Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 7, No. 3, July 2025, pp: 624-645 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Predicting Construction Costs with Machine Learning: A Comparative Study on Ensemble and Linear Models

Lifei Chen[1], Sew Sun Tiang[1], Kim Soon Chong[1], Abhishek Sharma[2], Tarek Berghout[3], Wei Hong Lim[1, *]

[1] Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur, 56000, Malaysia
[2] Department of Computer Science and Engineering, Graphic Era (Deemed to be) University, Dehradun 248002, India
[3] Laboratory of Automation and Manufacturing Engineering, University of Batna 2, Batna 05000, Algeria
Corresponding author: Wei Hong Lim. (e-mail: limwh@ucsiuniversity.edu.my), **Email Author(s)**: Lifei Chen (e-mail: 1002372862@ucsiuniversity.edu.my), Sew Sun Tiang (email: tiangss@ucsiuniversity.edu.my), Kim Soon Chong (email: ChongKS@ucsiuniversity.edu.my), Abhishek Sharma (email: abhishek15491@gmail.com),Tarek Berghout (email: t.berghout@univ-batna2.dz).

Abstract **Accurate prediction of construction costs plays a pivotal role in ensuring successful project delivery, influencing budget formulation, resource allocation, and financial risk management. However, traditional estimation methods often struggle to handle complex, nonlinear relationships inherent in construction datasets. This study proposes a process innovation by systematically evaluating six machine learning (ML) models, i.e., Ridge Regression, Lasso Regression, Elastic Net, K-Nearest Neighbors (KNN), XGBoost, and CatBoost, on a standardized RSMeans dataset comprising 4,477 real-world construction data points. The primary aim is to benchmark the predictive performance, generalizability, and stability of both linear and ensemble models in construction cost forecasting. Each model is subjected to rigorous hyperparameter tuning using grid search with 5-fold cross-validation. Performance is assessed using R^2 (coefficient of determination), RMSE (root mean squared error), and MBE (mean bias error), while confidence intervals are computed to quantify predictive uncertainty. Results indicate that linear models achieve modest accuracy ($R^2 \approx 0.83$), but struggle to model nonlinear interactions. In contrast, ensemble-based models significantly outperform , i.e., XGBoost and CatBoost achieve R^2 values of 0.988 and 0.987, respectively, RMSE values below 0.5, and near-zero MBE. Moreover, confidence interval visualization and feature importance analysis provide transparency and interpretability, enhancing the models practical applicability. Unlike prior studies that compare models in isolation, this work introduces a unified, interpretable framework and highlights the trade-offs between accuracy, overfitting, and deployment readiness. The findings have real-world implications for contractors, project managers, and cost engineers seeking reliable, data-driven decision support systems. In summary, this study present a scalable and robust ML-based framework that facilitate process innovation in construction cost estimation, paving the way for more intelligent, efficient, and risk-aware construction project management.**

Keywords **Construction Cost; Prediction; Machine Learning; Regression.**

I. Introduction

The success of a construction project highly depends on precise cost estimation. It not only helps effectively control the project budget but also provides reliable data to support decision-making [1]. The construction of the main stadium for the 2008 Beijing Olympic Games, known as Bird's Nest, exemplifies the critical role of precise cost estimation. The leading design teams from China and abroad collaborated to optimize the design, where the project costs are subdivided into specific categories such as steel structure, façade, and interior decoration. Each sub-project was allocated with detailed cost budgets, and throughout the project's execution, the project management software was used to monitor and analyze cost in real time to ensure they remained under control [2]. In contrast, the 2012 London Olympics serveas a cautionary example of the consequences of having poor cost estimation. Despite the extensive experience of London in hosting the Olympics, the scale and complexity of the 2012

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

Games, particularly in venue construction, encountered significant challenges. The project inaccurately estimated the costs required to preserve and renovate the historic buildings, resulting in repeated revisions of the construction plans [3]. These two extreme examples show that the accurate forecasting of construction costs is crucial to ensure project success.

