Journal of Science and Transport Technology
Journal homepage: https://jstt.vn/index.php/en

UNIVERSITY OF TRANSPORT TECHNOLOGY

An Interpretable Data-Driven Polynomial Regression Model for Early-Stage Prediction of Reinforced Concrete Floor Cost in Civil Buildings

Hong Ha Le[1], Dac Hoang Dao[2*]

[1]Faculty of Building and Industrial Construction, Hanoi University of Civil Engineering, 55 Giai Phong, Bach Mai Ward, Hanoi, Vietnam

[2]AMS4 Research Group (Advanced Materials and Smart Structure for Sustainable System), University of Transport Technology, 54 Trieu Khuc Street, Thanh Liet Ward, Hanoi, Vietnam

Article info

Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2026.en.6.3.216-238

***Corresponding author:**
Email address:
hoangdd@utt.edu.vn

Received: 10/04/2026
Received in Revised Form: 24/06/2026
Accepted: 05/07/2026

Abstract: Accurate early-stage cost estimation is essential for value engineering and preliminary decision-making in civil building projects. However, conventional estimating approaches often rely on detailed design information or simplified linear assumptions, which limit their usefulness at the conceptual stage. This study develops an interpretable data-driven polynomial regression framework for predicting reinforced concrete floor cost in civil buildings. A dataset of 4477 samples derived from RSMeans floor-system assemblies was used, including slab type, tributary area, superimposed load, formwork price, and concrete price as predictors. Four candidate polynomial regression models with degrees 1 to 4 were constructed and compared using a 70:30 training-testing split and 1000 Monte Carlo simulations based on the coefficient of determination, root mean square error, and mean absolute error. The results show that predictive performance improves consistently as the polynomial degree increases. Although the quartic model achieved the highest raw predictive accuracy, the cubic model was selected as the final model because it provided the most appropriate balance between predictive performance and interpretability. The selected model achieved R^2=0.965, RMSE = 8.764 USD/m², and MAE = 6.163 USD/m² on the training dataset, while the corresponding testing results were R^2=0.965, RMSE = 8.567 USD/m², and MAE = 6.058 USD/m². Feature-importance analysis, centered partial dependence plots, and contour plots further showed that formwork and concrete are the dominant cost drivers and that the response surface is strongly nonlinear and interaction-driven. The proposed framework provides a transparent and practically usable tool for early-stage floor cost prediction and engineering decision support.

Keywords: Construction cost prediction; Reinforced concrete floor cost; Polynomial regression; Interpretable machine learning; Monte Carlo simulation; Early-stage cost estimation.

1. Introduction

The construction industry is increasingly recognized as a data-intensive sector in which large volumes of information are generated across planning, design, procurement, and execution stages. However, the practical value of these data is still far from being fully exploited for decision support in early project phases [1–4]. Among the many managerial problems in construction, early-stage cost estimation remains one of the most critical because it directly affects design selection, feasibility assessment, budget planning, and investment decisions. This issue is particularly important in civil building projects, where multiple structural alternatives may satisfy the same functional requirements but differ significantly in cost, performance, and constructability [5–7].

In this context, value engineering has long been regarded as a systematic approach for identifying cost-efficient alternatives while maintaining technical adequacy and required project performance. The effectiveness of value engineering, however, depends strongly on the availability of sufficiently accurate preliminary cost information. If cost information is unreliable, then the basis for comparing design alternatives is weakened and the quality of early decision-making is reduced [5, 6]. This challenge is especially visible in reinforced concrete building systems, where designers and estimators must frequently choose among alternative slab systems, geometries, loading configurations, and material-price conditions under limited time and incomplete design information.

Traditional cost-estimating approaches are not fully adequate for this purpose. Detailed estimating methods usually rely on complete drawings, specifications, and quantity takeoff, which are time-consuming and difficult to apply during conceptual design. Comparative and parametric methods are faster, but they often assume simplified linear relationships between basic design variables and final cost [8, 9]. Such assumptions are frequently unrealistic in construction practice, where nonlinear interactions among geometry, structural type, loading, and market-dependent material prices are common. As a result, early-stage estimates based on conventional linear logic may lack robustness and may not provide a reliable foundation for systematic value-oriented decision-making [9, 10].

Against this background, machine learning has emerged as a promising alternative for construction cost estimation. Over the past two decades, a wide range of predictive models have been proposed for this purpose, including artificial neural networks, support vector machines, case-based reasoning systems, fuzzy systems, Bayesian approaches, and ensemble learning techniques [11]. For example, conceptual cost prediction in construction has been addressed using evolutionary fuzzy hybrid neural networks [10], principal component analysis combined with support vector regression [12], artificial neural networks for early design cost estimation [13], case-based reasoning at the preliminary stage [14], multistep-ahead estimation strategies [8], regression-based models for green and sustainable buildings [9, 15], support vector regression for residential and bridge-related estimation problems [16], artificial neural networks for sports-field and road-construction cost estimation [16, 17], hybrid regression–genetic algorithm models for storage tanks [18], artificial neural networks for engineering services [19], and hybrid artificial neural network–genetic algorithm models for power plant projects [20]. These studies collectively show that data-driven techniques can substantially improve predictive accuracy relative to overly simplified traditional approaches.

At the same time, the literature also reveals an important limitation. Many high-performing machine learning models are difficult to interpret and are therefore often treated by practitioners as black-box tools [4, 21, 22]. This lack of transparency is not a minor issue. In cost engineering, practitioners need more than accurate outputs; they also need to understand how the input variables influence predictions, whether the model behaves reasonably from an engineering perspective, and how predictions change under alternative design scenarios. Even when advanced ensemble methods such as gradient boosting and random forests offer strong predictive performance [23, 24], the trade-off between accuracy and interpretability remains a practical concern, particularly in early-stage applications where engineering judgment must still play a central role.

More recent studies have further emphasized that high predictive accuracy alone is not sufficient for practical construction cost estimation. Instead, predictive models should also provide explainable or interpretable outputs, using tools such as SHAP, interpretable machine learning frameworks, and uncertainty-aware prediction models to improve user trust and support engineering decision-making [25, 26].

This interpretability gap is highly relevant to reinforced concrete floor cost prediction. In the reference study by Chakraborty et al. [4], a hybrid boosting-based framework achieved excellent predictive performance on a mixed dataset of 4477 structural assembly cases and demonstrated the value of data-driven cost modeling for reinforced concrete floor systems. However, the same study also reinforces a broader point emphasized in the literature: the more sophisticated the predictive algorithm becomes, the more difficult it may be to directly communicate its logic to engineers and estimators [4, 21, 22]. Therefore, there is still clear value in developing simpler predictive structures that retain the ability to represent nonlinear effects and variable interactions while remaining analytically transparent.

The present study does not claim to introduce a new dataset; rather, it reuses the RSMeans-derived dataset reported by Chakraborty et al. [4] to address a different methodological question. While Chakraborty et al. [4] focused on maximizing predictive accuracy and uncertainty estimation through a hybrid LGBoost–NGBoost framework, the present study shifts the emphasis toward a compact and transparent polynomial regression structure. The additional scientific value lies in: (i) systematically comparing polynomial models with different degrees under repeated Monte Carlo evaluation; (ii) selecting a cubic model based on the balance between accuracy, model complexity, and interpretability; (iii) explaining the nonlinear cost response through feature importance, centered partial dependence plots, and contour-based interaction analysis; and (iv) translating the selected model into a simple engineering decision-support interface for early-stage floor cost prediction and target-cost scenario exploration.

Accordingly, the present study develops an interpretable data-driven polynomial regression model for early-stage prediction of reinforced concrete floor cost in civil buildings. The analysis is based on the same general problem setting addressed in prior machine-learning research on floor cost estimation, but it deliberately shifts the methodological emphasis toward transparency, structural simplicity, and practical interpretability. The model uses five input variables with direct engineering and economic meaning slab type, tributary area, superimposed load, formwork price, and concrete price to predict floor cost. Rather than relying on a highly complex black-box architecture, the study evaluates polynomial regression models with increasing degrees of nonlinear expansion in order to identify a model form that captures the essential nonlinear cost structure without sacrificing interpretability.

Other interpretable nonlinear modeling approaches, such as generalized additive models and regression splines, could also be considered for early-stage cost prediction. However, these methods usually emphasize smooth marginal effects of individual predictors and may require additional model structure to represent higher-order interactions among design type, tributary area, loading condition, and material-price variables. Polynomial regression was selected in this study because it provides an explicit analytical equation, directly incorporates both nonlinear and interaction terms, and remains straightforward to implement in an engineering decision-support tool. This makes it suitable for the objective of developing a compact, transparent, and practically usable prediction framework rather than a purely accuracy-oriented black-box model.

The objectives of this study are therefore threefold. First, it aims to establish a transparent polynomial prediction framework for reinforced concrete floor cost at the preliminary design stage. Second, it compares candidate polynomial models of different degrees under repeated Monte Carlo simulation in order to evaluate predictive performance and model stability in a rigorous manner. Third, it seeks to interpret the final selected model through descriptive analysis, feature-importance assessment, centered partial dependence plots, contour analysis, and engineering-oriented application. In doing so, the study contributes not merely another predictive model, but a practically usable and analytically explainable decision-support tool for early-stage design and cost evaluation in civil construction.

The remainder of this paper is organized as follows. Section 2 describes the dataset, variable framework, polynomial modeling strategy, and simulation-based evaluation design. Section 3 presents the comparative model results, the performance of the selected model, and the analytical interpretation of its behavior. The final section summarizes the main findings, contributions, and implications for future research and practical application.

2. Materials and Methods

2.1. Dataset, variables, and descriptive characteristics

In this study, construction cost is approached through the analysis of floor cost, which represents a major and structurally meaningful component of the overall structural cost of a building. Restricting the scope to floor cost is justified on two grounds. First, from a practical perspective, detailed information on total building cost is often fragmented and inconsistent across projects, whereas floor-related cost data are more standardized, more comparable, and more readily available in established construction cost databases such as RSMeans. Second, from a technical perspective, the floor system accounts for a substantial share of both construction volume and structural cost in medium- and high-rise civil buildings. It can therefore serve as a representative indicator for modeling cost variation in reinforced concrete structural systems. This scope is also consistent with the rationale emphasized in the reference study by Chakraborty et al., which used floor-system cost as a practical and analytically meaningful basis for early-stage structural cost prediction.

Based on this rationale, this section examines the dataset used for floor cost prediction, with emphasis on variable definition, descriptive statistics, and preliminary relationships among the study variables. These analyses provide the empirical foundation for subsequent model development and support the objective of building a predictive model that is both practically useful and analytically interpretable. The present study builds on the same general RSMeans-derived data environment reported in the previous machine-learning study, but reorients the modeling strategy toward a more transparent polynomial regression framework.

The dataset was derived from RSMeans Assemblies Books, a widely recognized source of construction cost information covering materials, labor, equipment, construction methods, and productivity. RSMeans is updated regularly to reflect changing material prices, labor conditions, equipment costs, and construction practices, which makes it a reliable source for comparative cost analysis. The data used in this study were compiled from floor-related structural assemblies for medium- and high-rise buildings and include observations collected over multiple years. The final database contains 4477 samples, which were previously reported by Chakraborty et al. [4] adapted here for the development of an interpretable polynomial prediction model.

