Hindawi
Complexity
Volume 2020, Article ID 8844367, 22 pages
https://doi.org/10.1155/2020/8844367

Research Article

Training and Testing Data Division Influence on Hybrid Machine Learning Model Process: Application of River Flow Forecasting

Hai Tao[1], **Ali Omran Al-Sulttani**[2], **Ameen Mohammed Salih Ameen**[2], **Zainab Hasan Ali**[3], **Nadhir Al-Ansari**[4], **Sinan Q. Salih**[5] and **Reham R. Mostafa**[6]

[1]*Computer Science Department, Baoji University of Arts and Sciences, Baoji, Shaanxi, China*
[2]*Department of Water Resources Engineering, College of Engineering, University of Baghdad, Baghdad, Iraq*
[3]*Civil Engineering Department, College of Engineering, University of Diyala, Baquba, Iraq*
[4]*Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, 97187 Lulea, Sweden*
[5]*Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam*
[6]*Information Systems Department, Faculty of Computers and Information Sciences, Mansoura University, Mansoura 35516, Egypt*

Correspondence should be addressed to Sinan Q. Salih; sinanq.salih@duytan.edu.vn

Received 17 August 2020; Revised 18 September 2020; Accepted 1 October 2020; Published 30 October 2020

Academic Editor: Shamsuddin Shahid

Copyright © 2020 Hai Tao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The hydrological process has a dynamic nature characterised by randomness and complex phenomena. The application of machine learning (ML) models in forecasting river flow has grown rapidly. This is owing to their capacity to simulate the complex phenomena associated with hydrological and environmental processes. Four different ML models were developed for river flow forecasting located in semiarid region, Iraq. The effectiveness of data division influence on the ML models process was investigated. Three data division modeling scenarios were inspected including 70%–30%, 80%–20, and 90%–10%. Several statistical indicators are computed to verify the performance of the models. The results revealed the potential of the hybridized support vector regression model with a genetic algorithm (SVR-GA) over the other ML forecasting models for monthly river flow forecasting using 90%–10% data division. In addition, it was found to improve the accuracy in forecasting high flow events. The unique architecture of developed SVR-GA due to the ability of the GA optimizer to tune the internal parameters of the SVR model provides a robust learning process. This has made it more efficient in forecasting stochastic river flow behaviour compared to the other developed hybrid models.

1. Introduction

The hydrological, environmental, and climatological processes related to different components of the hydrologic cycle such as rainfall, evaporation, infiltration, groundwater, and river flow are embedded with high nonlinearity, nonstationery, and redundancy [1, 2]. Mathematical models are generally used to address the different forms of nonlinearity and model different hydrological processes [3, 4]. A large number of hydrological models have been developed and successfully applied for forecasting different components of hydrological cycles [5]. Among them, the hydrological model for forecasting river flow has received the highest attention due to its enormous importance [6]. Being that river flow pattern is difficult to be fully understood due to the temporal and spatial changes in basin characteristics and variabilities in the rainfall-runoff process, univariate river flow simulation has become a trending topic in the field of hydrology [7]. River flow modeling in a particular climate environment (e.g., semiarid) is highly complicated as it is associated with regional climate and human interventions. Significant changes in climate have been witnessed across the globe in recent years. Streamflow time series is dynamic, complex, and presenting nonlinear and randomness phenomena. These characteristics make the forecasting process challenging for most of the hydrological researchers [8, 9].

Accurate long-term forecasting of river flow at monthly and yearly scale is very important for the planning and operation of water reservoir, agricultural and irrigation water management, estimation of catchment water balance, estimating minimum instream environmental flow, and other purposes [10, 11]. The accurate short-term (real-time) forecasting of river flow such as hourly or daily time step is important for flood and/or water scarcity forecasting in order to minimize and mitigate their effects on infrastructure and public health [12]. In addition, this river flow forecast makes it possible to predict the minimum instream environmental flow that is primarily used to sustain organisms' reproduction and growth and provide optimum habitat area [13].

Several developments have been recorded over the years in the application of machine learning (ML) models, artificial intelligence (AI) models, and big data mining technology for the solution of related hydrological engineering problems [14–18]. Being that these models do not depend on a physical meaning, they are suitable for solving problems, which are influenced by several complex factors, such as nonlinear natural processes and forecasting problems [19]. These models have also been found suitable for the solution of hydrological problems [20]. The ML models, unlike the physically based models, can mimic the actual trend of hydrological time series data by autonomously learning the various hydrological processes. However, ML models need a low level of expertise for the implementation and they can provide a fast solution compared to physically based models.

Despite the suitability of the ML models in hydrological studies, they are still prone to several drawbacks, such as prolonged computation time and insufficient feature extraction capability [21]. The recent interest in the ML models has exposed many other drawbacks of the classical ML models like artificial neural network (ANN), support vector regression (SVR), adaptive neurofuzzy inference (ANFIS), and random forest (RF) [22], which include trapping at local optima and gradient disappearance. Therefore, exploration of new robust and reliable versions of ML models for the modeling of various hydrological phenomena is always the motivation of hydrologists and soft computing scientists [20]. Recently, the new era of ML models is configured in the form of hybridized models where in integral of tuning parameter algorithms, it is conducted for solving the internal parameters using some bio-inspired or mathematical optimizers. The hybrid ML model has been emerged as the sought-after model due to its capability to overcome the drawbacks of standalone ML models [23]. It has been successfully applied in recent years for complex hydrological problems [24, 25].

The traditional ML models can build their learning processes using trial and error procedure that is associated with the possibility of the limited learning process. Hence, introducing the new optimization approaches can solve this problem and provide a reliable and robust learning mechanism. Risks associated with flooding can be reduced via accurate modeling of river flow time series dynamics. This can also enhance the capability of proper management of reservoirs during droughts [26]. The accurate forecasting of river flow time series should preferably be based on the existing long data with memory networks. Hybrid ML models as a robust methodology provide an excellent learning memory that could better model river flow patterns and provide better forecasting. During training the predictive model, the data is divided into the training and testing phase. A low ratio of training data may decrease the performance of the model, whereas the high ratio leads to overfitting. In both cases, the models get bad performance and unacceptable results. So, choosing the best ratio of data division is considered a challenging task in developing a machine learning model [27].

The main objective of the present study is to investigate the impact of training and testing data divisions on the process of several hybrid ML models including hybridized ANN and SVR with genetic algorithm (GA) and hybridized SVR and RF with the grid search algorithm. The development of the introduced models is investigated for river flow forecasting using historical data, which belongs to the Tigris River in semiarid climate of Iraq. The capacity of the developed model is examined to solve the complexity of river flow by using statistical metrics and graphical presentation.

The modeling procedure is structured based on different antecedent values of river flow and is defined as the matrix attributes for univariate modeling. Three data division scenarios for the training and testing dataset were inspected. The obtained results are discussed comprehensively and analysed comparatively to reveal the forecasting ability of different models. Thereafter, the forecasted river flow was used to estimate minimum instream environmental flow that is primarily used to sustain organisms' reproduction and growth and provide optimum habitat area.

2. Description of Study Areas and Data

The Tigris River is one of the largest rivers in Middle East. The total length of the river is about 1718 km which is shared by Turkey, Syria, and Iraq. About 85% of the total basin of Tigris River (253,000 km) lies in Iraq. The Tigris River along with the Euphrates River supplies the major share of total water required for irrigation, human use, and industrial purposes for several cities in Iraq, Turkey, and Iran counties. The climate of the basin is predominantly arid; however, semiaridity is the main characteristic of the river. The average rainfall in the basin is 216 mm with most of the rainfall occurring during winter (December to February) [28]. However, the rainfall concentration is varied from the north, middle, and south of Iraq [29]. The temperature varies from maximum 45°C during summer to minimum 10°C in winter [29]. The monthly river flow data of Tigris River for the period January 1991 to November 2010 was obtained from the USGS Data Series 540 for the present study [30]. The mean monthly discharge and the standard deviation of Tigris River flow at Baghdad Station are 411.35 m^3/s and 234.52 m^3/s, respectively [31]. The location of Tigris River in the map of Iraq is presented in Figure 1.

3. Data Division Scenarios and Input Combinations

In order to utilize the machine learning methods for forecasting, the observed river flow data was split into two sets

Figure 1: The location of the case study on Tigris River, Iraq.

(training and testing). Three data divisions were inspected in this research including 70%–30%, 80%–20, and 90%–10%. This is owing to the fact that ML models can behave differently based on the supplied dataset span for the learning process and testing phases [32].

The identification of the input parameters for the ML model's development is an essential step prior to the models' learning process. In this study, as the intended is the river flow forecasting, lead times were determined using the statistical approaches including the autocorrelation function (ACF) and the partial autocorrelation function (PACF). The suitable input parameters were decided based on the highly influential lead times toward the one-step-ahead forecasting. Valuable temporal pattern may exist in observed data which can be used as an input pattern for the development of the forecasting model. ACF can be used to capture information on the temporal patterns existing in time series. ACF provides information about the correlation between two separate points on the time series at different time leads and provides information about the repeating patterns in the time series. Thus, it tells how the past river flow influences the future river flow.

The ACF value ranges between 1 and −1; the value near to 1 indicates near-perfect correlation and the value near to

−1 indicates complete anticorrelation. However, the time series data for the river flow is time-independent, and therefore, the correlation between the pair of values depends only on the time differences between the two points without considering their position in the series. In order to distinguish the most appropriate lead times of the time series that notably or substantially might influence the forecasting, the reference value of ACF should be identified. In this study, the ACF values equal to or more than 0.5 were considered for the selection of the time-lag pattern [33, 34]. Figure 2 illustrates the ACF values for different lead times up to 20 time differences. Figure 2 shows that the ACF values for Tigris River for the lead times Q_{t-1} to Q_{t-5} are more than 0.2. In other words, five input combinations were constructed (Model I, Model II, . . . , Model V).

4. Machine Learning Models Overview

In this study, four different hybrid machine learning models were developed including ANN-GA, SVR-GA, Grid-SVR, and Grid-RF for monthly river flow forecasting located in semiarid region, Iraq. RapidMiner software was used to develop machine learning models. RapidMiner is an open-source, free, and flexible software implemented by Java language. The program has been used in data analysis, application design, and developing complex models [35]. The development of predictive models using the above hybrid ML models is described in the following sections.

4.1. ANN-GA. Inspired by the human neural network, ANN was proposed and developed to simulate the human brain during learning. With high computing power, ANN can outperform the performance of the human in some cases. ANN was applied for solving many regression, clustering, and classification tasks.

ANN, as shown in Figure 3, consists of three types of layers: input layer, hidden layer, and output layer [36]. Each layer consists of a set of nodes called artificial neurons that perform elementary calculations [37]. Weighted connections connect neurons in the successive layers. During the training procedure of ANN, weights are defined and updated with the aim of minimizing the error between the actual output and the computed output. ANN has the ability to produce output with reasonable accuracy [38], if it has gone through an effective learning phase.

The backpropagation neural network (BPNN) proposed by Rumelhart et al. [39] is one of the most popular learning algorithms. BP aims to optimize the network parameters by minimizing the least square error between actual and computed output.

Inspired by Darwin's theory of biological evolution, GA was developed as a heuristic method for finding the function's optimal value [40, 41]. It represents one of the most popular forms of an evolutionary algorithm used to solve different optimization problems [42, 43]. GA is initialized by generating a random population of individuals (solutions) and tries to optimize these individuals by applying three successive operations:

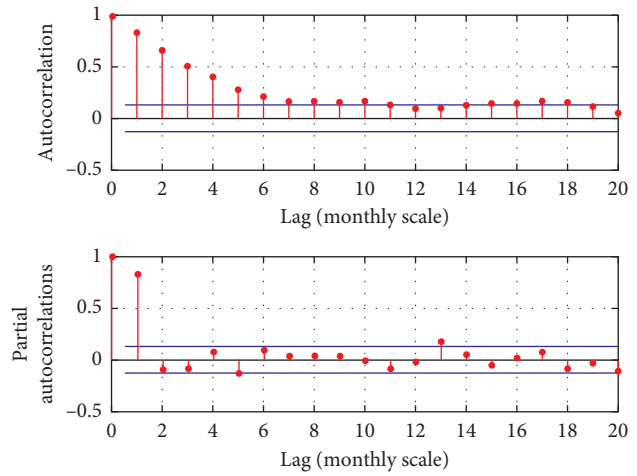

Figure 2: The statistics of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) for the studied time series river flow.

