



Hybrid machine learning approach for construction cost estimation: an evaluation of extreme gradient boosting model

Zainab Hasan Ali¹ · Abbas M. Burhan¹

Received: 17 March 2023 / Accepted: 1 April 2023 / Published online: 13 April 2023
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

Abstract

Estimating the project cost is an important process in the early stage of the construction project. Accurate cost estimation prevents major issues like cost deficiency and disputes in the project. Identifying the affected parameters to project cost leads to accurate results and enhances cost estimation accuracy. In this paper, extreme gradient boosting (XGBoost) was applied to select the most correlated variables to the project cost. XGBoost model was used to estimate construction cost and compared with two common artificial intelligence algorithms: extreme learning machine and multivariate adaptive regression spline model. Statistical indicators showed that XGBoost algorithm achieved the best performance with a coefficient of determination ($R^2=0.952$) and root mean square error (RMSE=590,609.782). Due to the reliability of XGBoost model, the presented approach can assist project managers in abstracting the influencing variables and estimating the cost of building projects. The findings of this study are helpful for the project's stockholder to decrease the errors of the estimated cost and take the appropriate decision in the early stage of the construction process.

Keywords Project cost estimation · Building project · Inflation · Extreme gradient boosting · Hybrid models

Introduction

The construction industry contributes to the social and economic wealth of developed and developing countries (Myers, 2016; Owusu-Manu et al., 2019). As a result, numerous researchers have studied the enhancement of the performance of construction projects (Al-Dhaheri & Burhan, 2022; Barnes, 1988; Bryde, 2008; Salim & Mahjoob, 2020; Wateridge, 1998). A successful performance means the construction project is finished within the three critical criteria: cost, duration, and quality (Abbas & Burhan, 2023; Mohammad et al., 2021; Pollack et al., 2018). Project quality can be controlled during the project's construction phases, while cost and duration need to estimate their amount at the beginning of a project (Azman et al., 2013). The project cost is considered a determination of project success and owners,

satisfaction due to its impact on their financial decisions (Huo et al., 2018; Matel et al., 2019). Estimating the project cost accurately helps decision-makers perform good feasibility studies and monitor the cash flows of construction projects (Shehu et al., 2014). Estimating construction cost is a complex problem characterized by incomplete information, risks, and uncertainties that lead to inaccurate results (Ahiaga-Dagbui & Smith, 2012; Fadhil & Burhan, 2022; Jing et al., 2019). Underestimated cost leads to cost overruns and financial problems for all parties in construction projects (Naghham Nawar Abbas & Burhan, 2022; Akintoye, 2000). To reduce these problems and achieve project objectives, several methods have been proposed in past studies to accurately estimate construction cost (Araba et al., 2021; Elhegazy et al., 2022; Sharma et al., 2021). Researchers in these studies have focused on two approaches, qualitative and quantitative methods. The qualitative method depends on expert opinion may lead to bias and inaccurate outcomes (Alex et al., 2010). Accordingly, recent studies have developed statistical approaches such as regression analysis (Al-Momani, 1996; Lowe et al., 2006) and artificial intelligence techniques for construction cost estimation (Shutian et al., 2017; Son et al., 2012). Several factors affect cost estimation, such as project characteristics and external economic

✉ Zainab Hasan Ali
zainabhasan222@gmail.com

Abbas M. Burhan
abbasm.burhan@coeng.uobaghdad.edu.iq

¹ Civil Engineering Department, College of Engineering, University of Baghdad, Baghdad, Iraq

parameters. Most of the studies have focused on project characteristics and ignore economic parameters. The reason behind this is there is no agreement among researchers on the impact of economic factors on the project cost, and there is little attention to incorporating these variables in the cost estimation process (Baloi & Price, 2003; Elhag et al., 2005; Gunduz & Maki, 2018; Zhao et al., 2020). This disagreement can be discussed using an inappropriate approach to investigate the influencing factors (Zhao et al., 2020). Consequently, there is a need to develop a suitable method to explore the impact of significant parameters of construction cost estimation (Wang et al., 2022).

According to a questionnaire survey and Among six influencing groups, Elhag et al. (2005) concluded that market conditions have gained the fourth rank in cost influencing factors (Elhag et al., 2005). In another study, the authors stated that economic variables have a high effect on the final cost of the construction project (Akinci & Fischer, 1998; Shane et al., 2009). In contrast, Hatamleh et al. (2018) indicated that market conditions had the least impact on cost estimation performance (Hatamleh et al., 2018). Several studies revealed that market conditions critically impact the cost estimation problem (Doloi, 2013; Iyer & Jha, 2005; Zhang et al., 2017). According to Zhao et al. (2019), market conditions have the most significant value among other affecting parameters. Wang et al. (2022) Stated that economic variables are more important than the project's parameters and have an essential role in improving cost estimation accuracy (Wang et al., 2022). One of the significant market conditions that affect cost estimation is inflation. Inflation significantly impacts the construction industry, especially the cost estimation process, due to its effects on material prices, labor wages, and equipment costs. These effects lead to problems among project parties and cost overruns (Musarat et al., 2021).

For the estimation process, several scholars used regression analysis as a popular method for cost estimation (Al-Momani, 1996; Lowe et al., 2006). The advantages of this method are its simplicity and the ability to produce simple results. However, this method has some drawbacks, such as it requires a defined mathematical expression and its inability to handle nonlinear relationships between input and output variables. In recent studies, soft computing algorithms have been used efficiently in construction management research and approved their ability to deal with complex systems and capture the nonlinear relationship between input and output parameters (Aljawder & Al-Karaghoul, 2022; Pan & Zhang, 2021). Artificial intelligence (AI) models help decision-makers capture historical data and deal with incomplete information in the early phases of the construction project (Almusawi & Burhan, 2020; Altaie & Borhan, 2018; Kaveh et al., 2008; Wang et al., 2022; Yaseen et al., 2020). Al-Momani (1996) Used a linear regression model

to estimate the cost of a construction project based on three project characteristics (Al-Momani, 1996). Artificial neural network (ANN) was used by Kaveh and Khalegi (1998) to estimate the compressive strength of concrete. The study revealed the capacity of ANN to predict plain and admixture concrete with accepted results (Kaveh & Khalegi, 1998). Another study investigated the improved neural network called counterpropagation neural net to analyze and optimize large-scale structures (Kaveh & Iranmanesh, 1998). The study showed the improved algorithm indicated better results than the traditional propagation neural network.

Three AI models named decision tree (DT), support vector machine (SVM), and ANN were developed to estimate construction cost in Turkey (Erdis, 2013). AI models were built based on 575 datasets collected from a public construction project and three input parameters, including the rate of price -cut, location, and duration of a construction project. Shutian et al. (2017) Used a Kalman filter with SVM model and multi-linear regression (MLR) to estimate construction cost in China (Shutian et al., 2017). The study showed that the presented methods are useful in estimating cost of building projects. A study by Mahalakshmi and Rajasekaran (2019) proposed an ANN model for 52 highway construction projects (Mahalakshmi & Rajasekaran, 2019). The study demonstrated that the ANN model with a back-propagation algorithm could predict construction cost with acceptable accuracy. Linear regression was hybridized with a random forest (RF) model to predict the labor cost of a BIM project (Huang & Hsieh, 2020). The authors concluded that the hybrid model effectively improves the prediction performance of labor cost in the BIM project. Three prediction models named multivariate adaptive regression spline (MARS), extreme learning machine (ELM), and partial least square regression (PLS) were applied to estimate the cost of field canal improvement (Shartooh Sharqi & Bhattarai, 2021). The researchers concluded that the MARS algorithm obtained the best cost prediction accuracy with high *R*-squared and low estimation error. The performance of Three AI models named RF, SVM and multi-linear regression (MLR) were investigated by Shoar et al. (2022) to predict cost overrun of engineering services of 95 construction projects. The study revealed that RF model performed better than the other two models in cost estimation.

To investigate the impact of influencing parameters on the construction cost problem, most researchers have used relative importance index, correlation statistics, structural equation modeling, and factor analysis (Cheng, 2014; Gunduz & Maki, 2018; Iyer & Jha, 2005). However, bias could occur in these techniques because the collected data depends on opinions and questionnaire surveys. Also, the traditional statistical methods used the linear correlation between input and output parameters, which led to an error in capturing the nonlinear relationship of the complex system. As a result,

mistakes are demonstrated in ranking influencing parameters and cost estimation results. It can be seen that traditional approaches cannot deal with the uncertainties and complexity of construction projects. Consequently, there is a necessity to develop an effective tool that can produce accurate cost estimation results. In recent years, a new AI algorithm called extreme gradient boosting (XGBoost) has been adopted to handle the complex nature of engineering problems. It is an efficient AI algorithm and has been used efficiently as a feature selector and a predictor by civil engineering researchers (Chakraborty et al., 2020; Chen & Guestrin, 2016; Falah et al., 2022; Tao et al., 2022).

The current research was done to investigate the efficiency of AI techniques in feature selection and prediction of construction cost estimation. Therefore, the research objectives are: (1) evaluate the ability of XGBoost, ELM, and MARS models in predicting construction cost estimation, and (2) examine the efficiency of XGBoost algorithm in selecting the influencing parameters of the cost estimation process incorporating inflation and project characteristics effects. This study contributes to the body of knowledge by helping decision-making identify and monitor the crucial parameters of cost estimation in a quantitative approach and enables project parties to compare the planning and estimated cost during the construction phase. The outcome of this study helps the project's stockholder decrease the errors in cost estimation and take the appropriate decision to reduce these defects.

