Hindawi
Complexity
Volume 2021, Article ID 8324272, 12 pages
https://doi.org/10.1155/2021/8324272

Research Article

Evaluation of Several Machine Learning Models for Field Canal Improvement Project Cost Prediction

Saadi Shartooh Sharqi[1] **and Aayush Bhattarai**[2]

[1]*Civil Engineering Department, Engineering College, University Of Anbar, Ramadi, Iraq*
[2]*Department of Mechanical and Aerospace Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Kathmandu, Nepal*

Correspondence should be addressed to Aayush Bhattarai; aayush@pcampus.edu.np

Received 16 June 2021; Revised 19 July 2021; Accepted 7 August 2021; Published 17 August 2021

Academic Editor: Mostafa Al-Emran

Project cost prediction is one of the key elements in the civil engineering activities development. Project cost is a highly sensitive component to diverse parameters and hence it is associated with complex trends that make it difficult to be predicted and fully understood. Due to the massive advancement of soft computing (SC) and Internet of things (IoT), the main research objective of the current study was initiative. Several machine learning (ML) models including extreme learning machine (ELM), multivariate adaptive regression spline (MARS), and partial least square regression (PLS) were adopted to predict field canal cost. Several essential predictors were used to develop the prediction network "the learning process" including the total length of the PVC pipeline, served area, geographical zone, construction year, and cost and duration of field canal improvement projects (FCIP) construction. Data were collected from the open source published literature. The modeling results evidenced the potential of the applied SC models in predicting the FCIP cost. In numerical magnitude evaluation, MARS model indicated the least value for the root mean square error (RMSE = 27422.7), mean absolute error (MAE = 19761.8), and mean absolute percentage error (MAPE = 0.05454) with Nash–Sutcliffe efficiency (NSE = 0.94), agreement index (MD = 0.89), and coefficient of determination (R^2 = 0.94), with best precision of prediction using all predictors, except geographical zone parameter in which less influence on the cost construction is presented. In general, the research outcome gave an informative primary cost initiative for cost civil engineering project.

1. Introduction

The scarcity of freshwater has been a global problem recently and expected to worsen in the future due to the increasing human population and decline in annual water allocation per capita [1, 2]. The present scenario portrays water unsustainability due to the drastic increase in water utilization (>6 folds) in the 20th century [3]. It is presently estimated that about 1.2 billion people globally have no access to a clean water supply [4]. Hence, several policies and projects are being implemented globally to ensure water sustainability. One of such projects aimed at water sustainability is the FCIP which aims at increasing the conveyance efficiency of field canals by about 25% via improvement of the field canals during irrigation processes in farmlands [5]. The project requires the construction of a burden PVC pipeline rather than relying on earthen field canals for the reduction of water seepage or losses during field operations [6]. FCIP is comprised of several simple components and structures which include concrete pain intakes for water collection from the source; water is channelled through the suction pipes to a plain concrete sump [7]. Water is first accumulated in the sump before being pumped by the pumping sets through the PVC pipelines by the irrigation valves. The FCIPs are comprised of civil works, mechanical components, and electrical components as the major components. The components of the civil works are the pump house, pipelines, suction pipes, intake, and sump structure while the mechanical components are the

irrigation valves, pump sets, and mechanical connections. The electrical boards and connections make up the electrical components of FCIPs [8].

The most interesting part of FCIPs is the cost estimation aspect that must be performed; manual cost estimation processes are time-consuming [9]. However, in some cases, scan be attained based on personal engineering and decision-makers' expertise. Cost estimation is highly associated with bias and inaccuracy and to overcome these issues of bias and inaccuracy during cost estimation [10]. Therefore, SC models have been proposed as the potential solution. In line with this, the aim of this work is to come up with a robust ML-based SC model for FCIP cost estimation. The proposed models are expected to help decision-makers and management engineers in making decisions from the perspective of the stockholders.

Literature review studies suggested that numerous researches have focused on the development of reliable regression and mathematical techniques that can be used for cost estimation in civil engineering projects [11–16]. The nagging problem in this domain still relates to the performance accuracy of these models as the predicted cost is required to be highly accurate before the conception of the project. The weighted ANN has been developed for unit cost prediction in highway projects by [16], while a parametric cost model was developed based on a questionnaire survey for the estimation of the final cost of pump stations by [17]. A fuzzy logic- (FL-) based parametric cost estimate model has been presented by [18], for the prediction of the cost of building projects in the Gaza Strip. The study by [19] presented a hybrid ANN-FL model for cost prediction of water infrastructure. The prediction of the unit cost of the highway project in Libya using the ANN model has been presented by [20] and the performance of the ANN model was excellent. A conceptual cost model for the German residential building project was developed by [21] using historical data for 75 residential projects sourced from the building cost information center. The use of ANN to determine the relevant parameters for cost prediction during tunnel construction in Greece was reported by [22] based on survey questionnaires. The survey was based on expert opinions and interviews in relation to the key cost drivers.

The reviewed literature suggests the need for intelligence models that are robust and capable of understanding the civil engineering complexity in more realistic manners. Several ML models have been reported recently, such as ANN [23], SVM [24], ANFIS [25], genetic programming [26], decision tree [27], and gradient boosting [28], and several others were reported in the latest review [29]. However, the fact remains that each of these models behaves differently in terms of prediction accuracy. Some existing models are also capable of providing accurate results interpretation; for instance, the variable coefficients of the regression models can explain the influence of each variable on the response of the model.

Numerous studies have focused on building projects without giving much attention to the conceptual cost of FCIPs. Hence, the attention of this study is on the pipeline construction projects which have not attracted appropriate research attention, especially on the provision of detailed model development steps in terms of sample size, multicollinearity, outliers, and singularity. For instance, the study by [16] only applied 14 and 4 cases for the training and validation of their neural network model. This may have elicited concerns about the sample size in this study as stated by [30]. The motivation of the current study was inspired from the exhibited literature on the prediction of the FCIP cost using newly explored machine learning models including ELM, MARS, and PLS. These models are proven to be advantageous as they have very quick learning speeds with good performances and are useful in capturing complicated data mapping in very high set of predictors which produces interpretative results [31–34]. Modeling structure was adopted based on the correlation statistic to identify the input predictors for the built ML models. Based on the reported modeling results, comprehensive comparative analytical aspects were reported and discussed.