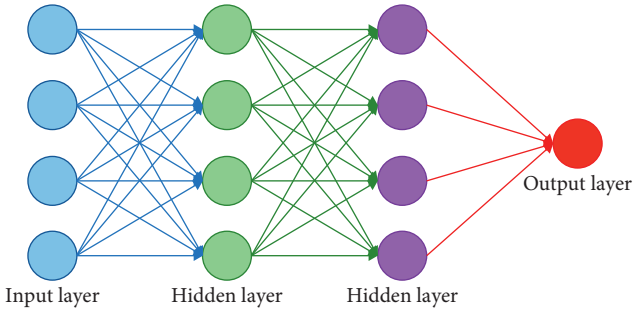

Figure 3: ANN architecture.

(i) Selection of the best individuals with high fitness values.

(ii) The crossover between two individuals to generate a new population.

(iii) The mutation performed by introducing random changes in individuals.

This procedure is repeated a certain number of times until the optimum values are obtained. In this section, the performance of BP was optimized using GA for tuning the parameters that affect the NN's performance. The training procedure for BP starts by using GA to perform a global search for network weight values. It refines an initially random set of weights to get a better estimate, and it is likely to be close to the global optimum [44, 45]. After that comes the role of BP in training in order to refine the solution provided by the GA to bring it to the optimum solution.

The general steps of ANN-GA can be summarized as follows. The flowchart of the GA-ANN method is illustrated in Figure 4. Initially, a feasible NN's topology was predefined through determining the number of neurons in the hidden layer [46, 47]. After that, steps to improve the performance of neural networks through GA algorithms begin as follows:

Step 1: initialize the random values for weight and bias (w_{ij} and b_i) according to initial network topology.

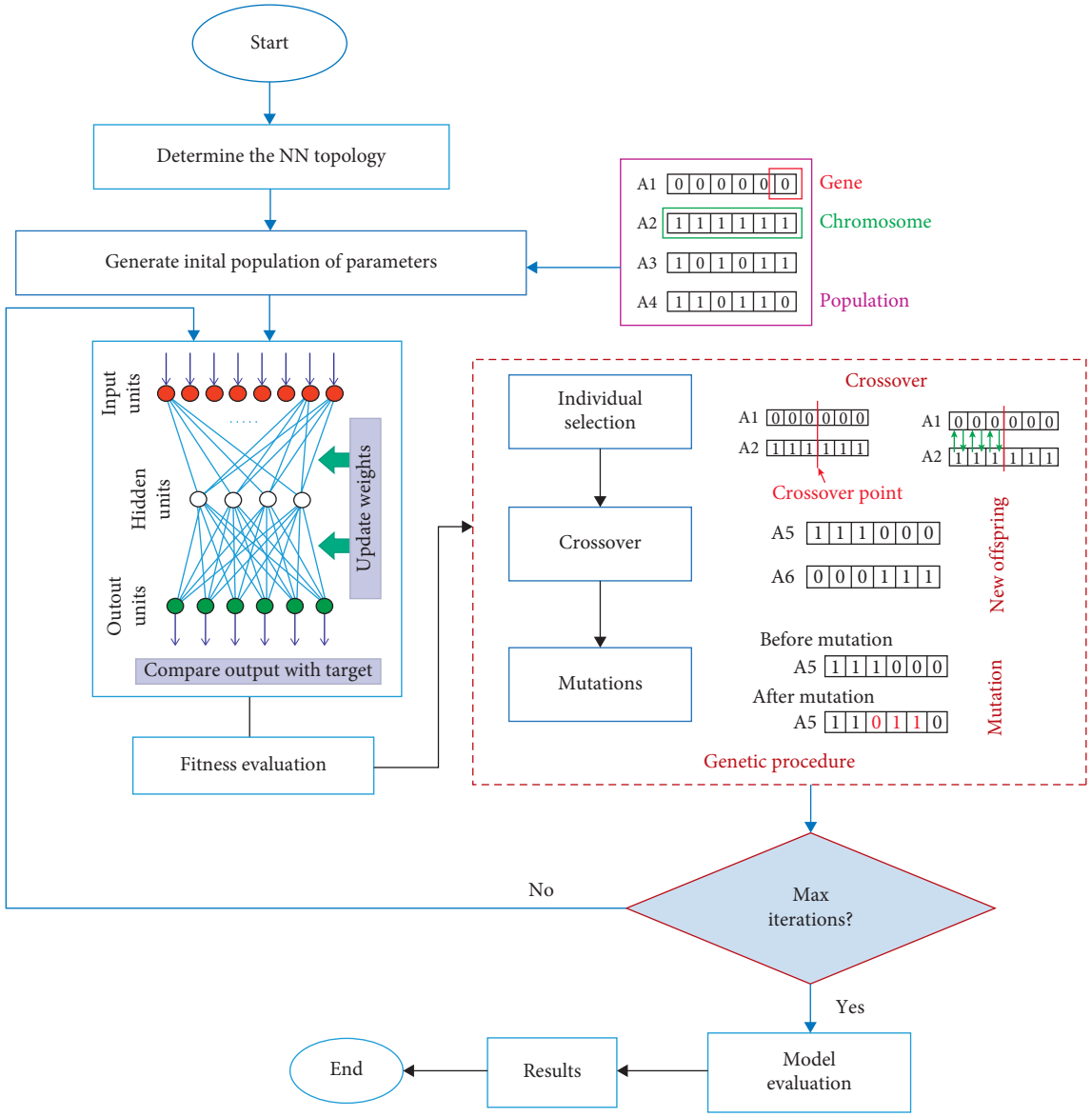

Figure 4: Flow diagram of ANN-GA model.

Step 2: generate an initial random population of parameters. Then perform the following steps until reaching the maximum number of iterations:

Step 2.1: calculate the fitness value by applying the fitness function on each individual in the population.
Step 2.2: update the network parameters (w_{ij} and b_i) based on the lowest error E.
Step 2.3: generate a new population through performing GA operations: selection, crossover, and mutation.

Step 3: obtain the optimal value of parameters from the last population

Step 4: train NN by PB training procedure (updating final weight and bias).

Step 5: evaluate the performance of NN using testing data. If the error is acceptable, stop and return the optimal model; otherwise, change the network topology (number of hidden layer's neurons) and return to step 1.

4.2. RF-Grid. Grid search is a traditional way for performing hyperparameters optimization for ML models. It is simply an exhaustive search method that sets up a grid of the possible values of the hyperparameters (Figure 5) and trains a model for each of the combinations [48]. In this method, all the possible combinations of the data are tried and tested using k-fold cross-validation technique.

Random forest (RF) is one of the most potent ensemble learning techniques developed by Breiman in 2001 [49] to solve different regression, classification, and clustering problem, and it exhibited excellent performance in many fields [50–52].

Despite the advantages of decision trees of its simplicity, ease of use, and interpretability [53, 54], it has many

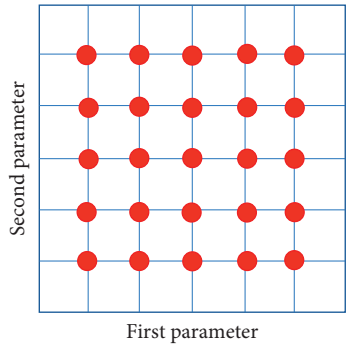

Figure 5: Grid search method.

limitations, such as their suboptimal performance and lack of robustness. Therefore, RF can overcome the limitations of traditional decision trees by combining the performance of many randomized, decorrelated decision trees to predict or classify the variable to a specific class. RF is a boosting technique in which it boosted the performance of a number of decision trees via a voting scheme. An example of RF is shown in Figure 6 in which the forest consists of n trees and a voter.

Regarding the main advantages of RF in regression tasks, it includes (i) bootstrap resampling, (ii) random features selection, and (iii) out-of-bag error estimation.

Suppose there are n trees $T_1(X), T_2(X), \ldots, T_n(X)$, where $X = x_1, x_2, \ldots, x_m$ is a m-dimension vector of inputs. The prediction value of each decision tree is $\widehat{Y}_i = T_i(x)$. While the final prediction output Y results from aggregating the outputs of all randomly generated trees. The aggregation process is accomplished in the classification task by taking the majority votes, while in regression task, it is accomplished by taking the average:

$$Y = \sum_{i=1}^{n} \widehat{Y}_i(x) = \frac{1}{n} \sum_{i=1}^{n} T_i(x). \quad (1)$$

The RF algorithm can be summarized as follows [55, 56]:

(i) Generate n number of trees by selecting randomly different bootstrap samples from the training data. The out-of-bag samples are the samples that are not selected.

(ii) For each bootstrap sample, grow a full decision tree to the maximum size without pruning. In splitting the nodes of the tree, a specific number of features were selected randomly instead of choosing all features (this refers to a random feature selection).

(iii) Repeat step 2 until forming a randomly generated forest consisting of n decision trees.

(iv) Predict the new data by applying the n trees and aggregate the results.

In this section, the grid search algorithm was used to optimize the RF algorithm by tuning the hyperparameters of it. The primary hyperparameters of RF that affect its performance are (1) the number of trees in the forest that must be generated before taking the maximum voting or average of predictions and (2) the maximum number of features to split in each node of the tree. The hybrid structure of RF and grid search was considered in the following steps (Figure 7):

(i) Step 1: define RF searching parameters range: maximum value, minimum value, and step size.

(ii) Step 2: build the grid search space on the coordinate system.

(iii) Step 3: build the RF models using all possible combinations of parameters and evaluate the performance of RF.

(iv) Step 4: return the multiple set of optimal parameters.

(v) Step 5: if the accuracy is satisfied, stop and return the optimal parameters; otherwise go to step 6.

(vi) Step 6: redefine the range of searching near the optimal parameters and reduce step size. Then, go to step 2.

(vii) Step 7: repeat steps 2–6 until the optimal hyperparameters values satisfying the accuracy were found.

(viii) Step 8: build RF forest model with the optimal parameters.

(ix) Step 9: predict the output value of data in testing set by RF model.

4.3. SVR-GA. Support vector regression (SVR) is introduced by Vapnik [57] as an extension of SVM for solving the regression problem. SVR is a very useful tool for prediction because of its ability to map nonlinear data space into a higher dimensional feature space [58].

Consider a learning dataset defined as $\{x_i, y_i\}$ $(i = 1, \ldots, n)$, where $x_i \in R^n$ is the input vector and $y_i \in R$ is its corresponding output vector. The main objective of SVR is to deduce the regression function $f(x)$ that describes the relationship between the input data x and the target output y with an error value that is less than epsilon ε deviation for all training data.

The SVR function can be written as follows [59]:

$$f(x) = w^T \varphi(x) + b, \quad (2)$$

where $f(x)$ is the computed output of x, $\varphi(x)$ is the nonlinear feature mapping function of inputs, and w and b are adjustable coefficients that represent the weight vector and intercept vector, respectively. The training of SVR is to find w and b values by minimizing the upper bound of the regression error.

Therefore, SVR is considered as an optimization problem that tries to make the regression function $f(x)$ as flat as possible by minimizing the value of w, which necessitates the minimization of Euclidean norm, that is, $\|w\|^2$.

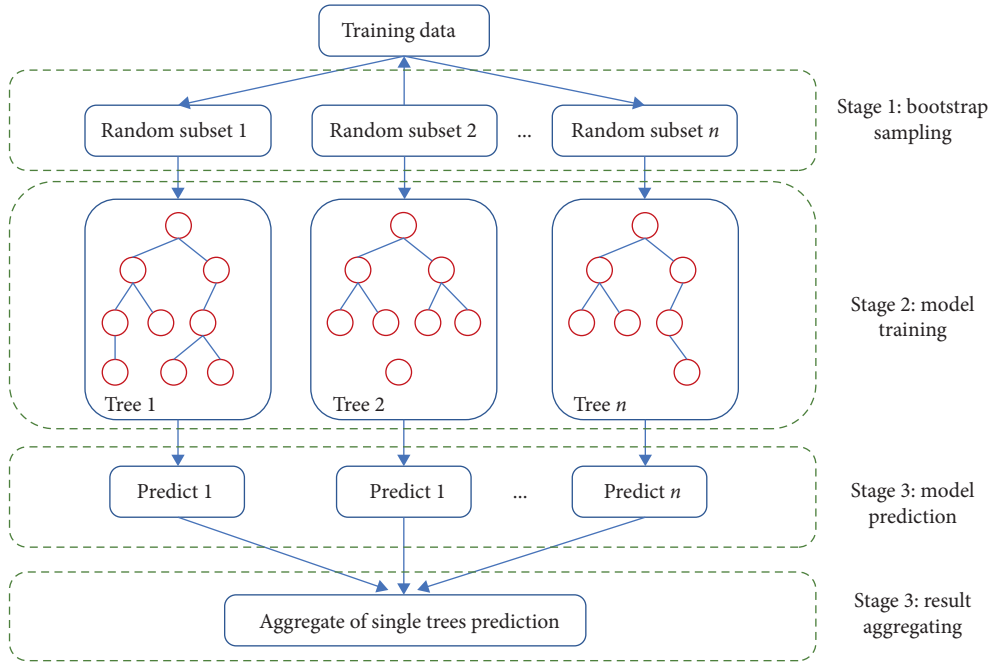

Figure 6: Example of random forest (RF) algorithm.