Construction cost dataset description

The dataset of the construction cost was gathered from building projects in Iraq. The data were collected using the survey of building documents for nineteen construction projects built for the period between 2016 and 2021. The collected data includes seven parameters named area of ground floor (GFA), total area of floor (TFA), duration (D), number of elevator (EN), floor number (FN), type of footing (FT), and

inflation (F). from the survey and reviewing of projects, documents, project characteristics were gathered while inflation information was taken from the open-source central Iraqi bank (<https://cbiraq.org/>). The statistical measures of the cost dataset, including minimum, maximum, mean, median, standard deviation, skewness, and kurtosis, are illustrated in Table 1. The statistical measures show that the mean number of project cost is 2177699 \$. The minimum and the maximum values of the duration are equal to 122 days and 787 days. It can be recognized that the value of kurtosis of the most gathered parameters is less than 3, which indicates that the collected data is normally distributed.

Methodology

Extreme gradient boosting (XGBoost)

Extreme gradient boost algorithm is a new development of a tree-based boosting model introduced as an algorithm that can fulfill the demand of prediction problems (Chen & Guestrin, 2016; Friedman, 2002). It is a flexible model, and its hyperparameters can be tuned using soft computing algorithms (Eiben & Smit, 2011; Probst et al., 2019). The most important reason behind the success of XGBoost is the algorithm's flexibility and ability to scale to billions of parameters in the distributed system. These properties make the algorithm more accurate and faster than the existing algorithm. Whereas, the traditional methods used trial and error and personal experience to choose the optimal parameters of the algorithm. Gradient boosting aims to produce more robust models by combining weak learners in an iterative process. In every iteration, the loss function can be reduced using the residual of the previous trees (Zhang et al., 2019). Every training tree can be modeled based on the residual of the previous predictors, and the new tree is added to the developed model for updating the residual value. XGBoost has proven successful results among tree models such as random forest, gradient boosting tree, and AdaBoost. The

Table 1 Statistical characteristics of the collected datasets

	Minimum	Maximum	Mean	Median	Std. deviation	Skewness	Kurtosis
GFA (m ²)	200	5320	1608	1382	1172.561	1.246	1.035
TFA (m ²)	344	9800	3456	2800	2480.113	0.879	- 0.152
FN	1	6	2.913	2	1.714	0.574	- 1.079
EN	0	4	0.989	1	1.124	0.939	- 0.057
FT (raft = 1; seperated = 2)	-	-	-	-	-	-	-
F (%)	0.1	7.5	2.758	1.9	2.386	0.651	- 0.931
D (days)	122	787	373.5	351.5	157.519	0.729	0.208
C (\$)	26,756	6,451,519	2,177,699	1,434,969	1,787,809	0.779	- 0.632

reason behind this effectivity is its ability to be scalable in all scenarios of prediction problems and the fast running of the system on a single machine.

The regularized objective loss function ‘ $f(L)$ ’ for L th in XGBoost model can be expressed as shown below:

$$f(L) = \sum_{i=1}^n l(y^{(i)}, \hat{y}_L^{(i)}) + \sum_{j=1}^L \Omega(f_j), \tag{1}$$

where n represents the number of observations; $\hat{y}_L^{(i)}$ is the estimation of observation i th for iteration L ; $l(-)$ represents the loss function; and Ω is the regularization term which is computed using the following expression:

$$\Omega(f) = \gamma N + \frac{1}{2} \lambda \sum_{j=1}^N \omega_j^2 \tag{2}$$

where N denotes the number of nodes in each leaf, and γ and λ are two symbols used to manage regularization.

The number of trees in XGBoost model is optimized using the following equation to produce the best results as follows below:

$$f(L) = \sum_{i=1}^n l(y^{(i)}, \hat{y}_{L-1}^{(i)} + f_L(x^{(i)})) + \Omega(f_L) \sum_{j=1}^{L-1} \Omega(f_j). \tag{3}$$

Furthermore, second-order Taylor expansion is applied for managing objective functions, as shown in the following equation:

$$f(L) = \sum_{i=1}^n \left[l(y^{(i)}, \hat{y}_{L-1}^{(i)}) + g_i \cdot f_L(x^{(i)}) + \frac{1}{2} h_i \cdot f_L(x^{(i)}) \right] + \Omega(f_L) + K, \tag{4}$$

where g_i is equal to $\partial_{\hat{y}_{L-1}} l(y^{(i)}, \hat{y}_{L-1}^{(i)})$ and represents the first-order derivatives of loss functions; h_i is $\partial_{\hat{y}_{L-1}}^2 l(y^{(i)}, \hat{y}_{L-1}^{(i)})$ and reflects the second-order derivatives of loss functions; K is a constant number.

To select input parameters, XGBoost model is considered a robust algorithm for such kinds of these problems. XGBoost efficiently builds boosting trees parallel to choose the essential parameters based on their weight (Friedman, 2002). gain, cover, and frequency are the popular approaches used by XGBoost for ranking evaluation. The gain evaluates the contribution of each feature in developing the prediction model. The cover revealed the number of the actual values for each feature and the frequency shows the number of features in the gradient boosted trees. The mathematical equation of ranking evaluation can be expressed as below:

$$N_v = \sum_{L=1}^L \sum_{l=1}^{X-1} I(V_L^l, v), \tag{5}$$

where L represents iterations, number, N is the nodes' number in each leaf, and (V_L^l) is the feature for the node l , and $I()$ is the indication term and (V_L^l, v) can be calculated using the following expression. The graphical scheme of XGBoost algorithm is presented in Fig. 1.

$$(V_L^l, v) = f(x) = \begin{cases} 1 & \text{if } V_L^l = v \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Extreme learning machine (ELM)

An extreme learning machine is a powerful ANN method characterized by simplicity and a non-iterative method for training a single-layer neural network (Kardani et al., 2021; Shi-fan et al., 2021). ELM algorithm can reach optimum performance more efficiently than the traditional ANN. A linear function has been used as an activation function for the input and output layer, and for the hidden layer, the method applied a sigmoid activation function (Hou et al., 2018). In the training process, ELM utilizes random weights for hidden neurons, and then it uses a Moore–Penrose Pseudo-inverse function to determine the weight in the output layer. This process makes the ELM model quickly and enables it to deal with many different transfer functions (Huang et al., 2004, 2006). The mathematical equation of the training ELM model is presented below:

$$\sum_{i=1}^N t_i - \tilde{t}_i = \sum_{i=1}^N \|t_i - \sum_{l=1}^L \beta_l \cdot g(w_l \cdot x_i + b_l)\| = 0, \tag{7}$$

where t_i represents the outcome vector and x_i refers to the input vector. Equation (7) can be written in the following expression:

$$H\beta = T, \tag{8}$$

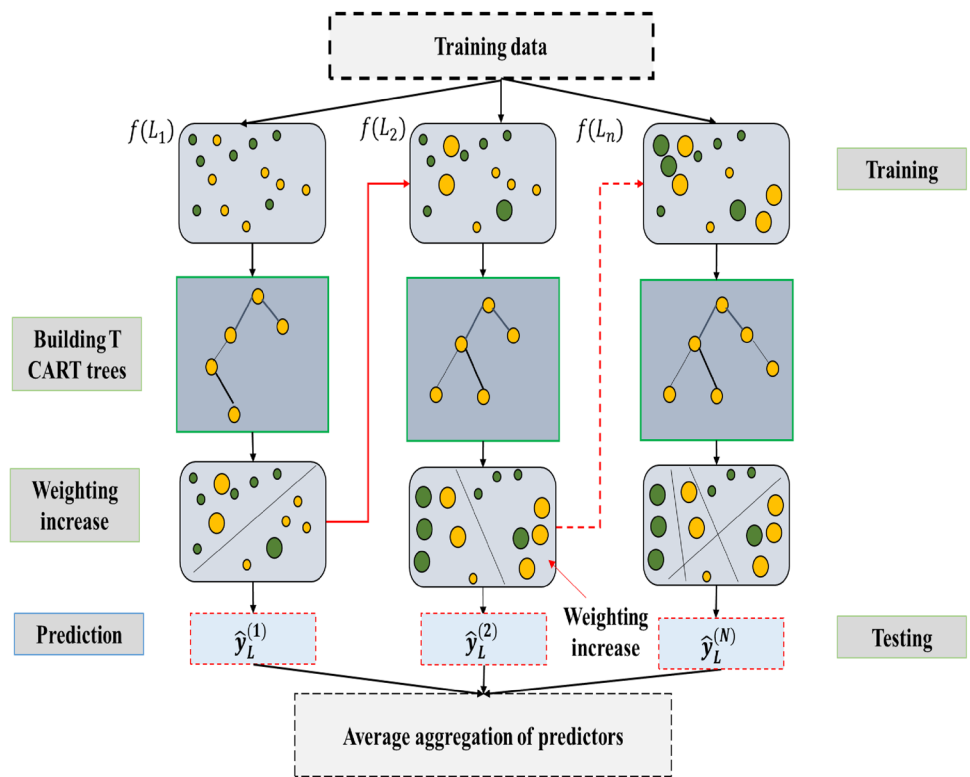
$$H = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_L \cdot x_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_L \cdot x_N + b_L) \end{bmatrix}_{N \times L},$$

$$\beta = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,m} \\ \vdots & \ddots & \vdots \\ \beta_{L,1} & \cdots & \beta_{L,m} \end{bmatrix}_{L \times m}, \tag{9}$$