2. Soft Computing Models

2.1. Extreme Learning Machine. ELM model is one of the new methods of training recently developed single-layer feedforward neural networks [35]. The traditional ELM, as shown in Figure 1, has one input layer, one hidden layer, and one output layer; each of these layers has a specific number of neurons. The linear function is generally selected as the activation function of the input and output layers of ELM while the sigmoid function is selected for the hidden layer [36]. The first step of the standard ELM is a random input weight and hidden biases determination, followed by the determination of the hidden weights using the Moore–Penrose generalized inverse method to achieve the optimal solution of the linear system [37]. The advantages of the ELM over the other gradient-based methods are its strong generalization capability, no parameter tuning, and fast learning; these have made ELM more popular in numerous engineering tasks [38–40]. Consider a training dataset with N samples; the first process is to linearly map the input vectors into an L-dimensional feature space via nonlinear transformation; the expression of the simulated values of the ELM model is as follows:

$$\widetilde{t}_i = \sum_{l=1}^{L} \beta_l \cdot g\left(w_l \cdot x_i + b_l\right), \quad i = 1, 2, \ldots, N, \tag{1}$$

where N represents the number of samples for training, $\widetilde{t}_i$ represents the output vectors that are associated with the input vector x_i; β_l stands for the weight vectors that connect the hidden neuron to the output layer; w_l is the weight vectors that connect the hidden neuron with the input layer; b_l is the bias; and g is the activation function.

In the ELM, the idea is that the classical single-layer ANN can approach all the samples with zero deviation as mathematically expressed in the relation:

$$\sum_{i=1}^{N} t_i - \widetilde{t}_t = \sum_{i=1}^{N} \left\| t_i - \sum_{l=1}^{L} \beta_l \cdot g\left(w_l \cdot x_i + b_l\right) \right\| = 0, \tag{2}$$

where t_i is the target output vector that is related to the input vector x_i. The reconstruction of the above expression gives the following:

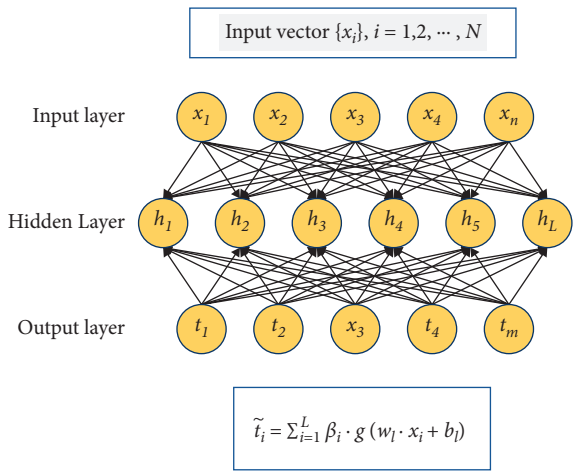

Figure 1: The extreme learning machine model paradigm.

$$H\beta = T, \tag{3}$$

where

$$H = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_L \cdot x_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_L \cdot x_N + b_L) \end{bmatrix}_{N\times L},$$

$$\beta = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,m} \\ \vdots & \ddots & \vdots \\ \beta_{L,1} & \cdots & \beta_{L,m} \end{bmatrix}_{L\times m}, \tag{4}$$

$$T = \begin{bmatrix} t_{1,1} & \cdots & t_{1,m} \\ \vdots & \ddots & \vdots \\ t_{N,1} & \cdots & t_{N,m} \end{bmatrix}_{N\times m},$$

where β is the weight of the matrix that connects the hidden and output layers; H is the hidden layer output matrix based on N samples; and T is the target output matrix based on N samples.

Assume that the hidden biases and input weights are constant; it implies that the model may be considered a special linear system in which H and T are equal to the matrixes of the known dependent and independent parameters, while β is considered the coefficient matrix that should be optimized. Hence, the least-squares solution of the represented linear system mentioned above can be derived as

$$\tilde{\beta} = H^{\dagger}T, \tag{5}$$

where $H^{\dagger}$ is the Moore–Penrose generalized inverse matrix of H.

2.2. Multivariate Adaptive Regression Spline Model. MARS algorithms are nonlinear-nonparametric flexible regression models that were first developed by [41] and have found application in many fields of engineering due to their robustness [42]. This model is built with three major components, which are the basis functions (BFs), the knots, and the spline function [43]. The role of the BFs is to capture the relationship between the predictands and the predicted variables, amounting max $(0, c - x)$ or max $(0, x - c)$, where x is the threshold value, while c is the input variable value. The knots also represent the function of the base and base endpoints. A regression model is developed for each node by applying a spline function that consists of 1 or more BFs, followed by the substitution of the principal predictors [44]. In the MARS model, the predicted value is based mainly on linear BF elements combination. The MARS model can be reviewed as follows: consider Y as the target variable and $X = (X1, X2, \ldots, XP)$ as the P input variable matrix; then, the equation of the MARS model can be as follows:

$$Y = f(X) = \beta_0 + \sum_{m=1}^{M} \beta_0 BF_m(X), \tag{6}$$

where β_0 is the initial fixed value; BF_m is the applied BF for the fitting of the MARS model; and M is the total number of BFs [45]. The two major phases of the MARS model are the selection phase (or forward search) and the reversal pruning phase, as seen in Figure 2. The forward phase or selection phase can be regarded as a set of optimum input parameters. A complicated over fitted model normally results from an excessive forward stepwise selection process due to a series of splits and such models cannot perform well predictively despite fitting the data perfectly. Hence, the backward procedure is normally applied to improve the predictive performance of the model by removing the unwanted variables that have been selected in the selection phase. The generalized cross-validation (GCV) is calculated as the deletion criterion as it is the basis for the backward pruning process [46, 47].

$$\mathrm{GCV}(M) = \frac{(1/N)\sum_{i=1}^{N}(O_i - f(x_i))^2}{(1 - (C(M)/N))^2}, \tag{7}$$

$$C(M) = (d + 1) \times M,$$

where O_i is the observed values; N is the number of data; $f(x_i)$ is the predicted values for pattern i; M is the number of BFs; and $C(M)$ is the penalty factor. In equation (7), the quantity of parameter d significantly impacts the procedure as it is the optimization cost of each BF; its range is $2 \le d \le 4$. The inclusion of several BFs can result in overfitting; therefore, it is important to omit some BFs during the pruning phase to enable the emergence of a well-fitted model with the least GCV value [48].