The dataset includes six variables: slab type, tributary area (m²), superimposed load (kN/m²), formwork price (USD/m²), concrete price (USD/m³), and floor cost (USD/m²). The variable slab type is encoded from 1 to 6 and represents six common reinforced concrete floor systems, namely: (1) one-way slab, (2) two-way slab, (3) cast-in-place flat slab, (4) cast-in-place flat plate, (5) cast-in-place multi-span joist slab, and (6) cast-in-place waffle slab. Although slab type is a categorical design variable, the numerical codes from 1 to 6 are used in this study only as compact identifiers of the six reinforced concrete floor-system alternatives, rather than as a continuous physical scale or an ordinal ranking of structural performance. The remaining variables describe the key geometric, loading, and cost-related characteristics of the floor system. This variable structure was retained because it captures both engineering and economic dimensions relevant to early-stage cost prediction.

Table 1 presents the descriptive statistics of the study variables. The encoded variable slab type has a mean of 3.590 and a standard deviation of 1.781, indicating substantial diversity in the floor systems represented. Tributary area has a mean of 70.896 m², with values ranging from 20.903 m² to 167.225 m², and exhibits moderate positive skewness (0.621), suggesting that larger supported areas are less frequent but still sufficiently represented in the dataset. Superimposed load has a mean of 5.127 kN/m² and ranges from 1.915 to 9.576 kN/m², with slight positive skewness (0.474). Formwork price has a mean of 79.957 USD/m² and a standard deviation of 24.500 USD/m², while concrete price has a mean of 125.199 USD/m³ and ranges from 61.084 to 184.664 USD/m³. The response variable, floor cost, has a mean of 180.951 USD/m², a standard deviation of 46.530 USD/m², and ranges from 82.882 to 320.226 USD/m². Its skewness value (0.185) indicates a mildly right-skewed distribution, implying that relatively high-cost cases exist but do not dominate the sample. Overall, Table 1 shows that the dataset covers a sufficiently wide range of floor-system configurations to support nonlinear predictive modeling.

In the RSMeans assembly-cost structure, the unit price of concrete expressed in USD/m³ is converted into floor cost expressed in USD/m² through the quantity and dimensional assumptions embedded in each standardized floor assembly.

Table 1. Descriptive statistics of the study variables used for floor cost prediction

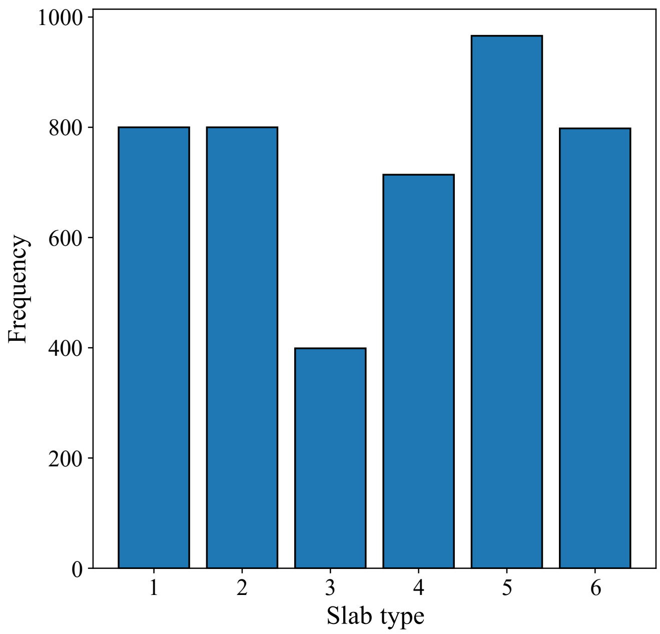
Statistic	Slab type	Tributary area (m²)	Superimposed load (kN/m²)	Formwork price (USD/m²)	Concrete price (USD/m³)	Floor cost (USD/m²)
Number of samples	4477	4477	4477	4477	4477	4477
Mean	3.59	70.896	5.127	79.957	125.199	180.951
Standard deviation	1.781	39.193	2.816	24.5	33.788	46.53
Minimum	1	20.903	1.915	45.103	61.084	82.882
25th percentile	2	37.161	1.915	63.292	96.054	144.021
Median	4	58.064	3.591	73.087	134.896	179.757
75th percentile	5	97.548	5.985	93.967	149.008	215.278
Maximum	6	167.225	9.576	147.998	184.664	320.226
Skewness	-0.129	0.621	0.474	0.838	-0.24	0.185

Note: Slab type is encoded as follows: 1 = one-way slab; 2 = two-way slab; 3 = cast-in-place flat slab; 4 = cast-in-place flat plate; 5 = cast-in-place multi-span joist slab; 6 = cast-in-place waffle slab.

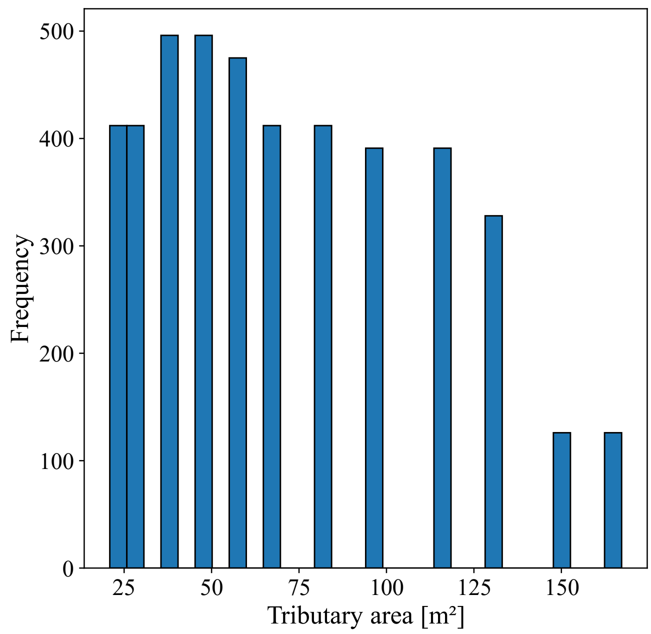
Fig. 1 illustrates the empirical distributions of the variables in the dataset. The frequency plot of slab type shows that type 5 occurs most frequently, whereas type 3 appears least often. Tributary area is moderately right-skewed, with most observations concentrated in the lower and middle ranges and relatively few cases above 150 m². Superimposed load exhibits several recurring design levels, which appear as distinct local peaks in the empirical distribution. Formwork price displays a visibly uneven distribution with concentration around several typical price bands, while concrete price also presents a multi-modal structure, indicating that the data reflect different pricing regimes rather than a single homogeneous market level. Finally, floor cost shows a broad distribution centered in the mid-range of the sample, with fewer observations at the very low and very high ends. These distributional features suggest that the dataset is heterogeneous enough to justify the use of a flexible nonlinear modeling framework.

Fig. 2 presents the Pearson correlation heatmap of the study variables and provides an initial view of their linear relationships. In general, slab type shows weak pairwise correlations with most of the remaining variables and nearly no direct linear correlation with floor cost. Tributary area shows a weak negative correlation with floor cost, suggesting that larger supported floor areas are, on average, associated with lower unit floor cost. By contrast, formwork price and concrete price show the strongest positive associations with the response variable, although these correlations remain moderate rather than dominant. Superimposed load displays only a very weak marginal linear relationship with floor cost.

These pairwise correlations should be interpreted cautiously. The Pearson correlation matrix only captures direct linear relationships between two variables at a time and does not reflect higher-order interactions or nonlinear effects. Therefore, a weak pairwise correlation does not imply that a variable is unimportant in prediction, especially when polynomial terms and interaction structures are introduced later in the modeling process. From an engineering perspective, all five predictors retained in this study slab type, tributary area, superimposed load, formwork price, and concrete price have clear physical or cost-related meaning in the design and construction of reinforced concrete floor systems. For this reason, all five input variables were preserved for subsequent model development, allowing the polynomial regression framework to capture both direct and interaction-based contributions to floor cost.



(a)



(b)

Fig. 1. Empirical distributions of the original study variables, including slab type, tributary area, superimposed load, formwork price, concrete price, and floor cost

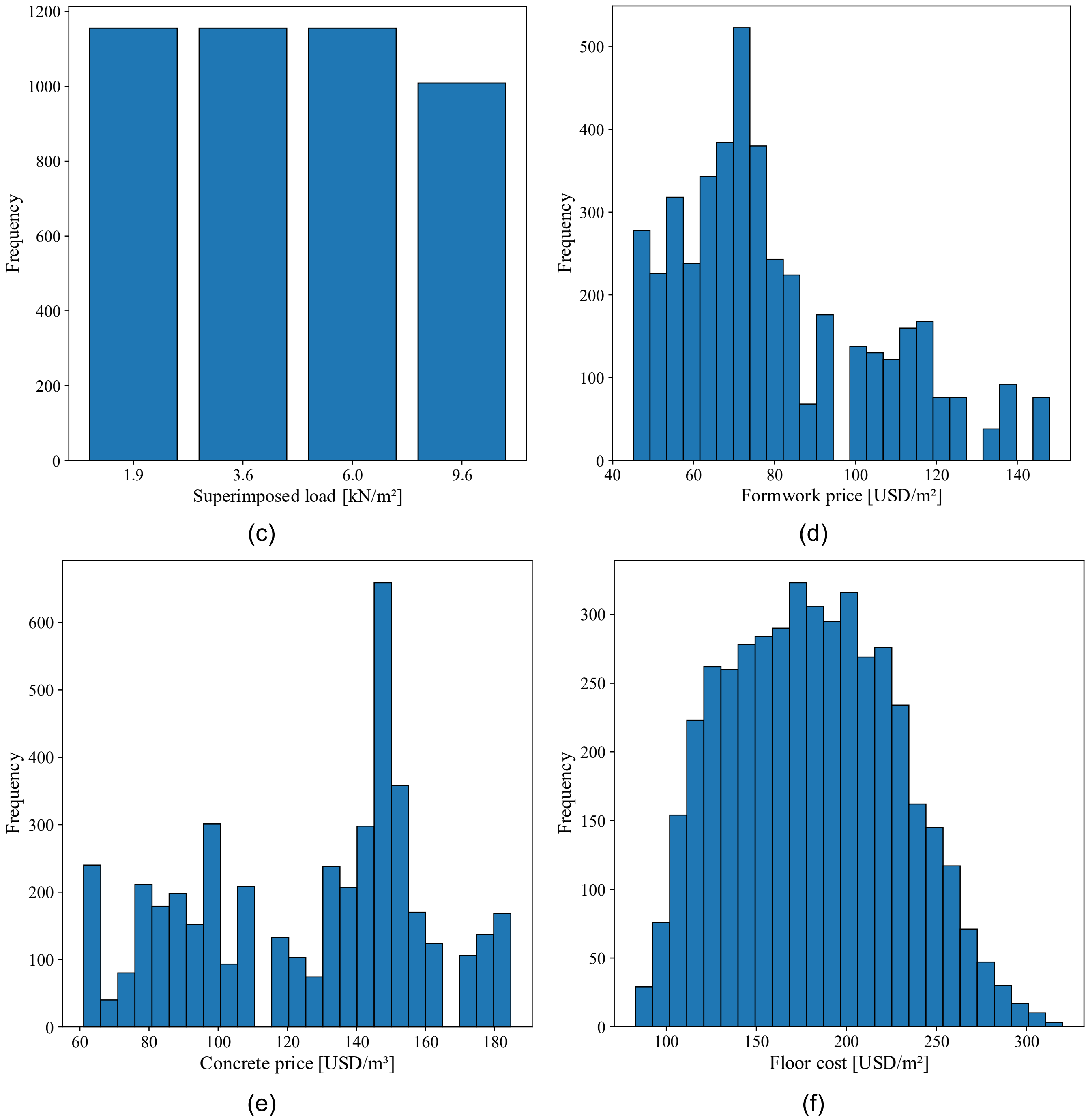


Fig. 1. (continued)

Taken together, Table 1 and Figs. 1–2 show that the dataset is not only diverse in its empirical distributions but also sufficiently rich in structure to support nonlinear predictive analysis. This empirical characterization motivates the use of polynomial regression models capable of representing nonlinear effects and variable interactions, as described in the following subsection.