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

where H represents the output of the hidden layer, β is a matrix that denotes the connection weights between the hidden and output layer, and T is the matrix of output predicted value depending on N training sets. To develop ELM model, the presented procedure is following: first, create random weights for the hidden layers, then generate H and

Fig. 1 Graphical scheme of XGBoost model

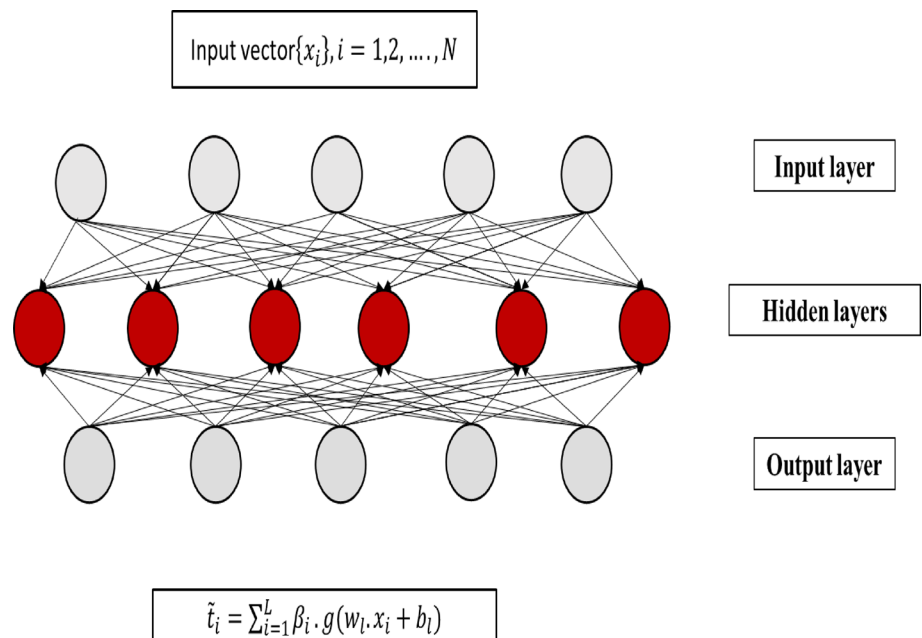


T , which represent the matrix of the hidden and output layer, and finally calculate the weight of the output layer using the below equation:

$$\tilde{\beta} = H^\dagger T, \tag{10}$$

where H^\dagger refers to the Moore–Penrose Pseudo-inverse function. The graphical scheme of ELM model is illustrated in Fig. 2.

Fig. 2 Graphical scheme of ELM model



Multivariate adaptive regression spline model (MARS)

MARS model is a nonlinear machine learning algorithm has been introduced to explore the nonlinearity of complex systems using piecewise segments (Friedman, 1991; Ikeagwuani, 2021; Naser et al., 2022). MARS model is a nonparametric method, and it is called a curve-based algorithm (Wu & Fan, 2019). The algorithm is similar to a tree-based model using the iterative approach in the learning process and selecting the critical features in the prediction problem. MARS model revealed better efficiency than other machine learning algorithms like ELM and SVM models (Guo et al., 2022; Shartooh Sharqi & Bhattarai, 2021; Wu & Fan, 2019). The concept of developing MARS model is as follows: At first, The MARS model changes the nonlinear regression model to a multiple linear regression model for the training dataset. Training data are divided into several groups to develop the linear regression model for each section. Each section has boundaries called the knots, which are identified using the adaptive regression algorithm. In each group of the divided data, the MARS model creates a basic function (BF) to represent the relationship between input and predicted parameters, as shown in the mathematical equation below:

$$BF = \max(0, x - t) = \begin{cases} x - t, & \text{if } x \geq t \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

where x is the value of the input variable and t represents the threshold value.

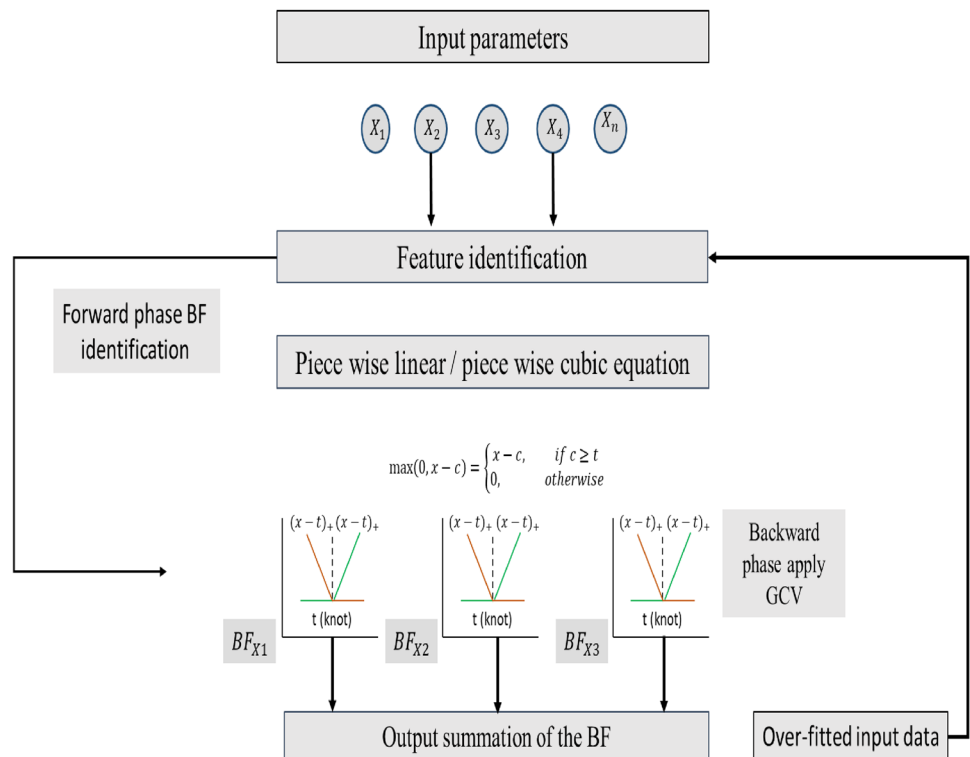
This process is called a forward phase, where the algorithm chooses the optimum input variables of predicted models. The final phase of the MARS model is called the backward phase. The algorithm eliminates unused parameters selected in the early phase to improve the prediction process's performance. The elimination of unnecessary parameters is achieved using a pruning algorithm based on generalized cross-validation (GCV), which is calculated as shown below:

$$GCV(M) = \frac{(1/N) \sum_{i=1}^N (O_i - f(x_i))^2}{(1 - C(M)/N)^2}, \tag{12}$$

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

where O_i represents the real value; N is the number of the dataset; $f(x_i)$ is the estimated value, M is the number of basic functions, and $C(M)$ represents the penalty factor. d ranges between 2 and 4 and represents the optimization cost of BFs. The final step in the MARS model is combining the BF function to get the predicted outcome of the developed model. Figure 3 provides the structure of MARS model.

Fig. 3 Systematic scheme of MARS model



Modeling process and performance evaluation

This study uses R programming language to develop the presented AI models. XGBoost was taken as a feature selector due to its ability to handle nonlinear and complex relationships. The libraries named xgboost, ggplot2, and Matrix were used to ease the selection process of input parameters. To run the XGBoost algorithm, xgboost function was used with max.depth (10), eta (0.3), nrounds (100), and xgb.importance functions were applied to illustrate the best input selection. For the prediction process, XGBoost's parameters were tuned using expand.grid function. The hyperparameters were tuned as follows: rounds set as 75:150 to determine the number of iterations; eta sets as 0.001, 0.01, 0.1 to control the learning rate of the algorithm; xgbtree the boosting method; max depth used as 5, 8, 10; gamma sets as 0, 1, 2; minchildweight (2); subsample (0.6); and colsamplebytree (0.8). For MARS model, two libraries named plotrix and earth were applied. The function expand.grid was used to control hypermeters of the algorithm such as degree and nprune; Degree sets as 1:8; and prune used as 1:100 with length.out (10). In the case of ELM model, the libraries kernlab, elmNNRcpp, and Matrix were applied. The ELM model parameters were set as nhid (100) to represent the number of hidden layers; actfun (sin) to control the activation function, and init.weights (uniform_positive) to choose the initial weight in the ELM. The integration of XGBoost with AI models is presented in Fig. 4.

Dataset was split into two phases, 70% for training and 30% for testing. The outcome of the AI models was evaluated by using statistical methods, including *r*-squared, mean absolute percentage error (MAPE), root mean square error

(RMSE), mean absolute error (MAE) (Shehu et al., 2014) as shown in the following equations:

$$R^2 = \left(\frac{\sum_{i=1}^N (y_p - \bar{y}_p) \cdot (y_a - \bar{y}_a)}{\sqrt{\sum_{i=1}^N (y_p - \bar{y}_p)^2 \sum_{i=1}^N (y_a - \bar{y}_a)^2}} \right)^2, \tag{13}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_p - y_a}{y_a} \right|, \tag{14}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_p - y_a)^2}{n}}, \tag{15}$$

$$MAE = \frac{\sum_{i=1}^N |y_p - y_a|}{N}, \tag{16}$$

where y_p and y_a represent the predicted and actual values of construction cost; \bar{y}_a is the average value of the actual data of construction cost, and N signifies the number of construction projects.