2.3. Partial Least Square Regression (PLS) Model. The first application of the PLS regression model was introduced over the literature by [49], and since then the model has been widely considered a new multivariate analysis technique in many fields [50, 51]. It combined the features of principal components, typical multiple regression, and linear regression analyses; hence, it is suitable for finding the solution to numerous problems, especially problems that cannot be solved using the conventional multiple regression methods and problems with multiple correlations [52]. The efficiency of PLS in such cases is based on its ability to decompose and

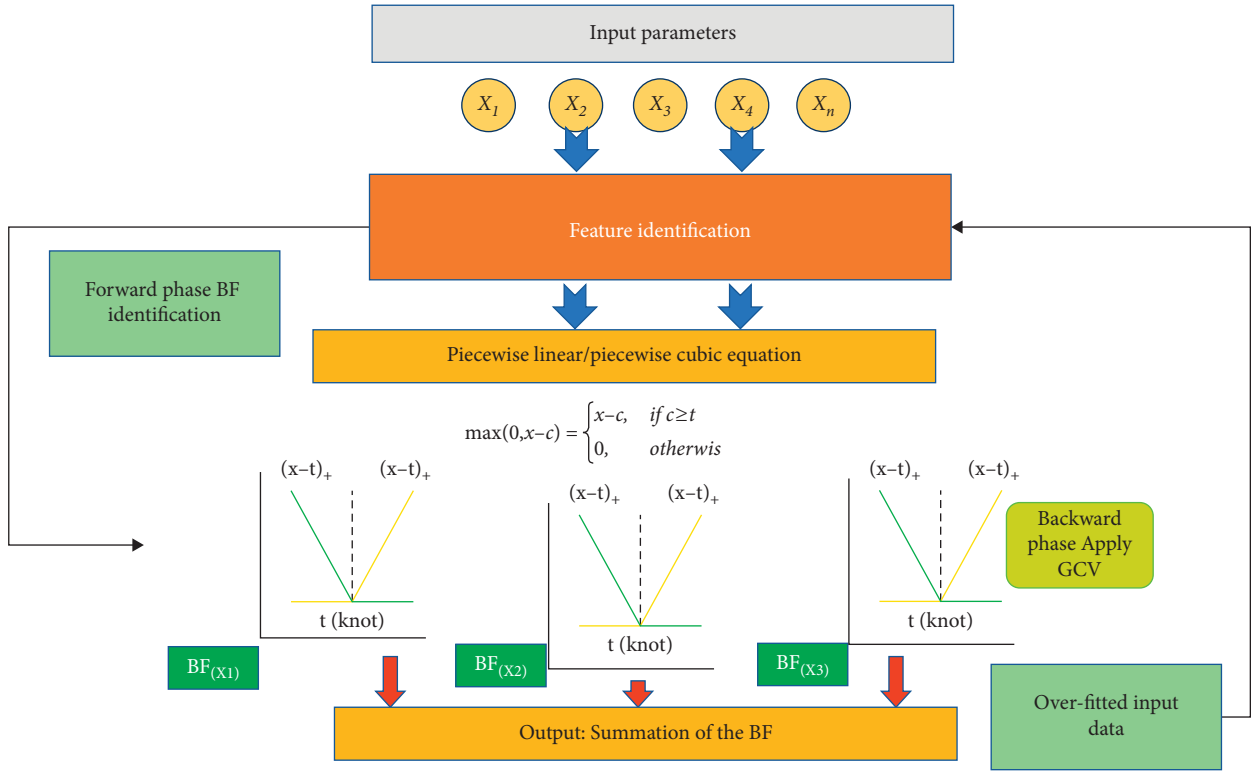

Figure 2: The systematic structure of the MARS model.

screen the variables that mostly explain the dependent variables [53]. The first step of the PLS method is to extract the new variable called the component which serves as the independent variable, followed by the determination and establishment of the linear relationship between the dependent and independent variables [54]. After calculating the coefficient using PLS, the next step is the construction of the regression equation of the dependent variable. The regression model developed by using the PLS method is represented as

$$y_m = a_{0m} + a_{1m}x_1 + \cdots + a_{Pm}x_p, \quad (m = 1, 2, \ldots, p), \tag{8}$$

where $x_1, \ldots, x_p$ represents the linear combinations of the remote sensing variables and $a_{0m}, a_{1m}, \ldots, a_{Pm}$ are the PLS-computed regression model parameters. A higher number of principal components in the established model by PLS translates to better model accuracy; however, an excessive number of principal components results in overfitting and higher error. Thus, the optimal number of principal components must be determined to achieve a balanced PLS model. The cross-validation method was used to calculate the sum of squared residuals in this study. The prediction ability of the resulting model is a function of the extent of predictive residual errors sum of square (PRESS) value. So, the optimal number of principal components can be determined based on the minimum PRESS value and this PRESS value can be calculated as

$$\text{PRESS} = \sum_{i=1}^{k} \left(y_i - y_{i,-1}\right)^2, \tag{9}$$

where y_i, $y_{i,-i}$ represent the measured value of the ith sample and the estimated value upon exclusion of the ith sample and k is the number of iterations for validation.

3. Case Study and Data Explanation

For the modeling purpose, datasets were collected from the open source of literature [7]. The datasets are explained the key cost derived from the FCIPs. The data were including $P1$, the served area; $P2$, the total length of the PVC pipeline; $P3$, irrigation valve number; $P4$, construction year; $P5$, geographical zone; and cost and duration of field canal improvement projects (FCIP) construction. The significance of the dataset is contributing to the best knowledge of irrigation authorities and decision makers to have a prior understanding on the FCIP cost. The biodata of the current research were collected from the survey conducted for Soltani Canal, Egypt. The quantitative costs are related to construction sites recorded between 2011 and 2018. The polyvinyl chloride (PVC) pipeline system is explained in Figure 3 with diameter ranging between 22.5 and 35 cm. The statistical properties of the dataset over the training and testing phases are reported in Tables 1 and 2. It is seen that all together of 228 data were taken for both training and testing phase. In Tables 1 and 2, the parameters that are collected for training and test phase are mean, standard error, median, mode, standard deviation, sample variance, kurtosis,

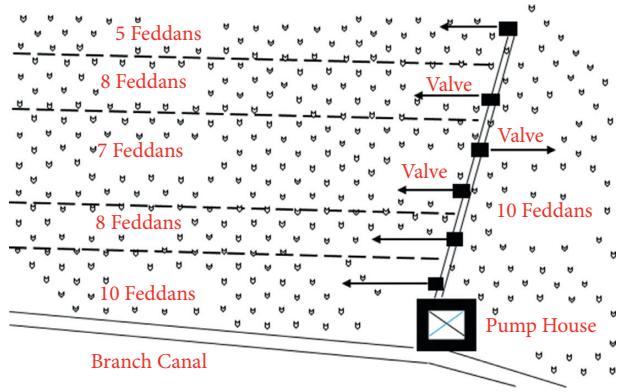

Figure 3: The Soltani Canal FCIP planning.

Table 1: The statistical properties of the dataset selected for the training modeling phase.