2.2. Polynomial regression models and research workflow

To develop an interpretable predictive framework for reinforced concrete floor cost, this study employed polynomial regression with progressively increasing model complexity. Starting from the original five input variables slab type, tributary area, superimposed load, formwork, and concrete four candidate models were constructed using polynomial feature expansion with degrees 1, 2, 3, and 4, respectively. The first-degree model corresponds to a standard linear specification, whereas the higher-degree models allow increasingly flexible nonlinear and interaction-based representations of the relationship between the predictors and floor cost. This modeling strategy makes it possible to examine how predictive performance evolves as the polynomial degree increases while preserving a transparent regression-based structure. The use of polynomial basis expansion is consistent with the general regression framework in which nonlinear relationships can be represented through transformed predictors while the model remains linear in the parameters.

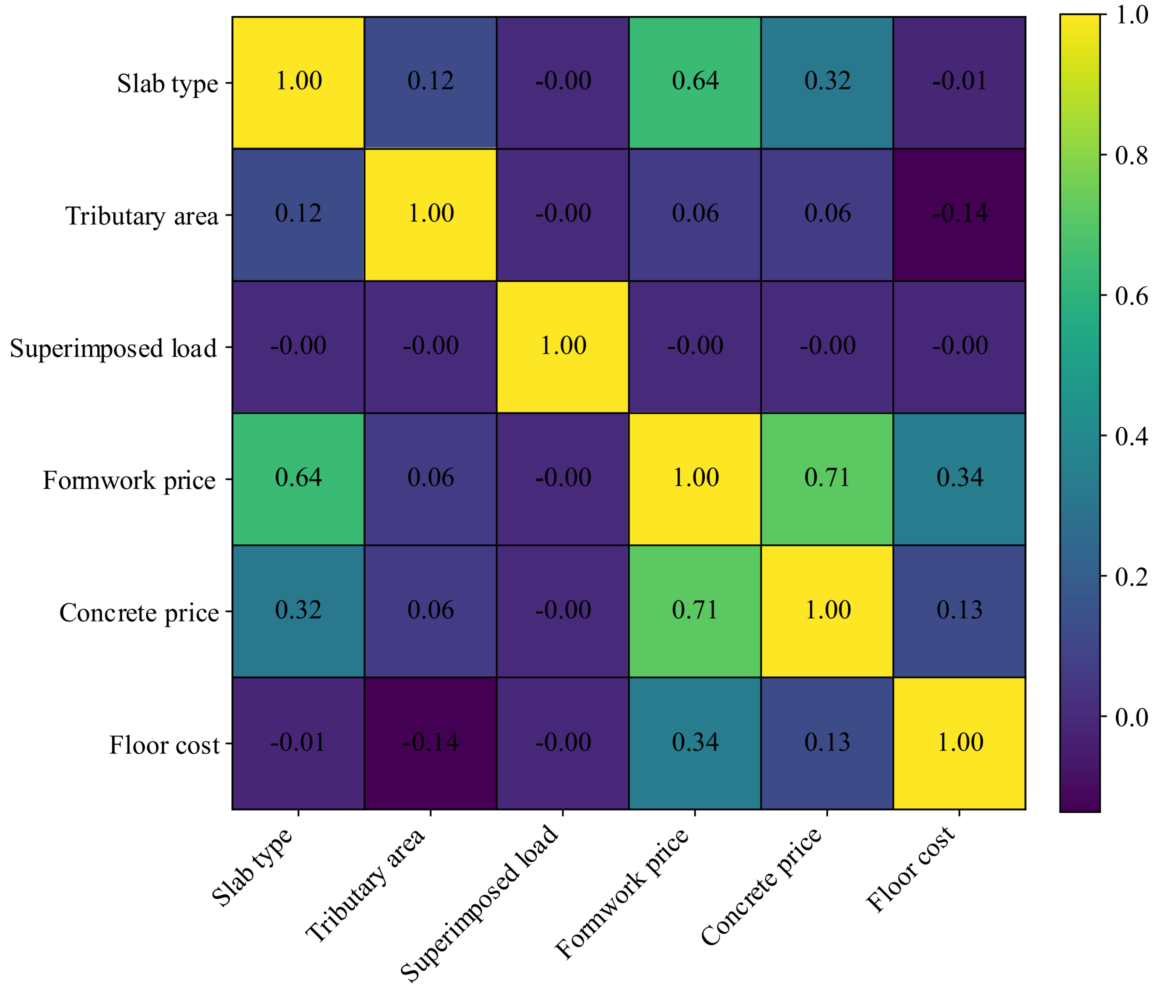


Fig. 2. Pearson correlation heatmap of the study variables used for floor cost prediction

Let the predictor vector be defined as:

$$x = (x_1, x_2, x_3, x_4, x_5) \tag{1}$$

where x_1 denotes slab type, x_2 denotes tributary area, x_3 denotes Superimposed load, x_4 denotes Formwork, and x_5 denotes Concrete. Accordingly, x_1 should be interpreted as a design-type indicator used to distinguish floor-system alternatives within the polynomial regression framework. In the general polynomial regression framework, the predicted floor cost $\hat{y}$ may be written as:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j z_j(x) \tag{2}$$

where $z_j(x)$ denotes the generated polynomial basis terms and β_j denotes the corresponding regression coefficients. In this representation, the model is nonlinear with respect to the original predictors but remains linear with respect to the unknown parameters, which allows estimation by ordinary least squares. For the first-degree model, the specification reduces to the conventional linear form

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{3}$$

For the second-degree model, the predictor space is expanded to include squared terms and pairwise interactions, yielding

$$\hat{y} = \beta_0 + \sum_{i=1}^{5} \beta_i x_i + \sum_{i=1}^{5} \beta_{ii} x_i^2 + \sum_{i<j}^{5} \beta_{ij} x_i x_j \tag{4}$$

For the third-degree model, cubic terms and mixed interaction terms are additionally introduced, so that the model can be expressed more generally as:

$$\begin{aligned} \hat{y} = \beta_0 + \sum_{i} \beta_i x_i + \sum_{i} \beta_{ii} x_i^2 + \sum_{i<j} \beta_{ij} x_i x_j \\ + \sum_{i \neq j} \beta_{iij} x_i^2 x_j + \sum_{i<j<k} \beta_{ijk} x_i x_j x_k \end{aligned} \tag{5}$$

The research workflow is summarized in Fig. 3. First, the RSMeans-derived dataset was prepared and standardized in terms of variable definition and units. Second, descriptive statistics and correlation analysis were carried out to characterize the data structure and justify the retention of all five predictors for subsequent modeling. Third, polynomial features were generated for each candidate degree and fitted using linear regression. Fourth, the predictive stability of the candidate models was evaluated through repeated Monte Carlo simulation. Finally, the best-performing model was selected by jointly considering predictive accuracy, stability, and interpretability, after which the selected model was further analyzed through error analysis, feature-importance assessment, centered partial dependence plots, contour plots, and engineering application through a graphical user interface.

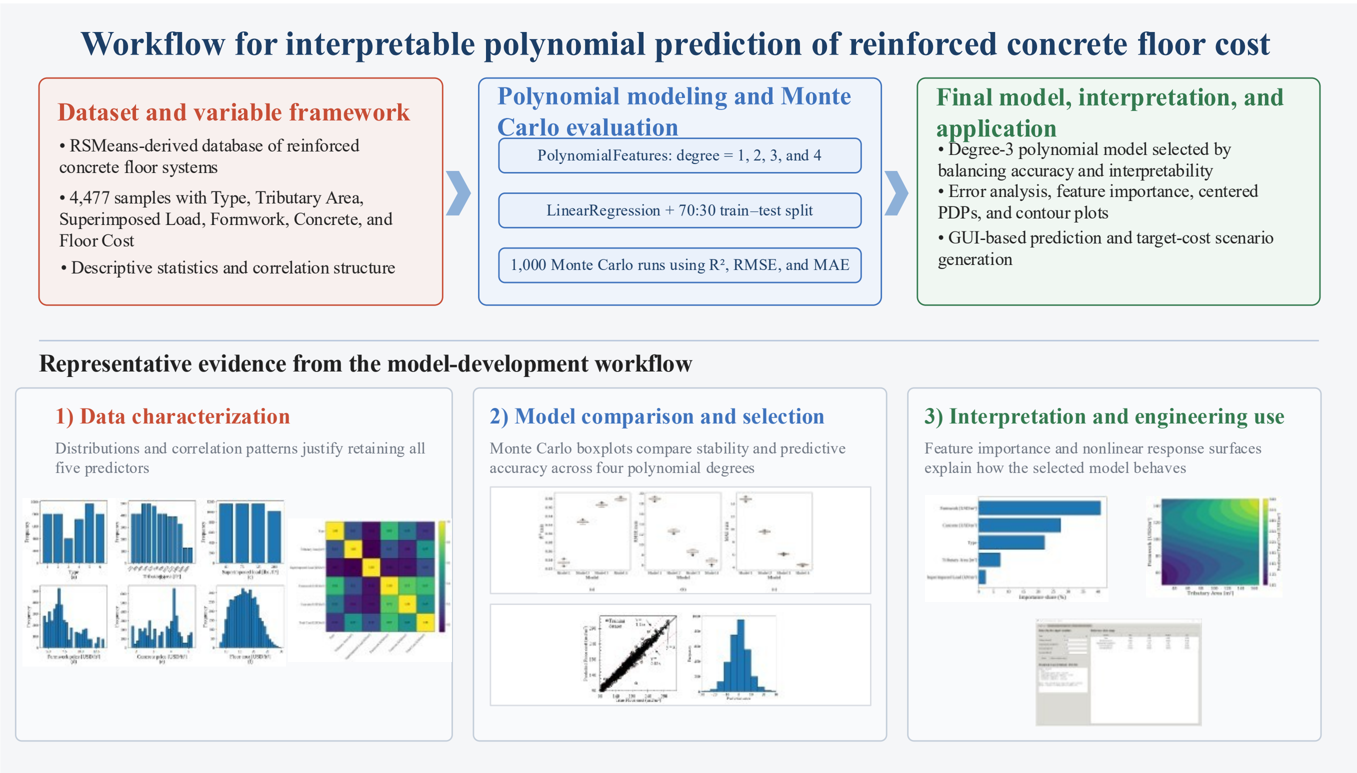


Fig. 3. Research workflow of the study, including dataset preparation, variable definition, polynomial feature generation, model training, Monte Carlo evaluation, model selection, and analytical interpretation

A one-hot encoding scheme could also be adopted for categorical representation; however, it would substantially increase the number of polynomial interaction terms and reduce the compactness of the proposed interpretable framework. Therefore, the present study retains the numerical design-type encoding and interprets its contribution cautiously.