The optimization problem that is used to identify the regression problem is given as follows:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 \\ &\text{subject to} \quad \begin{aligned} &y_i - w^T\varphi(x) - b \le \varepsilon \\ &-y_i + w^T\varphi(x) + b \le \varepsilon. \end{aligned} \end{aligned} \tag{3}$$

The preceding equation applies if there is function $f(x)$ which approximates all pairs of (x_i, y_i) with an accuracy of ε. Besides, some mistakes that violate the conditions above are introduced. The inaccessible limitations of the optimization problem are addressed by the slack variables ξ_i and ξ_i^*. Equation (2) can, therefore, be rewritten as explained as follows:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ &\text{subject to} \quad \begin{aligned} &y_i - w^T\varphi(x) - b \le \varepsilon + \xi_i \\ &-y_i + w^T\varphi(x) + b \le \varepsilon + \xi_i^*, \\ &\xi_i, \xi_i^* \ge 0, \quad i = 1, 2, \ldots, n, \end{aligned} \end{aligned} \tag{4}$$

where C is defined as a nonnegative constant which expresses the box constraint that is responsible for monitoring data points above the ε-insensitive error range and which is also helpful when avoiding overfitting problems [58]. There are several commonly used kernel types in SVR that convert the nonlinear function in equation (1) into higher dimensional space. The radial basis function (RBF) has been widely used in previous studies [5, 60]:

$$\begin{aligned} f(x) &= w^T K(x, x_i) + b, \\ K(x, x_i) &= \exp\left(-\frac{|x - x_i|^2}{2\gamma^2}\right), \end{aligned} \tag{5}$$

where $K(x, x_i)$ is the kernel function and γ is the width parameter.

The performance of the SVR model depends on the hyperparameter tuning of the model: the regularization parameter (C), the epsilon (ε), and the kernel parameter (γ). These parameter values affect the performance of the model incredibly, in which the influence of these parameter values is presented as follows. The value of the first parameter (i.e., regularization parameter, also called box constraints) is used to decide the penalty of the approximation function. It must not be very small or large. If it is too small (large), it will cause underfitting (overfitting). In contrast, the decision boundary's maximum margin is controlled by the insensitivity loss function (ε). Finally, the kernel function controls the ability to make the model for dealing with nonlinear function responsible for transforming the nonlinear function into a more suitable function [61, 62].

Many researches were conducted to tune the hyperparameters of SVR using manual or grid search [63, 64]. However, this approach's complexity is increased incredibly when the width of the search space is increased. Also, this approach does not always get the best hyperparameter values for the model. Other approaches have been inspired to overcome these limitations, that is, the genetic algorithm (GA). GA is considered one of the power optimization algorithms proposed by Holland in 1975 and inspired by Darwin's theory. In this section, GA is inspired and used to find the hyperparameter of SVR. The proceeding procedure

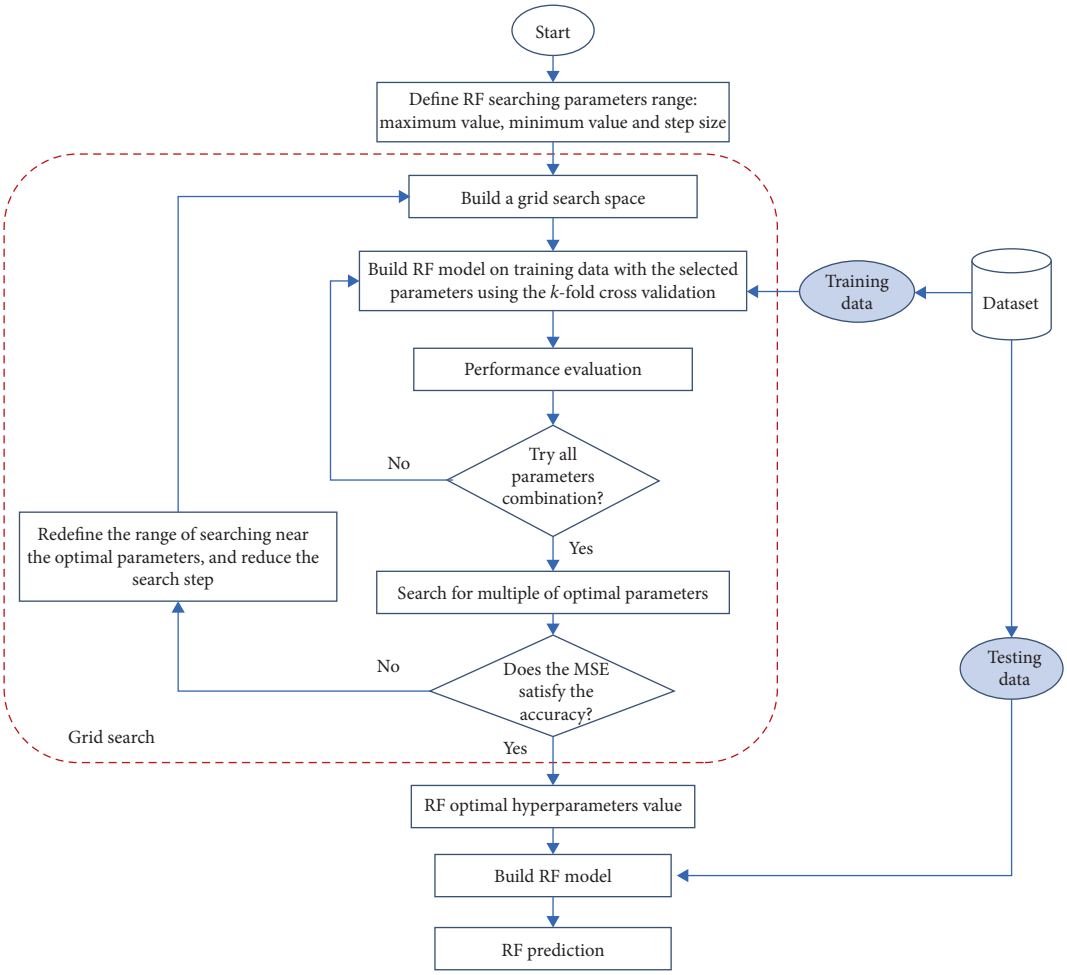

Figure 7: Flowchart of RF-grid model.

of optimizing SVR hyperparameters using GA was illustrated in Figure 8, and the steps involved are summarized as follows:

Step 1: initialize the SVR parameters (ε, C) and the kernel parameters γ. Code these parameters to create a chromosome directly.

Step 2: initialize the GA parameters randomly: population size, number of generations, mutation rate, and crossover rate.

Step 3: perform SVR model on training data using the k-fold cross validation.

Step 4: calculate the fitness value of each individual in the population according to the mean square error (MSE).

Step 5: generate new offspring parameters population through selection, crossover, and mutation.

Step 6: repeat steps 4-5 until ε, C, γ are satisfied with minimal error; otherwise, continue to optimize.

Step 7: output the optimal parameters found at the end of the generation. Train SVR model with these parameters

Step 8: predict the out of data in the testing dataset part by SVR model.

4.4. SVR-Grid. In this section, the grid search algorithm is used to obtain the optimal values of the SVR parameters (ε, C, and g). The grid search algorithm is based on trying all possible values of the parameters in a given space with a specified step distance. The cross-validation technique [65] derives the SVR model's parameters that improve its performance with the best accuracy.

The hybrid structure of SVR optimized by grid search is illustrated through a flowchart, as shown in Figure 9, and consists of the following steps:

Step 1: define the range of SVR searching parameters.

Step 2: initialize the values of the parameters and step distance.

Step 3: split dataset into two sets (training and testing sets)

Step 4: train SVR model using k-fold cross validation on the training dataset.

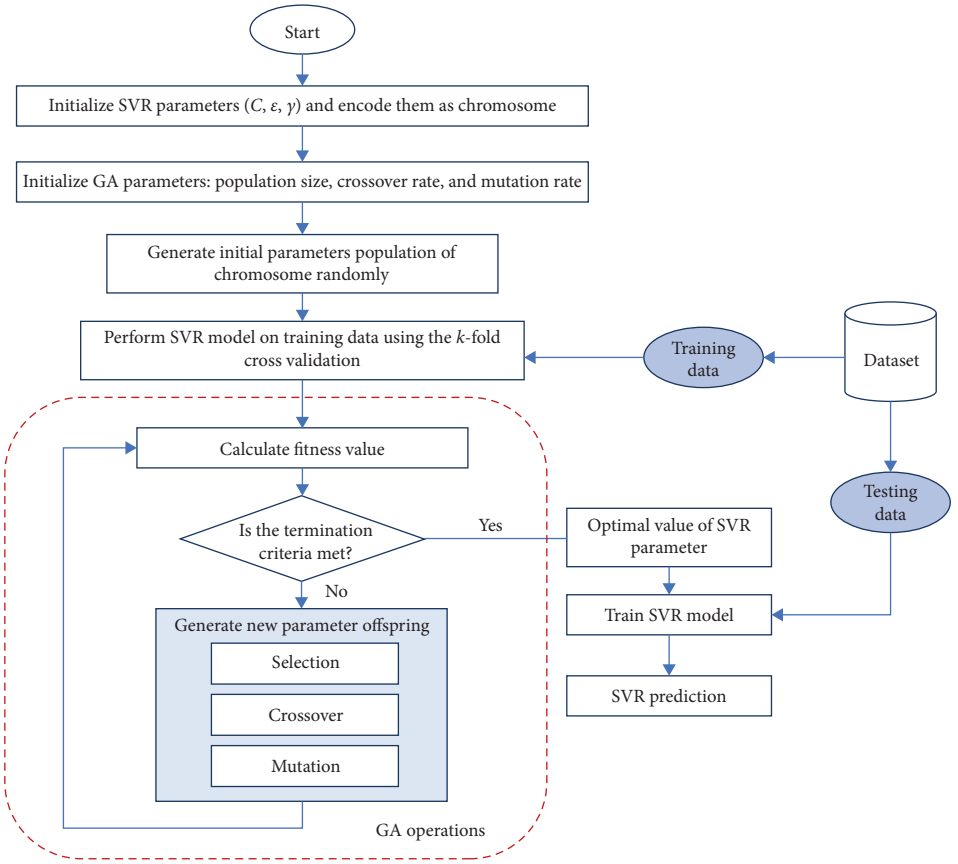

Figure 8: The flowchart of SVR-GA model.

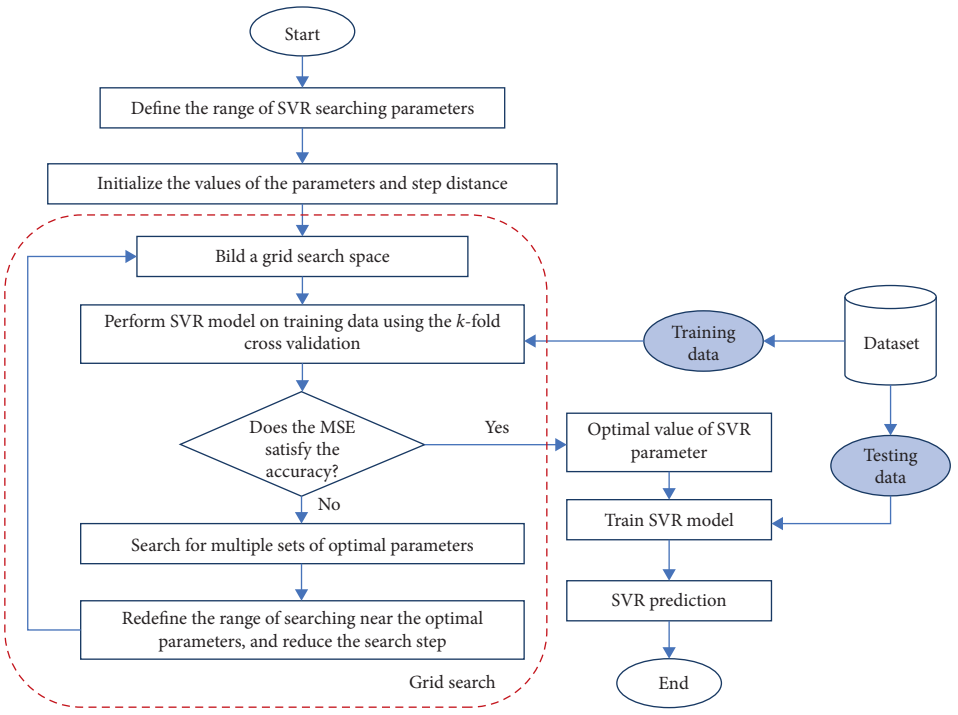

Figure 9: The flowchart of SVR-grid model.