Parameter	*C*	*D*	*P*1	*P*2	*P*3	*P*4	*P*5
Parameter name	Cost of FCIP	Duration of FCIP construction	Area served	Total length of PVC pipeline	Number of irrigation values	Construction year	Geographical zone
Unit	LE/FCIP	Day	Hectare	Meter	Number	Year	Zone
Mean	353463.30	76.34	49.41	813.83	8.18	2013.20	0.00
Standard error	7539.45	0.78	1.28	26.10	0.23	0.10	0.00
Median	320292.58	75.00	45.90	753.75	8.00	2014.00	0.00
Mode	514778.00	64.00	51.00	530.00	5.00	2014.00	0.00
Standard deviation	113843.24	11.71	19.31	394.03	3.52	1.48	0.00
Sample variance	12960283390.67	137.20	372.83	155258.84	12.36	2.18	0.00
Kurtosis	−0.22	1.90	−0.07	−0.24	2.72	−0.13	−1.18
Skewness	0.77	1.04	0.75	0.60	0.94	−1.05	−0.58
Range	518824.50	69.00	85.00	1956.45	26.00	5.00	0.00
Minimum	186825.98	58.00	19.00	119.00	1.00	2010.00	0.00
Maximum	705650.48	127.00	104.00	2075.45	27.00	2015.00	0.00
Sum	80589633.51	17405.00	11265.43	185554.07	1866.02	459010.00	0.00
Count	228.00	228.00	228.00	228.00	228.00	228.00	228.00
Confidence level (95.0%)	14856.26	1.53	2.52	51.42	0.46	0.19	0.00

skewness, range, minimum and maximum, sum, count, and confidence interval from *C* to *P*5. The mean value of FCIP cost is 353463.0 for training modeling phase whilst a mean cost of 352714.35 was taken for testing phase. From Tables 1 and 2, it can be seen that the duration for FCIP construction ranges from 58 days to 127 days in the training model dataset while it ranges from 59 days to 126 days in the testing model dataset. The datasets in both training model and testing model are well distributed and almost resemble a normal distribution, as for most of the datasets, the mean and median are very close to each other.

4. Application Results and Analysis

The feasibility of three machine learning models (ELM, MARS, and PLS) was evaluated to predict cost of FCIP construction. The models were built based on different input combinations, as reported in Table 3. Based on the correlation statistics, the input combinations were constructed as shown in Figure 4.

Based on the tabulated input parameters, it can be recognized that the total length of the PVC pipeline has the substantial correlation to the construction cost followed by the time duration, served area, irrigation valve number, and geographical zone.

Different statistical performance metrics including determination coefficient (R^2), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), Nash–Sutcliffe efficiency (NSE), and agreement index (MD) were calculated to validate the applied models statistically [55, 56].

$$R^2 = \left(\frac{\sum_{i=1}^{N} \left(y_p - \overline{y_p}\right) \cdot \left(y_o - \overline{y_o}\right)}{\sqrt{\sum_{i=1}^{N} \left(y_p - \overline{y_p}\right)^2 \sum_{i=1}^{N} \left(y_o - \overline{y_o}\right)^2}} \right)^2, \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_p - y_o\right)^2}, \tag{11}$$

Table 2: The statistical properties of the dataset selected for the testing modeling phase.

Parameter	C	D	$P1$	$P2$	$P3$	$P4$	$P5$
Parameter name	Cost of FCIP	Duration of FCIP construction	Area served	Total length of PVC pipeline	Number of irrigation values	Construction year	Geographical zone
Unit	LE/FCIP	Day	Hectare	Meter	Number	Year	Zone
Mean	352714.35	77.00	48.46	807.57	8.41	2013.23	1.29
Standard error	7469.77	0.81	1.24	27.65	0.26	0.09	0.05
Median	318652.92	76.00	45.68	720.82	7.46	2014.00	2.00
Mode	201587.44	66.00	51.00	630.00	5.00	2014.00	2.00
Standard deviation	112791.09	12.19	18.79	417.58	3.96	1.41	0.79
Sample variance	12721830134.86	148.62	352.92	174372.83	15.71	1.99	0.63
Kurtosis	0.05	2.09	0.21	0.06	6.34	0.12	−1.18
Skewness	0.88	1.22	0.84	0.76	1.72	−1.14	−0.58
Range	503217.10	67.00	86.00	1916.15	27.87	5.00	2.00
Minimum	198035.54	59.00	19.00	119.00	1.02	2010.00	0.00
Maximum	701252.64	126.00	105.00	2035.15	28.89	2015.00	2.00
Sum	80418872.53	17556.00	11049.13	184125.65	1917.50	459016.00	295.00
Count	228.00	228.00	228.00	228.00	228.00	228.00	228.00
Confidence level (95.0%)	14718.96	1.59	2.45	54.49	0.52	0.18	0.10

Table 3: The modeling input combinations for the adopted dataset.

M1	$P2$					
M2	$P2$	D				
M3	$P2$	D	$P1$			
M4	$P2$	D	$P1$	$P3$		
M5	$P2$	D	$P1$	$P3$	$P4$	
M6	$P2$	D	$P1$	$P3$	$P4$	$P5$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N}\left|y_p - y_0\right|}{N}, \tag{12}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_o - y_p\right|}{y_p} \times 100, \tag{13}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}\left(y_p - y_o\right)^2}{\sum_{i=1}^{N}\left(y_p - \overline{y_p}\right)^2}, \tag{14}$$

$$\mathrm{MD} = 1 - \frac{\sum_{i=1}^{N}\left(y_o - y_p\right)^j}{\sum_{i=1}^{N}\left(\left|y_p - \overline{y_o}\right| + \left|y_o - \overline{y_o}\right|\right)^j}, \tag{15}$$

where y_o and y_p are the observed and predicted values of the FCIP cost; $\overline{y_o}$ and $\overline{y_p}$ are the mean values of the observed and predicted values of the FCIP cost; N is the number of observations; and j is the exponent term.

Tables 4 and 5 report the statistical measures over the training and testing phases, respectively. In general, prediction performance of the models indicated less accuracy by using few predictors. However, MARS model exhibited better predictability performance over both the training and testing phases. It has been noticed that the maximum determination coefficient was achieved for model M6 ($R^2 = 0.94$) with a minimum RMSE of 28458.17 in the training phase while 27422.7 in the testing phase using all the

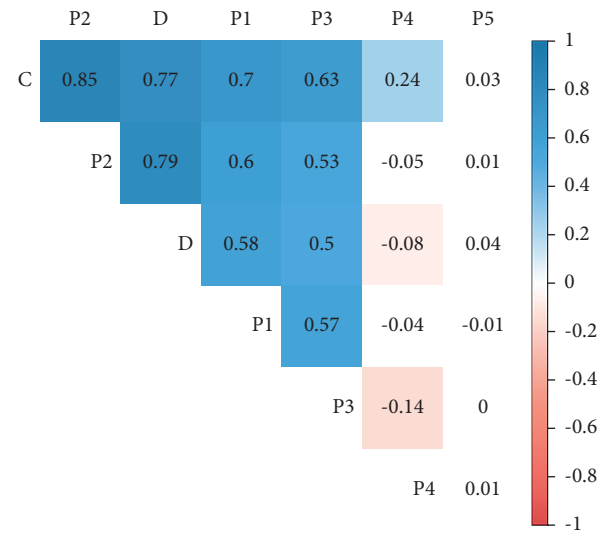

Figure 4: The correlation matrix between the predictors and the FCIP cost.

predictor parameters, excluding the geographical zone in which less influence on the cost phenomena was revealed when compared to ELM and PLS whose coefficient of determination (R^2) maxed out at 0.90 with RMSE of 36011.43 and 36013.16, respectively, for model M6 in the training phase. Similarly, in testing phase ELM and PLS, coefficient of determination (R^2) maxed out at 0.89 with RMSE of 37141.8 and 37140.3 for model M6. In addition, it is seen that the ratio of the MSE and the potential error which is denoted by MD is 0.89 for MARS M6 model on both cases, i.e., training and testing phases.

