Hindawi
Complexity
Volume 2022, Article ID 4285328, 18 pages
https://doi.org/10.1155/2022/4285328

Research Article

Developing an Integrative Data Intelligence Model for Construction Cost Estimation

Zainab Hasan Ali,[1] **Abbas M. Burhan**,[1] **Murizah Kassim**,[2,3] **and Zainab Al-Khafaji**[4]

[1]*Civil Engineering Department, College of Engineering, University of Baghdad, Baghdad, Iraq*
[2]*Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, Shah Alam 40450, Selangor, Malaysia*
[3]*School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Shah Alam 40450, Selangor, Malaysia*
[4]*Building and Construction Techniques Engineering Department, Al-Mustaqbal University College, Babylon 51001, Iraq*

Correspondence should be addressed to Zainab Al-Khafaji; zainab.sattar@mustaqbal-college.edu.iq

Received 15 July 2022; Accepted 1 September 2022; Published 29 September 2022

Academic Editor: Haitham Abdulmohsin Afan

Construction cost estimation is one of the essential processes in construction management. Project cost is a complex engineering problem due to various factors affecting the construction industry. Accurate cost estimation is important in construction management and significantly impacts project performance. Artificial intelligence (AI) models have been effectively implemented in construction management studies in recent years owing to their capability to deal with complex problems. In this research, extreme gradient boosting is developed as an advanced input selector algorithm and coupled with three AI models, including random forest (RF), artificial neural network (ANN), and support vector machine (SVM) for cost estimation. Datasets were gathered based on a survey conducted on 90 building projects in Iraq. Statistical indicators and graphical methods were used to evaluate the developed models. Several input predictors were used, and XGBoost highlighted inflation as the most crucial parameter. The results indicated that the best prediction was attained by XGBoost-RF using six input parameters, with r-squared and the mean absolute percentage error equal to 0.87 and 0.25, respectively. The comparison results revealed that all AI models showed good prediction performance when applied to datasets affected by more than two parameters. The outcomes of this research revealed an optimistic strategy that can help decision makers select the influencing parameters in the early phases of project management. Also, developing a prediction model with high precision results can assist the project's estimators in decreasing the errors in the cost estimation process.

1. Introduction

The construction industry is complex and comprises parties such as owners, contractors, and consultants [1, 2]. The construction sector affects the global economy of countries, so many studies have explored methods to improve the performance of construction projects [3–5]. Due to its global impact, several scholars measure the project performance's success to understand it better. The construction project's success can be measured by achieving the project within the estimated cost, duration, and specifications [6, 7]. Implementing a project with successful performance is challenging with growing awareness of the environment and customer requirements changing [8]. The construction industry is dynamic and complex since it needs the implementation of successful project management strategies [9, 10]. Unsuccessful strategies cause cost and schedule overruns, leading to undesirable results and reducing customer satisfaction [11–13].

Consequently, there is a need to develop effective strategies and methods to mitigate risks and uncertainties in a construction project [14, 15]. The construction project management sector mainly includes the initial design phase, detailed design phase, construction cost estimation, project bidding phase, construction phase, and final delivery after completion [16–18]. Cost is an essential criterion of project

performance due to its impact on feasibility studies and choosing design alternatives [19]. Cost studies describe and evaluate the costs of buildings and other construction projects [20]. These studies seek to maximize the project's revenue by using the available resources. Cost estimation is essential in construction and significantly impacts project management [21]. Accuracy of cost estimation is considered a necessary factor of project success during various phases of project construction [22, 23]. Accurate cost estimation affects the project's profitability, owner's satisfaction, and financial decision [24]. Inaccurate cost estimation leads to problems like cost overruns and project parties' disputes [25, 26]. Several studies on cost management indicated that accurate cost estimation affects the profitability of construction projects at the tender phase and is a vital part of project survival [27, 28]. They also showed that establishing accurate cost estimation ensures the contract's profit at the tendering phase.

Cost estimation is performed by a coordination role of tender managers and technical experts called estimators. The existing cost estimation methods require complete information about construction projects and are costly and time consuming [29, 30]. During the tendering phase, estimators have little information; therefore, they depend on their knowledge and expertise to attain the estimated cost [31]. The level of knowledge is different among estimators, which can affect the accuracy of the cost estimation process [13]. Many studies have used statistical and artificial intelligence techniques to prevent these problems and accurately estimate construction costs [13, 32, 33]. Several studies performed the regression method as a traditional technique for cost estimation [34–37]. This method is easy and has achieved simple results. However, this technique cannot handle a complex system's nonlinear relationship among parameters. Recently, computer-aided algorithms have been successfully conducted in construction management studies. AI models can handle the complexity and nonlinearity of construction projects and help the project's parties understand the uncertainties and incomplete information at the early stage of the construction process [38–41].

The capacity of artificial neural network (ANN) and support vector machine (SVM) models in estimating the cost and duration of road projects was analysed [42]. The comparison analysis showed that SVM has higher accuracy and fewer errors than the ANN model. Two researchers presented the ordinary least square regression (OLSR) for construction cost forecasting of the Pune region in India [43, 44]. The model was applied over a 12-year prediction period and attained 91%–97% accuracy. The SVM model's capability in estimating the residential building's conceptual cost was studied by the authors in [45]. The results showed that the model achieved low mean absolute percentage errors, with values ranging between 7% and 8.19%. Another study used SVM to estimate the cost of bridge construction [46]. The study showed that the SVM model could estimate the cost during the initial process of project construction. The model obtained the highest performance results with a correlation coefficient equal to 0.974. An ANN model was trained with a backpropagation algorithm for early performance estimation of buildings in India [47]. The study revealed the ability of the ANN model in cost prediction and its importance to the financial investors in the construction project. The precision of construction cost prediction was improved by integrating a genetic algorithm (GA) with the ANN model [48]. The study reported that the developed model attained a high predictive performance equal to 0.9471 and assisted project managers at the beginning of the project. Principal component analysis (PCA) and particle swarm optimization (PSO) were combined with the SVM model to predict cost of substation projects [49]. The developed model was compared with PCA-SVM and PSO-SVM, and the study demonstrated that the integration of three algorithms achieved better prediction outputs than other models, which can help decision makers in substation projects. Twenty AI techniques for construction cost estimation of field canal improvement projects were compared by the author in [50]. The author concluded that the extreme gradient boosting (XGBoost) algorithm gained the best prediction results with r-squared equal to 0.929. Three AI methods, namely, multilayer perceptron (MLP), radial basis function neural network (RBFNN), and general regression neural network (GRNN), were developed for cost prediction of road projects [51]. The study showed that GRNN gained better prediction results than other models, with R^2 equal to 0.9595. The study also indicated that ANN obtained fewer prediction errors and could handle limited information during the early phases of the construction process. The study by the authors in [52] used three machine learning algorithms: MLP, GRNN, and RBFNN, with a process-based method for cost prediction at the early stage of project management. The study confirmed that the GRNN algorithm provided better outcomes than other models, which can help project managers predict construction costs in the contracting phase, where several input parameters are unknown at this stage. The labor cost of the BIM project was explored by integrating a simple linear regression (SLR) with the random forest (RF) model [53]. The study demonstrated the effectiveness of the hybrid model in the cost estimation process. An optimization algorithm called PSO was integrated with the ANN model to improve the performance of the cost prediction of high-rise residential buildings [54]. The study concluded that the PSO-ANN model has higher prediction precision and generalization than the single ANN algorithm.

The ability of RF, SVM, and multilinear regression (MLR) to predict the cost overrun of high-rise buildings' engineering services was examined [55]. The authors showed that the RF model achieved better prediction results than the other two AI models. Based on the reported studies, AI algorithms can be applied successfully for cost estimation. These studies indicated that the performance of these algorithms is affected by the algorithm's structure and the abstraction of input variables. The correlation statistics, factor analysis, and relative importance index are the popular methods used by past studies [56–59].

In these studies, data were collected based on personal opinions and expert surveys, leading to bias in the existing approaches. Moreover, the current method explores the

Table 1: Descriptive measures of the cost data for the training phase.

	Minimum	Maximum	Mean	Median	Std. deviation	Skewness	Kurtosis
Ground floor area (m^2)	200	5320	1622	1370	1196.252	1.127	0.466
Total floor area (m^2)	355	9088	3135	2525	2214.788	1.079	0.466
Floor number	1	6	2.762	2	1.711	0.651	−1.052
Elevator number	0	4	0.794	0	1.002	1.073	0.349
Footing type (raft = 1; separated = 2)	—	—	—	—	—	—	—
Inflation (%)	0.1	7.5	2.484	1.8	2.161	0.775	−0.656
Duration (days)	122	731	357.6	348	137.171	0.715	0.652
Cost ($)	26756	6451519	1975260	1321630	1623076	0.903	−0.312

linear relationship between input and output predictors. Based on that, exploring an advanced method that can investigate the complex system of cost estimation parameters is very important to achieve accurate results [38]. Furthermore, integrating a new feature selection method with AI algorithms is significant for construction management engineering to get accurate prediction performance [60]. Recently, the XGBoost algorithm has been explored as an advanced version of the feature selection approach in engineering problems. XGBoost is a recent version of gradient boosting and has been applied effectively as an input selection method by civil engineering scholars [61, 62]. The research scope of applying AI algorithms in construction management is still limited, so exploring a new cost estimation method is the motivation of this research.

Developing effective predictive models that can achieve accurate performance is a vital issue in the early phases of engineering management. The construction industry in Iraq has a special issue due to the risky conditions and exceptional political circumstances that had happened in this region [2]. The instability of political and economic conditions has a massive impact on the performance of construction projects. Due to economic and political circumstances, most constructed projects have failed to be completed within the specified budget [63]. Thus, developing an integrative model using AI algorithms in cost management can improve cost performance by evaluating, controlling, and monitoring the project cost under uncertain conditions.

The current study is achieved by integrating the XGBoost algorithm with three AI algorithms, namely, RF, SVM, and ANN. XGBoost was used to select the influencing parameters of the prediction process, and then, the AI model used these parameters. The attained results are analysed and discussed by using statistical and visualization methods. The output of this study can assist a project manager in selecting the influencing parameters at the early stage of a construction project. Also, introducing an integrated model with high precision results helps project estimators reduce errors in the cost estimation phase.