Results and discussion

Statistical evaluation

In this study, XGBoost was applied as a robust algorithm for prediction and input selection. The results of feature combinations of construction cost prediction are presented in Table 2. It can be seen that the most correlated variable to cost estimation is inflation. The second input combination is

Fig. 4 Processing phases of the applied models

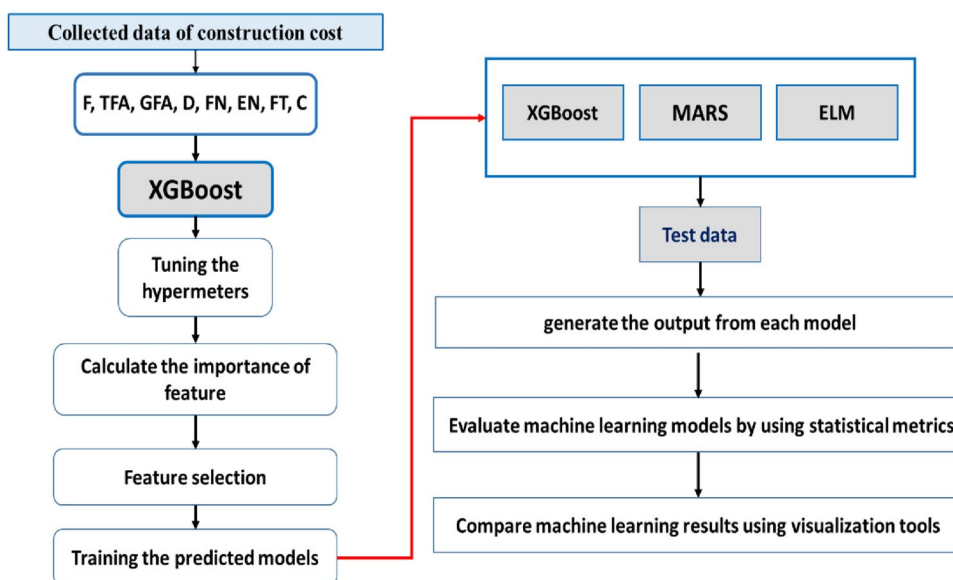


Table 2 Feature combinations selected by XGBoost

Models	Feature combinations
M1	$C = F$
M2	$C = F, TFA$
M3	$C = F, TFA, GFA$
M4	$C = F, TFA, GFA, D$
M5	$C = F, TFA, GFA, D, FN$
M6	$C = F, TFA, GFA, D, FN, EN$
M7	$C = F, TFA, GFA, D, FN, EN, FT$

the inflation and total floor area. The most correlated parameters in the third combination are inflation, total area of floor, and area of ground floor. The rest of feature combinations are reported in Table 2.

Statistical indicators of the introduced machine learning algorithms for the training/ testing phase are presented in Tables 3 and 4. The results showed that XGBoost model achieved an outstanding performance for the training phase, where all input combinations attained R^2 more than 0.8. ELM model illustrated a good enhancement in prediction accuracy for all model combinations when increasing the

Table 3 Performance measurements of the applied models for the training division

	R^2	RMSE	MAE	MAPE
<i>XGBoost model</i>				
M1	0.83881	673,221.9041	501,543.3362	0.73382
M2	0.84867	626,349.0281	444,491.725	0.51656
M3	0.95058	383,449.4023	246,906.7032	0.43818
M4	0.96473	325,415.6342	212,591.1021	0.33242
M5	0.95901	327,553.6062	226,749.5562	0.38002
M6	0.97822	268,500.0294	166,301.5715	0.22701
M7	0.97853	270,774.5197	170,766.6061	0.24218
<i>ELM model</i>				
M1	0.77669	760,883.3087	586,218.6132	0.83649
M2	0.79993	720,190.1008	517,265.5828	0.4951
M3	0.84081	642,433.1251	484,766.3293	0.54611
M4	0.85456	614,644.2939	455,212.7321	0.51885
M5	0.84242	639,167.0795	483,191.602	0.56233
M6	0.85526	612,580.7163	467,137.5367	0.70894
M7	0.83845	647,826.2112	482,758.1373	0.72025
<i>MARS model</i>				
M1	0.72866	838,810.8081	656,991.3137	0.75273
M2	0.72267	870,423.9231	612,268.7411	0.86023
M3	0.77317	770,308.2773	515,229.9608	0.75458
M4	0.81012	703,924.2808	476,741.9313	0.75793
M5	0.83006	665,282.635	409,738.5077	0.64087
M6	0.79683	728,222.516	499,940.0069	0.78483
M7	0.78041	758,228.0846	501,818.4464	0.78978

Table 4 Performance measurements of the applied models for the testing division

	R^2	RMSE	MAE	MAPE
<i>XGBoost model</i>				
M1	0.66322	1,141,949.801	736,321.8604	0.64118
M2	0.8554	726,844.6512	550,674.9322	0.33461
M3	0.86972	859,466.6779	560,469.2227	0.24263
M4	0.90535	705,627.2557	402,050.3197	0.17322
M5	0.95216	590,609.7821	332,157.171	0.0875
M6	0.88614	906,337.0054	638,437.6939	0.27955
M7	0.8712	842,285.5216	564,264.8955	0.23297
<i>ELM model</i>				
M1	0.62558	1,200,899.141	821,912.9956	0.62408
M2	0.8137	828,387.7434	635,262.9907	0.42241
M3	0.8377	777,982.897	562,190.5526	0.33338
M4	0.84264	758,649.9922	526,958.7476	0.32656
M5	0.86005	732,387.351	476,386.9595	0.26184
M6	0.79751	857,455.4889	637,808.5125	0.32674
M7	0.76792	920,453.6042	675,721.6844	0.33527
<i>MARS model</i>				
M1	0.61958	1,265,364.669	910,970.8102	0.55136
M2	0.7787	907,643.8118	649,858.8521	0.25386
M3	0.86203	730,717.4588	573,230.6743	0.3102
M4	0.82683	839,332.8152	575,758.8835	0.30234
M5	0.77935	1,004,475.6	650,489.3751	0.3262
M6	0.7266	1,024,789.776	649,903.1701	0.27644
M7	0.69415	1,117,299.421	682,735.8779	0.30479

number of input variables. For all the developed AI models, the best results were attained by XGBoost-M6, where $R^2=0.97822$, $RMSE=268,500.0294$, $MAE=166,301.5715$, and $MAPE=0.22701$. For the testing division, XGBoost model shows a noticeable performance of cost prediction for all combinations with r -squared more than 0.8 except M1, where R^2 reduces to 0.66. The best prediction accuracy was achieved by XGBoost-M5, where $R^2=0.95216$, $RMSE=590,609.7821$, $MAE=332,157.171$, and $MAPE=0.0875$. MARS model indicated less prediction performance than the other AI models with r -squared less than 0.78 except M3 and 4, which attained $R^2 > 0.8$. the best performance was attained by MARS-M3 with $R^2=0.86203$ and $RMSE=730,717.4588$. for ELM model, the best combination was revealed using five input variables where ELM-M5 achieved $R^2=0.86005$ and $MAPE=0.26184$.

Figures 5, 6, and 7 depict the scatter diagram for the applied models for the testing part. It can be seen that XGBoost model revealed an excellent improvement in the prediction performance when increases the input variables and the best results were gained by XGBoost-M5 with $R^2=0.95$. For ELM model, the developed algorithms indicated good prediction accuracy for all input combinations