The model performances were assessed using graphical presentations such as scatter plots and Taylor diagram. Figure 5 shows the scatter plots between the actual observations and the predicted values. Among the three applied prediction models, MARS model is indicated as the best

Table 4: Prediction performance results over the training phase.

	R^2	RMSE	MAE	MAPE	Nash	MD
MARS model						
*M*1	0.76	57097.21	40090.04	0.11	0.76	0.78
*M*2	0.79	52443.11	35922.94	0.10	0.79	0.80
*M*3	0.83	46957.45	35377.61	0.10	0.83	0.81
*M*4	0.87	41427.33	31383.75	0.09	0.87	0.83
*M*5	0.94	28458.17	20444.55	0.06	0.94	0.89
*M*6	0.94	28458.17	20444.55	0.06	0.94	0.89
ELM model						
*M*1	0.72	61144.00	42644.28	0.12	0.72	0.76
*M*2	0.75	58309.91	40861.50	0.12	0.75	0.77
*M*3	0.79	52776.65	40372.48	0.11	0.79	0.78
*M*4	0.81	50951.98	39296.32	0.11	0.81	0.79
*M*5	0.90	36019.94	24352.69	0.07	0.90	0.87
*M*6	0.90	36011.43	24362.29	0.07	0.90	0.87
PLS model						
*M*1	0.72	61182.31	43650.76	0.12	0.72	0.75
*M*2	0.73	59634.10	42063.22	0.12	0.73	0.76
*M*3	0.79	52816.73	40059.10	0.11	0.79	0.78
*M*4	0.81	50951.98	39297.60	0.11	0.81	0.79
*M*5	0.90	36026.97	24406.13	0.07	0.90	0.87
*M*6	0.90	36013.16	24393.67	0.07	0.90	0.87

Table 5: Prediction performance results over the testing phase.

	R^2	RMSE	MAE	MAPE	Nash	MD
MARS model						
*M*1	0.75562	55636.4	39175.3	0.11317	0.75561	0.77964
*M*2	0.78542	52169	36042.4	0.10461	0.78513	0.79856
*M*3	0.81853	47949.3	36928.8	0.10698	0.81848	0.79537
*M*4	0.86669	41152.6	32861.1	0.09655	0.86629	0.82096
*M*5	0.94125	27422.7	19761.8	0.05454	0.94063	0.89374
*M*6	0.94125	27422.7	19761.8	0.05454	0.94063	0.89374
ELM model						
*M*1	0.70389	61662	43434.8	0.12204	0.69981	0.75368
*M*2	0.72687	59271.2	41879.9	0.11948	0.72264	0.76358
*M*3	0.77521	53666.1	41776.6	0.11991	0.77262	0.76784
*M*4	0.80047	50554.1	40050.8	0.11642	0.79822	0.77885
*M*5	0.894	37239.8	25507.3	0.06922	0.89051	0.86168
*M*6	0.89463	37141.8	25376.6	0.06887	0.89109	0.86241
PLS model						
*M*1	0.6997	62852.2	45250.3	0.12747	0.68811	0.74448
*M*2	0.71411	60628.3	43182.2	0.12358	0.70979	0.75411
*M*3	0.77381	53854.1	41631.9	0.11941	0.77102	0.76864
*M*4	0.80047	50554.3	40051.8	0.11643	0.79822	0.77884
*M*5	0.8941	37229.5	25541	0.06928	0.89057	0.86148
*M*6	0.89464	37140.3	25418.7	0.06898	0.89109	0.86217

identical match with high correlation value. On the other hand, Figure 6 shows the Taylor diagram map in which the prediction models were evaluated based on the distance coordination in accordance with multiple statistical metrics (i.e., standard deviation, RMSE, and correlation value).

5. Discussion

Various studies have been conducted to estimate a reliable parametric cost model, but there is no available study carried out for FCIP [5]. However, prediction of cost is not new; a simplex optimization of ANN weights was used to create a model for estimating the unit cost of highway projects with a mean absolute percentage error (MAPE) of 1% [16]. Another study used a combination of ANN and fuzzy logic to create a high-precision cost prediction model for water infrastructure based on the sum of squares of mistakes. During the validation phase, the researchers produced multiple prediction models with perceptions ranging from 4.6 percent to 0.6 percent