A key advantage of this workflow is that it preserves the interpretability of a regression-based model while still allowing nonlinear response patterns to emerge from the data. Compared with more opaque machine learning algorithms, the polynomial regression framework used here enables direct examination of model structure, generated terms, and engineering meaning. This is particularly important for early-stage design and value-engineering applications, where a predictive model is expected not only to provide accurate estimates but also to remain understandable to engineers and decision-makers.

2.3. Model evaluation, simulation design, and model interpretation tools

Model performance was evaluated using a repeated random holdout strategy. The full dataset was randomly divided into training and testing subsets using a 70:30 split, and the same evaluation framework was applied consistently across all four candidate polynomial models. This training dataset and testing dataset design was selected to assess predictive performance on both fitted data and unseen samples, thereby supporting a more reliable evaluation of model generalization. In the final selected model implementation, the split was performed with test_size = 0.3 and random_state = 967, which ensured reproducibility of the reported prediction results.

To assess model stability beyond a single random split, this study employed Monte Carlo simulation with 1000 repeated runs. In each run, the candidate model was re-evaluated and its predictive performance was recorded on both the training and testing datasets. This repeated random sub-sampling design makes it possible to evaluate not only average model accuracy but also the consistency and robustness of performance under repeated sampling variation. In the present study, cumulative mean Monte Carlo curves and Monte Carlo boxplots were later used to examine convergence, dispersion, and comparative stability across the four polynomial degrees.

Three performance indicators were used throughout the evaluation process: the coefficient of determination (R^2), the root mean square error (RMSE), and the mean absolute error (MAE). The coefficient of determination is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2} \tag{6}$$

where y_i is the observed floor cost, $\hat{y}_i$ is the predicted floor cost, and $\overline{y}_i$ is the mean observed floor cost. The root mean square error is computed as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{7}$$

and the mean absolute error is given by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{8}$$

In general, higher R^2 values and lower RMSE and MAE values indicate better predictive performance. The combined use of these three indicators allows the models to be assessed from complementary perspectives, including goodness-of-fit, sensitivity to large deviations, and average absolute prediction error.

In addition to predictive accuracy, model interpretability was explicitly considered in the analytical framework of this study. To support the interpretation of the selected final model, one-dimensional and two-dimensional partial dependence analysis was employed. Following the classical formulation introduced by Friedman, the partial dependence function for a set of features x_S is defined as the average model response over the marginal distribution of the remaining predictors x_C :

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_S, x_C^{(i)}\right) \tag{9}$$

where $\hat{f}(\cdot)$ is the fitted prediction function, S denotes the subset of features of interest, and C denotes its complement. This formulation provides a way to visualize how the fitted model responds to changes in one variable or a pair of variables while averaging out the influence of the remaining predictors.

For a single predictor x_j, the one-dimensional partial dependence function can be written as

$$\hat{f}_j(x_j) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_j, x_{-j}^{(i)}\right) \quad (10)$$

where $x_{-j}^{(i)}$ represents the observed values of all other predictors for sample i. This one-dimensional partial dependence plot reveals the average marginal shape of the predictor–response relationship. In the present study, centered one-dimensional partial dependence plots were used, so that the vertical axis reflects deviation from the average predicted floor cost rather than the absolute model output. The centered version may be written as:

$$c\hat{f}_j(x_j) = \hat{f}_j(x_j) - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \quad (11)$$

where $\hat{y}_i$ denotes the fitted prediction for sample i. This centering improves interpretability by making positive and negative deviations from the model-average baseline visually explicit.

For a pair of predictors (x_j, x_k), the two-dimensional partial dependence function is defined as:

$$\hat{f}_{jk}(x_j, x_k) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_j, x_k, x_{-(j,k)}^{(i)}\right) \quad (12)$$

where $x_{-(j,k)}^{(i)}$ denotes the remaining predictors for sample i. In practice, this function can be visualized as a response surface or contour map, making it possible to examine whether the combined influence of two predictors is approximately additive or exhibits nonlinear interaction structure. In this study, two-dimensional contour plots were later used to interpret how the selected polynomial model responds jointly to pairs of quantitative variables such as Tributary Area and Formwork, Formwork and Concrete, and Superimposed Load with either Formwork or Concrete.

It should be noted, however, that partial dependence plots are descriptive model-interpretation tools rather than formal causal estimators. Their interpretation is most reliable when the predictor combinations explored by the plots are well supported by the observed data. When predictors are strongly correlated, partial dependence analysis may involve unrealistic combinations of values and should therefore be interpreted cautiously. For this reason, the present study uses PDPs and contour plots as complementary interpretive tools alongside feature-importance analysis, residual analysis, and engineering judgment, rather than as standalone evidence.

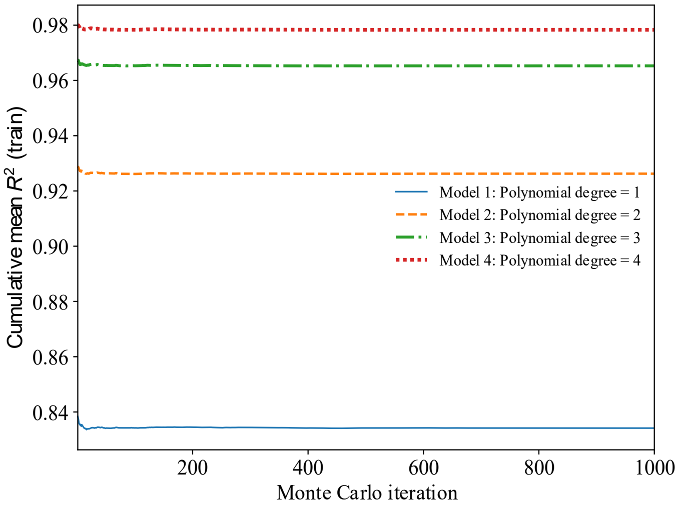
Finally, the comparative evaluation of the four candidate polynomial models was based on two complementary dimensions. First, cumulative mean Monte Carlo curves were used to examine whether the estimated performance of each model stabilized as the number of iterations increased. Second, the full Monte Carlo distributions of R^2, RMSE, and MAE were compared using boxplots on both training and testing datasets. In parallel, model complexity was explicitly taken into account through the number of generated polynomial terms. Consequently, the final model was selected not only on the basis of predictive accuracy but also with regard to structural simplicity, interpretability, and engineering usability. This evaluation principle is central to the present study, because the objective is to develop a predictive framework that is both accurate and practically understandable in early-stage civil-building cost analysis.

3. Results and discussion

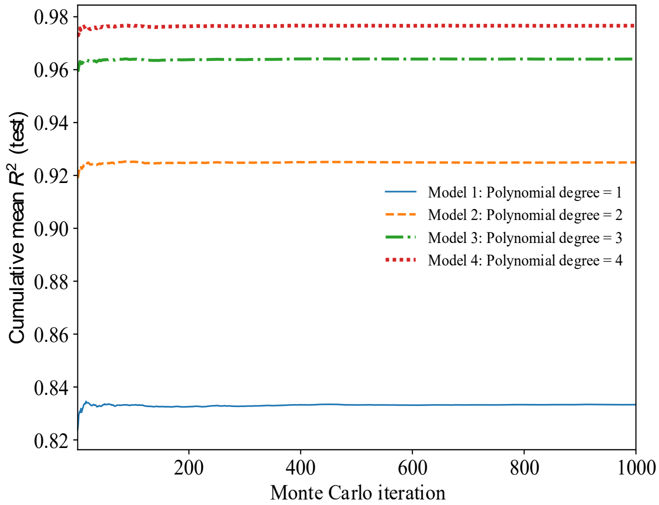
3.1. Comparative performance of polynomial regression models

Based on the evaluation framework described above, the comparative results of the four polynomial regression models are presented in the following section

Fig. 4 shows the convergence of the cumulative mean Monte Carlo R^2 values for the four polynomial regression models on the training and testing datasets. In both cases, the curves stabilize progressively as the number of Monte Carlo iterations approaches 1000, indicating that the simulation size is sufficient to provide stable estimates of model performance. More importantly, the relative ranking of the models remains consistent throughout the simulation process: predictive performance improves monotonically as the polynomial degree increases from 1 to 4. This pattern is observed on both the training and testing datasets, suggesting that higher-order polynomial expansion captures additional nonlinear structure in the data rather than merely fitting noise.



(a)



(b)

Fig. 4. Convergence of the cumulative mean Monte Carlo R^2 values for the four polynomial regression models: (a) training dataset and (b) testing dataset

Table 2. Comparison of the mean Monte Carlo performance metrics of the four polynomial regression models on the training and testing datasets

Model	Mean value of training dataset			Mean value of testing dataset		
	R^2	RMSE (USD /m^2)	MAE (USD /m^2)	R^2	RMSE (USD /m^2)	MAE (USD /m^2)
Polynomial degree = 1	0.834	18.940	14.706	0.833	18.992	14.746
Polynomial degree = 2	0.926	12.630	9.588	0.925	12.743	9.676
Polynomial degree = 3	0.965	8.662	6.102	0.964	8.810	6.218
Polynomial degree = 4	0.978	6.842	4.425	0.977	7.087	4.602
Chakraborty et al. [4] hybrid LGBoost–NGBoost baseline	NR	NR	NR	0.990	5.382	NR

Note: NR = not reported.

The average Monte Carlo results are summarized in Table 2. Model 1, corresponding to first-degree polynomial regression, provides the weakest performance, with mean R^2 values of 0.834 and 0.833 for the training and testing datasets, respectively, together with the largest RMSE and MAE values. Model 2 substantially improves predictive accuracy, reaching mean R^2 values of 0.926 for training and 0.925 for testing. Model 3 yields a further marked improvement, achieving mean R^2 values of 0.965 and 0.964, while reducing RMSE to 8.662 USD/m² for training and 8.810 USD/m² for testing. Model 4 achieves the strongest raw predictive accuracy overall, with mean R^2 values of 0.978 and 0.977 and the lowest RMSE and MAE among all candidate models.