Fig. 5 Scatter plot graph of XGBoost model over the testing phase

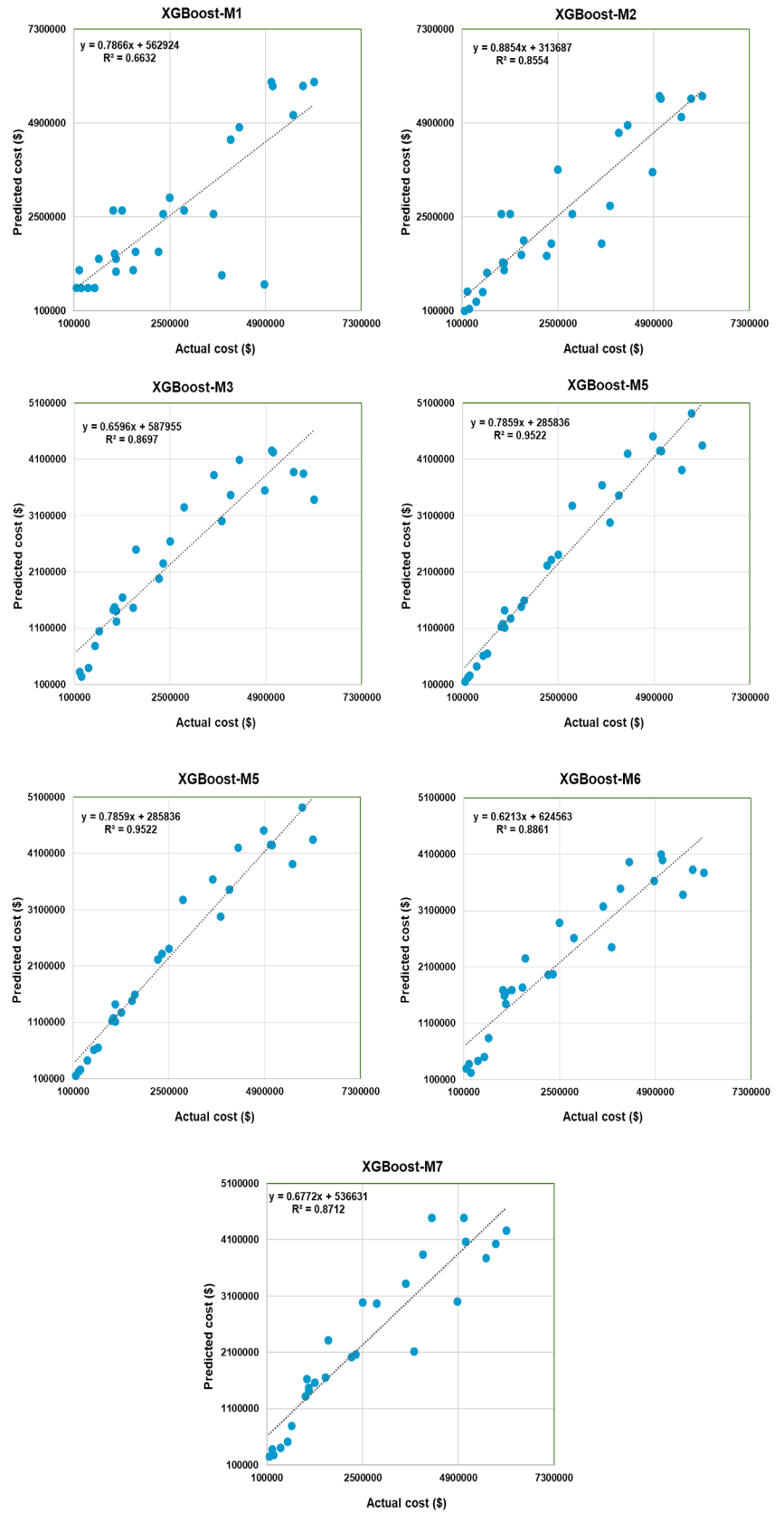


Fig. 6 Scatter plot graph of ELM model over testing phase

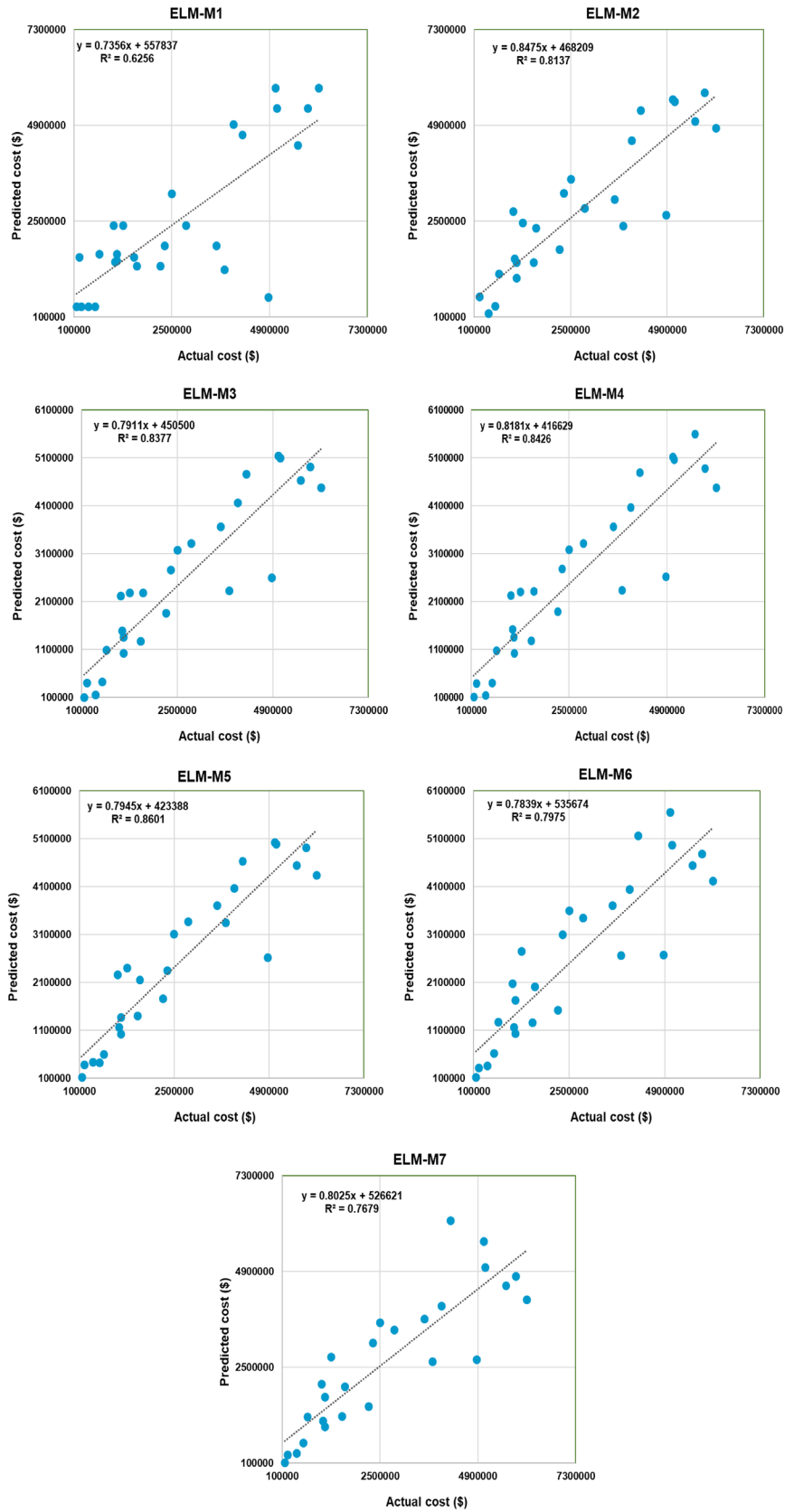
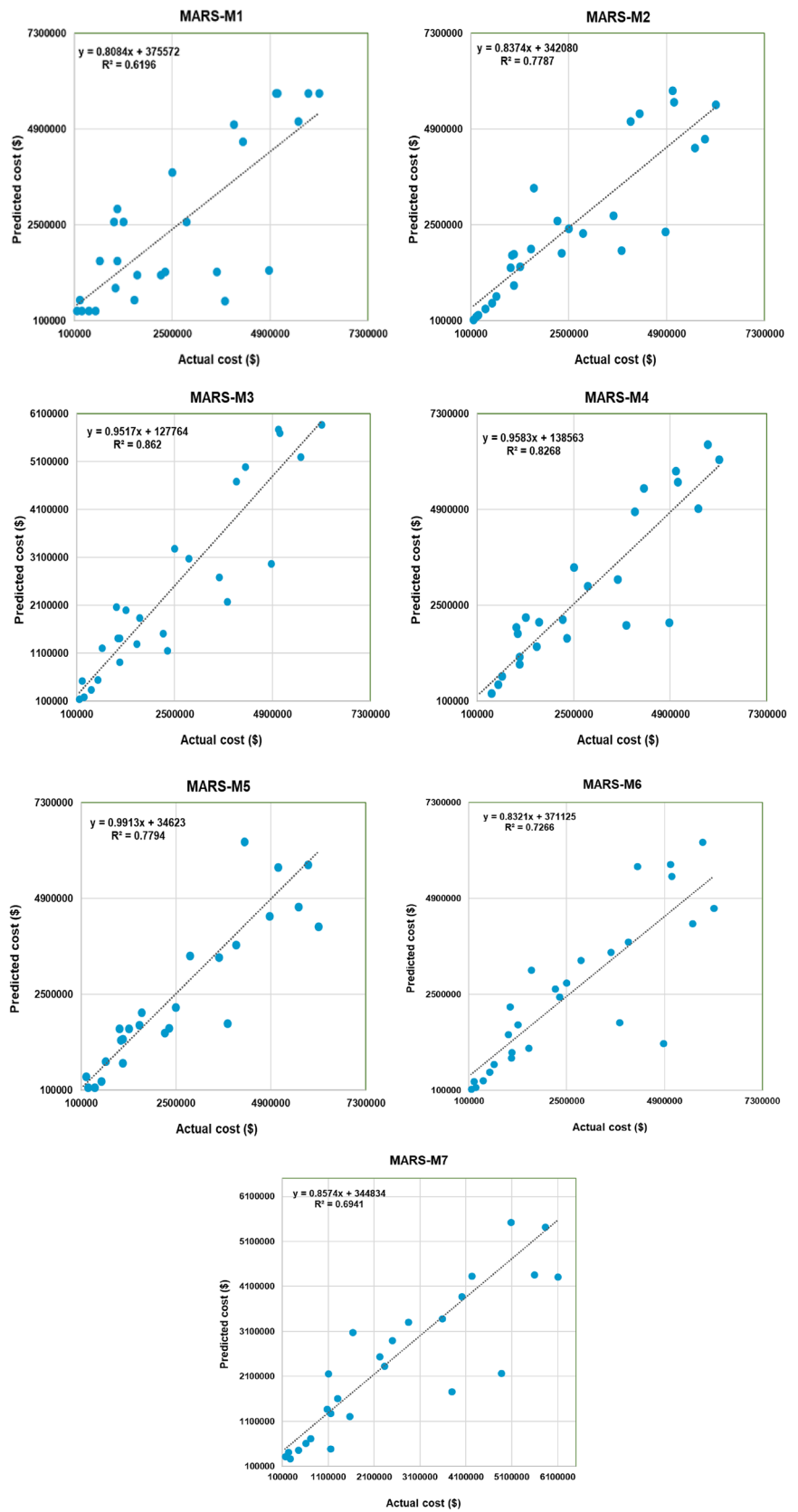