Step 5: if the MSE satisfies the accuracy, then select the sets of parameters with the minimum MSE; otherwise, continue the optimize and go to step 5.

Step 6: search for multiple sets of optimal parameters.

Step 7: redefine the range of searching near the optimal parameters and reduce the search step; then go to step 4.

5. Data Analysis and Results

5.1. Model Performance Evaluation Using Statistical Indices. The effectiveness of the proposed modeling techniques was examined by comparing the forecasted river flow with the observed river flow data. It should be noted that the data used in this investigation was continuous and without any missing value.

The performance of the models in forecasting river flow one month ahead was forecasted and evaluated using several statistical metrics. Five statistical metrics were used to measure the performance of the predictive models in forecasting river flow during model testing, namely, Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R^2) [66, 67]. These statistical metrics were widely used by researchers to evaluate the performance of predictive models in hydrological and machine learning models [68]. These measurements are considered a good indicator of the accuracy and robustness of the model:

$$\begin{aligned}
\mathrm{ME} &= \frac{\sum_{i=1}^{N} y_o - y_p}{N},\\
\mathrm{RMSE} &= \sqrt{\sum_{i=1}^{N} \frac{\left(y_o - y_p\right)^2}{N}},\\
\mathrm{MAE} &= \frac{\sum_{i=1}^{N} \left|y_o - y_p\right|}{N},\\
\mathrm{MPE} &= \frac{1}{N}\sum_{i=1}^{N} \frac{y_o - y_p}{y_o},\\
\mathrm{MAPE} &= \frac{1}{N}\sum_{i=1}^{N} \frac{\left|y_o - y_p\right|}{y_o},\\
R^2 &= \left(\frac{\sum_{i=1}^{N}\left(y_o - \ddot{y}_o\right)\sum_{i=1}^{N}\left(y_p - \ddot{y}_p\right)}{\sqrt{\sum_{i=1}^{N}\left(y_o - \ddot{y}_o\right)^2 \sum_{i=1}^{N}\left(y_p - \ddot{y}_p\right)^2}}\right)^2,
\end{aligned} \tag{6}$$

where N is the length of the testing data set. y_o and y_p are the actual and forecasted river flow data. $\ddot{y}_o$ and $\ddot{y}_p$ are the mean values of the actual and forecasted river flow data.

Tables 1–3 present the statistical performance indicators (i.e., ME, RMSE, MAE, MPE, MAPE, and R^2) for the five constructed input combinations, training and testing phases, and the three data division scenarios. The tables showed that all the four hybrid ML models (ANN-GA, SVR-GA, Grid-SVR, and Grid-RF) are performed in a general good performance. However, they are varied from one input combination to another. That totally depends on the historical data memory provided by the lead time "antecedent river flow values." A notable enhancement is achieved using the hybridized SVR-GA which collaborates with the findings of several other studies established over the literature within hydrological engineering [69–72]. It is observed that the data division plays an essential role in the learning process of the developed ML models. Apparently, increasing the span of the training phase contributes to model predictability enhancement. In quantitative terms, the best results of forecasting are attained for the SVR-GA with (RMSE = 0.04, MAE = 0.03, and $R^2 = 0.95$). The SVR-GA model indicated boosting in forecasting accuracy; although, the other ML models obtained a reasonable prediction accuracy. This observation approved the capacity of the hybrid SVR-GA model to solve the complexity of river flow located in semiarid environment. The Coefficient of Determination (R^2) was achieved more than 0.90 for almost all the input combinations over the using the SVR-GA model although the performance of the SVR-GA model over the training phase was not superior (Table 3). This can explain the feasibility of the SVR-GA to be more potent.

5.2. Models Graphical Evaluation. Two different graphical presentations are hereby presented for the evaluation of the performance of the proposed models. The actual and the forecasted values of the river flow for Tigris River are presented using scatter plots and Taylor diagram. The scatter plots for the three modeling scenarios of data division (70%–30%, 80%–20, and 90%–10%) and the four developed ML models (ANN-GA, SVR-GA, Grid-SVR, and Grid-RF) are illustrated in Figures 10–12. It can be clearly observed that the 90%–10% data division scenario using the SVR-GA model attained the best match between the observed and forecasted river based on four-month antecedent values. For this particular data division and based on this lead time, the observed and the forecasted values by all the models were found to follow the similar trends. However, the matching between the observed and the forecasted values was found to be the best for SVR-GA model. To assess the efficacy of the models, it was justifiable to investigate the linear relationship between the time series of observed and the forecasted river flows for the testing period. The uniformity plot of river flow forecasted by SVR-GA attained the maximum determination coefficient ($R^2 = 0.96$).

Another graphical presentation that was investigated for the applied predictive models was Taylor diagram [73] (Figures 13–15). It is a distinguished two-dimension graphical presentation that accompanied three statistical metrics including RMSE, correlation, and standard deviation. In harmony with the presented statistical results and the scatter plot presentation, using four-month lead time predictors for the one month ahead river flow, the SVR-GA model indicated the best forecasting value owing to the location of the model results on the Taylor map for the 90%–10% data division scenario.

Table 1: The statistical performance metrics for the developed hybrid ML models over the 70–30% modeling data division scenario.

	ME	RMSE	MAE	MPE	MAPE	R^2
Model I						
			Training phase			
ANN-GA	59.32	256.43	175.66	−1261.38	1549.29	0.63
SVR-GA	10.34	247.05	154.56	−862.91	1285.20	0.68
Grid-RF	−2.33	123.24	77.09	−264.95	527.544	0.92
Grid-SVR	−15.35	254.68	159.67	−576.68	1198.01	0.58
			Testing phase			
ANN-GA	66.05	205.10	163.48	−1632.74	1886.53	0.69
SVR-GA	7.83	157.47	124.27	−806.19	1270.95	0.78
Grid-RF	−3.25	138.48	108.65	−380.57	977.75	0.80
Grid-SVR	−5.68	128.32	91.09	−338.22	751.28	0.91
Model II						
			Training phase			
ANN-GA	−33.47	331.87	211.77	−1370.87	1938.01	0.46
SVR-GA	−10.54	235.56	142.34	−664.55	1183.01	0.72
Grid-RF	−7.59	116.75	71.18	205.18	490.95	0.93
Grid-SVR	−5.96	234.96	141.50	−516.18	1057.58	0.72
			Testing phase			
ANN-GA	13.75	194.68	137.36	−744.31	1242.38	0.56
SVR-GA	11.80	133.54	108.60	−506.76	941.08	0.79
Grid-RF	−14.79	160.42	119.92	−335.94	936.55	0.73
Grid-SVR	42.39	136.07	106.04	−782.93	1098.75	0.77
Model III						
			Training phase			
ANN-GA	−44.17	245.56	150.87	−218.53	921.87	0.65
SVR-GA	−9.11	236.37	143.78	−815.63	1334.09	0.72
Grid-RF	7.27	125.03	83.15	−454.79	921.87	0.91
Grid-SVR	−33.94	273.67	174.09	−862.27	1445.52	0.64
			Testing phase			
ANN-GA	21.40	164.3153	124.08	−856.11	1274.74	0.80
SVR-GA	27.02	142.074	115.56	−804.35	1138.05	0.87
Grid-RF	0.97	143.7111	110.28	−400.61	958.01	0.82
Grid-SVR	32.84	152.1099	105.33	−836.43	1173.84	0.38
Model IV						
			Training phase			
ANN-GA	21.40	164.36	124.08	−856.11	1274.74	0.67
SVR-GA	27.02	142.07	115.57	−804.35	1138.05	0.69
Grid-RF	9.78	143.711	110.28	−400.61	958.01	0.94
Grid-SVR	32.84	152.11	105.34	−836.43	1173.84	0.75
			Testing phase			
ANN-GA	21.40	164.31	124.09	−856.11	1274.74	0.79
SVR-GA	27.02	142.07	115.56	−804.35	1138.05	0.82
Grid-RF	9.78	143.71	110.28	−400.61	958.01	0.76
Grid-SVR	32.84	152.11	105.34	−836.43	1173.84	0.76
Model V						
			Training phase			
ANN-GA	−2.65	212.08	159.06	−318.12	1537.58	0.63
SVR-GA	−10.60	238.59	132.55	−583.22	1113.42	0.66
Grid-RF	−21.21	132.55	79.53	−397.65	662.75	0.9
Grid-SVR	−23.86	185.57	106.04	−371.14	901.34	0.83
			Testing phase			
ANN-GA	26.51	185.57	132.55	1166.44	1564.09	0.83
SVR-GA	18.56	159.06	106.04	−715.77	1086.91	0.83
Grid-RF	2.65	132.55	106.04	−450.67	954.36	0.67
Grid-SVR	−7.95	106.04	79.53	−212.08	715.77	0.75

The minimum instream environmental flow for the Tigris River was calculated on the basis of the SVR-GA model's mean forecast flow obtained during the dry season. The minimum environmental flow was 215.46 $m^3\ s^{-1}$ (based on 19.22 percent of the average flow 1121 $m^3\ s^{-1}$). This flow refers to the minimum instream environmental flow required to maintain the organisms in the river [13].

Table 2: The statistical performance metrics for the developed hybrid ML models over the 80–20% modeling data division scenario.

	ME	RMSE	MAE	MPE	MAPE	R^2
Model I						
			Training phase			
ANN-GA	−38.54	252.65	154.75	−363.45	1060.36	0.65
SVR-GA	−9.52	244.25	154.08	−766.23	1293.02	0.69
Grid-RF	32.83	256.13	170.43	−877.83	1365.77	0.65
Grid-SVR	−17.71	243.98	151.04	−531.12	1127.40	0.66
			Testing phase			
ANN-GA	27.193	199.25	154.61	−394.37	767.53	0.84
SVR-GA	38.94	198.88	157.71	−585.79	952.09	0.72
Grid-RF	28.12	209.33	168.85	−577.06	1064.34	0.76
Grid-SVR	6.38	187.98	145.17	−531.11	779.82	0.89
Model II						
			Training phase			
ANN-GA	11.19	244.34	152.65	−2214.68	1275.26	0.66
SVR-GA	16.59	237.26	148.64	−2134.11	2628.70	0.685
Grid-RF	14.91	251.78	152.43	−736.97	1441.17	0.66
Grid-SVR	−11.58	228.31	135.41	−1862.16	2420.91	0.72
			Testing phase			
ANN-GA	21.89	154.13	118.15	−682.55	1102.56	0.80
SVR-GA	33.23	134.97	113.16	−1104.83	1417.38	0.83
Grid-RF	26.17	140.25	109.19	−578.82	1000.16	0.65
Grid-SVR	6.38	187.98	145.17	−531.12	779.825	0.75
Model III						
			Training phase			
ANN-GA	30.35	247.37	163.25	−1009.14	1386.72	0.67
SVR-GA	3.59	235.55	147.26	−761.82	1255.21	0.69
Grid-RF	−2.70	155.95	99.25	−536.86	855.20	0.86
Grid-SVR	−31.80	268.65	169.28	−800.78	1380.50	0.62
			Testing phase			
ANN-GA	93.78	154.59	124.46	−1037.85	1178.38	0.45
SVR-GA	23.51	142.13	112.42	−723.81	1071.39	0.80
Grid-RF	35.81	126.91	102.84	−662.98	966.92	0.81
Grid-SVR	18.62	135.36	96.18	−800.78	1176.49	0.54
Model IV						
			Training phase			
ANN-GA	−113.06	269.44	171.97	432.71	994.68	0.61
SVR-GA	−6.04	236.42	149.41	−800.34	1333.56	0.66
Grid-RF	−2.23	130.28	78.18	−436.18	686.68	0.90
Grid-SVR	−54.99	303.65	191.10	−950.77	1615.86	0.51
			Testing phase			
ANN-GA	−72.62	139.88	115.74	423.35	1070.50	0.71
SVR-GA	12.64	113.96	92.45	−420.64	749.46	0.92
Grid-RF	−26.20	153.04	118.39	−217.54	862.89	0.81
Grid-SVR	33.19	125.64	97.16	−926.32	1208.60	0.65
Model V						
			Training phase			
ANN-GA	−24.92	230.48	141.42	−642.60	1227.55	0.72
SVR-GA	−15.24	230.22	143.09	−631.34	1192.60	0.71
Grid-RF	−5.14	142.65	83.66	−317.06	592.59	0.89
Grid-SVR	−22.75	176.60	106.58	−330.21	860.40	0.83
			Testing phase			
ANN-GA	−38.02	149.35	119.27	−53.29	793.33	0.64
SVR-GA	23.04	138.42	105.01	−616.43	1093.90	0.74
Grid-RF	9.44	145.23	95.98	−469.15	896.69	0.44
Grid-SVR	17.41	133.36	94.31	−723.81	1214.13	0.55

Table 3: The statistical performance metrics for the developed hybrid ML models over the 90–10% modeling data division scenario.