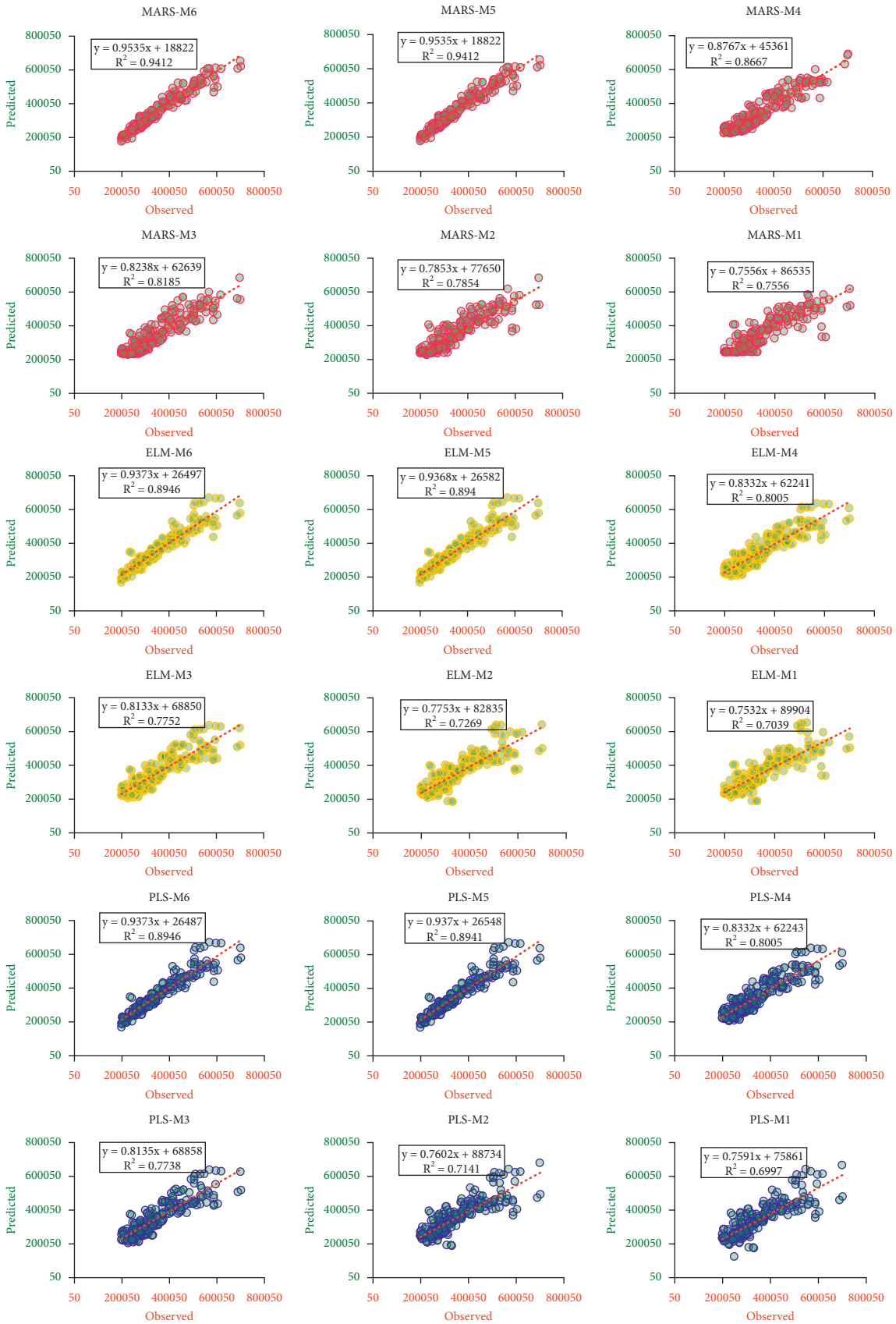

Figure 5: The scatter plots between the observed and predicted values of the cost over the testing phase for all tested input combinations and applied predictive models.

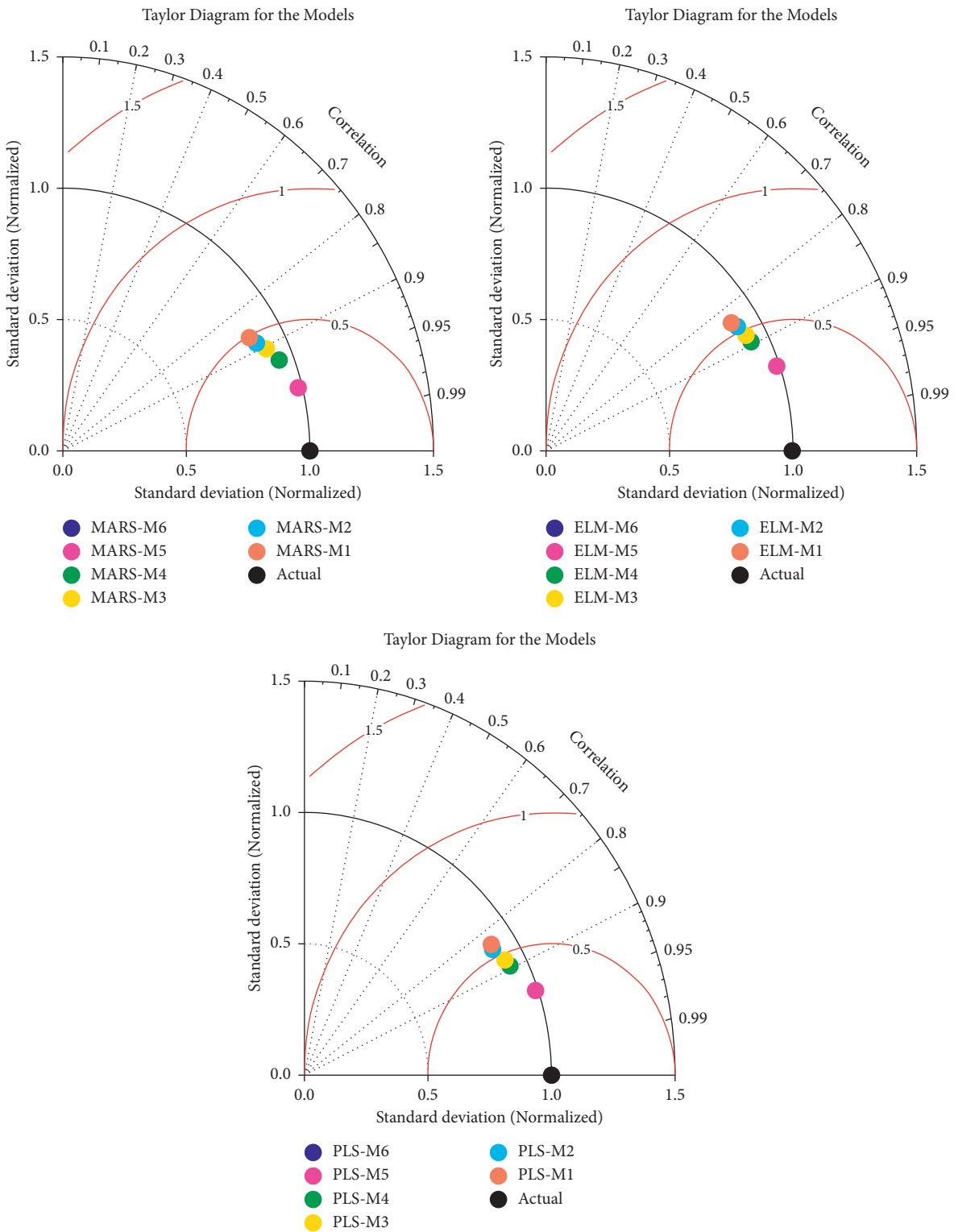

Figure 6: The Taylor diagram for the adopted modeling scenarios and the established machine learning models over the testing phase.

[19]. Furthermore, by varying the ANN structure, training function, and training algorithm until an optimum model was found, a researcher built a prediction model with a MAPE of 1.4 percent for the unit cost of the highway project in Libya [20].It is seen that, in this study, the value of MAPE for MARS model M6 in both training and testing phases ranges from 5% to 6% when compared with other models.