2. Construction Cost Data Explanation

This study used public building projects in Iraq as a case study for the modelling process and used their dataset. These projects are managed by the Iraqi government, so developing an accurate precision model can help decision makers by producing precious results. The information in the dataset was collected from the survey conducted on 90 construction projects between 2016 and 2021. The information was gathered from historical records of construction projects, including project drawings, bills of quantities, and project schedules. The dataset includes information on ground floor area (GFA), total floor area (TFA), floor number (FN), elevator number (EN), footing type (FT), inflation (F), duration (D), and construction cost (C). The inflation data were collected from the Iraqi central bank (https://cbiraq.org/). Tables 1 and 2 show the descriptive measures of the cost data for the training-testing phase. The statistical descriptions include minimum, maximum, mean, median, standard deviation, skewness, and kurtosis. For the training process, the mean value of construction cost is 1623076 $, while for the testing phase, the mean value is 2571431 $. The maximum and minimum values of project duration are 731 days and 122 days for the training phase, while for the testing phase, they are 150 days and 787 days. It can be seen that the datasets for the training and testing phases are well distributed. It is near the normal distribution because the mean and median values of the datasets are close to each other.

3. Method Overview

3.1. Extreme Gradient Boosting (XGBoost). XGBoost is an advanced version of tree-based boosting modelling introduced by Chen and Guestrin [62], which is applied effectively in input selection problems [64, 65]. The boosting algorithm's concept uses an iteration process for learning the functional relationship between the target and predictor values [66]. Through this iterative process, the individual trees are trained sequentially on the residual output of the previous trees to reduce the training errors [67]. The algorithm uses a cache-aware structure and a regularized method for boosting learning. The mathematical expression of prediction can be shown as follows:

$$\widehat{Y} = \phi(X) = \frac{1}{n}\sum_{k=1}^{n} f_k(X), \tag{1}$$

where $\widehat{Y}$ is the predicted value of the target, X represents the input variable, K is the value ranging between 1 and n, f_k is the function between input and output variables, and n is the number of trained functions by boosting trees. The loss

Table 2: Descriptive measures of the cost data for the testing phase.

	Minimum	Maximum	Mean	Median	Std. deviation	Skewness	Kurtosis
Ground floor area (m^2)	239	3940	1490	1394	882.414	0.991	0.713
Total floor area (m^2)	344	9800	4086	3840	2692.433	0.384	−1.068
Floor number	1	6	3.222	3	1.625	0.483	−1.089
Elevator number	0	4	1.37	1	1.182	0.501	−0.936
Footing type (raft = 1; separated = 2)	—	—	—	—	—	—	—
Inflation (%)	0.1	7.5	3.319	2.1	2.674	0.297	−1.474
Duration (days)	150	787	404.7	364	177.036	0.468	−0.770
Cost ($)	160147	6104878	2571431	2215402	1940152	0.409	−1.345

function in XGBoost must be minimized to train several functions f_k, as shown in the following expression:

$$\mathscr{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k),$$
$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{2}$$

where $\mathscr{L}(\phi)$ is the regularized function, i represents the loss function measurement between $\hat{y}_i$ (prediction value) and y_i (actual value), Ω is a regularization term that prevents the building of additional trees in the model from decreasing overfitting and error. γ is the leaf's complexity, T is the leaves' number in the tree model, λ is the penalty parameter, and w is the score vector on the leaves.

In the input selection process, the main aim of XGBoost is to produce the feature importance of input variables [68]. According to Hastie et al. [69], the algorithm uses gain, frequency, and cover to calculate feature importance. The gain method calculates the role of each feature in the model's development. Frequency is the weight representing the occurrence number for each feature in the boosted trees. The cover method shows the number of samples related to each feature. XGBoost uses the following expression to calculate the importance of the feature:

$$N_v = \sum_{L=1}^{L} \sum_{l=1}^{X=1} I\left(V_L^l, v\right),$$
$$\left(V_L^l, v\right) = f(x) = \begin{cases} 1\, if\, V_L^l = v, \\ 0, \text{otherwise}, \end{cases} \tag{3}$$

where L is the tree' number, N is the number of a node for each leaf, (V_L^l) represents the feature of node l, and I is the indicator function.

3.2. Random Forest (RF). Random forest (RF) is an ensemble algorithm introduced by Breiman [70] and based on combining multiple decision trees to produce a robust prediction model. It has been used effectively for classification and regression problems in several areas of construction management [71–73]. The RF model can deal with many input variables and work efficiently with outliers and unbalanced datasets. The algorithm reduces overfitting results and performs accurately with simple computation processes [74]. The RF model uses bootstrap and random space techniques to improve the predictive model's performance [75, 76]. In the RF model, the algorithm uses a bootstrap method to choose new training sets from the original data randomly, and these new data will be utilized to develop a regression tree (n_{tree}). The number of split m_{try} for each node in the regression tree is computed using a stochastic random space technique.

The modelling steps to develop the RF model are as follows: First, we generate a new training dataset using a bootstrap algorithm where two-thirds of the original data (in bag data) are used to train the developed model. After that, several regression trees were built based on bootstrap samples, and these regression trees were used to develop the RF model. The RF model is created by training a sequence of regression trees. The variance between the trees can be measured by randomly choosing the optimal number of attributes based on maximum depth values. These computations increase the ability of the RF model to reduce the errors in prediction results. The RF model is built by training a sequence of regression trees. Finally, the algorithm collects the output value for each tree and calculates the final prediction using the average method [77]. The mathematical calculation of the RF technique is expressed as follows:

$$y' = \frac{1}{n_{\text{tree}}} \sum_{i=1}^{n_{\text{tree}}} f_i(x), \tag{4}$$

where y' represents the prediction value of the RF model, n_{tree} is the number of regression trees, and $f_i(x)$ is the regression tree model based on input value (x). The schematic diagram of the RF algorithm is described in Figure 1.

3.3. Artificial Neural Network (ANN). Artificial algorithms such as ANN and other machine learning algorithms have been introduced recently. These models have been characterized by their capabilities in handling complex datasets and producing accurate results [78]. The artificial neural network is a mathematical expression, building its components by emulating the biological structure of the human brain [79]. The main element of the ANN model is the series of connected layers called neurons. Several types of ANN exist; scholars commonly apply a feedforward ANN with a backpropagation algorithm [80]. The popularity of the backpropagation algorithm in ANN applications is gained by its capacity to learn ANN networks based on a supervised learning algorithm [81]. In this method, the error in prediction results is computed by comparing the predicted values with actual variables. The weights in the ANN model are updated by backpropagation to reduce the expected error

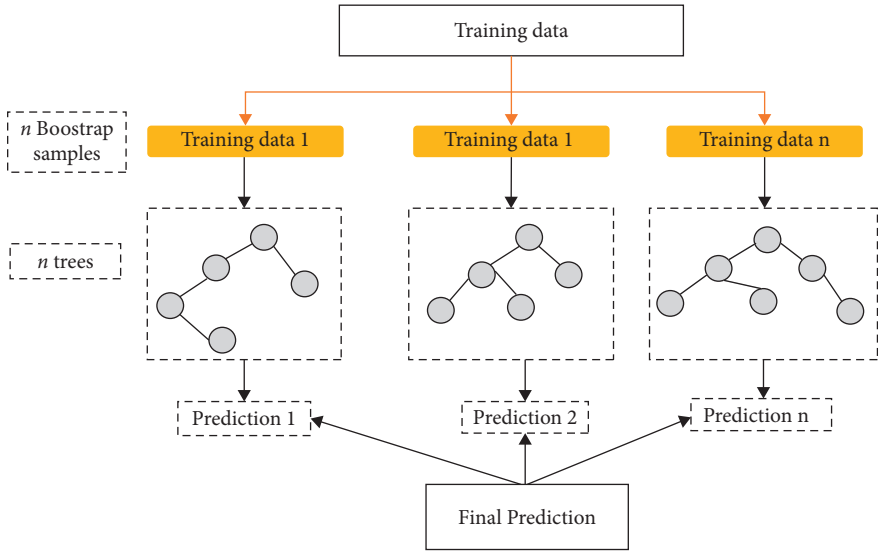

Figure 1: Schematic diagram of the RF algorithm.

to the allowable value. Feedforward ANN consists of input, hidden, and output layer components. In this type of ANN, moving information is performed directly from the input to the output neurons without returning in the reversed direction. The number of neurons in the input and output layer is based on the number of input and output variables in the ANN model. In the hidden layer, neurons nonlinearly transform input variables into the output layer [82]. Hidden layers in the ANN algorithm can be expressed mathematically as follows:

$$v_i = \left(1 + \exp\left(-1 \times \sum_{i=1}^{1} x_i w_{ij}\right)\right)^{-1}, \tag{5}$$

where v_i represents the hidden layer, x_i is the input parameters, and w_{ij} refers to the weight between input and hidden layers. The value of the output layer can be computed as follows:

$$y = \left(1 + \exp\left(-1 \times \sum_{j=1}^{1} v_j w_{ij}\right)\right)^{-1}. \tag{6}$$

The design of an ANN network requires identifying the number of neurons and hidden layers. According to previous studies, the best prediction results can be achieved using one or two layers [83, 84]. The optimum input variables can attain the best performance during the training process. The relationship between input and output variables is designed to improve prediction performance by training ANN. For each repeated process, the biases and weights are modified by the algorithm to reduce the error measures between the original and predicted values. The errors of the expected value can be presented as follows:

$$\text{Error} = 0.5(d - y)^2, \tag{7}$$

where d represents the actual value and y is the predicted value achieved by the ANN model. The presentation of the ANN model is shown in Figure 2.