	ME	RMSE	MAE	MPE	MAPE	R^2
Model I						
			Training phase			
ANN-GA	−43.82	250.85	156.29	−521.03	1181.23	0.64
SVR-GA	−8.58	234.82	148.27	−629.26	1166.91	0.67
Grid-RF	14.09	234.43	151.84	−822.16	1297.23	0.69
Grid-SVR	−15.84	233.79	146.80	−656.04	1215.01	0.69
			Testing phase			
ANN-GA	10.08	106.21	82.72	−312.59	684.22	0.88
SVR-GA	16.46	100.58	81.27	−954.43	1332.15	0.92
Grid-RF	18.98	84.577	71.86	−275.80	622.07	0.79
Grid-SVR	16.46	100.58	81.27	−954.43	1332.15	0.92
Model II						
			Training phase			
ANN-GA	−71.45	244.98	150.23	−44.06	970.63	0.66
SVR-GA	5.20	229.71	147.46	−813.07	1290.19	0.69
Grid-RF	−3.26	138.91	84.65	−404.59	718.67	0.89
Grid-SVR	−7.78	221.91	134.74	−627.86	1156.57	0.73
			Testing phase			
ANN-GA	−27.24	110.15	79.18	19.78	657.83	0.61
SVR-GA	44.15	106.22	85.60	−754.08	977.85	0.91
Grid-RF	16.26	116.19	91.95	−757.04	1163.95	0.77
Grid-SVR	40.57	75.62	64.78	−710.36	894.09	0.71
Model III						
			Training phase			
ANN-GA	−28.96	279.03	173.82	−901.58	1455.28	0.55
SVR-GA	7.97	231.47	148.60	−839.83	1316.29	0.70
Grid-RF	−6.04	143.03	89.34	−459.07	769.44	0.88
Grid-SVR	−42.45	266.44	165.71	−734.56	1350.49	0.59
			Testing phase			
ANN-GA	29.79	147.39	116.93	−889.74	1306.29	0.71
SVR-GA	48.075	88.21	72.67	−703.78	841.28	0.67
Grid-RF	−20.47	100.62	79.93	−18.08	592.86	0.71
Grid-SVR	−7.54	96.34	76.34	−483.47	931.94	0.75
Model IV						
			Training phase			
ANN-GA	−26.51	238.59	159.06	−503.69	1219.46	0.66
SVR-GA	−2.65	212.08	132.55	−715.77	1245.97	0.66
Grid-RF	−10.60	132.55	79.53	−371.14	689.26	0.89
Grid-SVR	−31.81	238.59	185.57	−477.18	1272.48	0.61
			Testing phase			
ANN-GA	−13.55	140.20	116.22	−583.19	1233.32	0.86
SVR-GA	−14.73	100.78	81.585	−214.02	670.30	0.96
Grid-RF	16.86	85.65	76.07	−375.48	739.69	0.83
Grid-SVR	17.86	82.19	62.39	−675.72	751.89	0.63
Model V						
			Training phase			
ANN-GA	−74.23	238.59	159.06	−291.61	1139.93	0.66
SVR-GA	−13.25	212.08	132.55	−715.77	1245.97	0.70
Grid-RF	−10.60	132.55	79.53	−397.65	662.75	0.90
Grid-SVR	−15.91	185.57	106.04	−556.71	1086.91	0.79
			Testing phase			
ANN-GA	7.953	106.04	79.53	−1113.42	1537.58	0.82
SVR-GA	10.604	79.53	53.02	−238.59	636.24	0.93
Grid-RF	−15.906	106.04	79.53	−212.08	715.77	0.50
Grid-SVR	5.302	87.48	106.04	−291.61	768.79	0.74

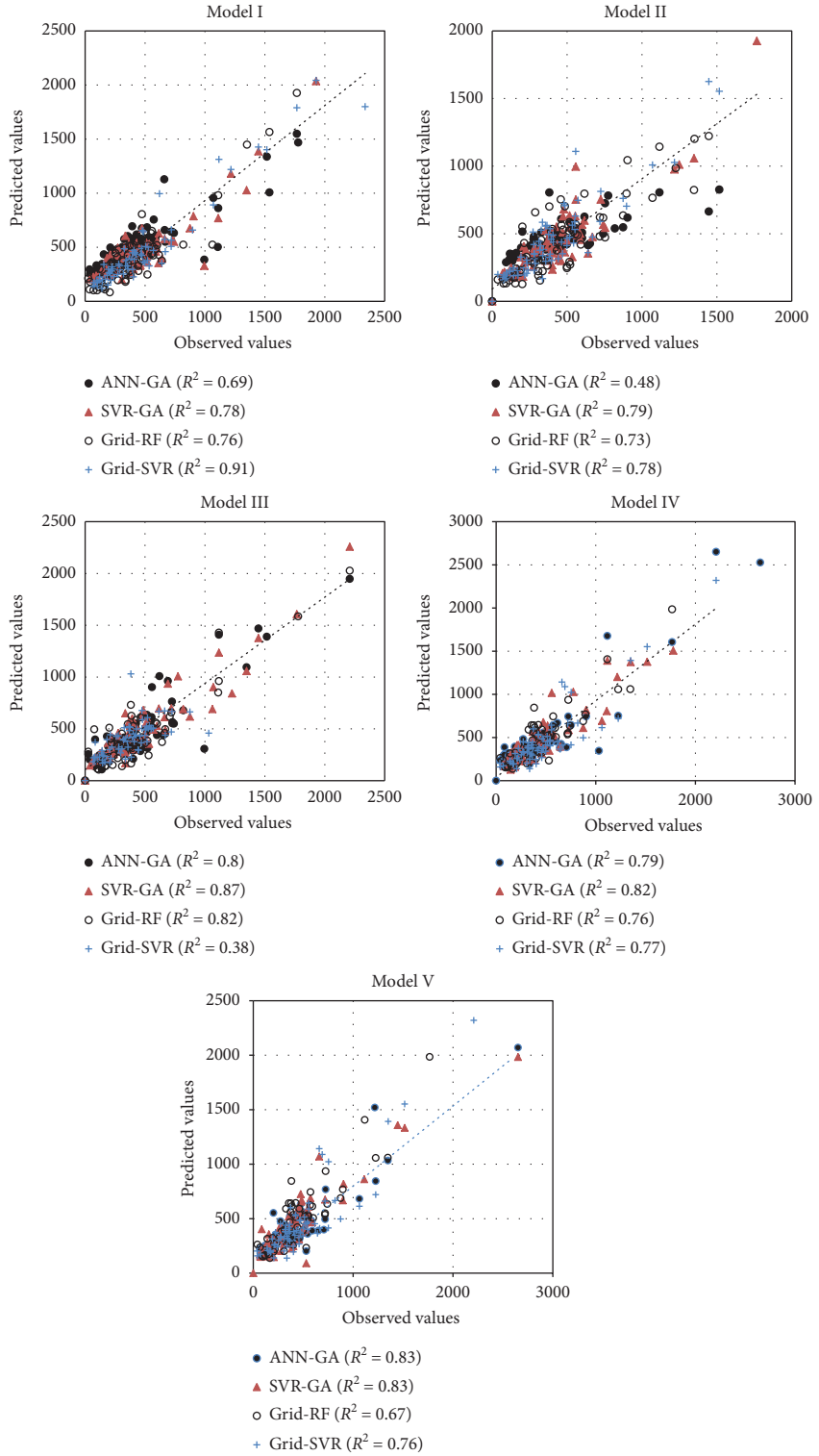

Figure 10: The scatter plot between the observed and forecasted river flow using the developed ML models and for the modeling scenario data division (70%–30%).

6. Discussion and Possible Future Research Directions

The results indicate that the proposed hybrid ML models can provide high accuracy in forecasting river flows for the studied Tigris River where the variability of river flow is less. Among all the developed four models, it was noticed that SVR-GA was superior to the other models. The model revealed the ability to solve complex process related to engineering problem. SVR-GA achieved a high Coefficient

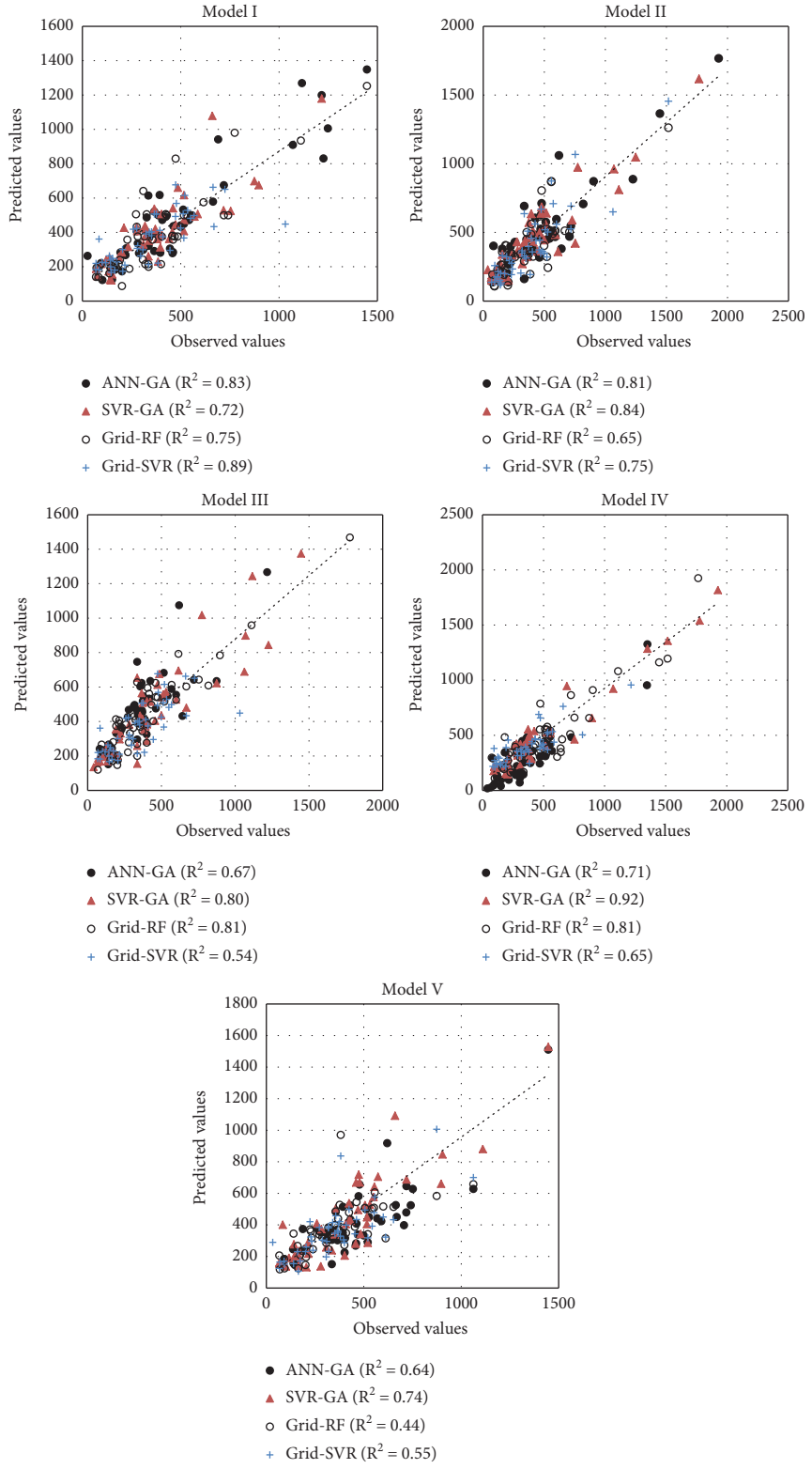

Figure 11: The scatter plot between the observed and forecasted river flow using the developed ML models and for the modeling scenario data division (80%–20%).

of Determination for almost all the input combinations for streamflow forecasting. It can be concluded that SVR-GA has the potential to deal with dynamics and chaotic environment with high accuracy in forecasting process.