Fig. 7 Scatter plot graph of MARS model over testing phase



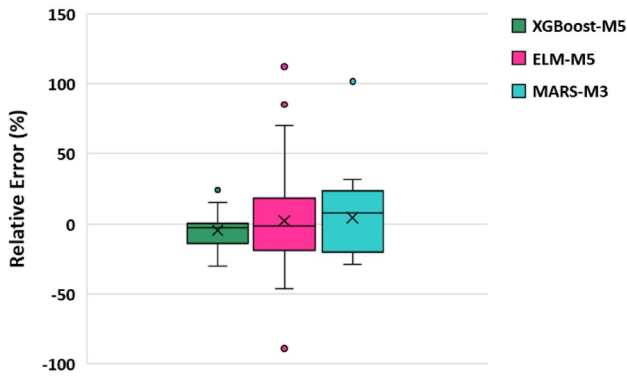


Fig. 8 Relative error plot for the developed models over the testing phase

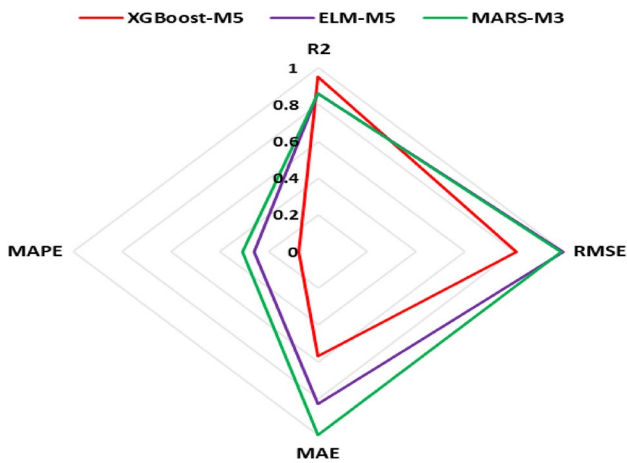
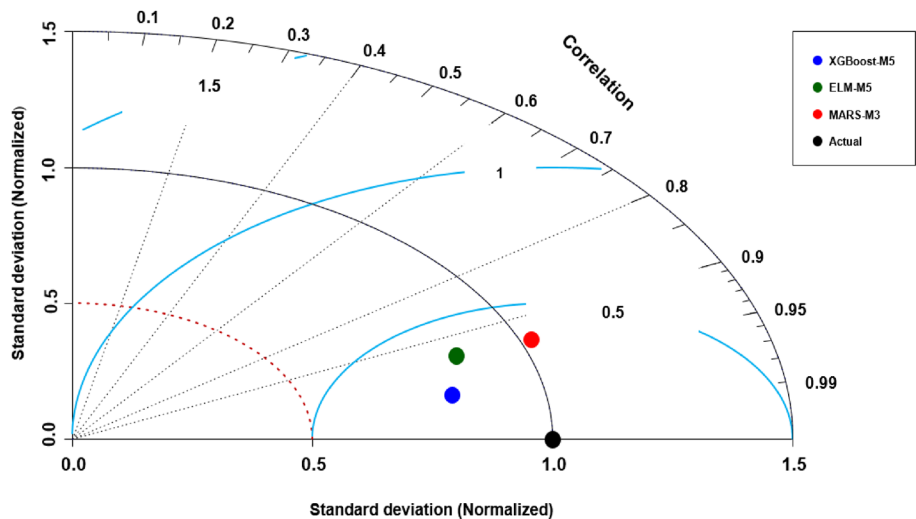


Fig. 9 Spider plot for the presented AI models over testing phase

except M1 where R^2 is less than 0.7 as shown in Fig. 6. MARS model shows less prediction accuracy than the other applied models and MARS-M3 has gained the best

Fig. 10 Taylor plot for the applied algorithms over the testing phase



performance with a coefficient of determination equal to 0.862 as depicted in Fig. 7.

Box plot and spider plot are another graphical tools were used to illustrate the performance metrics of the introduced AI algorithms as presented in Figs. 8 and 9. Figure 8 indicated that the minimum relative error was reflected by XGBoost-M5 followed by MARS-M3 and ELM-M5. The graphical results showed that XGBoost-M5 model presented a significant reduction in residual error and minimum positive error without a negative outlier. ELM-M5 revealed the maximum negative error with one outlier point using five input variables. Figure 9 shows the comparison results of AI models using statistical metrics in the form of spider plot. The findings showed that XGBoost-M5 gained the highest r -squared and lowest performance errors than the other models. Figure 9 also showed that although MARS-M3 and ELM-M5 have an equal R^2 , MARS-M3 attained higher absolute error than ELM-M5.

Figure 9 was constructed to demonstrate the relationship between the developed models and actual cost based on 3 statistical matrices (i.e., RMSE, correlation, standard deviation) was illustrated by Taylor diagram (Taylor, 2001) as depicted in Fig. 10. The developed diagram showed that the closest position to the actual point was achieved by XGBoost-M5 with a correlation coefficient maxed out 0.95. The visualization results revealed that XGBoost model achieved the nearest distance to the observation points than ELM and MARS models, which refers to the efficiency of the XGBoost approach in cost estimation problems.

Validation against past studies

Over the past years, numerous studies have been conducted on cost estimation. A study by Juszczak (2018) evaluated the ability of SVM model in the construction cost estimation of the residential building. The researcher revealed

that the presented model attained low MAPE with a range between 7% and 8.19. In another study, SVM model was also investigated for cost estimation of bridge construction (Juszczak, 2019). The developed model showed its ability in cost estimation with a coefficient of determination equal to 0.94. Genetic algorithm (GA) was coupled with ANN by Hashemi et al. (2019) to enhance the results of estimated cost. The study concluded that ANN-GA achieved good prediction performance with an accuracy equal to 0.94. XGBoost model was compared with twenty predicted models to measure the estimated cost of field canal projects by (Elmousalami, 2020). The researcher indicated that the presented model achieved high prediction accuracy with $R^2=0.929$. It can be recognized that previous studies made good efforts to enhance the effectiveness of cost estimation process; however, they have focused on exploring project characteristics and ignoring other factors such as economic factors. Also, they demonstrated little attention to feature selection algorithms and hybrid models. This study is different from the previous studies by (1) investigating both project characteristics and economic factors represented by inflation, (2) using a recent AI algorithm (XGBoost) for feature selection and prediction and validating it with ELM and MARS models, and (3) the hybrid model achieved an excellent prediction accuracy using only five input parameters with R^2 equal to 0.952.

Discussion

Applying hybrid models in complex prediction processes like cost estimation enhances prediction accuracy and reduces estimation error. The results of XGBoost in the input selection process revealed that the most correlated variables to cost estimation is inflation. This result is agreed with Wang et al. (2022) who concluded that economic factors are more important than project characteristics. Analysis of the predicted models showed that all AI algorithms are able to estimate construction cost because all the developed models achieved an acceptable prediction performance. XGBoost model exhibited excellent performance in cost estimation using five input variables (i.e., inflation, total area of floor, area of ground floor, duration, floor number), where $R^2=0.952$ and $MAPE=0.087$ as reported in Table 4. The poorest accuracy was achieved using one input parameter, where XGBoost attained $R^2 < 70$ as illustrated in Fig. 5. For ELM model, the best statistical indicators were gained using five input parameters where $RMSE=732,387.351$ and $MAE=476,386.9595$. The scatter plot diagram revealed that applying models on datasets with 2–5 input parameters achieved high prediction outcomes with R^2 more than 0.8 as shown in Fig. 6. The lowest performance was revealed by ELM-M1 with $R^2=0.625$ and high mean absolute error as shown in Table 4. MARS model attained good prediction

accuracy using three variables with a coefficient of determination maxed out 0.8 depicted in Table 4. The comparison results revealed that XGBoost model outperformed ELM and MARS models with R^2 more than 0.9 using four and five input variables. Spider plot revealed that XGBoost algorithm gained excellent performance metrics and outperformed ELM and MARS models as shown in Fig. 9. The visualization results showed the reliability of XGBoost model in cost estimation by having the least residual error and nearest distance to the actual point as demonstrated in Figs. 8 and 10.

Conclusion

Estimating the construction cost accurately is an important issue in construction management studies. This study introduces XGBoost model as an input selector and a predictor to enhance cost estimation accuracy. XGBoost model was compared with two well-known AI algorithms named ELM and MARS models. The study was conducted based on datasets collected from nineteen construction projects. Tabulated metrics and graphical schemes were constructed to examine the applied AI models. The feature selection results revealed that inflation is the most correlated parameter to project cost followed by project characteristics. The comparison analysis between the predictive models showed that all the developed models exhibited efficient predictability when the number of input parameters increased. The tabulated results showed that XGBoost model gained an excellent performance in all input combinations with r -squared maxed out 0.8 except M1 where the coefficient of determination was reduced to 0.663. The study revealed that incorporating inflation with project characteristics enhances the accuracy of the estimated cost. The study found that XGBoost gains the highest prediction results using five input variables. Furthermore, the study showed that XGBoost model provided an excellent capacity in feature selection and prediction processes within a complex cost estimation system. This study focuses on the impact of project characteristics and inflation on the cost estimation modeling of building projects. Studying the impact of the other influencing variables can enhance the accuracy of the cost estimation process. For future study, more affected variables like the characteristics of the client and other stakeholders can be explored to increase cost estimation accuracy. Also, other recent versions of AI models like deep neural network can be investigated to reduce the error of construction cost estimation.