3.4. Support Vector Machine (SVM). Support vector machine is a supervised machine learning algorithm developed as a method that uses a hyperplane to divide the data and measure the nearest position between the external point and the hyperplane [85]. SVM is a popular algorithm commonly used by scholars to improve the estimated process of engineering problems [86, 87]. The algorithm simulates the errors between actual and predicted parameters by measuring the distance from the SVM margin. The SVM model can be expressed mathematically as follows:

$$M = \{x_1, y_1), x_2, y_2), \ldots, x_n y_n). \tag{8}$$

M represents the training dataset and x and y are the input and output parameters. SVM uses the following function to learn the dataset during the training phase:

$$f(x) = w \cdot \varphi(x) + b, \tag{9}$$

where w represents the weight indicator, $\varphi(x)$ is the nonlinear function of input parameter x, and b is a scalable term. The standard error of the prediction process is minimized using the following equation:

$$\begin{aligned} &\min\varphi(w, \xi) = 0.5w^2 + C\left(\sum_{i=1}^{n} \xi_i\right)\varphi(x_i) + b, \\ &\text{subjected to } M_i\left(w^M x_i + b\right) \geq \varepsilon + \xi_i, \xi \geq 0, \end{aligned} \tag{10}$$

where ξ refers to the slag variable, C is a penalty variable controlling the error between regularization and empirical prediction, and ε is the function corresponding to the accuracy of the training process. The SVM model can be optimized by using Lagrange multipliers and optimum generic functions as shown in the following expression:

$$f(x, a_i, a_i^*) = \sum_{i=1}^{n} (a_i - a_i^*) K(x, x_i) + b, \tag{11}$$

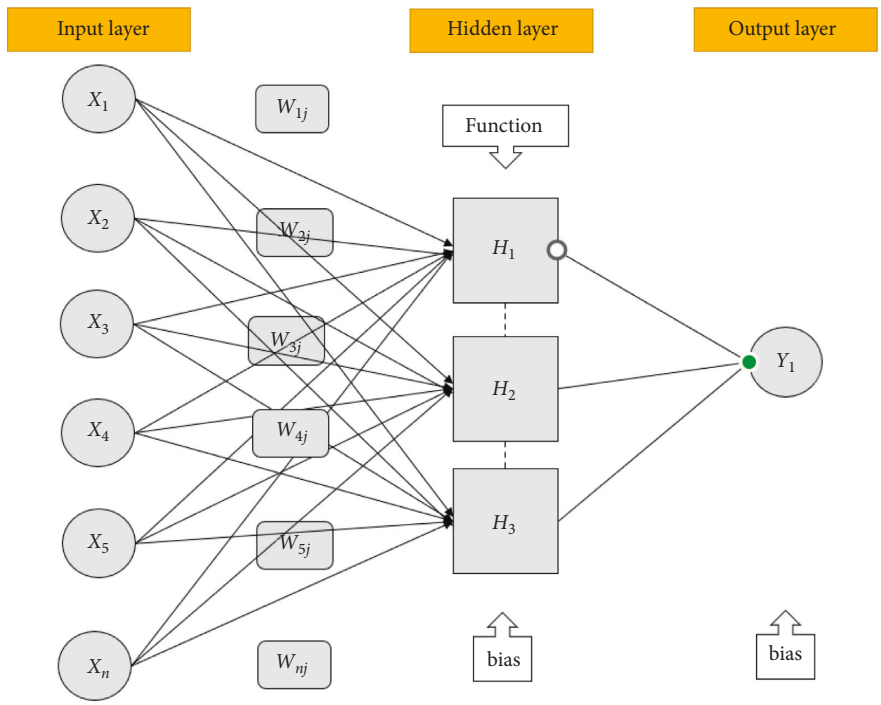

Figure 2: Structure diagram of the ANN model.

where $K(x \text{ and } x_i)$ represents the kernel function. The main feature of the SVM algorithm in regression problems is that it correlates input and output parameters using a nonlinear relationship. The SVM model's kernel function helps the algorithm generate nonlinear mappings in high-dimensional space. The SVM model has four kernel functions: linear, sigmoid, polynomial, and radial basis function (RBF) [87]. The RBF kernel function is simple, effective, and reliable and has been used in several complex studies [88]. The RBF nonlinear equation was used in this study, and the kernel function was defined depending on three parameters: C, ε, and γ where the optimal values of this equation can be reached using the trial and error method. The illustrative diagram of the SVM algorithm is depicted in Figure 3.

4. Model Development and Performance Assessment

In the present study, the construction cost estimation of the building projects in Iraq was explored. First, input selection was made to abstract the appropriate features for the prediction process. Due to the complex nature of cost estimation, the XGBoost algorithm was developed to choose the most important parameters. XGBoost was integrated with popular AI algorithms, namely, RF, ANN, and SVM. The hybrid model was developed using the R programming language (version 4.1.1). Three libraries called XGBoost, Matrix, and ggplot2 were applied to construct the XGBoost algorithm. The best results of feature selection were attained using the xgb importance function. For the SVM model, library (dplyr), library (caret), library (ggplot2), and library (kernlab) were used. Function train control was applied to control the following parameters: method (cv), number (), and set as 5-fold cross-validation. The radial basis function was applied in this study using the svmRadial function. The RF algorithm was designed using the library (ranger) and ranger function. The designed parameters are as follows: num.trees set as 200, mtry equal to 3, and min.node size used as 3. In the case of the ANN model, library neuralnet with one hidden layer and resilient backpropagation were applied to enhance model prediction. The performance of the developed model was assessed by using several statistical evaluators, including coefficient of determination (R^2), error measures (i.e., mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE)), Nash–Sutcliffe efficiency (Nash) and Willmott's index (WI) [89, 90]. The process of the presented AI models is shown in Figure 4.

5. Result and Discussion

5.1. Results Analysis. In this research, the ability of three AI algorithms, namely, RF, ANN, and SVM, was examined to estimate the cost of construction projects. The authors introduced the XGBoost model to determine the best combination of input parameters. The developed models were built based on different combinations extracted from the XGBoost model. Input combinations were constructed using the important parameters selected by the XGBoost algorithm, as shown in Figure 5. Figure 5 shows the relative importance values of input features using the XGBoost algorithm. The XGBoost results show that the inflation parameter (F) is the most important cost estimation, followed by the total and ground floor area. The results also indicated that the elevator number and the footing type gained the least significant scores by XGBoost. To determine the impact

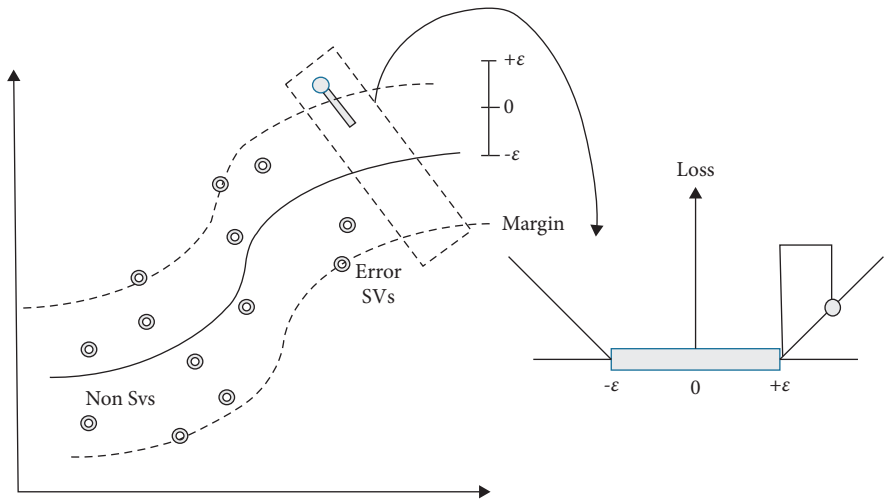

Figure 3: Illustrative diagram of the SVM algorithm.

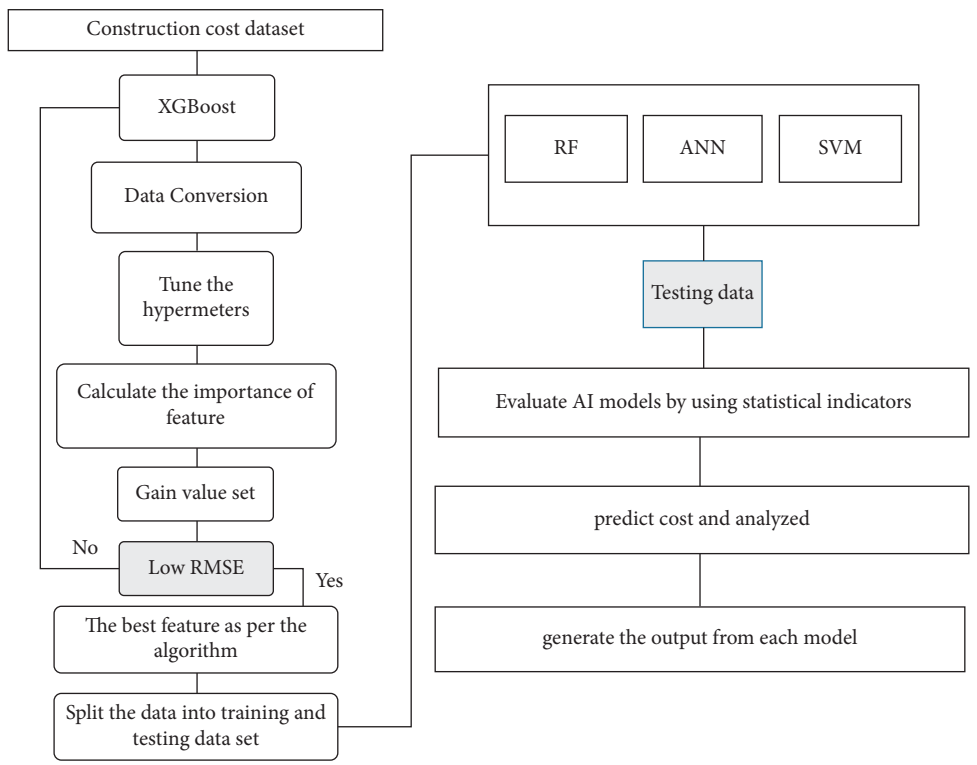

Figure 4: The proposed AI model for construction cost estimation.

of input variables on the performance of predictive models, several combinations were constructed and tested by each model. Seven models were developed for each algorithm (Model I, Model II, Model III, ..., Model VII), including variables from one to seven input parameters.