The Chakraborty et al. [4] baseline is included for reference based on the original study. Their final hybrid LGBoost–NGBoost model reported testing (R^2 = 0.99), RMSE = 0.5 USD/ft², and MBE = −0.009 USD/ft². The RMSE value was converted to 5.382 USD/m² for consistency with the unit system used in the present study. MAE was not reported in the original study. The benchmark reported by Chakraborty et al. [4] achieves higher raw predictive accuracy than the proposed cubic polynomial model, with a testing (R^2) of 0.990 and an RMSE of 5.382 USD/m² after unit conversion. However, this baseline is reported for reference only because it was not re-evaluated under the same repeated Monte Carlo simulation protocol used in the present study. Moreover, the hybrid LGBoost–NGBoost model is an ensemble boosting-based model with a more complex internal structure, whereas the proposed cubic polynomial model retains an explicit analytical form and allows direct examination of nonlinear and interaction terms. Therefore, the objective of this study is not to outperform the hybrid boosting baseline in raw accuracy, but to provide a transparent and interpretable alternative with competitive predictive performance for early-stage engineering decision support.

However, predictive accuracy alone is not sufficient to justify the choice of the final model. Table 3 shows that model complexity increases sharply with polynomial degree. The number of generated terms rises from 5 in Model 1 to 20 in Model 2, 55 in Model 3, and 125 in Model 4. This growth is substantial. While the quartic model produces the best numerical performance, it also introduces a much larger number of polynomial terms, making direct interpretation more difficult and reducing analytical transparency. In contrast, Model 3 retains a considerably simpler structure while still achieving strong predictive performance.

Table 3. Comparison of polynomial model complexity in terms of polynomial degree and number of generated terms

Model	Polynomial degree	Linear terms	Quadratic terms	Cubic terms	Quartic terms	Total number of terms
Model 1	1	5	0	0	0	5
Model 2	2	5	15	0	0	20
Model 3	3	5	15	35	0	55
Model 4	4	5	15	35	70	125

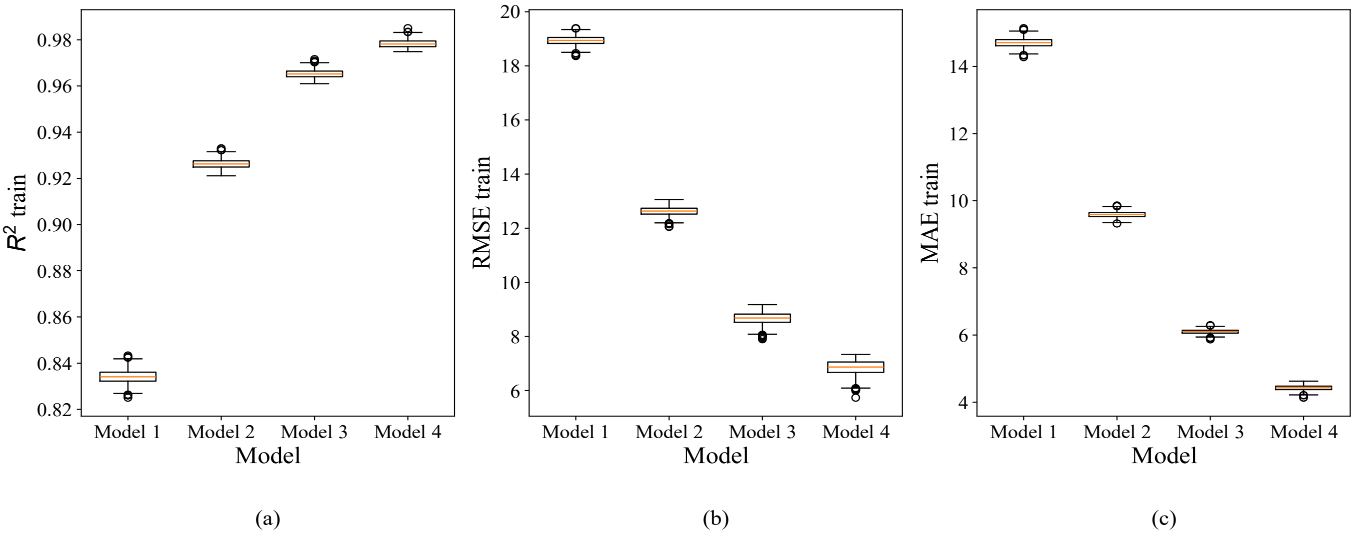


Fig. 5. Boxplot comparison of Monte Carlo performance on the training dataset for the four polynomial regression models: (a) R^2, (b) RMSE, and (c) MAE.

Figs. 5 and 6 reinforce this interpretation by presenting the full Monte Carlo distributions of R2, RMSE, and MAE for the training and testing datasets. The boxplots show that higher-degree models generally achieve both better central performance and tighter distributions. Nevertheless, the improvement from degree 3 to degree 4 is visibly smaller than the improvement from lower-order models, whereas the increase in structural complexity is substantial. For this reason, the third-degree polynomial regression model was selected as the final model, as it provides the most appropriate balance between predictive performance and interpretability.

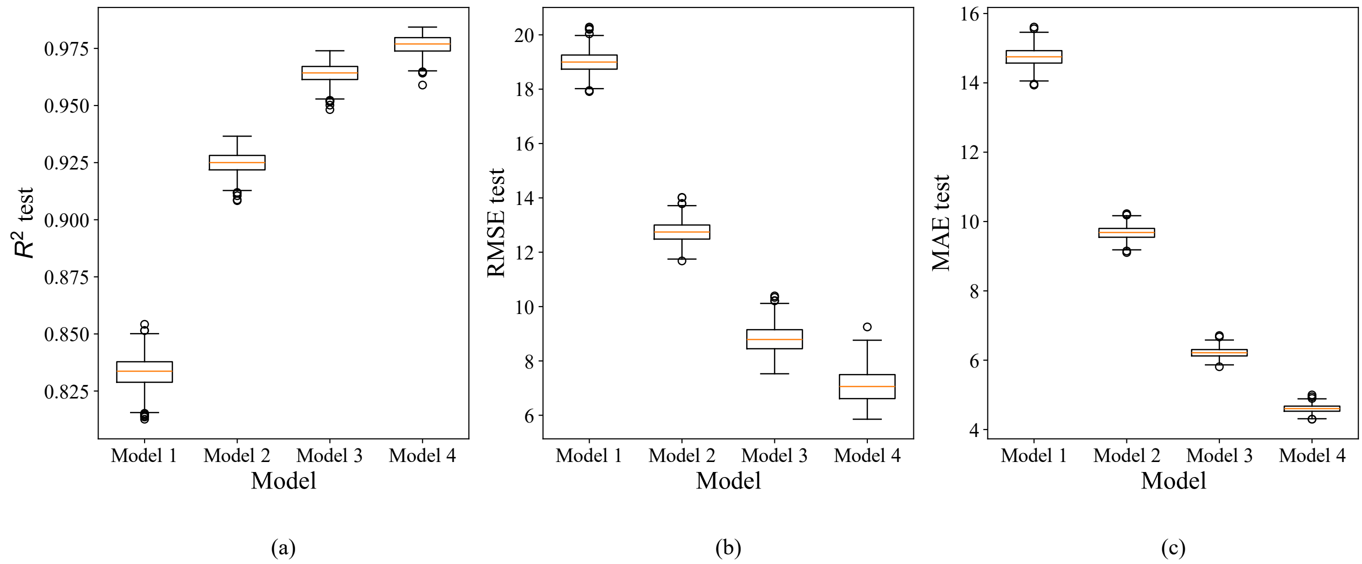


Fig. 6. Boxplot comparison of Monte Carlo performance on the testing dataset for the four polynomial regression models: (a) R^2, (b) RMSE, and (c) MAE.

Although the selected cubic model contains 55 generated polynomial terms, its interpretability should not be understood as direct manual reading of every individual coefficient. Rather, interpretability in this study refers to the explicit polynomial model structure and to response-pattern interpretation through feature importance, centered partial dependence plots, and contour-based interaction analysis.

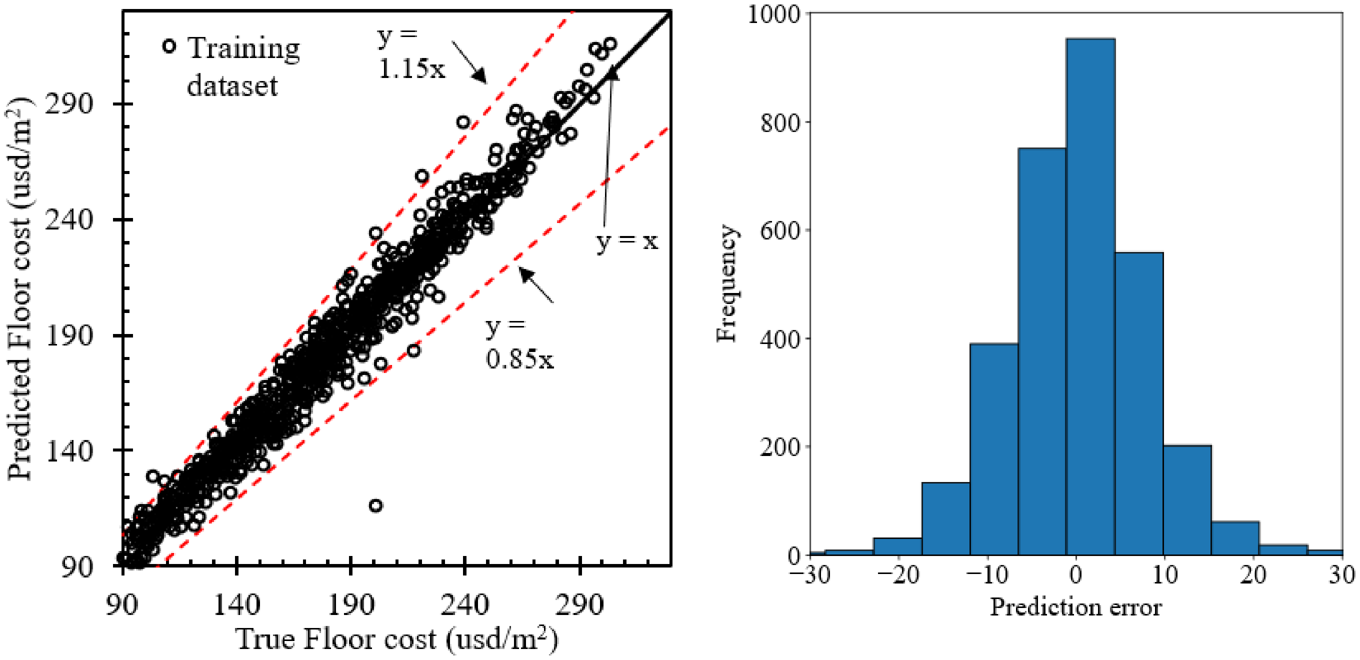
3.2. Performance of the selected model

To improve the interpretability of the selected cubic polynomial model at the variable and response-pattern levels, rather than relying solely on coefficient-level inspection, additional analyses were performed using feature importance, centered partial dependence plots, and contour-based interaction visualizations.