Acknowledgements The authors would like thank the university of Baghdad, college of engineering for providing their technical support to this paper.

Author contributions The authors collectively introduced the concept of the study, developed research methodology, generated tables and

figures, prepared the original draft, reviewed and edited the revisions prior to submission.

Funding The research has no financial support for the publication of this article.

Data availability Data will be available upon the requests from the corresponding author.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

References

- Abbas, N. N., & Burhan, A. M. (2022). Investigating the causes of poor cost control in Iraqi construction projects. *Engineering, Technology and Applied Science Research*, 12(1), 8075–8079.
- Abbas, N. N., & Burhan, A. M. (2023). Evaluation of the current status of the cost control processes in Iraqi construction projects. *Journal of Engineering*, 29(1), 128–144.
- Ahiaga-Dagbui, D. D., & Smith, S. D. (2012). Neural networks for modelling the final target cost of water projects. In *Proceedings 28th annual ARCOM conference* (pp. 307–316). <http://hdl.handle.net/1842/6550>
- Akinci, B., & Fischer, M. (1998). Factors affecting contractors' risk of cost overburden. *Journal of Management in Engineering*, 14(1), 67–76. [https://doi.org/10.1061/\(ASCE\)0742-597X](https://doi.org/10.1061/(ASCE)0742-597X)
- Akintoye, A. (2000). Analysis of factors influencing project cost estimating practice. *Construction Management and Economics*, 18(1), 77–89.
- Al-Dhaheri, S. A. M., & Burhan, A. M. (2022). Evaluation of construction and demolition waste recycling sites within Iraq. *Innovative Infrastructure Solutions*, 7(2), 143.
- Alex, D. P., Al Hussein, M., Bouferguene, A., & Fernando, S. (2010). Artificial neural network model for cost estimation: City of Edmonton's water and sewer installation services. *Journal of Construction Engineering and Management*, 136(7), 745–756.
- Aljawder, A., & Al-Karaghoul, W. (2022). The adoption of technology management principles and artificial intelligence for a sustainable lean construction industry in the case of Bahrain. *Journal of Decision Systems*, 2022, 1–30.
- Al-Momani, A. H. (1996). Construction cost prediction for public school buildings in Jordan. *Construction Management and Economics*, 14(4), 311–317. <https://doi.org/10.1080/014461996373386>
- Almusawi, H. T., & Burhan, A. M. (2020). Developing a model to estimate the productivity of ready mixed concrete batch plant. *Journal of Engineering*, 26(10), 80–93. <https://doi.org/10.31026/j.eng.2020.10.06>
- Altaie, M., & Borhan, A. M. (2018). Using neural network model to estimate the optimum time for repetitive construction projects in Iraq. *Association of Arab Universities Journal of Engineering Sciences*, 25(5), 100–114.
- Araba, A. M., Memon, Z. A., Alhawat, M., Ali, M., & Milad, A. (2021). Estimation at completion in civil engineering projects: Review of regression and soft computing models. *Knowledge-Based Engineering and Sciences*, 2(2), 1–12.
- Azman, M. A., Abdul-Samad, Z., & Ismail, S. (2013). The accuracy of preliminary cost estimates in Public Works Department (PWD) of Peninsular Malaysia. *International Journal of Project Management*, 31(7), 994–1005. <https://doi.org/10.1016/j.ijproman.2012.11.008>
- Baloi, D., & Price, A. D. F. (2003). Modelling global risk factors affecting construction cost performance. *International Journal of Project Management*, 21(4), 261–269. [https://doi.org/10.1016/S0263-7863\(02\)00017-0](https://doi.org/10.1016/S0263-7863(02)00017-0)
- Barnes, M. (1988). Construction project management. *International Journal of Project Management*, 6(2), 69–79. [https://doi.org/10.1016/0263-7863\(88\)90028-2](https://doi.org/10.1016/0263-7863(88)90028-2)
- Bryde, D. (2008). Perceptions of the impact of project sponsorship practices on project success. *International Journal of Project Management*, 26(8), 800–809. <https://doi.org/10.1016/j.ijproman.2007.12.001>
- Chakraborty, D., Elhegazy, H., Elzarka, H., & Gutierrez, L. (2020). A novel construction cost prediction model using hybrid natural and light gradient boosting. *Advanced Engineering Informatics*, 46, 101201. <https://doi.org/10.1016/j.aei.2020.101201>
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd Acm Sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).
- Cheng, Y.-M. (2014). An exploration into cost-influencing factors on construction projects. *International Journal of Project Management*, 32(5), 850–860. <https://doi.org/10.1016/j.ijproman.2013.10.003>
- Doloi, H. (2013). Cost overruns and failure in project management: Understanding the roles of key stakeholders in construction projects. *Journal of Construction Engineering and Management*, 139(3), 267–279. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000621](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000621)
- Eiben, A. E., & Smit, S. K. (2011). Parameter tuning for configuring and analyzing evolutionary algorithms. *Swarm and Evolutionary Computation*, 1(1), 19–31. <https://doi.org/10.1016/j.swevo.2011.02.001>
- Elhag, T. M. S., Boussabaine, A. H., & Ballal, T. M. A. (2005). Critical determinants of construction tendering costs: Quantity surveyors' standpoint. *International Journal of Project Management*, 23(7), 538–545. <https://doi.org/10.1016/j.ijproman.2005.04.002>
- Elhegazy, H., Chakraborty, D., Elzarka, H., Ebid, A. M., Mahdi, I. M., AboulHaggag, S. Y., & Abdel Rashid, I. (2022). Artificial intelligence for developing accurate preliminary cost estimates for composite flooring systems of multi-storey buildings. *Journal of Asian Architecture and Building Engineering*, 21(1), 120–132. <https://doi.org/10.1080/13467581.2020.1838288>
- Elmousalami, H. H. (2020). Artificial intelligence and parametric construction cost estimate modeling: State-of-the-art review. *Journal of Construction Engineering and Management*, 146(1), 3119008. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001678](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001678)
- Erdis, E. (2013). The effect of current public procurement law on duration and cost of construction projects in Turkey. *Journal of Civil Engineering and Management*, 19(1), 121–135. <https://doi.org/10.3846/13923730.2012.746238>
- Fadhil, G. A., & Burhan, A. M. (2022). Developing crisis management system for construction projects in Iraq. *Journal of Engineering*, 28(1), 33–51.
- Falah, M. W., Hussein, S. H., Saad, M. A., Ali, Z. H., Tran, T. H., Ghoniem, R. M., & Ewees, A. A. (2022). Compressive strength prediction using coupled deep learning model with extreme gradient boosting algorithm: environmentally friendly concrete incorporating recycled aggregate. *Complexity*. <https://doi.org/10.1155/2022/5433474>
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 1–67. <https://doi.org/10.1097/MD.00000000000011870>
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4), 367–378.