Tables 3 and 4 show the statistical measurements of the presented computer-aided algorithms for the training and testing phase. The tabulated measurements revealed that both the XGBoost-RF and XGBoost-ANN models have excellent performance for the training phase when using more than two input variables. The XGBoost-SVM model achieved less predictive performance than the other AI models for all input combinations, and the best accuracy was attained by using three input parameters. The best results for the training phase are demonstrated by XGBoost-ANN–Model V with $R^2 = 0.97551$, RMSE = 253464.6776, MAE = 151999.7328, MAPE = 0.40876, Nash = 0.97522, and WI = 0.99376. For the testing phase, the results revealed that increasing the number of input variables leads to increased predictive accuracy for all AI models. RF outperformed SVM and ANN models, and the

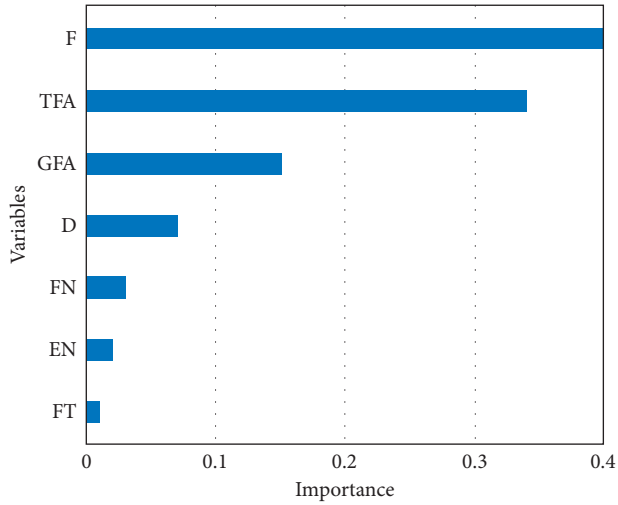

Figure 5: Feature selection of the XGBoost algorithm.

Table 3: Performance measures of AI models for training data.

	R^2	RMSE	MAE	MAPE	Nash	WI
XGBoost-RF model						
Model I	0.76081	788890.3191	604981.6715	0.99102	0.75994	0.93074
Model II	0.70272	879643.0411	589663.4386	0.63845	0.70154	0.90834
Model III	0.91381	472714.8598	339706.5592	0.43244	0.91381	0.97704
Model IV	0.92352	446670.1333	328896.3663	0.44469	0.92304	0.97921
Model V	0.96647	323123.8558	208083.017	0.36537	0.95973	0.98879
Model VI	0.93993	394961.8018	274148.5711	0.41215	0.95603	0.98446
Model VII	0.90717	495125.5653	337773.447	0.43415	0.90544	0.97555
XGBoost-ANN model						
Model I	0.69691	860530.7166	839959.9524	0.77566	0.61666	0.83414
Model II	0.88621	582676.712	290912.5011	0.26372	0.86904	0.96865
Model III	0.95504	341388.8802	224162.0108	0.37779	0.95505	0.98838
Model IV	0.95569	338911.6945	214402.0251	0.40876	0.95569	0.98856
Model V	0.97551	253464.6776	151999.7328	0.40876	0.97522	0.99376
Model VI	0.95603	337627.8199	226863.17	0.41546	0.71921	0.91102
Model VII	0.70773	1108463.447	872191.1237	0.95867	0.52607	0.75929
XGBoost-SVM model						
Model I	0.71722	838810.8081	666081.6149	0.63946	0.71436	0.91413
Model II	0.89679	527105.4427	333897.5918	0.40414	0.89283	0.97276
Model III	0.80479	717164.8809	466107.817	0.64501	0.80161	0.93934
Model IV	0.80151	735499.0773	491505.0568	0.65881	0.79134	0.93388
Model V	0.8031	715627.4657	451267.6691	0.53051	0.80246	0.94242
Model VI	0.72157	853201.115	542534.255	0.60776	0.79545	0.93821
Model VII	0.80124	727702.5318	450836.0521	0.54879	0.79574	0.93678

best performance was achieved by XGBoost-RF-Model VI with $R^2 = 0.87211$, RMSE = 693311.4488, MAE = 424619.6505, MAPE = 0.25539, Nash = 0.86739, and WI = 0.962557.

The developed AI models are also evaluated by graphical presentations like scatter plots, box plots, and Taylor diagrams. Figures 6–8 illustrate scatter plots for the testing phase for the three hybrid models (XGBoost-RF, XGBoost-ANN, and XGBoost-SVM). The XGBoost-RF model exhibits a good prediction with R^2 greater than 0.81 for all combinations except for Model I, where R^2 decreases to 0.6215. In the ANN algorithm, the developed model performs well with R^2 maxed out at 0.83 when increasing the number of input combinations for models V, VI, and VII, as shown in Figure 6. The SVM model shows an enhancement in the prediction accuracy when the number of combinations increases, except for Models VI and VII, where R^2 reduces to 0.7579.

Figure 9 shows a box plot presentation to illustrate the residual error between the observed and estimated values of cost estimation. The results showed that XGBoost-RF-Model III and XGBoost-RF-Model V gained the minimum residual plot with an error value of less than 50%. For the ANN model, the minimum positive error was attained by Models V and VII, while Models V and VI gained the minimum

Table 4: Performance measures of AI models for testing data.

	R^2	RMSE	MAE	MAPE	Nash	WI
XGBoost-RF model						
Model I	0.62154	1227246.543	874941.6745	0.55214	0.58448	0.88937
Model II	0.66971	831959.7023	619292.1689	0.35191	0.80904	0.94221
Model III	0.87382	804900.462	582701.9977	0.30825	0.82126	0.9404
Model IV	0.8553	859581.3363	591577.1935	0.2804	0.79615	0.92993
Model V	0.85961	823500.3632	554104.6337	0.25399	0.81291	0.93815
Model VI	0.87211	693311.4488	424619.6505	0.25539	0.86739	0.962557
Model VII	0.86511	706201.1783	461238.6876	0.26279	0.86241	0.961966
XGBoost-ANN model						
Model I	0.59486	1297368.569	919557.1617	0.56905	0.53564	0.88069
Model II	0.75537	1150901.062	763471.3537	0.33141	0.63458	0.9049
Model III	0.73428	1004876.227	578230.6713	0.28253	0.72142	0.9241
Model IV	0.7481	979906.0497	523097.5798	0.28778	0.7351	0.92872
Model V	0.85042	750698.034	459356.0909	0.26579	0.84453	0.95641
Model VI	0.84828	864172.2326	570128.7514	0.23589	0.79398	0.93062
Model VII	0.83202	836314.9646	545942.9388	0.23004	0.80704	0.93977
XGBoost-SVM model						
Model I	0.64182	1158278.86	803076.3523	0.80044	0.62987	0.89511
Model II	0.77873	946666.831	713338.1417	0.38731	0.75276	0.92379
Model III	0.81801	879551.0025	631843.2588	0.33659	0.78658	0.93023
Model IV	0.8256	856049.4301	586251.8811	0.26047	0.79783	0.93494
Model V	0.83321	826462.8309	542359.3988	0.23651	0.81156	0.94219
Model VI	0.78213	907434.4113	623458.3053	0.24091	0.77283	0.93306
Model VII	0.75791	958666.6949	633398.3941	0.24339	0.74646	0.92462

negative error. In the XGBoost-SVM model, combinations with 2 to 7 input parameters show a reduction in the error value, and the minimum error was achieved by Models III, IV, and V with a residual error of less than 50%. The maximum residual error was demonstrated by Model I for all the developed models with a negative error value of less than 85%.

Another graphical method (i.e., the Taylor diagram) was constructed to evaluate the developed models based on correlation and standard deviation [91]. Figure 10 illustrates the Taylor diagram for the three AI algorithms with different input combinations for tested data. Based on the constructed Taylor diagram, the XGBoost-RF model attained the nearest position to the actual cost using three and six input parameters (i.e., Model III and Model VI). The rest of the combinations of the RF model demonstrate good prediction performance, except for Model I, which attained less correlation value than the other combinations.

XGBoost-ANN and XGBoost-SVM models achieved the best performance using five input variables (i.e., Model V). XGBoost-ANN attained the nearest distance to the actual cost using three models (i.e., Model V, Model VI, and Model VII), whereas the remaining models achieved the farthest distance with a correlation value of less than 0.9. For XGBoost-SVM, only three models gained the best performance using three, four, and five input parameters, whereas the poorest performance was achieved using one input variable.

5.2. Validation against Previous Studies. To confirm the ability of the introduced AI models in cost estimation, it is necessary to validate these models against the developed AI models in past studies. An ANN algorithm was developed to estimate the construction cost of highway projects in India [92]. The study showed that the ANN model could estimate construction cost with an R of 0.94. In another study, the developed approach gained a correlation coefficient of 0.97 for the cost prediction of bridge construction by using an SVM model with 27 input variables [46]. The investigation of three AI algorithms called multivariate adaptive regression spline (MARS), extreme learning machine (ELM), and partial least square regression (PLS) was performed to estimate the construction cost of field canal projects [93]. According to the reported results, the MARS model attained the best results with an $R^2 = 0.94$ and five input parameters. Previous studies also reported hybrid models as effective models for cost estimation, such as ANN-GA and RF-SLR [48, 53]. The previous studies gained an acceptable performance in cost estimation. However, they only focused on single models and gave little attention to hybrid models. Also, they developed AI models based on all input parameters. This study combined the XGBoost algorithm with AI models to enhance the cost estimation accuracy. It can be noticed that XGBoost-RF achieved good estimation performance with r-squared ranging from 0.87 to 0.91 for both testing and training phases using only three input variables.