In the current study, a type of forecasting is based on univariate modeling procedure where only river flow historical data was intercepted in the model development. In such case, it is suggested to use other variables such as

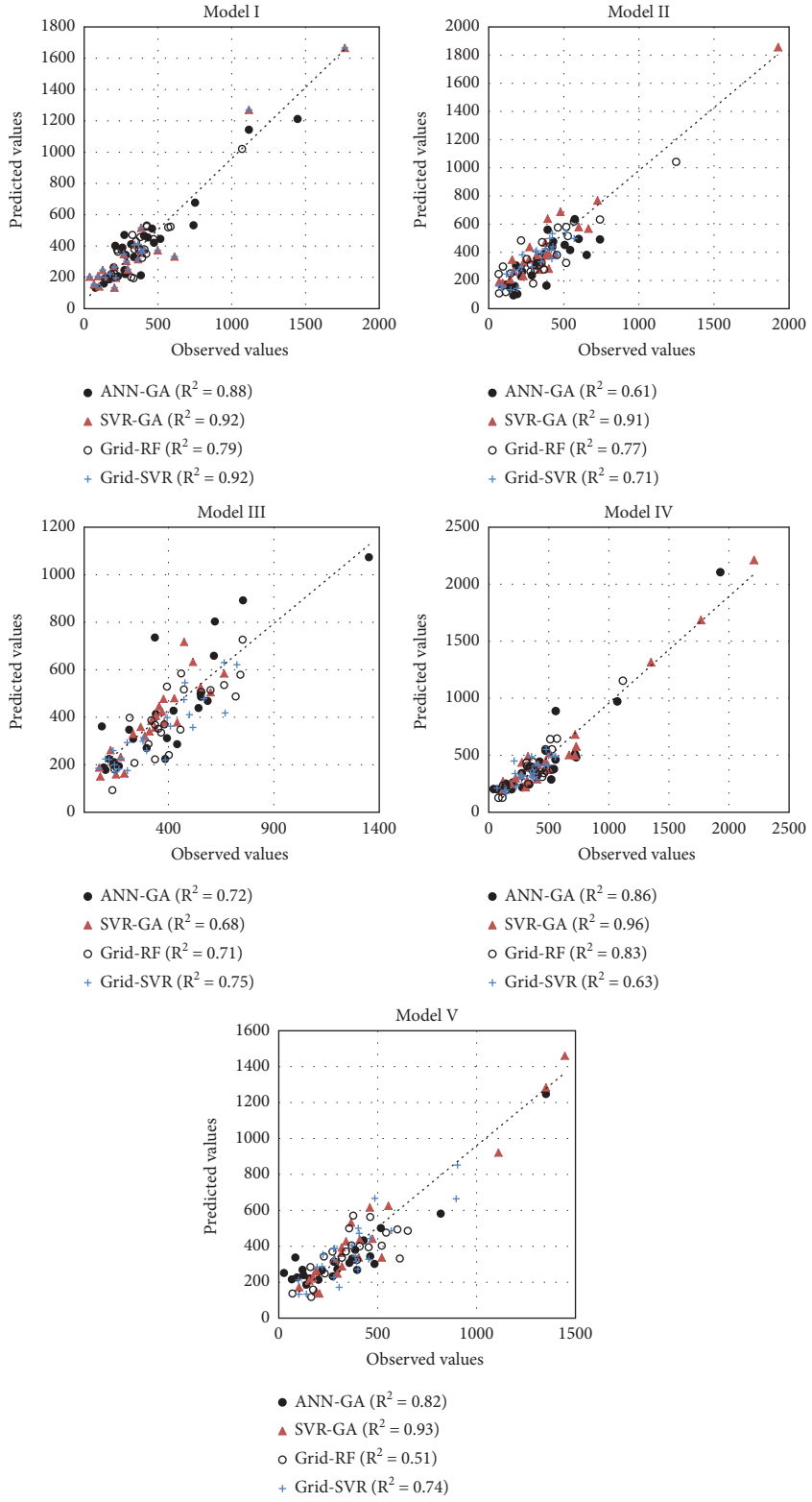

Figure 12: The scatter plot between the observed and forecasted river flow using the developed ML models and for the modeling scenario data division (90%–10%).

rainfall, humidity, temperature, or even evaporation rate to have a strong relationship with river flow. However, it is worth highlighting that the proposed models demonstrated an efficient soft computing model to capture the actual trend of the river flow time series. This is highly essential for several water and environmental engineering applications and particularly for management and monitoring of flooding and mitigations events.

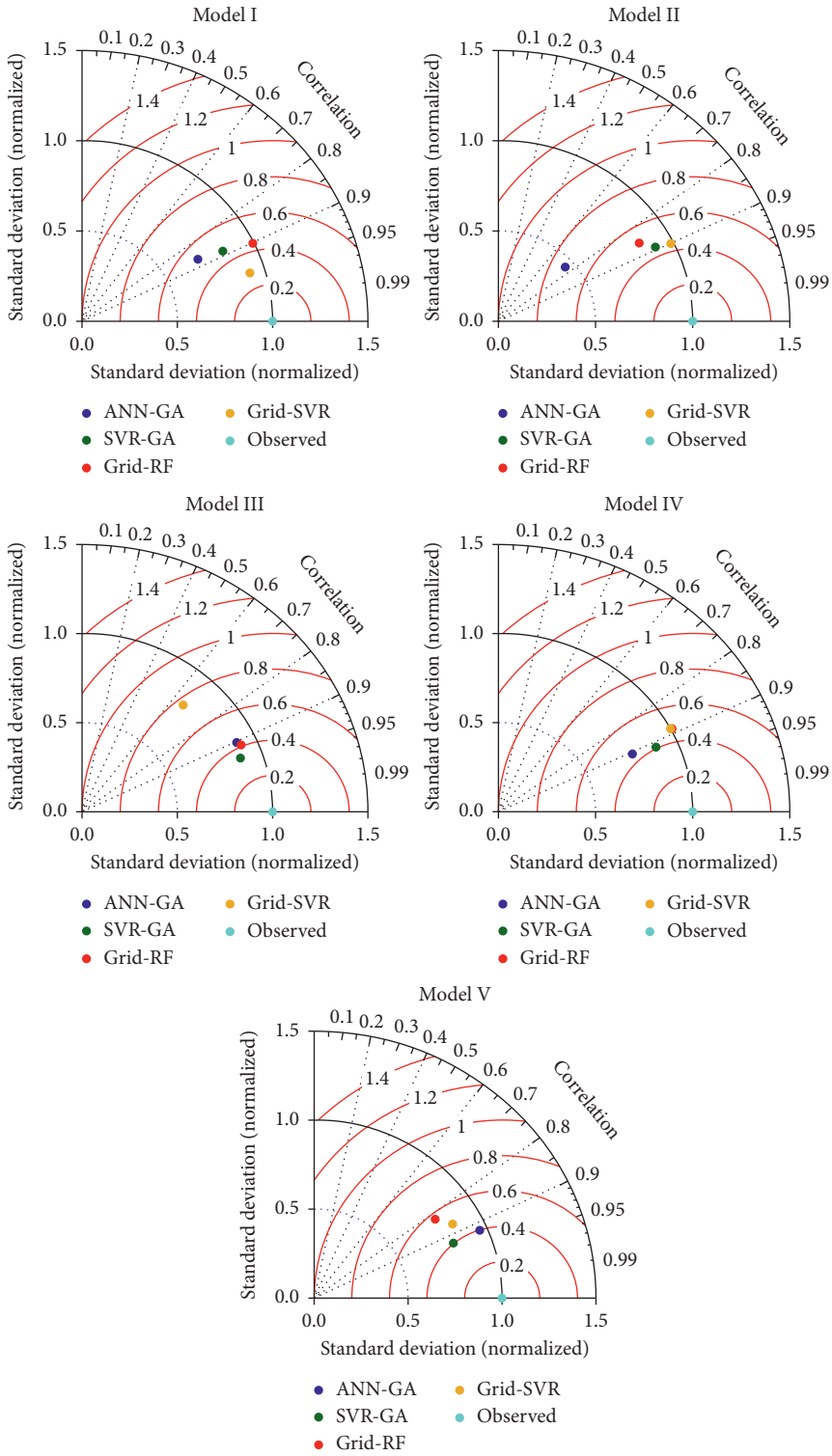

Figure 13: The Taylor diagram of the developed ML models and for the modeling scenario data division (70%–30%).

Based on the reported results, it was observed that using 20-year river flow data is sufficient for the development of the forecasting model. However, the length of the data span used for the modeling learning process has a considerable effect on the accuracy of model performance. Therefore, this is the essential finding of the hypotheses data division scenarios on the capacity of the machine learning models. Indeed, serious attention should be given in selecting the length of data for training the models. Selecting the length of data in an appropriate way reduces underfitting and helps the modeler to choose the best size of training data. This is due to the fact that the training stage should experience the majority of river flow

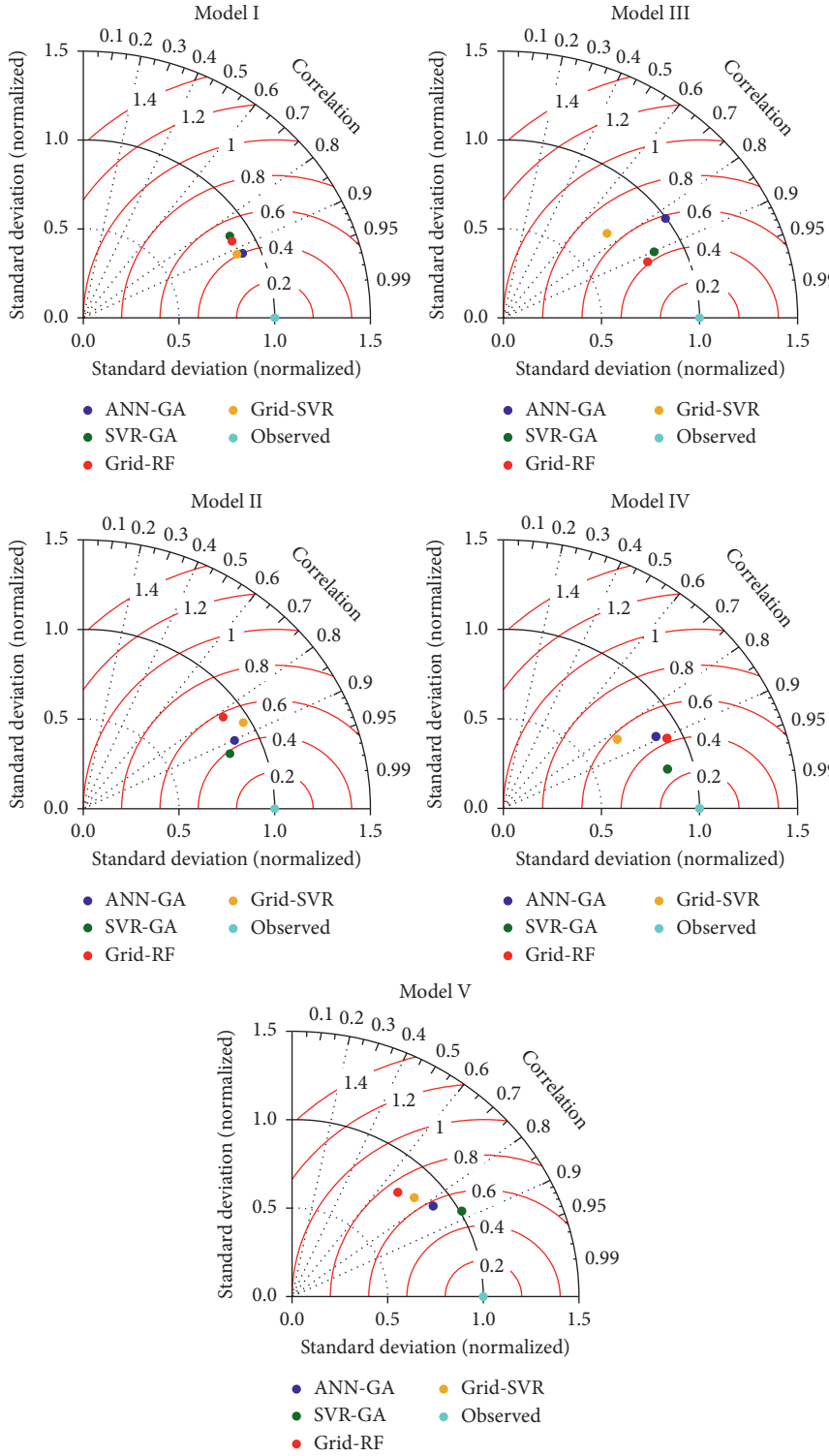

Figure 14: The Taylor diagram of the developed ML models and for the modeling scenario data division (80%–20%).

patterns to allow the models in the testing session to forecast river flow with an acceptable level of accuracy. The present study indicates that 20-year river flow data is enough to provide an acceptable accuracy in forecasting river flow.