Based on the comparative assessment above, the third-degree polynomial regression model was retained as the final predictive model. It should be noted that the performance values reported in Table 2 are Monte Carlo mean values computed over 1000 repeated random partitions. By contrast, the results reported in this subsection correspond to one representative and reproducible 70:30 split into training and testing datasets of the selected cubic model, implemented with a fixed random state of 967. Therefore, the small numerical differences between Table 2 and Section 3.2 reflect the difference between simulation-averaged performance and single-split representative performance.

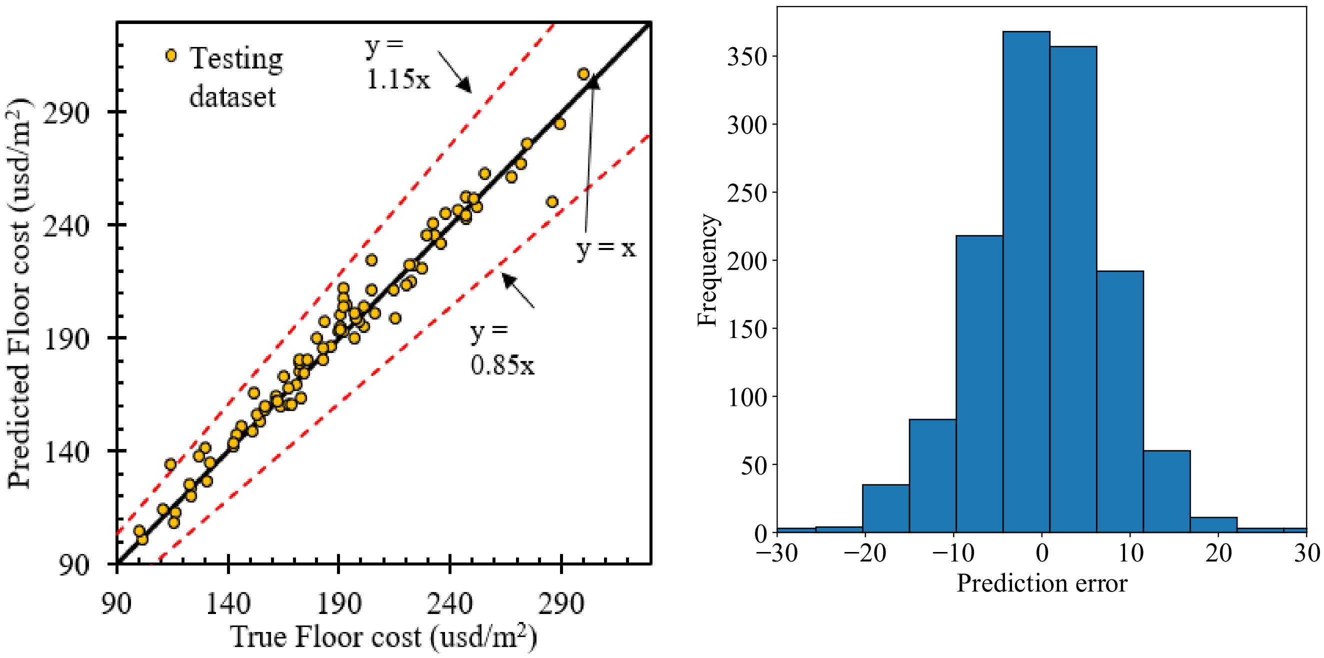
This model combines strong predictive accuracy with a still-manageable level of structural complexity, making it particularly suitable for engineering interpretation and practical use. According to the manuscript results, the selected model achieved R^2=0.965, RMSE = 8.764 USD/m², and MAE = 6.163 USD/m² on the training dataset, while the corresponding testing results were R^2=0.965, RMSE = 8.567 USD/m², and MAE = 6.058 USD/m². The close agreement between training and testing performance indicates that the selected model generalizes well and does not exhibit strong signs of overfitting.

The slightly lower RMSE on the testing dataset should be interpreted in light of the deterministic and assembly-based nature of the RSMeans-derived data. Unlike noisy project-level cost records collected from heterogeneous construction sites, the present dataset is generated from standardized structural assembly cost information, where floor cost is governed by repeated and regular combinations of design type, tributary area, loading level, and unit cost variables. Therefore, a random testing subset may exhibit slightly lower prediction error than the training subset without indicating data leakage. This interpretation is also supported by the repeated Monte Carlo evaluation, which shows stable and consistent model performance across random partitions.



(a) (b)

Fig. 7. Predictive performance of the selected third-degree polynomial regression model on the training dataset: (a) observed versus predicted floor cost and (b) histogram of prediction errors



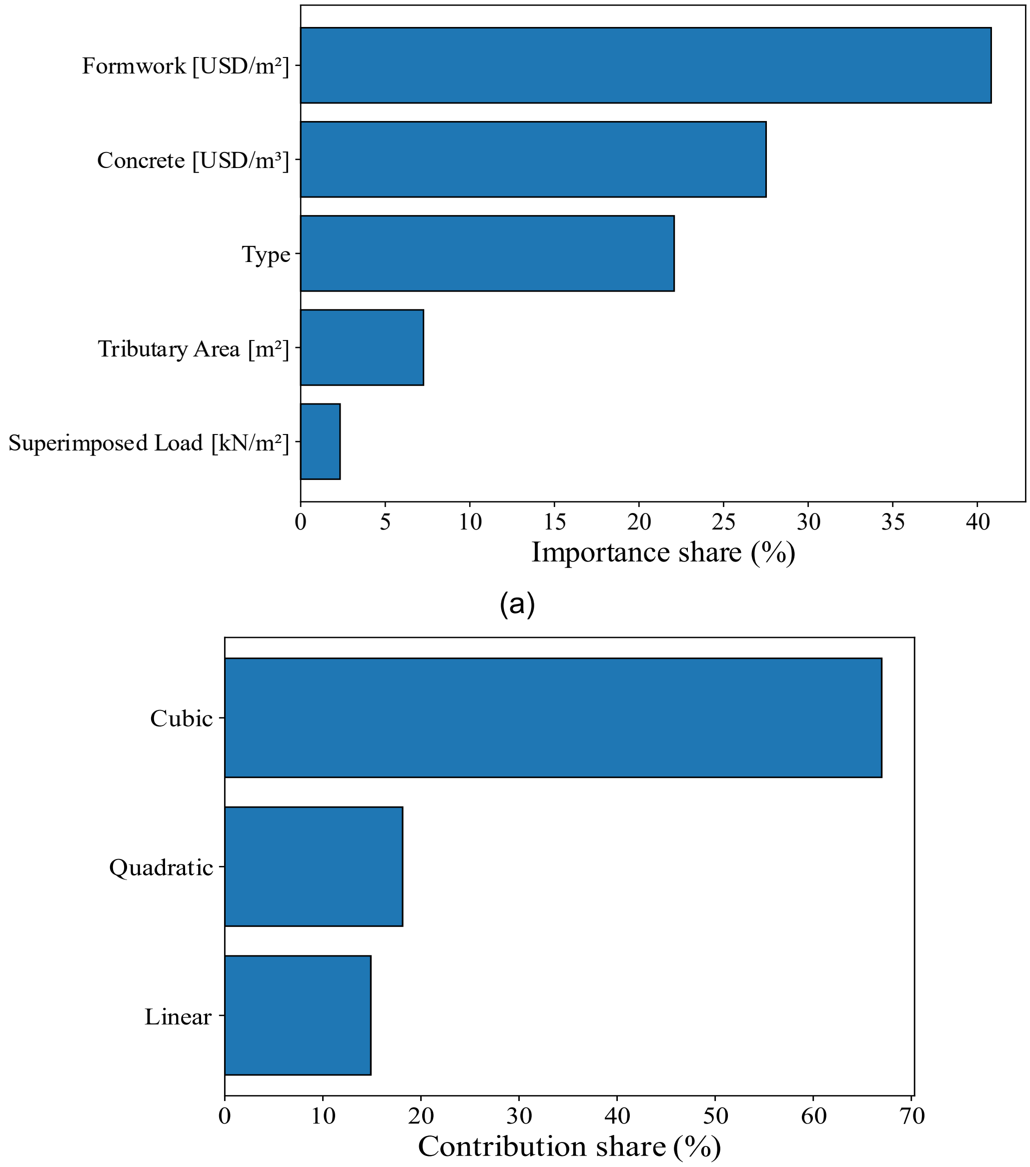
(a) (b)

Fig. 8. Predictive performance of the selected third-degree polynomial regression model on the testing dataset: (a) observed versus predicted floor cost and (b) histogram of prediction errors

Fig. 7 presents the predictive performance of the selected model on the training dataset. The observed-versus-predicted scatter plot shows that most data points lie close to the $y=x$ reference line, with limited dispersion around the tolerance lines y=0.85x and y=1.15x. This indicates that the model captures the overall response pattern of the training data with good fidelity. The accompanying histogram of prediction errors is centered near zero and displays a relatively compact distribution, which suggests that the fitted model is approximately unbiased and that large residuals are not dominant.

A similar pattern is observed in Fig. 8 for the testing dataset. The predicted values remain closely aligned with the observed values, and the error histogram again shows a concentration of residuals around zero, with only a small proportion of larger deviations. Taken together, these results indicate that the selected cubic model is capable of capturing the dominant nonlinear cost structure of reinforced concrete floor systems while preserving stable performance on previously unseen data.

3.3. Analytical interpretation of the final model



(b)

Fig. 9. Feature-importance analysis of the selected third-degree polynomial regression model: (a) variable importance based on mean absolute contribution and (b) contribution shares of the linear, quadratic, and cubic components

To improve the interpretability of the selected cubic polynomial model, additional analyses were performed at the variable, component, and response-pattern levels. Fig. 9 shows the feature-importance analysis based on the mean absolute contribution of each predictor. The results indicate that formwork price is the most influential variable, followed by concrete price and slab type, whereas tributary area and superimposed load contribute less strongly to the overall predictive structure. This ranking is consistent with the engineering nature of the problem, in which direct cost drivers such as formwork and concrete are expected to exert the strongest influence on floor cost.

To examine the average marginal influence of the quantitative predictors, centered one-dimensional partial dependence plots were used, as shown in Fig. 10. These plots indicate that the relationship between each predictor and floor cost is not purely linear. Formwork and concrete exhibit the strongest positive influence on the predicted response, although their effects are not strictly monotonic across the entire observed range. Tributary area displays a weaker but still non-negligible nonlinear effect, while the marginal influence of superimposed load is comparatively small. Because the plots are centered relative to the model-average predicted floor cost, they allow positive and negative deviations from the average response level to be interpreted directly, thereby improving the readability of the predictor–response relationships.