5.3. Discussion. Using an AI approach in complex construction projects is highly recommended to get accurate estimation process results and simulate the nonlinear relationships between input and output parameters. Using XGBoost as an advanced input selector revealed that inflation, total floor area, and ground floor area are the most important variables in cost estimation. The comparison

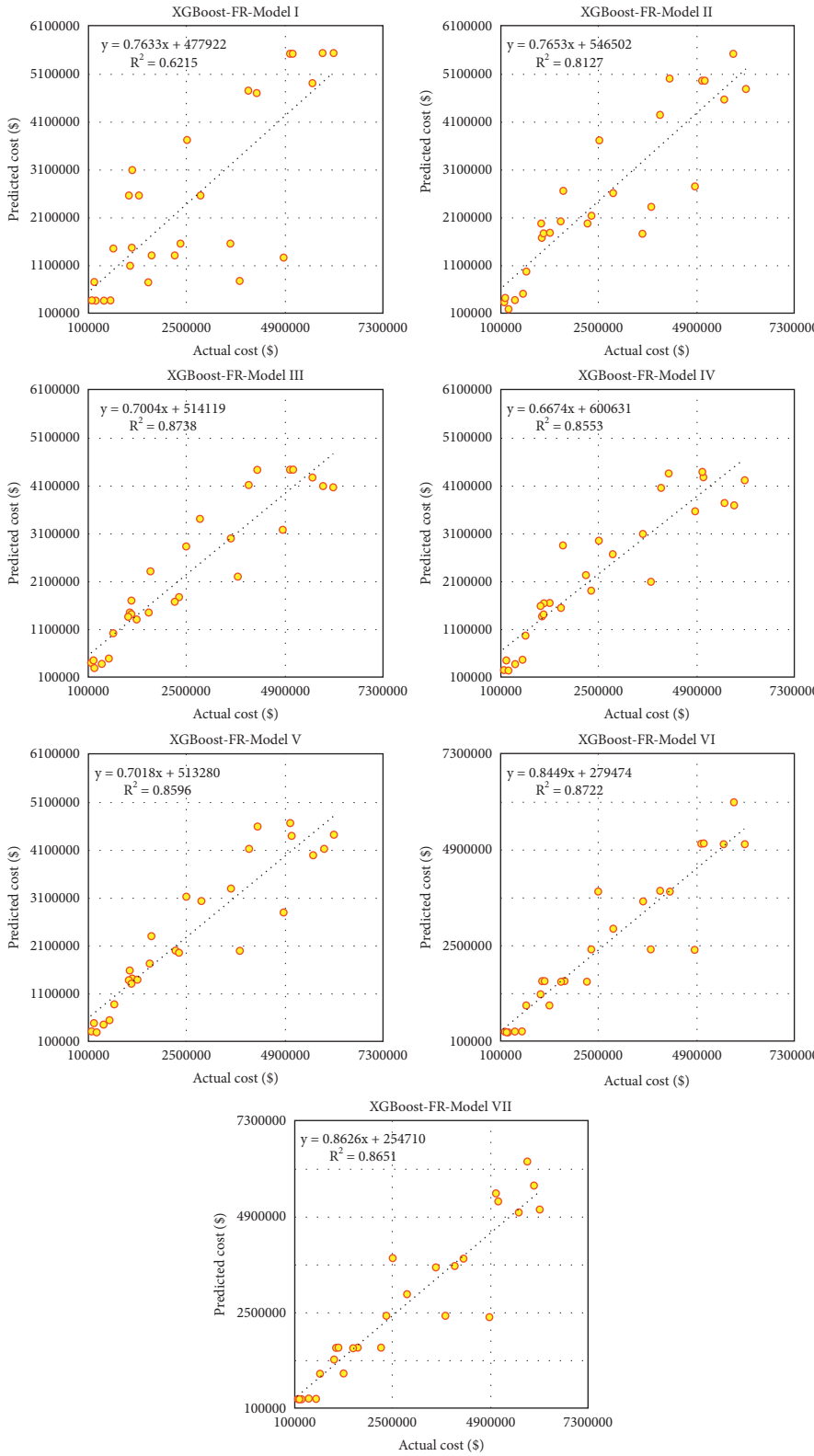

Figure 6: Scatter plot presentation of the XGBoost-RF model for tested data.

results between AI models showed the ability of the developed algorithms to predict construction costs because all models attained good predictive performance except for models that used one input variable. The XGBoost-RF model showed a significant enhancement in the prediction process using only three input parameters where $R^2 = 0.87$ and MAPE = 0.308 and minimum negative error, as shown in Table 4 and Figure 9. Applying an RF model with six input

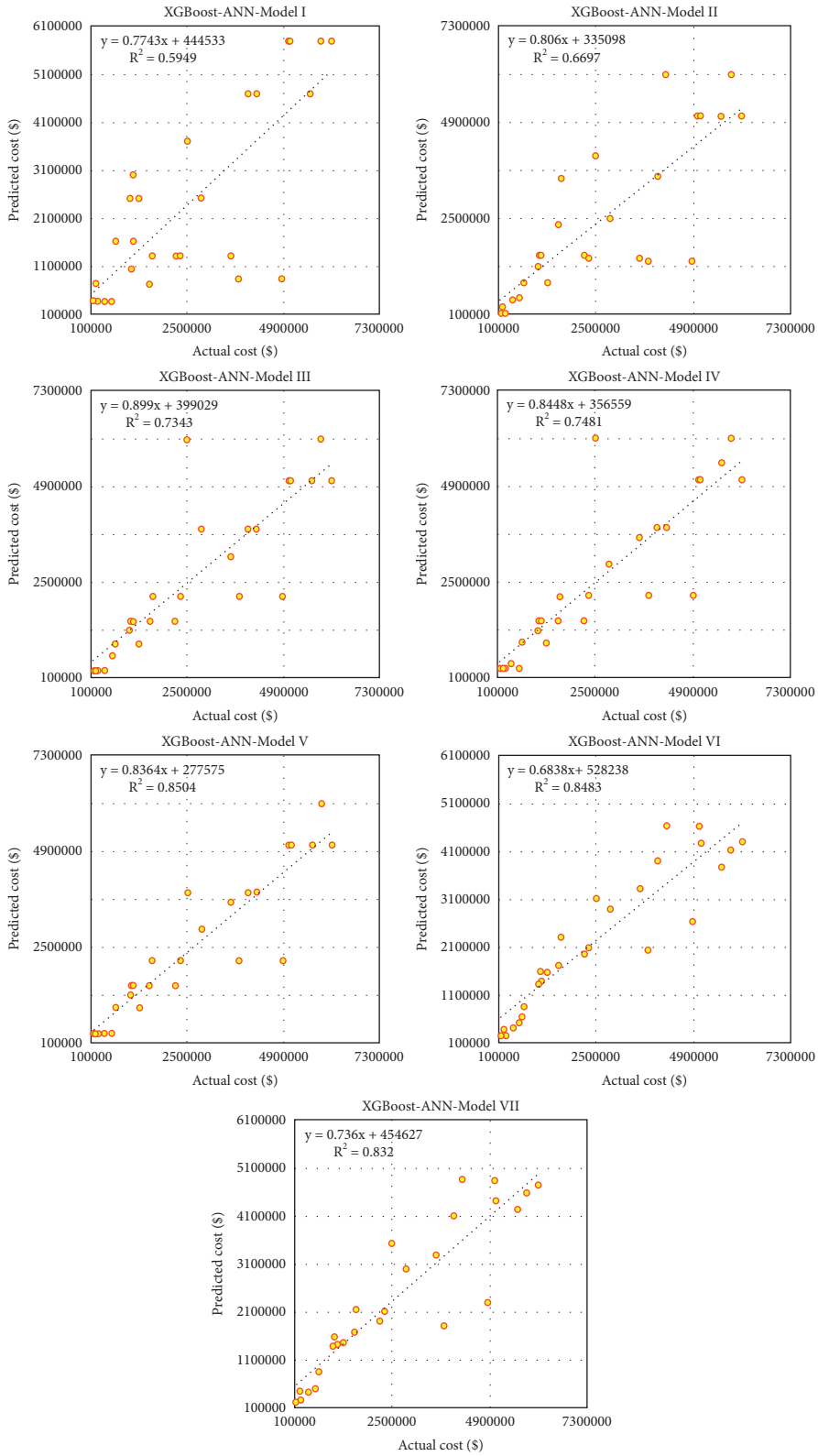

Figure 7: Scatter plot presentation of the XGBoost-ANN model for tested data.

parameters (i.e., XGBoost-RF-Model VI) led to reducing MAPE to 0.25 and producing a residual plot with no outlier points.

The poor performance was achieved by the RF algorithm in Model I and Model II, where the gained R^2 was less than 0.7 and the residual error was high, as illustrated in Figures 8

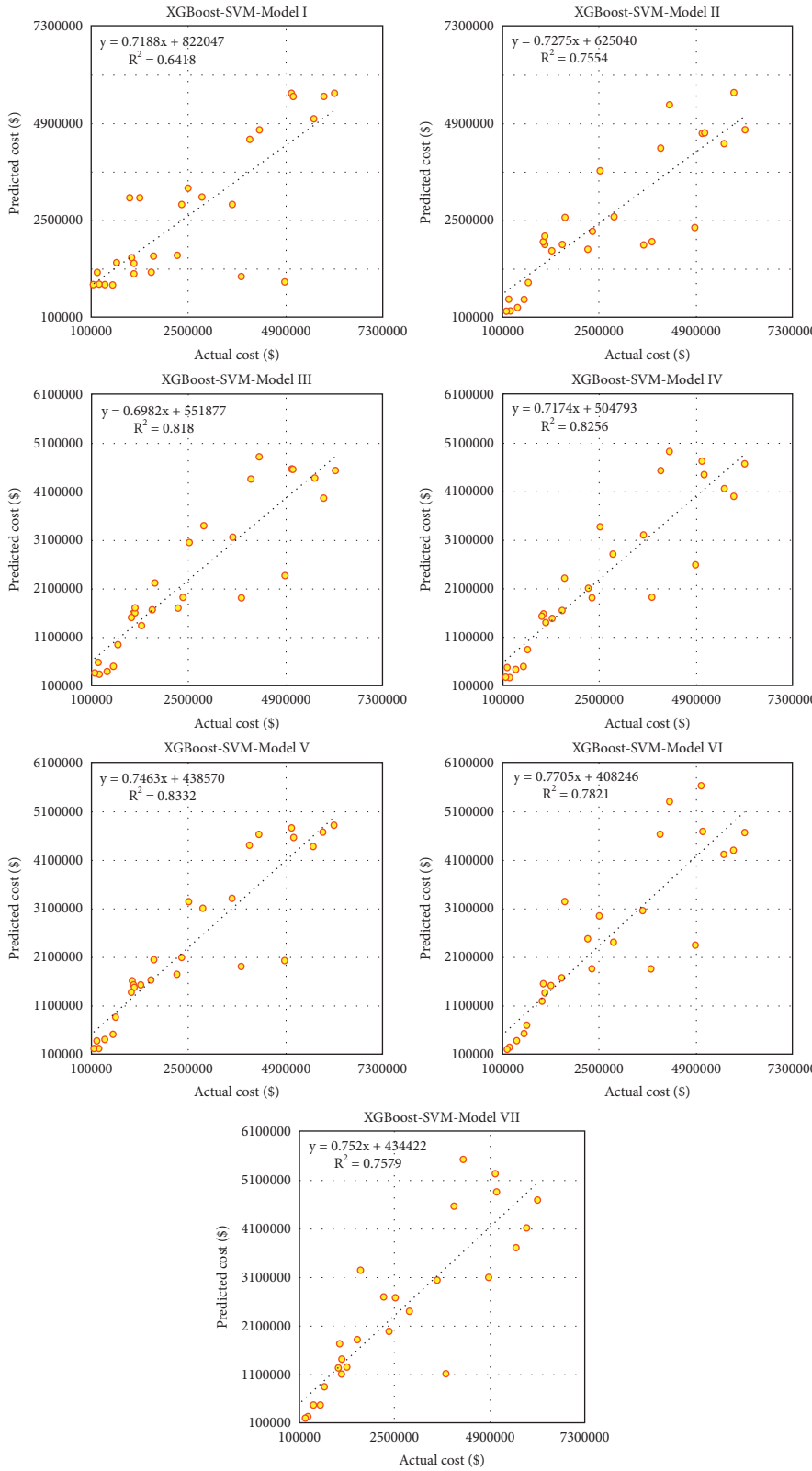

Figure 8: Scatter plot presentation of the XGBoost-SVM model for tested data.