Another significant aspect which can improve the predictability of the applied predictive models is the optimum selection of the lead times correlated to the targeted variable. Mutual information (MI) statistical approach potentially can be integrated as a prior stage of the forecasting model development process to abstract the highly associated information. The approach is based on the information theory and the notion of entropy [74].

It is worth highlighting that there is a need to extract the highly correlated features (the correlated lead times)

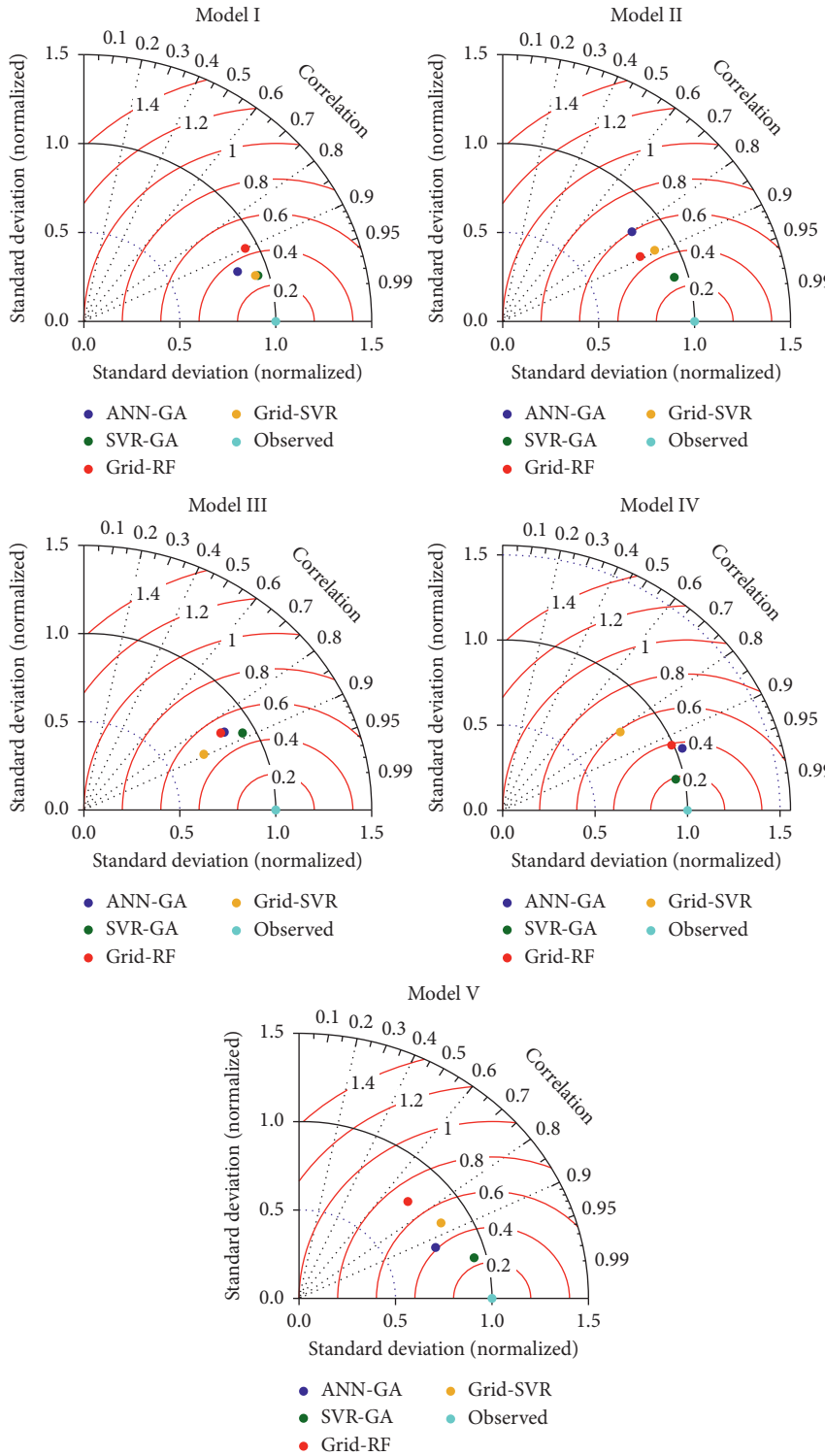

Figure 15: The Taylor diagram of the developed ML models and for the modeling scenario data division (90%–10%).

for the development of similar developed hybrid machine learning. Hence, the deep learning model has the advantage of deploying the hidden layers to automatically extract the features. The hydrological process is linked to typical time-sequential data [75] and therefore, the forecasting of hydrological time series is based on a fixed number of previous steps selected based on correlation. For such a case, deep learning model can be trained to learn time-varying or sequential patterns by facilitating time delay units through feedback connections for the selection of antecedent values as input. The suitability of the deep learning model for hydrological application relies on its capability of providing accurate and timely time-series prediction in the systems.

7. Conclusion

The motivation for the current research was to forecast monthly river flow in semiarid environment. Four hybrid ML models (i.e., ANN-GA, SVR-GA, Grid-SVR, and Grid-RF) were developed for this purpose. Three data division modeling scenarios (i.e., 70%–30%, 80%–20, and 90%–10%) were proposed and inspected for the model's performance predictability. The architecture of the applied ML models was established based on several antecedent values of the river flow in accordance with the correlation analysis. The performance of the models was assessed using a number of numerical skill indicators and graphical presentations. In general, the results demonstrated that the SVR-GA model achieved the highest ability in forecasting monthly river flow with significant accuracy. Therefore, it is possible to improve the river flow forecasting ability using the proposed hybrid machine learning model. In addition, the results indicated that building the predictive based on 90%–10% training-testing dataset attained better prediction capability. The results indicated that using 20 years of river flow data is sufficient for the development of the forecasting model. The study concluded that the size of the training data has a significant effect on the accuracy of the predicted model. The study revealed that the data division has an important role in the learning process of the developed ML models. The results demonstrated that increasing the span of the training phase can enhance the accuracy of model performance. The current research is possible to be further extended for the forecasting enhancement of river flow by including more information on river flow patterns through the inclusion of climate parameters such as rainfall, humidity, and temperature as an input. The models in the present study were developed for the forecasting of only one-step-ahead river flow. However, multiple-month ahead forecasting models are important for water resources planning and management. Although the success of SVR-GA model in forecasting one-step-ahead river flow indicates its capability in longer time-step ahead river flow forecasting, it is still necessary to examine the ability of SVR-GA model in multiple-month ahead forecasting. The study recommended using mutual information (MI) statistical approach as a prior stage of the forecasting model development process to extract the highly associated information.

Data Availability

The data used in the study are available upon request from the corresponding author.

Conflicts of Interest

The authors have no conflicts of interest to declare.

Acknowledgments

This study was funded by theLulea University of Technology. In addition, the authors acknowledge the support received from the Key Research and Development Program in Shaanxi Province (2020GY-078).

References

[1] J. Quilty and J. Adamowski, "A stochastic wavelet-based data-driven framework for forecasting uncertain multiscale hydrological and water resources processes," *Environmental Modelling & Software*, vol. 130, Article ID 104718, 2020.

[2] K. Lee, H. Gao, M. Huang, J. Sheffield, and X. Shi, "Development and application of improved long-term datasets of surface hydrology for Texas," *Advances in Meteorology*, vol. 2017, Article ID 8485130, 13 pages, 2017.

[3] M. Lefebvre and F. Bensalma, "An application of filtered renewal processes in hydrology," *International Journal of Engineering Mathematics*, vol. 2014, Article ID 593243, 9 pages, 2014.

[4] F. B. Hamzah, F. MohdHamzah, S. F. Mohd Razali, O. Jaafar, and N. AbdulJamil, "Imputation methods for recovering streamflow observation: a methodological review," *Cogent Environmental Science*, vol. 6, no. 1, 2020.

[5] Z. M. Yaseen, O. Kisi, and V. Demir, "Enhancing long-term streamflow forecasting and predicting using periodicity data component: application of artificial intelligence," *Water Resources Management*, vol. 30, no. 12, pp. 4125–4151, 2016.

[6] Z. M. Yaseen, S. R. Naganna, Z. Sa'adi et al., "Hourly river flow forecasting: application of emotional neural network versus multiple machine learning paradigms," *Water Resources Management*, vol. 34, no. 3, pp. 1075–1091, 2020.

[7] Z. Yaseen, W. H. M. W. Mohtar, A. M. S. Ameen et al., "Implementation of univariate paradigm for streamflow simulation using hybrid data-driven model: case study in tropical region," *IEEE Access*, vol. 7, pp. 74471–74481, 2019.

[8] A. H. El-Shafie and M. S. El-Manadely, "An integrated neural network stochastic dynamic programming model for optimizing the operation policy of Aswan High Dam," *Hydrology Research*, vol. 42, no. 1, pp. 50–67, 2011.

[9] H. Singh and A. Sankarasubramanian, "Systematic uncertainty reduction strategies for developing streamflow forecasts utilizing multiple climate models and hydrologic models," *Water Resources Research*, vol. 50, no. 2, pp. 1288–1307, 2014.

[10] M. Fu, T. Fan, Z. Ding, S. Q. Salih, N. Al-ansari, and Z. M. Yaseen, "Deep learning data-intelligence model based on adjusted forecasting window scale: application in daily streamflow simulation," *IEEE Access*, vol. 8, pp. 32632–32651, 2020.

[11] T. Asefa, M. Kemblowski, M. McKee, and A. Khalil, "Multi-time scale stream flow predictions: the support vector machines approach," *Journal of Hydrology*, vol. 318, no. 1–4, pp. 7–16, 2006.

[12] P. A. Kagoda, J. Ndiritu, C. Ntuli, and B. Mwaka, "Application of radial basis function neural networks to short-term streamflow forecasting," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 35, no. 13-14, pp. 571–581, 2010.

[13] F. Li, Q. Cai, X. Fu, and J. Liu, "Construction of habitat suitability models (HSMs) for benthic macroinvertebrate and their applications to instream environmental flows: a case study in Xiangxi River of Three Gorges Reservior region, China," *Progress in Natural Science*, vol. 19, no. 3, p. 359, 2009.

[14] C. Shen, "A transdisciplinary review of deep learning research and its relevance for water resources scientists," *Water Resources Research*, vol. 54, no. 11, pp. 8558–8593, 2018.

[15] R. C. Mamat, A. M. Samad, A. Kasa, S. F. M. Razali, A. Ramli, and M. B. H. C. Omar, "Slope stability prediction of road

embankment on soft ground treated with prefabricated vertical drains using artificial neural network," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 236–243, 2020.

[16] Z. Sheikh Khozani, M. J. S. Safari, A. Danandeh Mehr, and W. H. M. Wan Mohtar, "An ensemble genetic programming approach to develop incipient sediment motion models in rectangular channels," *Journal of Hydrology*, vol. 584, Article ID 124753, 2020.