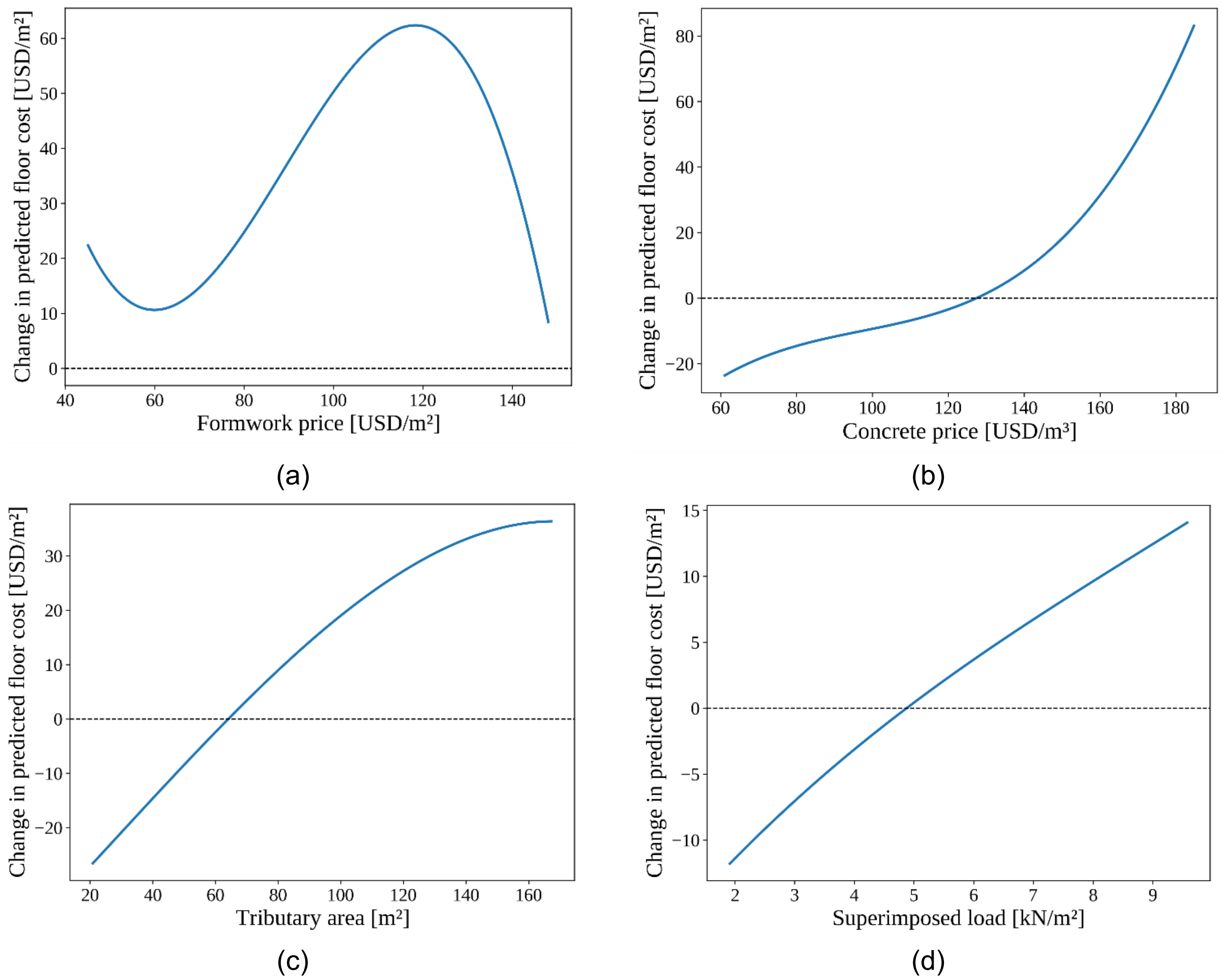


Fig. 10. Centered one-dimensional partial dependence plots of the quantitative variables for the selected third-degree polynomial regression model: (a) formwork price, (b) concrete price, (c) tributary area, and (d) superimposed load

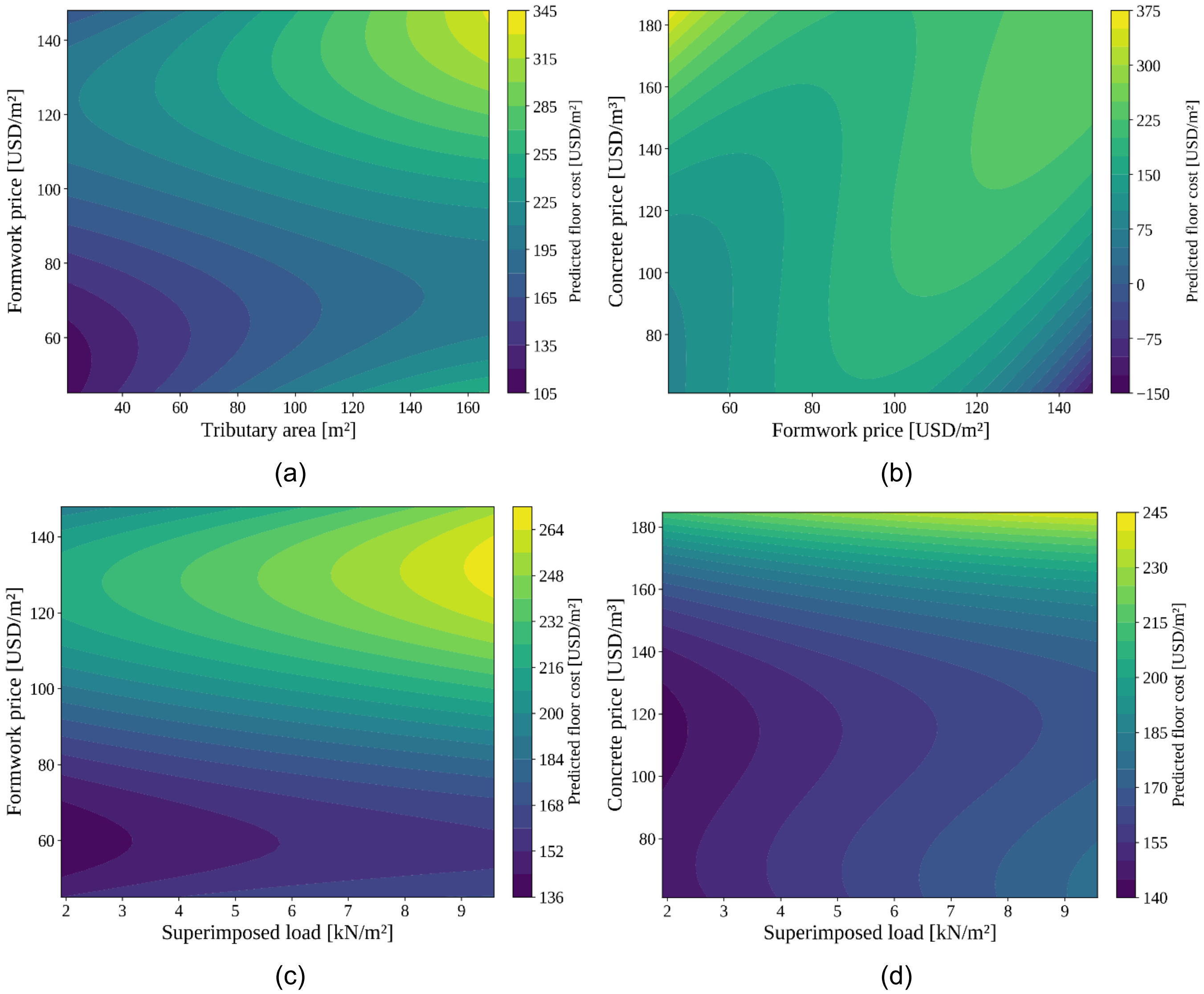


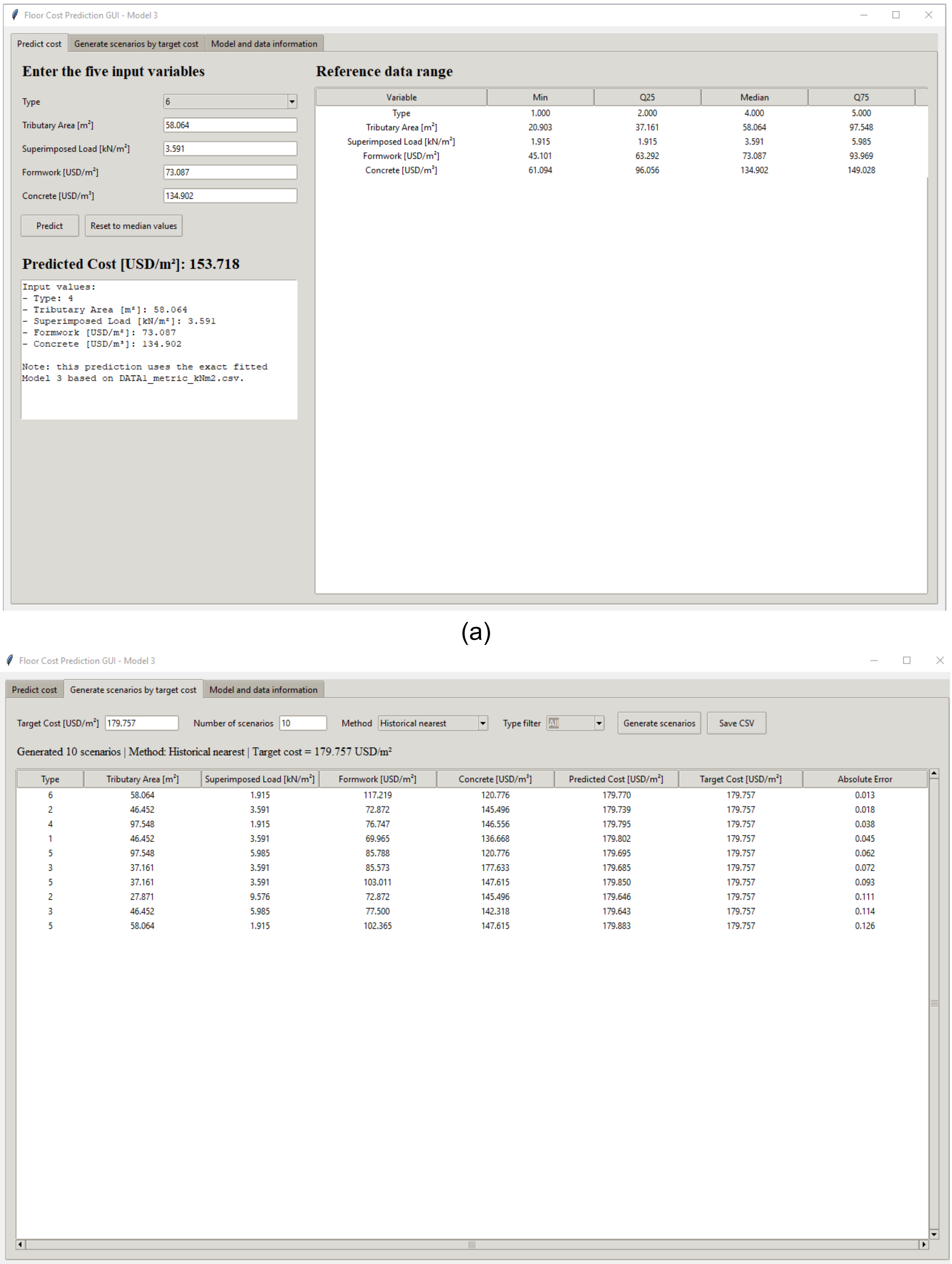
Fig. 11. Contour plots showing the pairwise nonlinear effects of key quantitative variables on predicted floor cost, with the remaining variables fixed at their median values: (a) tributary area × formwork price, (b) formwork price × concrete price, (c) superimposed load × formwork price, and (d) superimposed load × concrete price

The interaction structure of the selected model is further illustrated through the contour plots in Fig. 11. These plots show that the combined effects of tributary area and formwork, formwork and concrete, superimposed load and formwork, and superimposed load and concrete are nonlinear and vary across the predictor space. In particular, the contour patterns indicate that the effect of one variable depends on the level of another, which is consistent with the inclusion of interaction and higher-order polynomial terms in the selected model. From an engineering perspective, these surfaces help explain why a cubic polynomial model is needed: the behavior of floor cost cannot be adequately represented by a purely linear or weakly nonlinear formulation. Instead, the fitted response is shaped by multiple interaction mechanisms among geometric, loading, and cost-related variables.