and 9. For the ANN algorithm, the model increased its prediction performance by increasing the number of input parameters, and the best R^2 was achieved by XGBoost-ANN-Model V with RMSE equal to 750698.034, as reported in Table 4. The XGBoost-ANN-Model revealed the poorest results of Model I with high residual errors and the farthest distance to actual cost (see Figures 9 and 10). In the case of the SVM model, three models (i.e., Model III, Model IV, and

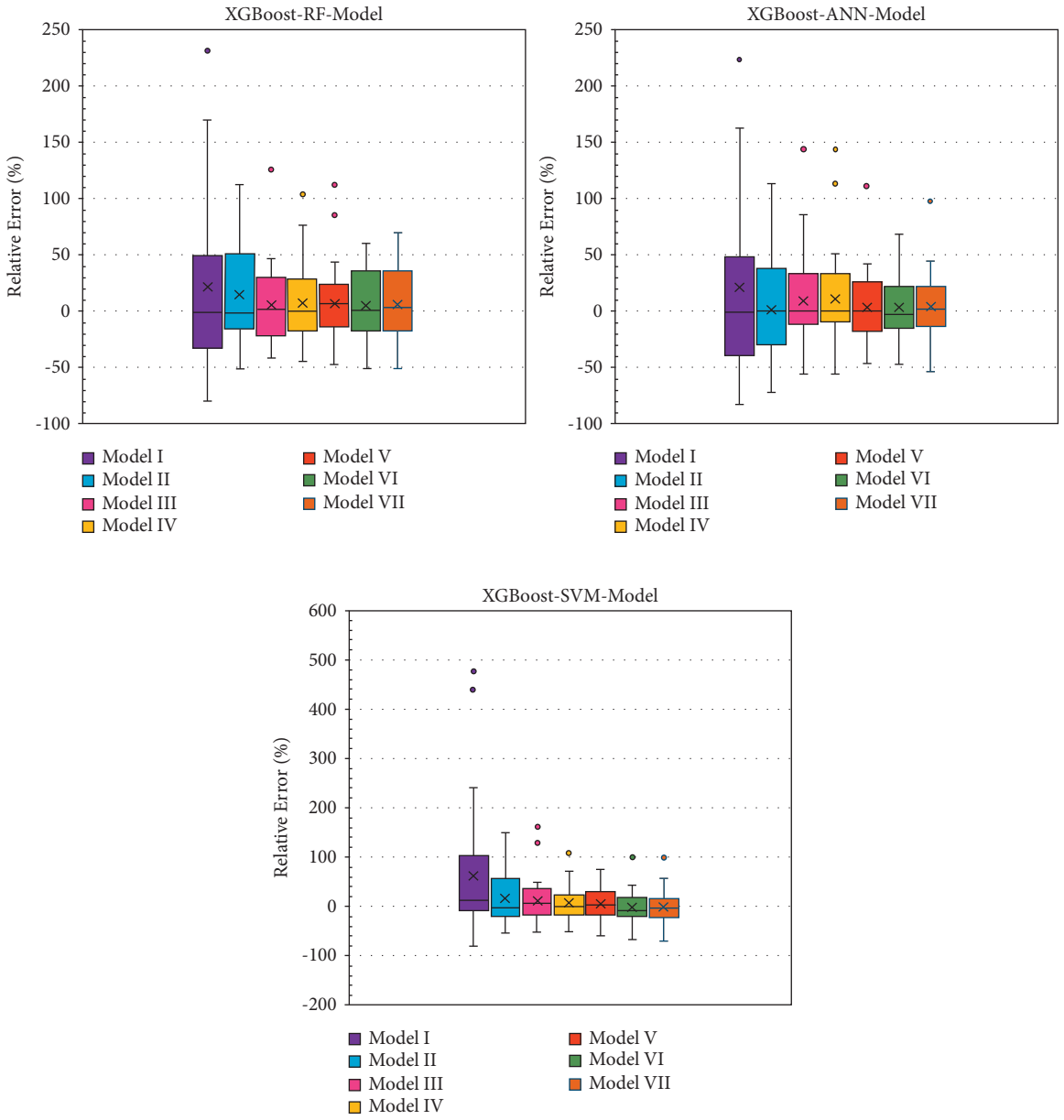

Figure 9: Box plot presentation of the developed AI models for the testing phase.

Model V) illustrated good performance with r-squared maxed out at 0.8, as depicted in Figure 7. The other combinations of the XGBoost-SVM model achieved good predictive accuracy with R^2 greater than 0.7, except for Model I that showed the lowest correlation coefficient and farthest position to the observed value, as shown in Figures 8 and 10. Based on the evaluation results, all AI models exhibited good performance when the number of input variables was increased in the estimation process. RF and SVM models performed better than ANN when using a few input variables, especially when applied to one input parameter. The comparison results revealed that integrating XGBoost with AI models enhanced the prediction accuracy by selecting the appropriate parameters for the modelling process. The integrated XGBoost algorithm with AI models revealed that using three input parameters, i.e. inflation, the total floor area, and the ground floor area, is necessary to get the accurate performance of cost estimation. The results revealed that increasing input variables from three to six reduced the error percentage and increased modelling efficiency. The results also showed that tree-based models outperformed classical models in their ability to handle complex models based on a few input variables. The reported results indicated that the RF model was able to understand the complex nature of construction cost estimation. The integrated XGBoost algorithm with AI models revealed the robustness of using predictive models when limited input variables are known by the project's stockholders. These results indicate the ability of the developed model to be used under uncertain circumstances. For future studies, other advanced selection algorithms like GA can be tested to simulate the complex behaviour among modelling parameters. Also,

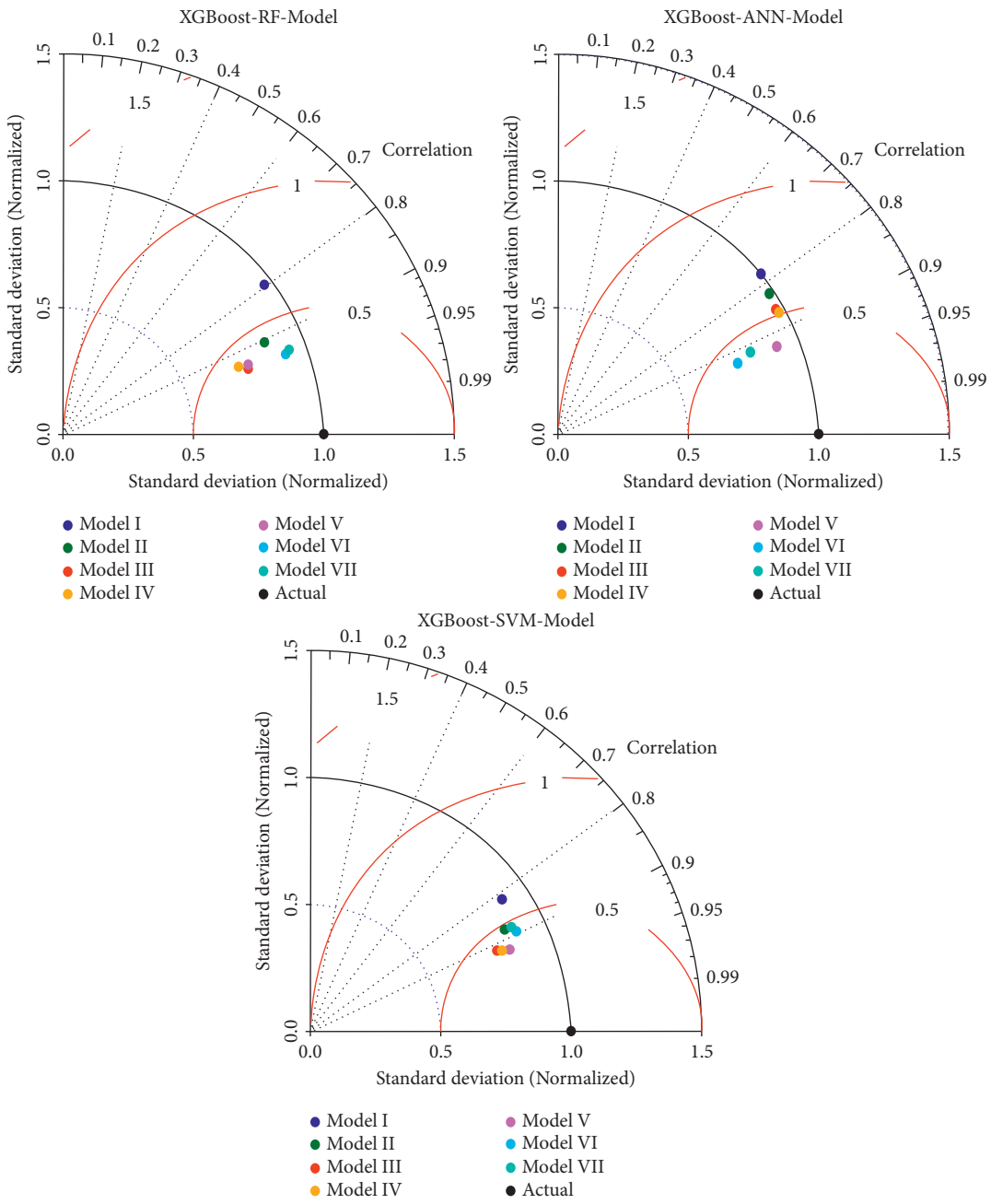

Figure 10: Taylor diagram of the developed AI models for the testing phase.

recent algorithms such as deep neural networks can be integrated with input selector algorithms to get low errors and more accurate results [94, 95].