[17] Z. S. Khozani, K. Khosravi, B. T. Pham et al., "Determination of compound channel apparent shear stress: application of novel data mining models," *Journal of Hydroinformatics*, vol. 21, no. 5, pp. 798–811, 2019.

[18] Z. M. Yaseen, I. Ebtehaj, S. Kim, and H. Sanikhani, "Novel hybrid data-intelligence model for forecasting monthly rainfall with uncertainty analysis," *Water*, vol. 11, no. 3, 2019.

[19] H. Zhang, H. A. Loáiciga, F. Ren, Q. Du, and D. Ha, "Semi-empirical prediction method for monthly precipitation prediction based on environmental factors and comparison with stochastic and machine learning models," *Hydrological Sciences Journal*, vol. 65, no. 11, p. 1928, 2020.

[20] Z. M. Yaseen, S. O. Sulaiman, R. C. Deo, and K.-W. Chau, "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," *Journal of Hydrology*, vol. 569, pp. 387–408, 2019.

[21] V. Nourani, A. Hosseini Baghanam, J. Adamowski, and O. Kisi, "Applications of hybrid wavelet-artificial intelligence models in hydrology: a review," *Journal of Hydrology*, vol. 514, pp. 358–377, 2014.

[22] H. R. Maier, Z. Kapelan, J. Kasprzyk et al., "Evolutionary algorithms and other metaheuristics in water resources: current status, research challenges and future directions," *Environmental Modelling & Software*, vol. 62, pp. 271–299, 2014.

[23] M. Ehteram, S. Q. Salih, and Z. M. Yaseen, "Efficiency evaluation of reverse osmosis desalination plant using hybridized multilayer perceptron with particle swarm optimization," *Environmental Science and Pollution Research*, vol. 27, no. 13, p. 15278, 2020.

[24] G. Zhao, B. Pang, Z. Xu, and L. Xu, "A hybrid machine learning framework for real-time water level prediction in high sediment load reaches," *Journal of Hydrology*, vol. 581, Article ID 124422, 2020.

[25] S. Shamshirband, S. Hashemi, H. Salimi et al., "Predicting standardized streamflow index for hydrological drought using machine learning models," *Engineering Applications of Computational Fluid Mechanics*, vol. 14, no. 1, p. 339, 2020.

[26] M. B. Kia, S. Pirasteh, B. Pradhan, A. R. Mahmud, W. N. A. Sulaiman, and A. Moradi, "An artificial neural network model for flood simulation using GIS: Johor river basin, Malaysia," *Environmental Earth Sciences*, vol. 67, no. 1, p. 251, 2012.

[27] Y. Shao and R. S. Lunetta, "Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 70, pp. 78–87, 2012.

[28] Z. M. Yaseen, H. Faris, and N. Al-Ansari, "Hybridized extreme learning machine model with Salp Swarm algorithm: a novel predictive model for hydrological application," *Complexity*, vol. 2020, Article ID 8206245, 14 pages, 2020.

[29] S. O. Sulaiman, J. Shiri, H. Shiralizadeh, O. Kisi, and Z. M. Yaseen, "Precipitation pattern modeling using cross-station perception: regional investigation," *Environmental Earth Sciences*, vol. 77, no. 19, p. 709, 2018.

[30] D. K. Saleh, *Stream Gage Descriptions and Streamflow Statistics for Sites in the Tigris River and Euphrates River Basins, Iraq*, US Department of the Interior, US Geological Survey, Reston, VA, USA, 2010.

[31] Z. M. Yaseen, S. M. Awadh, A. Sharafati, and S. Shahid, "Complementary data-intelligence model for river flow simulation," *Journal of Hydrology*, vol. 567, pp. 180–190, 2018.

[32] O. Kisi, S. Heddam, and Z. M. Yaseen, "The implementation of univariable scheme-based air temperature for solar radiation prediction: new development of dynamic evolving neural-fuzzy inference system model," *Applied Energy*, vol. 241, pp. 184–195, 2019.

[33] M. Firat, "Comparison of artificial intelligence techniques for river flow forecasting," *Hydrology and Earth System Sciences*, vol. 12, no. 1, pp. 123–139, 2008.

[34] K. Mohammadi, H. R. Eslami, and R. Kahawita, "Parameter estimation of an ARMA model for river flow forecasting using goal programming," *Journal of Hydrology*, vol. 331, no. 1-2, pp. 293–299, 2006.

[35] R. Burget, J. Karasek, Z. Smékal, V. Uher, and O. Dostal, "Rapidminer image processing extension: a platform for collaborative research," in *Proceedings of the 33rd International Conference on Telecommunication and Signal Processing TSP*, Vienna, Austria, 2010.

[36] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, p. 533, 1986.

[37] S. Zaza and M. Al-Emran, "Mining and exploration of credit cards data in UAE," in *Proceedings of the 2015 5th International Conference on E-Learning, ECONF 2015*, Manama, Bahrain, 2015.

[38] S. Naganna, P. Deka, M. Ghorbani, S. Biazar, N. Al-Ansari, and Z. Yaseen, "Dew point temperature estimation: application of artificial intelligence model integrated with nature-inspired optimization algorithms," *Water*, vol. 11, p. 742, 2019.

[39] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation*, Institute for Cognitive Science, University of California, San Diego, CA, USA, 1985.

[40] M. Kubat, "The genetic algorithm," *An Introduction to Machine Learning*, Springer, Berlin, Germany, 2017.

[41] I. Arpaci, S. Alshehabi, M. Al-Emran et al., "Analysis of twitter data using evolutionary clustering during the COVID-19 pandemic," *Computers, Materials & Continua*, vol. 65, no. 1, p. 193, 2020.

[42] S. Q. Salih, A. Sharafati, I. Ebtehaj et al., "Integrative stochastic model standardization with genetic algorithm for rainfall pattern forecasting in tropical and semi-arid environments," *Hydrological Sciences Journal*, vol. 65, 2020.

[43] M. E. A. B. Seghier, B. Keshtegar, K. F. Tee, T. Zayed, R. Abbassi, and N. T. Trung, "Prediction of maximum pitting corrosion depth in oil and gas pipelines," *Engineering Failure Analysis*, vol. 112, Article ID 104505, 2020.

[44] M. Jahandideh-Tehrani, G. Jenkins, and F. Helfer, "A comparison of particle swarm optimization and genetic algorithm for daily rainfall-runoff modelling: a case study for southeast Queensland, Australia," *Optimization and Engineering*, 2020.

[45] M. Zounemat-Kermani, E. Matta, A. Cominola et al., "Neurocomputing in surface water hydrology and hydraulics: a review of two decades retrospective, current status and

future prospects," *Journal of Hydrology*, vol. 588, Article ID 125085, 2020.

[46] M. E. A. Ben Seghier, H. Carvalho, B. Keshtegar, J. A. F. O. Correia, and F. Berto, "Novel hybridized adaptive neuro-fuzzy inference system models based particle swarm optimization and genetic algorithms for accurate prediction of stress intensity factor," *Fatigue & Fracture of Engineering Materials & Structures*, vol. 43, no. 11, pp. 2653–2667, 2020.

[47] S. H. Mai, M. E. A. B. Seghier, P. L. Nguyen, J. Jafari-Asl, and D.-K. Thai, "A hybrid model for predicting the axial compression capacity of square concrete-filled steel tubular columns," *Engineering with Computers*, pp. 1–18, 2020.

[48] A. A. Saa, M. Al-Emran, and K. Shaalan, "Mining student information system records to predict students' academic performance," *Advances in Intelligent Systems and Computing*, Springer, Berlin, Germany, 2020.

[49] L. Breiman, *Random Forest.Pdf*, Kluwer Academic Publishers, Dordrecht, Netherlands, 2001.

[50] C. Qi, W. Zhou, X. Lu, H. Luo, B. T. Pham, and Z. M. Yaseen, "Particulate matter concentration from open-cut coal mines: a hybrid machine learning estimation," *Environmental Pollution*, vol. 263, Article ID 114517, 2020.

[51] Z. M. Yaseen, S. Naghshara, S. Q. Salih, S. Kim, A. Malik, and M. A. Ghorbani, "Lake water level modeling using newly developed hybrid data intelligence model," *Theoretical and Applied Climatology*, vol. 141, no. 3-4, p. 1285, 2020.

[52] M. Ali, R. Prasad, Y. Xiang, and Z. M. Yaseen, "Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts," *Journal of Hydrology*, vol. 584, Article ID 124647, 2020.

[53] Y. Tang, S. Krasser, Y. He, W. Yang, and D. Alperovitch, "Support vector machines and random forests modeling for spam senders behavior analysis," in *Proceedings of the 2008 IEEE Global Telecommunications Conference*, New Orleans, LA, USA, 2008.

[54] I. Arpaci, M. Al-Emran, M. A. Al-Sharafi, and K. Shaalan, "A novel approach for predicting the adoption of smartwatches using machine learning algorithms," *Studies in Systems, Decision and Control*, Springer, Berlin, Germany, 2021.

[55] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[56] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93–104, 2012.

[57] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, Berlin, Germany, 2013.

[58] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK, 2000.

[59] Z. M. Yaseen, R. C. Deo, A. Hilal et al., "Predicting compressive strength of lightweight foamed concrete using extreme learning machine model," *Advances in Engineering Software*, vol. 115, pp. 112–125, 2018.

[60] S. Belaid and A. Mellit, "Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate," *Energy Conversion and Management*, vol. 118, pp. 105–118, 2016.

[61] T. T. M. Tiyasha, T. M. Tung, and Z. M. Yaseen, "A survey on river water quality modelling using artificial intelligence models: 2000–2020," *Journal of Hydrology*, vol. 585, Article ID 124670, 2020.

[62] S. K. Bhagat, T. M. Tung, and Z. M. Yaseen, "Development of artificial intelligence for modeling wastewater heavy metal removal: state of the art, application assessment and possible future research," *Journal of Cleaner Production*, vol. 250, Article ID 119473, 2019.

[63] K. S. Sajan, V. Kumar, and B. Tyagi, "Genetic algorithm based support vector machine for on-line voltage stability monitoring," *International Journal of Electrical Power & Energy Systems*, vol. 73, p. 200, 2015.

[64] K.-P. Wu and S.-D. Wang, "Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space," *Pattern Recognition*, vol. 42, no. 5, pp. 710–717, 2009.

[65] Q. Huang, J. Mao, and Y. Liu, "An improved grid search algorithm of SVR parameters optimization," in *Proceedings of the 2012 International Conference on Communication Technology Proceedings ICCT*, Chengdu, China, 2012.

[66] K. Khosravi, L. Mao, O. Kisi, Z. M. Yaseen, and S. Shahid, "Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile," *Journal of Hydrology*, vol. 567, pp. 165–179, 2018.

[67] O. Kisi and Z. M. Yaseen, "The potential of hybrid evolutionary fuzzy intelligence model for suspended sediment concentration prediction," *CATENA*, vol. 174, pp. 11–23, 2019.

[68] D. R. Legates and G. J. McCabe, "Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation," *Water Resources Research*, vol. 35, no. 1, pp. 233–241, 1999.

[69] M. Zounemat-Kermani, A. Mahdavi-Meymand, M. Alizamir, S. Adarsh, and Z. M. Yaseen, "On the complexities of sediment load modeling using integrative machine learning: application of the great river of Loíza in Puerto Rico," *Journal of Hydrology*, vol. 585, Article ID 124759, 2020.

[70] Y. Tikhamarine, A. Malik, D. Souag-Gamane, and O. Kisi, "Artificial intelligence models versus empirical equations for modeling monthly reference evapotranspiration," *Environmental Science and Pollution Research*, vol. 27, no. 24, p. 30001, 2020.

[71] Y. Seo, S. Kim, and V. P. Singh, "Physical interpretation of river stage forecasting using soft computing and optimization algorithms," *Advances in Intelligent Systems and Computing*, Springer, Berlin, Germany, 2016.

[72] B. Yaghoubi, S. A. Hosseini, and S. Nazif, "Monthly prediction of streamflow using data-driven models," *Journal of Earth System Science*, vol. 128, no. 6, 2019.

[73] K. E. Taylor, "Summarizing multiple aspects of model performance in a single diagram," *Journal of Geophysical Research: Atmospheres*, vol. 106, no. D7, pp. 7183–7192, 2001.

[74] C. E. Shannon, "The mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[75] N. Rusk, "Deep learning," *Nature Methods*, vol. 13, no. 1, p. 35, 2015.