6. Conclusion and Remarks

The prediction of cost related to civil engineering project is considered as vital topic to be studied comprehensively. In this study, couple of machine learning models including extreme learning machine (ELM), multivariate adaptive regression spline (MARS), and partial least square regression (PLS) were developed to predict field canal improvement project (FCIP) cost. For the purpose of the modeling development, datasets related to irrigation projects were collected from the open source published literature. Input combinations were initiated based on the total length of the PVC pipeline, served area, geographical zone, construction year, and cost and duration of FCIP construction. The prediction results showed that MARS and ELM models were presented positively in comparison with the PLS model. However, MARS model reported the superior results. Also, the research finding exhibited that all the predictors are substantial toward the cost calculation with almost no influence for the geographical zone of the pipeline network.

Nomenclature

ANFIS: Adaptive neuro-fuzzy inference system
MD: Agreement index
ANN: Artificial neural network
BF: Basic function
R^2: Determination coefficient
ELM: Extreme learning machine
FCIP: Field canal improvement project
FL: Fuzzy logic
GCV: Generalized cross validation
IoT: Internet of things
NSE: Nash–Sutcliffe efficiency
ML: Machine learning
MAE: Mean absolute error
MAPE: Mean absolute percentage error
MARS: Multivariate adaptive regression spline
PLS: Partial least square regression
PVC: Polyvinyl chloride
PRESS: Predictive residual errors sum of square
RMSE: Root mean square error
SC: Soft computing
SVM: Support vector machine.

Data Availability

The data used in this study can be provided upon request from the authors.

Conflicts of Interest

The authors report no conflicts of interest.

References

[1] M. T. Rock, "Freshwater use, freshwater scarcity, and socioeconomic development," *The Journal of Environment & Development*, vol. 7, no. 3, pp. 278–301, 1998.

[2] J. Awange, "Global freshwater resources," *Lake Victoria Monitored from Space*, Springer, Berlin, Germany, pp. 3–19, 2021.

[3] H. S. Salem, Y. Yihdego, and H. H. Muhammed, "The status of freshwater and reused treated wastewater for agricultural irrigation in the Occupied Palestinian Territories," *Journal of Water and Health*, vol. 19, no. 1, pp. 120–158, 2021.

[4] D. P. Loucks and E. Van Beek, *Water Resource Systems Planning and Management: An Introduction to Methods, Models, and Applications*, Springer, Berlin, Germany, 2017.

[5] H. H. ElMousalami, A. H. Elyamany, and A. H. Ibrahim, "Predicting conceptual cost for field canal improvement projects," *Journal of Construction Engineering and Management*, vol. 144, Article ID 4018102, 2018.

[6] H. H. Elmousalami, "Artificial intelligence and parametric construction cost estimate modeling: State-of-the-art review," *Journal of Construction Engineering and Management*, vol. 146, no. 1, 2020.

[7] H. H. Elmousalami, "Data on field canals improvement projects for cost prediction using artificial intelligence," *Data in Brief*, vol. 31, Article ID 105688, 2020.

[8] H. G. Radwan, "Sensitivity analysis of head loss equations on the design of improved irrigation on-farm system in Egypt," *International Journal of Advanced Research in Science*, vol. 2, pp. 1–9, 2013.

[9] L. Sabol, "Challenges in cost estimating with building information modeling," *IFMA World Work*, pp. 1–16, 2008.

[10] H. D. S. Budiono, G. Kiswanto, and T. P. Soemardi, "Method and model development for manufacturing cost estimation during the early design phase related to the complexity of the machining processes," *International Journal of Technology*, vol. 5, no. 2, pp. 183–192, 2014.

[11] R. Sonmez and B. Ontepeli, "Predesign cost estimation of urban railway projects with parametric modeling," *Journal of Civil Engineering and Management*, vol. 15, no. 4, pp. 405–409, 2009.

[12] S.-H. Ji, M. Park, and H.-S. Lee, "Cost estimation model for building projects using case-based reasoning," *Canadian Journal of Civil Engineering*, vol. 38, no. 5, pp. 570–581, 2011.

[13] T. Hegazy and O. Moselhi, "Elements of cost estimation: a survey in Canada and the unit," *Cost Engineering*, vol. 37, p. 27, 1995.

[14] F. M. S. Al-Zwainy and T. H. Neran, "Investigation and evaluation of the cost estimation methods of Iraqi communication projects," *International Journal of Engineering and Management Research*, vol. 5, pp. 41–48, 2015.

[15] D. A. Hollar, W. Rasdorf, M. Liu, J. E. Hummer, I. Arocho, and S. M. Hsiang, "Preliminary engineering cost estimation model for bridge projects," *Journal of Construction Engineering and Management*, vol. 139, no. 9, pp. 1259–1267, 2013.

[16] T. Hegazy and A. Ayed, "Neural network model for parametric cost estimation of highway projects," *Journal of Construction Engineering and Management*, vol. 124, no. 3, pp. 210–218, 1998.

[17] M. M. Marzouk and R. M. Ahmed, "A case-based reasoning approach for estimating the costs of pump station projects," *Journal of Advanced Research*, vol. 2, no. 4, pp. 289–295, 2011.

[18] N. I. El-Sawalhi, "Modelling the parametric construction project cost estimate using fuzzy logic," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, p. 2, 2012.

[19] D. D. Ahiaga-Dagbui, O. Tokede, S. D. Smith, and S. Wamuziri, "A neuro-fuzzy hybrid model for predicting final cost of water infrastructure projects," in *Proceedings of*

the 29th Annual ARCOM Conference, pp. 181–190, SD Smith DD Ahiaga-Dagbui, Cambridge, UK, September 2013.

[20] E. Elbeltagi, O. Hosny, R. Abdel-Razek, and A. El-Fitory, "Conceptual cost estimate of Libyan highway projects using artificial neural network," *International Journal of Engineering Research in Africa*, vol. 4, pp. 56–66, 2014.

[21] C. Stoy, S. Pollalis, and O. Dursun, "A concept for developing construction element cost models for German residential building projects," *International Journal of Project Organisation and Management*, vol. 4, no. 1, pp. 38–53, 2012.

[22] K. Petroutsatou, E. Georgopoulos, S. Lambropoulos, and J. P. Pantouvakis, "Early cost estimating of road tunnel construction using neural networks," *Journal of Construction Engineering and Management*, vol. 138, no. 6, pp. 679–687, 2012.

[23] C. L. C. Roxas and J. M. C. Ongpeng, "An artificial neural network approach to structural cost estimation of building projects in the Philippines," in *Proceedings of the DLSU Research Congress*, Manila, Philippines, February 2014.

[24] I. Peško, V. Mučenski, M. Šešlija et al., "Estimation of costs and durations of construction of urban roads using ANN and SVM," *Complexity*, vol. 2017, Article ID 2450370, 2017.