- Gunduz, M., & Maki, O. L. (2018). Assessing the risk perception of cost overrun through importance rating. *Technological and Economic Development of Economy*, 24(5), 1829–1844. <https://doi.org/10.3846/20294913.2017.1321053>
- Guo, D., Chen, H., Tang, L., Chen, Z., & Samui, P. (2022). Assessment of rockburst risk using multivariate adaptive regression splines and deep forest model. *Acta Geotechnica*, 17(4), 1183–1205. <https://doi.org/10.1007/s11440-021-01299-2>
- Hashemi, S. T., Ebadati, E. O. M., & Kaur, H. (2019). A hybrid conceptual cost estimating model using ANN and GA for power plant projects. *Neural Computing and Applications*, 31(7), 2143–2154. <https://doi.org/10.1007/s00521-017-3175-5>
- Hatamleh, M. T., Hiyassat, M., Sweis, G. J., & Sweis, R. J. (2018). Factors affecting the accuracy of cost estimate: Case of Jordan. *Engineering, Construction and Architectural Management*, 25(1), 113–131. <https://doi.org/10.1108/ECAM-10-2016-0232>
- Hou, M., Zhang, T., Weng, F., Ali, M., Al-Ansari, N., & Yaseen, Z. M. (2018). Global solar radiation prediction using hybrid online sequential extreme learning machine model. *Energies*, 11(12), 3415.
- Huang, G. -B., Zhu, Q. -Y., & Siew, C. -K. (2004). Extreme learning machine: a new learning scheme of feedforward neural networks. In *2004 IEEE International joint conference on neural networks (IEEE Cat. No. 04CH37541)* (Vol. 2, pp. 985–990).
- Huang, C.-H., & Hsieh, S.-H. (2020). Predicting BIM labor cost with random forest and simple linear regression. *Automation in Construction*, 118, 103280. <https://doi.org/10.1016/j.autcon.2020.103280>
- Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1–3), 489–501. <https://doi.org/10.1016/j.neucom.2005.12.126>
- Huo, T., Ren, H., Cai, W., Shen, G. Q., Liu, B., Zhu, M., & Wu, H. (2018). Measurement and dependence analysis of cost overruns in megatransport infrastructure projects: Case study in Hong Kong. *Journal of Construction Engineering and Management*, 144(3), 5018001.
- Ikeagwuani, C. C. (2021). Estimation of modified expansive soil CBR with multivariate adaptive regression splines, random forest and gradient boosting machine. *Innovative Infrastructure Solutions*, 6(4), 1–16. <https://doi.org/10.1007/s41062-021-00568-z>
- Iyer, K. C., & Jha, K. N. (2005). Factors affecting cost performance: evidence from Indian construction projects. *International Journal of Project Management*, 23(4), 283–295. <https://doi.org/10.1016/j.ijproman.2004.10.003>
- Jing, W., Naji, H. I., Zehawi, R. N., Ali, Z. H., Al-Ansari, N., & Yaseen, Z. M. (2019). System dynamics modeling strategy for civil construction projects: The concept of successive legislation periods. *Symmetry*, 11(5), 1–18. <https://doi.org/10.3390/sym11050677>
- Juszczyk, M. (2018). Residential buildings conceptual cost estimates with the use of support vector regression. *MATEC Web of Conferences*, 196, 4090.
- Juszczyk, M. (2019). On the search of models for early cost estimates of bridges: An SVM-based approach. *Buildings*, 10(1), 2. <https://doi.org/10.3390/buildings10010002>
- Kardani, N., Bardhan, A., Samui, P., Nazem, M., Zhou, A., & Armaghani, D. J. (2021). A novel technique based on the improved firefly algorithm coupled with extreme learning machine (ELM-IFF) for predicting the thermal conductivity of soil. *Engineering with Computers*. <https://doi.org/10.1007/s00366-021-01329-3>
- Kaveh, A., & Khalegi, A. (1998). Prediction of strength for concrete specimens using artificial neural networks. *Advances in Engineering Computational Technology*, 165–171.
- Kaveh, A., Gholipour, Y., & Rahami, H. (2008). Optimal design of transmission towers using genetic algorithm and neural networks. *International Journal of Space Structures*, 23(1), 1–19.
- Kaveh, A., & Iranmanesh, A. (1998). Comparative study of back-propagation and improved counterpropagation neural nets in structural analysis and optimization. *International Journal of Space Structures*, 13(4), 177–185.
- Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting construction cost using multiple regression techniques. *Journal of Construction Engineering and Management*, 132(7), 750–758. [https://doi.org/10.1061/\(ASCE\)0733-9364\(2006\)132:7\(750\)](https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(750))
- Mahalakshmi, G., & Rajasekaran, C. (2019). Early cost estimation of highway projects in India using artificial neural network. *Sustainable Construction and Building Materials*, 25, 659–672. https://doi.org/10.1007/978-981-13-3317-0_59
- Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T., & Voordijk, H. (2019). An artificial neural network approach for cost estimation of engineering services. *International Journal of Construction Management*. <https://doi.org/10.1080/15623599.2019.1692400>
- Mohammad, K. H., Ali, N. S., & Najm, B. M. (2021). Assessment of the cost and time impact of variation orders on construction projects in Sulaimani governorate. *Journal of Engineering*, 27(2), 106–125. <https://doi.org/10.31026/j.eng.2021.02.08>
- Musarat, M. A., Alaloul, W. S., & Liew, M. S. (2021). Impact of inflation rate on construction projects budget: A review. *Ain Shams Engineering Journal*, 12(1), 407–414. <https://doi.org/10.1016/j.asej.2020.04.009>
- Myers, D. (2016). *Construction economics: A new approach*. Routledge.
- Naser, A. H., Badr, A. H., Henedy, S. N., Ostrowski, K. A., & Imran, H. (2022). Application of Multivariate Adaptive Regression Splines (MARS) approach in prediction of compressive strength of eco-friendly concrete. *Case Studies in Construction Materials*, 17, e01262.
- Owusu-Manu, D.-G., Edwards, D. J., Mohammed, A., Thwala, W. D., & Birch, T. (2019). Short run causal relationship between foreign direct investment (FDI) and infrastructure development. *Journal of Engineering, Design and Technology*, 17(6), 1202–1221. <https://doi.org/10.1108/JEDT-04-2019-0100>
- Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in Construction*, 122, 103517.
- Pollack, J., Helm, J., & Adler, D. (2018). What is the iron triangle, and how has it changed? *International Journal of Managing Projects in Business*, 11(2), 527–547. <https://doi.org/10.1108/IJMPB-09-2017-0107>
- Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, 20(1), 1934–1965.
- Salim, M. S., & Mahjoob, A. M. R. (2020). Integrated project delivery (IPD) method with BIM to improve the project performance: A case study in the Republic of Iraq. *Asian Journal of Civil Engineering*, 21, 947–957.
- Shane, J. S., Molenaar, K. R., Anderson, S., & Schexnayder, C. (2009). Construction project cost escalation factors. *Journal of Management in Engineering*, 25(4), 221–229. [https://doi.org/10.1061/\(ASCE\)0742-597X\(2009\)25:4\(221\)](https://doi.org/10.1061/(ASCE)0742-597X(2009)25:4(221))
- Sharma, S., Ahmed, S., Naseem, M., Alnumay, W. S., Singh, S., & Cho, G. H. (2021). A survey on applications of artificial intelligence for Pre-Parametric Project Cost and Soil Shear-Strength estimation in construction and geotechnical engineering. *Sensors*, 21(2), 463.
- ShartoohSharqi, S., & Bhattarai, A. (2021). Evaluation of several machine learning models for field canal improvement project cost prediction. *Complexity*, 2021, 12. <https://doi.org/10.1155/2021/8324272>

- Shehu, Z., Endut, I. R., Akintoye, A., & Holt, G. D. (2014). Cost overrun in the Malaysian construction industry projects: A deeper insight. *International Journal of Project Management*, *32*(8), 1471–1480.
- Shi-fan, Q., Jun-kun, T., Yong-gang, Z., Li-jun, W., Ming-fei, Z., Jun, T., & Qing, H. (2021). Settlement prediction of foundation pit excavation based on the GWO-ELM model considering different states of influence. *Advances in Civil Engineering*, *2021*, 1–11.
- Shoar, S., Chileshe, N., & Edwards, J. D. (2022). Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: Application of random forest regression. *Journal of Building Engineering*, *50*, 104102.
- Shutian, F., Tianyi, Z., & Ying, Z. (2017). Prediction of construction projects' costs based on fusion method. *Engineering Computations*, *43*(7), 2396–2408. <https://doi.org/10.1108/EC-02-2017-0065>
- Son, H., Kim, C., & Kim, C. (2012). Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables. *Automation in Construction*, *27*, 60–66.
- Tao, H., Awadh, S. M., Salih, S. Q., Shafik, S. S., & Yaseen, Z. M. (2022). Integration of extreme gradient boosting feature selection approach with machine learning models: Application of weather relative humidity prediction. *Neural Computing and Applications*, *34*(1), 515–533. <https://doi.org/10.1007/s00521-021-06362-3>
- Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183–7192. <https://doi.org/10.1029/2000JD900719>
- Wang, R., Asghari, V., Cheung, C. M., Hsu, S.-C., & Lee, C.-J. (2022). Assessing effects of economic factors on construction cost estimation using deep neural networks. *Automation in Construction*, *134*, 104080. <https://doi.org/10.1016/j.autcon.2021.104080>
- Wateridge, J. (1998). How can IS/IT projects be measured for success? *International Journal of Project Management*, *16*(1), 59–63. [https://doi.org/10.1016/S0263-7863\(97\)00022-7](https://doi.org/10.1016/S0263-7863(97)00022-7)
- Wu, L., & Fan, J. (2019). Comparison of neuron-based, kernel-based, tree-based and curve-based machine learning models for predicting daily reference evapotranspiration. *PLoS ONE*, *14*(5), e0217520. <https://doi.org/10.1371/journal.pone.0217520>
- Yaseen, Z. M., Ali, Z. H., Salih, S. Q., & Al-Ansari, N. (2020). Prediction of risk delay in construction projects using a hybrid artificial intelligence model. *Sustainability*, *12*(4), 1514.
- Zhang, S., Bogus, S. M., Lippitt, C. D., & Migliaccio, G. C. (2017). Estimating location-adjustment factors for conceptual cost estimating based on nighttime light satellite imagery. *Journal of Construction Engineering and Management*, *143*(1), 4016087. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001216](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001216)
- Zhang, H., Qiu, D., Wu, R., Deng, Y., Ji, D., & Li, T. (2019). Novel framework for image attribute annotation with gene selection XGBoost algorithm and relative attribute model. *Applied Soft Computing*, *80*, 57–79. <https://doi.org/10.1016/j.asoc.2019.03.017>
- Zhao, L., Mbachu, J., & Liu, Z. (2020). Identifying significant cost-influencing factors for sustainable development in construction industry using structural equation modelling. *Mathematical Problems in Engineering*. <https://doi.org/10.1155/2020/4810136>
- Zhao, L., Wang, B., Mbachu, J., & Liu, Z. (2019). New Zealand building project cost and its influential factors: A structural equation modelling approach. *Advances in Civil Engineering*, *2019*, 1362730. <https://doi.org/10.1155/2019/1362730>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.