6. Conclusions

Developing a reliable predictive model is an essential issue in construction cost estimation. In this research, the XGBoost algorithm was used to select the correlated parameters of the modelling process and hybridized with three AI models, namely, RF, ANN, and SVM, to estimate the construction cost. Datasets were collected based on the survey of 90 building projects constructed between 2016 and 2021. The results showed that the most correlated variables selected by XGBoost are inflation, the total floor area, and the ground floor area. For prediction performance, all AI models showed good reliability in the prediction process when applied to input variables of more than 2. The XGBoost-RF model revealed a high correlation coefficient in all combinations except for Model I, where R is below the acceptable performance level. The graphical evaluation showed that XGBoost-RF-Model VI gained the best performance with an r-squared of more than 0.8 and low residual error. The results also indicated that tree models could deal with complex systems and get accurate results based on the limited number of input variables. More input parameters should be investigated for future direction, and the GA

algorithm can be explored to generate significant feature selection. Also, a new deep learning algorithm can be presented to enhance the capability of predictive models.

Data Availability

The used data in the current research study can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to thank the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia, for the support in publishing this paper. In addition, the authors would like to thank Al-Mustaqbal University College (MUC-E-0122) for providing technical support for this research. This research was funded by Universiti Teknologi MARA. The authors would also like to thank the support received by the University of Baghdad.

References

[1] P. V. Ingle and G. Mahesh, "Construction project performance areas for Indian construction projects," *International Journal of Construction Management*, vol. 22, no. 8, pp. 1443–1454, 2020.

[2] W. Jing, H. I. Naji, R. N. Zehawi, Z. H. Ali, N. Al-Ansari, and Z. M. Yaseen, "System dynamics modeling strategy for civil construction projects: the concept of successive legislation periods," *Symmetry*, vol. 11, no. 5, p. 677, 2019.

[3] D. Myers, *Construction Economics: A New Approach*, Routledge, London, 2016.

[4] D.-G. Owusu-Manu, D. J. Edwards, A. Mohammed, W. D. Thwala, and T. Birch, "Short run causal relationship between foreign direct investment (FDI) and infrastructure development," *Journal of Engineering, Design and Technology*, vol. 17, no. 6, pp. 1202–1221, 2019.

[5] D. Bryde, "Perceptions of the impact of project sponsorship practices on project success," *International Journal of Project Management*, vol. 26, no. 8, pp. 800–809, 2008.

[6] J. Pollack, J. Helm, and D. Adler, "What is the Iron Triangle, and how has it changed?" *International Journal of Managing Projects in Business*, vol. 11, no. 2, pp. 527–547, 2018.

[7] A. M. Araba, Z. A. Memon, M. Alhawat, M. Ali, and A. Milad, "Estimation at completion in civil engineering projects: review of regression and soft computing models," *Knowledge-Based Engineering and Sciences*, vol. 2, no. 2, pp. 1–12, 2021.

[8] M. R. Haas, "Knowledge gathering, team capabilities, and project performance in challenging work environments," *Management Science*, vol. 52, no. 8, pp. 1170–1184, 2006.

[9] S. Hwang, "Dynamic regression models for prediction of construction costs," *Journal of Construction Engineering and Management*, vol. 135, no. 5, pp. 360–367, 2009.

[10] K. H. Mohammad, N. S. Ali, and B. M. Najm, "Assessment of the cost and time impact of variation orders on construction projects in Sulaimani governorate," *Journal of Engineering*, vol. 27, no. 2, pp. 106–125, 2021.

[11] A. Akintoye, "Analysis of factors influencing project cost estimating practice," *Construction Management & Economics*, vol. 18, no. 1, pp. 77–89, 2000.

[12] T. Huo, H. Ren, and W. Cai, G. Q. Shen, Measurement and dependence analysis of cost overruns in megatransport infrastructure projects: case study in Hong Kong," *Journal of Construction Engineering and Management*, vol. 144, no. 3, Article ID 5018001, 2018.

[13] E. Matel, F. Vahdatikhaki, S. Hosseinyalamdary, T. Evers, and H. Voordijk, "An artificial neural network approach for cost estimation of engineering services," *International Journal of Construction Management*, vol. 22, no. 7, pp. 1274–1287, 2019.

[14] D. D. Ahiaga-Dagbui and S. D. Smith, "Neural networks for modelling the final target cost of water projects," *Proceedings of the 28th Annu. ARCOM Conf*, pp. 307–316, Edinburgh, UK, September 2012.

[15] A. Jaafari, "Management of risks, uncertainties and opportunities on projects: time for a fundamental shift," *International Journal of Project Management*, vol. 19, no. 2, pp. 89–101, 2001.

[16] S. Demirkesen and B. Ozorhon, "Impact of integration management on construction project management performance," *International Journal of Project Management*, vol. 35, no. 8, pp. 1639–1654, 2017.

[17] S. Banihashemi, M. R. Hosseini, H. Golizadeh, and S. Sankaran, "Critical success factors (CSFs) for integration of sustainability into construction project management practices in developing countries," *International Journal of Project Management*, vol. 35, no. 6, pp. 1103–1119, 2017.

[18] S. Kim, S. Chang, and D. Castro-Lacouture, "Dynamic modeling for analyzing impacts of skilled labor shortage on construction project management," *Journal of Management in Engineering*, vol. 36, no. 1, Article ID 4019035, 2020.

[19] Z. Shehu, I. R. Endut, A. Akintoye, and G. D. Holt, "Cost overrun in the Malaysian construction industry projects: a deeper insight," *International Journal of Project Management*, vol. 32, no. 8, pp. 1471–1480, 2014.

[20] D. R. Al-Tawal, M. Arafah, and G. J. Sweis, "A model utilizing the artificial neural network in cost estimation of construction projects in Jordan," *Engineering Construction and Architectural Management*, vol. 28, no. 9, pp. 2466–2488, 2020.

[21] M. O. Sanni-Anibire, R. Mohamad Zin, and S. O. Olatunji, "Developing a preliminary cost estimation model for tall buildings based on machine learning," *International Journal of Management Science and Engineering Management*, vol. 16, no. 2, pp. 134–142, 2021.

[22] A. Enshassi, S. Mohamed, and I. Madi, "Factors affecting accuracy of cost estimation of building contracts in the Gaza Strip," *Journal of Financial Management of Property and Construction*, vol. 10, no. 2, 2005.

[23] I. Attarzadeh and S. H. Ow, "Proposing a new software cost estimation model based on artificial neural networks," *2nd International Conference on Computer Engineering and Technology*, vol. 3, pp. V3–V487, 2010.

[24] J. Ahn, S. H. Ji, S. J. Ahn et al., "Performance evaluation of normalization-based CBR models for improving construction cost estimation," *Automation in Construction*, vol. 119, Article ID 103329, 2020.

[25] M. Alzara, J. Kashiwagi, D. Kashiwagi, and A. Al-Tassan, "Using PIPS to minimize causes of delay in Saudi Arabian construction projects: university case study," *Procedia Engineering*, vol. 145, pp. 932–939, 2016.

[26] A. Enshassi, S. Mohamed, and M. Abdel-Hadi, "Factors affecting the accuracy of pre-tender cost estimates in the Gaza

Strip," *Journal of Construction in Developing Countries*, vol. 18, no. 1, 2013.

[27] A. Akintoye and E. Fitzgerald, "A survey of current cost estimating practices in the UK," *Construction Management & Economics*, vol. 18, no. 2, pp. 161–172, 2000.

[28] I. ElSawy, H. Hosny, and M. A. Razek, "A neural network model for construction projects site overhead cost estimating in Egypt," arXiv Prepr, 2011, https://arXiv1106.1570.

[29] L. Sabol, *Challenges in Cost Estimating with Building Information Modeling*, pp. 1–16, IFMA world Work, Washington, DC, 2008.

[30] M. Barakchi, O. Torp, and A. M. Belay, "Cost estimation methods for transport infrastructure: a systematic literature review," *Procedia Engineering*, vol. 196, pp. 270–277, 2017.

[31] M.-Y. Cheng, H.-C. Tsai, and E. Sudjono, "Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry," *Expert Systems with Applications*, vol. 37, no. 6, pp. 4224–4231, 2010.

[32] O. Alshboul, A. Shehadeh, G. Almasabha, and A. S. Almuflih, "Extreme gradient boosting-based machine learning approach for green building cost prediction," *Sustainability*, vol. 14, no. 11, p. 6651, 2022.

[33] K. Upreti, U. K. Singh, R. Jain, K. Kaur, and A. K. Sharma, "Fuzzy logic based support vector regression (SVR) model for software cost estimation using machine learning," in *ICT Systems and Sustainability*, pp. 917–927, Springer, Berlin, Germany, 2022.

[34] A. H. Al-Momani, "Construction cost prediction for public school buildings in Jordan," *Construction Management & Economics*, vol. 14, no. 4, pp. 311–317, 1996.

[35] D. J. Lowe, M. W. Emsley, and A. Harding, "Predicting construction cost using multiple regression techniques," *Journal of Construction Engineering and Management*, vol. 132, no. 7, pp. 750–758, 2006.

[36] F. Shutian, Z. Tianyi, and Z. Ying, "Prediction of construction projects' costs based on fusion method," *Engineering Computations*, vol. 34, no. 7, pp. 2396–2408, 2017.

[37] H. Son, C. Kim, and C. Kim, "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*, vol. 27, pp. 60–66, 2012.

[38] R. Wang, V. Asghari, C. M. Cheung, S.-C. Hsu, and C.-J. Lee, "Assessing effects of economic factors on construction cost estimation using deep neural networks," *Automation in Construction*, vol. 134, Article ID 104080, 2022.

[39] Z. M. Yaseen, Z. H. Ali, S. Q. Salih, and N. Al-Ansari, "Prediction of risk delay in construction projects using a hybrid artificial intelligence model," *Sustainability*, vol. 12, no. 4, p. 1514, 2020.

[40] H. T. Almusawi and A. M. Burhan, "Developing a model to estimate the productivity of ready mixed concrete batch plant," *Journal of Engineering*, vol. 26, no. 10, pp. 80–93, 2020.

