932254>
- Schölkopf B, Smola AJ (2002) *Learning with Kernels: Support Vector Machines, Optimization, and Beyond*. Adaptive computation and machine learning. MIT Press, Cambridge, Regularization, p 626
- Serrano-Finetti E, Aliau-Bonet C, López-Lapeña O, Pallàs-Areny R (2019) Cost-effective autonomous sensor for the long-term monitoring of water electrical conductivity of crop fields. *Comput Electron Agric* 165:104940. <https://doi.org/10.1016/j.compag.2019.104940>
- Shahid ES, Ehteshami M (2015) Application of artificial neural networks to estimating DO and salinity in San Joaquin River basin. *Desalination Water Treat* 57:4888–4897. <https://doi.org/10.1080/19443994.2014.995713>
- Sharafati A, Asadollah SBHS, Hosseinzadeh M (2020) The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. *Process Saf Environ Prot* 140:68–78. <https://doi.org/10.1016/J.PSEP.2020.04.045>
- Shi B, Wang P, Jiang J, Liu R (2018) Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies. *Sci Total Environ* 610:1390–1399. <https://doi.org/10.1016/j.scitotenv.2017.08.232>
- Sparks DL (2003) *The Chemistry of Saline and Sodic Soils*. In: Sparks DL (ed) *Environmental Soil Chemistry*. Elsevier, Amsterdam, pp 285–300. <https://doi.org/10.1016/B978-012656446-4/50010-4>
- Suykens JAK, Van Gestel T, De Brabanter J et al (2002) *Least Squares Support Vector Machines*. World Scientific, Singapore. <https://doi.org/10.1142/5089>
- Tiyasha A, Tung TM, Yaseen ZM (2021) Deep Learning for Prediction of Water Quality Index Classification: Tropical Catchment Environmental Assessment. *Nat Resour Res* 30:4235–4254. <https://doi.org/10.1007/S11053-021-09922-5>
- Tokachichu J, Gaddam TRD (2021) Performance analysis of a transmission line connected with UPFC designed with three level cascaded H bridge inverter with generalized SVM technique using PI, FUZZY LOGIC, ANN and ANFIS controllers. *Mater Today Proc*. <https://doi.org/10.1016/j.matpr.2021.07.338>
- Valyon J, Horvath G (2007) Extended Least Squares LS-SVM. *World Acad Sci Eng Technol* 3:234–242
- Yang R, Yang S, Lin Y et al (2021) Miniature microplasma carbon optical emission spectrometry for detection of dissolved oxygen in water. *Microchem J* 171:106862. <https://doi.org/10.1016/J.MICROC.2021.106862>
- Yang X, Zhang H, Zhou H (2014) A Hybrid Methodology for Salinity Time Series Forecasting Based on Wavelet Transform and NARX Neural Networks. *Arab J Sci Eng* 39:6895–6905. <https://doi.org/10.1007/S13369-014-1243-Z>
- Ye Q, Yang X, Chen C, Wang J (2019) River Water Quality Parameters Prediction Method Based on LSTM-RNN Model. In: *Proc 31st Chinese Control Decis Conf CCDC*. IEEE, p 3024–3028. <https://doi.org/10.1109/CCDC.2019.8832885>
- Zhang Y, Gao X, Smith K et al (2019) Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. *Water Res* 164:114888. <https://doi.org/10.1016/J.WATRES.2019.114888>
- Zhang Y, Wu L, Deng L, Ouyang B (2021) Retrieval of Water Quality Parameters from Hyperspectral Images Using a Hybrid Feedback Deep Factorization Machine Model. *Water Res* 204:117618. <https://doi.org/10.1016/J.WATRES.2021.117618>
- Zhu S, Heddiam S (2019) Modelling of Maximum Daily Water Temperature for Streams: Optimally Pruned Extreme Learning Machine (OPELM) versus Radial Basis Function Neural Networks (RBFNN). *Environ Process* 6:789–804. <https://doi.org/10.1007/S40710-019-00385-8>