Taken together, the feature-importance analysis, centered partial dependence plots, and contour surfaces demonstrate that the selected third-degree polynomial model is not only predictive but also analytically interpretable. This is a major advantage over more opaque machine learning methods, as it allows the model outputs to be linked back to meaningful engineering patterns rather than being treated as isolated numerical predictions.

For transparency and reproducibility, the complete fitted equation and coefficient table of the selected cubic polynomial model are provided in the Supplementary Material.

3.4. Application engineering



(a)

(b)

Fig. 12. Engineering application of the selected third-degree polynomial regression model through the developed graphical user interface: (a) direct floor cost prediction from five input variables and (b) target-cost scenario generation

To demonstrate the practical applicability of the selected third-degree polynomial regression model, the final predictive framework was implemented in a graphical user interface (GUI), as shown in Fig. 12. The interface can be downloaded free of charge from the following link: https://github.com/dachoangdao-UTT/floor-cost-prediction. The GUI supports two complementary functions: direct floor cost prediction from user-defined inputs and target-cost scenario generation for preliminary design exploration.

Fig. 12(a) illustrates the direct prediction mode. In this mode, the user enters the five input variables retained in the final model slab type, tributary area, superimposed load, formwork price, and concrete price and the system immediately returns the predicted floor cost. This function is useful when a designer or estimator already has a tentative floor configuration and needs a rapid cost estimate at the conceptual stage. For example, when a project team is comparing alternative slab systems under fixed loading and price conditions, the interface allows the user to modify the slab type or geometric parameters and observe the resulting cost change instantly. This makes the model operational as a fast-screening tool during early-stage value-engineering analysis.

Fig. 12(b) shows the target-cost scenario mode. In this mode, the user specifies a desired floor cost, and the interface generates combinations of the five input variables that are consistent with that target under the fitted cubic model. This is particularly useful when the practical question is not only “What is the cost of this design?” but also “What design conditions are likely to satisfy a target budget?” For example, if a project team seeks to keep the floor cost near a predefined budget threshold, the interface can suggest feasible combinations of slab type, tributary area, loading condition, and material prices that move the prediction toward that target. This extends the model from a prediction tool to a decision-support tool for preliminary engineering planning.

From an engineering management perspective, the main value of the GUI lies in its combination of speed, transparency, and usability. The selected third-degree polynomial model is sufficiently accurate for early-stage screening, yet still interpretable enough to remain consistent with the analytical results presented earlier, including feature importance, partial dependence plots, and contour-based interaction analysis. Therefore, the developed interface provides a practical bridge between predictive modeling and real engineering application. It should, however, be viewed as a support tool for preliminary decision-making rather than a substitute for detailed final-stage cost estimation.

4. Conclusion

This study developed an interpretable polynomial regression framework for early-stage prediction of reinforced concrete floor cost in civil buildings using a dataset of 4477 RSMeans-based samples. Four candidate polynomial models with degrees 1 to 4 were compared under a repeated Monte Carlo evaluation design with 1000 simulations. Although the quartic model achieved the highest raw predictive accuracy, the third-degree polynomial model was selected as the final model because it provided the most suitable balance between predictive performance and interpretability.

The selected cubic model showed strong and stable predictive performance on both the training and testing datasets. In addition, the analytical interpretation results demonstrated that formwork and concrete are the dominant cost drivers, while the centered partial dependence plots and contour plots confirmed that the response surface is nonlinear and interaction-driven. These findings indicate that an interpretable polynomial structure can capture the essential cost behavior of reinforced concrete floor systems without relying on a black-box model.

From a practical perspective, the proposed model can support rapid screening of alternative reinforced concrete floor systems during preliminary design and value-engineering analysis. The developed graphical user interface further demonstrates how the model can be translated into a usable engineering decision-support tool for direct cost prediction and target-cost scenario exploration. Future research should extend the framework through external validation using independent project datasets, comparison with alternative categorical encoding strategies, incorporation of additional design and market-related variables, and integration with digital design environments such as BIM-based cost assessment platforms.

Supplementary Material

The full list of the 55 fitted polynomial terms and their coefficients is presented in File excel.

Declaration of Generative AI

During the preparation of this work, the authors used CHATGPT in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

References

[1] M. Bilal, L.O. Oyedele, J. Qadir, K. Munir, S.O. Ajayi, O.O. Akinade, H. A. Owolabi, H.A. Alaka, M. Pasha. (2016). Big data in the construction industry: a review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, 30(3), 500–521. https://doi.org/10.1016/j.aei.2016.07.001

[2] C. Zhou, L.Y. Ding, M.J. Skibniewski, H. Luo, H.T. Zhang. (2018). Data based complex network modeling and analysis of shield tunneling performance in metro construction. *Advanced Engineering Informatics*, 38, 168–186. https://doi.org/10.1016/j.aei.2018.06.011

[3] M. Bilal, L.O. Oyedele. (2020). Guidelines for applied machine learning in construction industry—a case of profit margins estimation. *Advanced Engineering Informatics*, 43, 101013. https://doi.org/10.1016/j.aei.2019.101013

[4] D. Chakraborty, H. Elhegazy, H. Elzarka, L. Gutierrez. (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

[5] B.J. Gardner, D.D. Gransberg, J.A. Rueda. (2016). Stochastic conceptual cost estimating of highway projects to communicate uncertainty using bootstrap sampling. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 3(3), 05016002. https://doi.org/10.1061/AJRUA6.0000895

[6] M. Ghosh, G. Kabir, M.A.A. Hasin. (2017). Project time–cost trade-off: a Bayesian approach to update project time and cost estimates. *International Journal of Management Science and Engineering Management*, 12(3), 206–215. https://doi.org/10.1080/17509653.2016.1255857

[7] H.H. Elmousalami. (2019). Artificial intelligence and parametric construction cost estimate modeling: state-of-the-art review. *Journal of Construction Engineering and Management*, 146(1), 03119008. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001678

[8] O. Dursun, C. Stoy. (2016). Conceptual estimation of construction costs using the multistep ahead approach. *Journal of Construction Engineering and Management*, 142(9), 04016038. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001150

[9] O.S. Alshamrani. (2017). Construction cost prediction model for conventional and sustainable college buildings in North America. *Journal of Taibah University for Science*, 11(2), 315–323. https://doi.org/10.1016/j.jtusci.2016.01.004

[10] M.-Y. Cheng, H.-C. Tsai, E. Sudjono. (2010). Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in

construction industry. *Expert Systems with Applications*, 37(6), 4224–4231. https://doi.org/10.1016/j.eswa.2009.11.080

[11] X. Wang. (2017). Application of fuzzy math in cost estimation of construction project. *Journal of Discrete Mathematical Sciences and Cryptography*, 20(4), 805–816. https://doi.org/10.1080/09720529.2017.1358864

[12] H. Son, C. Kim, C. Kim. (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. https://doi.org/10.1016/j.autcon.2012.05.013

[13] N. El-Sawalhi, O. Shehatto. (2014). A neural network model for building construction projects cost estimating. *KICEM Journal of Construction Engineering and Project Management*, 4, 9. https://doi.org/10.6106/JCEPM.2014.4.4.009

[14] K. Zima. (2015). The case-based reasoning model of cost estimation at the preliminary stage of a construction project. *Procedia Engineering*, 122, 57–64. https://doi.org/10.1016/j.proeng.2015.10.007

[15] O.S. Alshamrani. (2020). Initial cost forecasting model of mid-rise green office buildings. *Journal of Asian Architecture and Building Engineering*, 19(6), 613–625. https://doi.org/10.1080/13467581.2020.1778005

[16] M. Juszczyk. (2018). Residential buildings conceptual cost estimates with the use of support vector regression. *MATEC Web of Conferences*, 196, 04090. https://doi.org/10.1051/matecconf/201819604090

[17] M. Juszczyk, A. Leśniak, K. Zima. (2018). ANN based approach for estimation of construction costs of sports fields. *Complexity*, 2018, 7952434. https://doi.org/10.1155/2018/7952434

[18] V. Arabzadeh, S.T.A. Niaki, V. Arabzadeh. (2018). Construction cost estimation of spherical storage tanks: artificial neural networks and hybrid regression-GA algorithms. *Journal of Industrial Engineering International*, 14, 747–756. https://doi.org/10.1007/s40092-017-0240-8

[19] E. Matel, F. Vahdatikhaki, S. Hosseinyalamdary, T. Evers, H. Voordijk. (2022). An artificial neural network approach for cost estimation of engineering services. *International Journal of Construction Management*, 22(7), 1274–1287. https://doi.org/10.1080/15623599.2019.1692400

[20] S.T. Hashemi, O.M. Ebadati E., H. Kaur. (2019). A hybrid conceptual cost estimating model using ANN and GA for power plant projects. *Neural Computing and Applications,* 31, 2143–2154. https://doi.org/10.1007/s00521-017-3175-5

[21] S.M. Lundberg, S.-I. Lee. (2017). A unified approach to interpreting model predictions. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768-4777. arXiv:1705.07874.* https://doi.org/10.48550/arXiv.1705.07874

[22] C. Molnar. (2022). Interpretable machine learning: a guide for making black box models explainable. https://doi.org/10.1145/3511299

[23] J.H. Friedman. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. DOI: 10.1214/aos/1013203451

[24] L. Breiman. (2001). Random forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[25] C. Meng, D. Qu, X. Duan. (2024). Cost estimation of metro construction projects using interpretable machine learning. *Journal of Computing in Civil Engineering*, 38(6), 04024038. https://doi.org/10.1061/JCCEE5.CPENG-6018

[26] L. Chen, C. Xu, W.H. Lim, A. Sharma, S.S.

Tiang, K.S. Chong, E.-S. M. El-kenawy, A.A. Alhussan, M.M. Eid, D.S. Khafaga. (2025). Transparent and reliable construction cost prediction using advanced machine learning and explainable AI. *Engineering Science and Technology, an International Journal*, 70, 102159. https://doi.org/10.1016/j.jestch.2025.102159