[41] H. Sanikhani, R. C. Deo, P. Samui et al., "Survey of different data-intelligent modeling strategies for forecasting air temperature using geographic information as model predictors," *Computers and Electronics in Agriculture*, vol. 152, pp. 242–260, 2018.

[42] I. Peško, V. Mucenski, M. Seslija et al., "Estimation of costs and durations of construction of urban roads using ANN and SVM," *Complexity*, vol. 2017, p. 13, 2017.

[43] S. S. Arage and N. V Dharwadkar, "Cost estimation of civil construction projects using machine learning paradigm," in *Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 594–599, Palladam India, February 2017.

[44] S. S. Arage and N. V. Dharwadkar, "Prediction and estimation of civil construction cost using linear regression and neural network," *International Journal of Intelligent Systems Design and Computing*, vol. 2, no. 1, pp. 28–44, 2018.

[45] M. Juszczyk, "Residential buildings conceptual cost estimates with the use of support vector regression," *MATEC Web of Conferences*, vol. 196, p. 04090, 2018.

[46] M. Juszczyk, "On the search of models for early cost estimates of bridges: an SVM-based approach," *Buildings*, vol. 10, no. 1, p. 2, 2019.

[47] V. Chandanshive and A. R. Kambekar, "Estimation of building construction cost using artificial neural networks," *J. Soft Comput. Civ. Eng.*vol. 3, no. 1, pp. 91–107, 2019.

[48] S. T. Hashemi, O. M. Ebadati E, and H. Kaur, "A hybrid conceptual cost estimating model using ANN and GA for power plant projects," *Neural Computing & Applications*, vol. 31, no. 7, pp. 2143–2154, 2019.

[49] T. Lin, T. Yi, C. Zhang, and J. Liu, "Intelligent prediction of the construction cost of substation projects using support vector machine optimized by particle swarm optimization," *Mathematical Problems in Engineering*, vol. 2019, Article ID 7631362, 10 pages, 2019.

[50] 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.

[51] K. Tijanić, D. Car-Pušić, and M. Šperac, "Cost estimation in road construction using artificial neural network," *Neural Computing & Applications*, vol. 32, no. 13, pp. 9343–9355, 2020.

[52] D. Car-Pusic, S. Petruseva, V. Zileska Pancovska, and Z. Zafirovski, "Neural network-based model for predicting preliminary construction cost as part of cost predicting system," *Advances in Civil Engineering*, vol. 2020, Article ID 8886170, 13 pages, 2020.

[53] C.-H. Huang and S.-H. Hsieh, "Predicting BIM labor cost with random forest and simple linear regression," *Automation in Construction*, vol. 118, Article ID 103280, 2020.

[54] D. Ye, "An algorithm for construction project cost forecast based on particle swarm optimization-guided BP neural network," *Scientific Programming*, vol. 2021, p. 8, 2021.

[55] S. Shoar, N. Chileshe, and J. D. Edwards, "Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: application of random forest regression," *Journal of Building Engineering*, vol. 50, Article ID 104102, 2022.

[56] Y.-M. Cheng, "An exploration into cost-influencing factors on construction projects," *International Journal of Project Management*, vol. 32, no. 5, pp. 850–860, 2014.

[57] M. Gunduz and O. L. Maki, "Assessing the risk perception of cost overrun through importance rating," *Technological and Economic Development of Economy*, vol. 24, no. 5, pp. 1829–1844, 2018.

[58] K. C. Iyer and K. N. Jha, "Factors affecting cost performance: evidence from Indian construction projects," *International Journal of Project Management*, vol. 23, no. 4, pp. 283–295, 2005.

[59] L. Zhao, J. Mbachu, and Z. Liu, "Identifying significant cost-influencing factors for sustainable development in construction industry using structural equation modelling," *Mathematical Problems in Engineering*, vol. 2020, p. 16, 2020.

[60] H. A. Afan, M. F. Allawi, A. El-Shafie et al., "Input attributes optimization using the feasibility of genetic nature inspired

algorithm: application of river flow forecasting," *Scientific Reports*, vol. 10, no. 1, pp. 4684–4715, 2020.

[61] D. Chakraborty, H. Elhegazy, H. Elzarka, and L. Gutierrez, "A novel construction cost prediction model using hybrid natural and light gradient boosting," *Advanced Engineering Informatics*, vol. 46, Article ID 101201, 2020.

[62] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco, California, USA, August 2016.

[63] H. Al-Ageeli and A. S. Alzobaee, "The most influential factor on the stumble and failure of the governmental projects," *Journal of Engineering*, vol. 22, no. 3, pp. 93–110, 2016.

[64] H. Tao, M. Habib, I. Aljarah, H. Faris, H. A. Afan, and Z. M. Yaseen, "An Intelligent Evolutionary Extreme Gradient Boosting Algorithm Development for Modeling Scour Depths under Submerged Weir," *Inf. Sci. (Ny)*, vol. 570, 2021.

[65] H. Tao, S. M. Awadh, S. Q. Salih, S. S. Shafik, and Z. M. Yaseen, "Integration of extreme gradient boosting feature selection approach with machine learning models: application of weather relative humidity prediction," *Neural Computing & Applications*, vol. 34, no. 1, pp. 515–533, 2021.

[66] H. Tao, S. Salih, A. Y. Oudah et al., "Development of new computational machine learning models for longitudinal dispersion coefficient determination: case study of natural streams, United States," *Environmental Science and Pollution Research*, vol. 29, no. 24, pp. 35841–35861, 2022.

[67] S. Alves Basilio and L. Goliatt, "Gradient boosting hybridized with exponential natural evolution strategies for estimating the strength of geopolymer self-compacting concrete," *Knowledge-Based Engineering and Sciences*, vol. 3, no. 1, pp. 1–16, 2022.

[68] Y. Wang and X. S. Ni, "A XGBoost Risk Model via Feature Selection and Bayesian Hyper-Parameter Optimization," arXiv Prepr, 2019, https://arXiv1901.08433.

[69] K. Nordhausen, "The elements of statistical learning: data mining, inference, and prediction, second edition by trevor Hastie, robert tibshirani, jerome friedman," *International Statistical Review*, vol. 77, no. 3, p. 482, 2009.

[70] L. Breiman, "Using iterated bagging to debias regressions," *Machine Learning*, vol. 45, no. 3, pp. 261–277, 2001.

[71] J. Zhou, X. Shi, K. Du, X. Qiu, X. Li, and H. S. Mitri, "Feasibility of random-forest approach for prediction of ground settlements induced by the construction of a shield-driven tunnel," *International Journal of Geomechanics*, vol. 17, no. 6, Article ID 4016129, 2017.

[72] Y. Zhang, A. Javanmardi, Y. Liu et al., "How does experience with delay shape managers' making-do decision: random forest approach," *Journal of Management in Engineering*, vol. 36, no. 4, Article ID 4020030, 2020.

[73] X. Liu, Y. Song, W. Yi, X. Wang, and J. Zhu, "Comparing the random forest with the generalized additive model to evaluate the impacts of outdoor ambient environmental factors on scaffolding construction productivity," *Journal of Construction Engineering and Management*, vol. 144, no. 6, Article ID 4018037, 2018.

[74] S. K. Bhagat, T. Tiyasha, A. Kumar et al., "Integrative artificial intelligence models for Australian coastal sediment lead prediction: an investigation of in-situ measurements and meteorological parameters effects," *Journal of Environmental Management*, vol. 309, Article ID 114711, 2022.

[75] B. Leo, "Bagging Predictors," *Mach. Learn*, vol. 24, 1996.

[76] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[77] M. A. Khan, S. A. Memon, F. Farooq, M. F. Javed, F. Aslam, and R. Alyousef, "Compressive strength of fly-ash-based geopolymer concrete by gene expression programming and random forest," *Advances in Civil Engineering*, vol. 2021, p. 17, 2021.

[78] B. Salman and M. M. Kadhum, "Predicting of load carrying capacity of reactive powder concrete and normal strength concrete column specimens using artificial neural network," *Knowledge-Based Eng. Sci.*vol. 3, no. 1, pp. 45–53, 2022.

[79] A. De Fenza, A. Sorrentino, and P. Vitiello, "Application of artificial neural networks and probability ellipse methods for damage detection using lamb waves," *Composite Structures*, vol. 133, pp. 390–403, 2015.

[80] C. Yan, Z. Yousef, S. A. Alireza et al., "A Review Study of Application of Artificial Intelligence in Construction Management and Composite Beams," vol. 39, no. 6, 2021.

[81] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[82] H. Adeli, "Neural networks in civil engineering: 1989–2000," *Computer-Aided Civil and Infrastructure Engineering*, vol. 16, no. 2, pp. 126–142, 2001.

[83] R. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, vol. 4, no. 2, pp. 4–22, 1987.

[84] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation*, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[85] V. N. Vapnik, *The Nature of Statistical Learning Theory*, vol. 8, no. 6, 2000.

[86] G. Zhang, Z. H. Ali, M. S. Aldlemy et al., "Reinforced concrete deep beam shear strength capacity modelling using an integrative bio-inspired algorithm with an artificial intelligence model," *Engineering with Computers*, vol. 38, 2020.

[87] S. Raghavendra N and P. C. Deka, "Support vector machine applications in the field of hydrology: a review," *Applied Soft Computing*, vol. 19, pp. 372–386, 2014.

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

[89] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and statistics for engineers and scientists*, Vol. 5, Macmillan, New York, 1993.

[90] 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.

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

[92] G. Mahalakshmi and C. Rajasekaran, "Early cost estimation of highway projects in India using artificial neural network," *Lecture Notes in Civil Engineering*, vol. 25, pp. 659–672, 2019.

[93] S. Shartooh Sharqi and A. Bhattarai, "Evaluation of several machine learning models for field canal improvement project cost prediction," *Complexity*, vol. 2021, p. 12, 2021.
[94] J. A. Khalaf, A. A. Majeed, M. S. Aldlemy et al., "Hybridized deep learning model for perfobond rib shear strength connector prediction," *Complexity*, vol. 2021, p. 21, 2021.
[95] M. W. Falah, S. H. Hussein, M. A. Saad et al., "Compressive strength prediction using coupled deep learning model with extreme gradient boosting algorithm: environmentally friendly concrete incorporating recycled aggregate," *Complexity*, vol. 2022, p. 22, 2022.