[25] D. Jumas, F. A. Mohd-Rahim, N. Zainon, and W. P. Utama, "Improving accuracy of conceptual cost estimation using MRA and ANFIS in Indonesian building projects," *Built Environment Project and Asset Management*, vol. 8, no. 4, 2018.

[26] K. J. Kim and K. Kim, "Preliminary cost estimation model using case-based reasoning and genetic algorithms," *Journal of Computing in Civil Engineering*, vol. 24, no. 6, pp. 499–505, 2010.

[27] A. Mahmoodzadeh, M. Mohammadi, A. Daraei, H. Farid Hama Ali, A. Ismail Abdullah, and N. Kameran Al-Salihi, "Forecasting tunnel geology, construction time and costs using machine learning methods," *Neural Computing & Applications*, vol. 33, no. 1, pp. 321–348, 2021.

[28] H. H. Elmousalami, "Comparison of artificial intelligence techniques for project conceptual cost prediction: a case study and comparative analysis," *IEEE Transactions on Engineering Management*, vol. 68, pp. 183–196, 2020.

[29] S. T. Hashemi, O. M. Ebadati, and H. Kaur, "Cost estimation and prediction in construction projects: a systematic review on machine learning techniques," *SN Applied Sciences*, vol. 2, pp. 1–27, 2020.

[30] S. B. Green, "How many subjects does it take to do a regression analysis," *Multivariate Behavioral Research*, vol. 26, no. 3, pp. 499–510, 1991.

[31] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," *International Journal of Machine Learning and Cybernetics*, vol. 2, no. 2, pp. 107–122, 2011.

[32] G.-B. Guang-Bin Huang, H. Hongming Zhou, X. Xiaojian Ding, and R. Rui Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2012.

[33] Y. Miche, M. Van Heeswijk, P. Bas, O. Simula, and A. Lendasse, "TROP-ELM: A double-regularized ELM using LARS and Tikhonov regularization," *Neurocomputing*, vol. 74, no. 16, pp. 2413–2421, 2011.

[34] W. Zhang and A. T. C. Goh, "Multivariate adaptive regression splines and neural network models for prediction of pile drivability," *Geoscience Frontiers*, vol. 7, no. 1, pp. 45–52, 2016.

[35] N.-Y. Nan-Ying Liang, G.-B. Guang-Bin Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1411–1423, 2006.

[36] M. Hou, T. Zhang, F. Weng, M. Ali, N. Al-Ansari, and Z. Yaseen, "Global solar radiation prediction using hybrid online sequential extreme learning machine model," *Energies*, vol. 11, no. 12, p. 3415, 2018.

[37] L. Zhang and D. Zhang, "Evolutionary cost-sensitive extreme learning machine," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 12, pp. 3045–3060, 2017.

[38] G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in extreme learning machines: a review," *Neural Networks*, vol. 61, pp. 32–48, 2015.

[39] S. Sekhar Roy, R. Roy, and V. E. Balas, "Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of MARS and ELM," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 4256–4268, 2018.

[40] M. Ali and R. Prasad, "Significant wave height forecasting via an extreme learning machine model integrated with improved complete ensemble empirical mode decomposition," *Renewable and Sustainable Energy Reviews*, vol. 104, pp. 281–295, 2019.

[41] J. H. Friedman, "Multivariate adaptive regression splines," *Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[42] F. Yerlikaya-Özkurt, İ. Batmaz, and G.-W. Weber, "A review and new contribution on conic multivariate adaptive regression splines (CMARS): A powerful tool for predictive data mining," *Springer Proceedings in Mathematics & Statistics*, vol. I, pp. 695–722, 2014.

[43] M. A. Sahraei, H. Duman, M. Y. Çodur, and E. Eyduran, "Prediction of transportation energy demand: Multivariate adaptive regression splines," *Energy*, vol. 224, Article ID 120090, 2021.

[44] G. Zheng, W. Zhang, H. Zhou, and P. Yang, "Multivariate adaptive regression splines model for prediction of the liquefaction-induced settlement of shallow foundations," *Soil Dynamics and Earthquake Engineering*, vol. 132, Article ID 106097, 2020.

[45] Q.-S. Xu, M. Daszykowski, B. Walczak et al., "Multivariate adaptive regression splines-studies of HIV reverse transcriptase inhibitors," *Chemometrics and Intelligent Laboratory Systems*, vol. 72, no. 1, pp. 27–34, 2004.

[46] A. A. H. Alwanas, A. A. Al-Musawi, S. Q. Salih, H. Tao, M. Ali, and Z. M. Yaseen, "Load-carrying capacity and mode failure simulation of beam-column joint connection: Application of self-tuning machine learning model," *Engineering Structures*, vol. 194, pp. 220–229, 2019.

[47] L. Wang, C. Wu, X. Gu, H. Liu, G. Mei, and W. Zhang, "Probabilistic stability analysis of earth dam slope under transient seepage using multivariate adaptive regression splines," *Bulletin of Engineering Geology and the Environment*, vol. 79, no. 6, pp. 2763–2775, 2020.

[48] R. C. Deo, O. Kisi, and V. P. Singh, "Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5 tree model," *Atmospheric Research*, vol. 184, pp. 149–175, 2017.

[49] H. Abdi, "Partial least square regression (PLS regression)," *Encyclopedia of Social Science Research Methods*, vol. 6, pp. 792–795, 2003.

[50] H. Abdi and L. J. Williams, "Partial least squares methods: partial least squares correlation and partial least square regression," *Methods in Molecular Biology*, vol. 930, pp. 549–579, 2013.

[51] J. Hulland, "Use of partial least squares (PLS) in strategic management research: a review of four recent studies," *Strategic Management Journal*, vol. 20, no. 2, pp. 195–204, 1999.

[52] L. Kaufmann and J. Gaeckler, "A structured review of partial least squares in supply chain management research," *Journal of Purchasing and Supply Management*, vol. 21, no. 4, pp. 259–272, 2015.

[53] B. D. Liengaard, P. N. Sharma, G. T. M. Hult et al., "Prediction: coveted, yet forsaken? Introducing a cross-validated predictive ability test in partial least squares path modeling," *Decision Sciences*, vol. 52, no. 2, pp. 362–392, 2021.

[54] H. M. Lin, M. H. Lee, J. C. Liang, H. Y. Chang, P. Huang, and C. C. Tsai, "A review of using partial least square structural equation modeling in e-learning research," *British Journal of Educational Technology*, vol. 51, no. 4, pp. 1354–1372, 2020.

[55] R. Walpole, R. Myers, S. Myers, and K. Ye, "Probability and Statistics for Engineers and Scientist," *Pearson*, Boston, MA, USA, 2011.

[56] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling," *Journal of Hydrology*, vol. 377, no. 1-2, pp. 80–91, 2009.