Despite the importance of accurately estimating construction costs in project management, it is often challenging to attain high precision due to the wide range of complex and interrelated factors influencing the overall cost. Uncertainties in external factors, such as economic fluctuations, natural disasters, changes in government policies, and social aspects, can directly impact construction costs [4]. In addition, the inherent complexity of construction projects, including complex construction methods, frequent design changes , and concealed works, can further increase the work volume and cost [5]. Inexperienced estimators who lack a deep understanding of the cost components and influencing factors, may also make mistakes due to subjective judgements. Hence, accurately determining construction cost is a challenging task that demands the careful consideration of various factors [6]. Improving cost prediction accuracy is essential for successful project delivery, and it can be achieved through the advancement of prediction techniques and the strengthening of management studies.

Some of the widely used traditional construction cost prediction approaches include the parametric method, the analogical method, the bill of quantities method, expert judgment method, and the empirical estimation method. The parametric method [7] typically involves developing a mathematical model based on a statistical relationship between historical data and costs, which is then utilized to predict new project costs based on the current parameters. This simple approach is and practical for predicting construction costs for similar projects. Analogical method [8] compares the current project with those of similar historical projects, then adjusts the cost predictions based on their differences. This method can achieve relatively accurate cost estimation by considering specific project characteristics. The Bill of Quantities method [9] requires the preparation of a detailed bill based on design drawings, calculation of the cost of each component using market prices, and summarization of the total project cost. Expert judgment method [10] estimates the project costs by gathering experts' insights from related fields, enabling the comprehensive consideration of uncertainties and yielding reliable predictions. Empirical estimation method [11] is suitable for rapid assessment because it can provide quick and simple estimation based on the experience of engineers.

Despite the popularity of these conventional methods in the past, they tend to encounter several limitations with the increasing complexity in the construction industry. First, these methods are heavily reliant on historical data, and they tend to suffer from significant reductions in prediction accuracy if the historical data is lacking or the project types differ substantially. Second, these methods are exposed to the risk of human error as they require extensive human intervention for data processing and model development. Moreover, these traditional prediction methods are less adaptable to the changes of the evolving construction industry, with the continuous emergence of new materials and techniques. Finally, these conventional methods tend to struggle in accurately capturing the actual costs for large and complex projects, leading to their questionable reliability in the forecasts.

The rise of Industry 4.0, characterized by the integration of technologies like IoT, Big Data, Robotics and Automation, 5G Communication, and Artificial Intelligence, is reshaping multiple facets of modern society. Among these technologies, Machine Learning, a core component of AI, is increasingly being utilized in both daily life improvements and sophisticated scientific advancements. In daily life, major e-commerce platforms, music services, and video streaming sites leverage ML to recommend products, music, and videos based on users' historical behavior and preferences, thereby improving the user experience [12]. In healthcare, ML aids in more accurate disease diagnosis, tumor detection through medical image analysis, and the development of personalized treatment plans based on patients' genomic and medical history data [13]. In agriculture, ML empowers autonomous agricultural robots to perform tasks such as sowing, fertilizing, and weeding with precision [14]. Moreover, ML is also being utilized in other sectors such as automation and non-destructive testing.

The rapid advancement of ML technology has also revitalized the construction industry. Increasingly, scholars are exploring the applications of ML in architectural design, construction, management, and other related areas, with the potential enhance both efficiency and quality significantly. ML can offer promising prospect for construction section by assisting the engineers in: (a) structural design optimization [15] via identification of optimal solutions to meet performance requirements, (b) structural health monitoring [16] via early detection of structural damage to prevent the catastrophic failures and (c) predicting the remaining useful life of structure to inform maintenance and reinforcement strategies [17]. Apart from these critical roles in structural analysis, ML also has a profound impact on building materials as well,

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

where it can be used to accurately predict mechanical properties (e.g., strength and elastic modulus) by analyzing material composition and preparation processes [18]. Additionallu, ML can be helpful in the creation of advanced construction materials by predicting the longevity of these materials and evaluating their characteristics under different environmental scenarios. In the context of material recycling, the integration of ML with image processing can facilitate the accurate sorting of construction waste, thus fostering industrial sustainability [19].

Given the promising data processing capability, ML showcases its broad applications in construction cost prediction. By harnessing and analyzing the extensive historical construction project data, ML can be used to construct more accurate prediction models for project cost control. Numerous ML algorithms, such as Linear Regression, Decision Trees, Random Forests, and Neural Networks, have been employed to tackle the cost prediction problems of different complexity levels. The Random Forest [20] excels at handling high-dimensional and noisy data, while Neural Networks [21] are effective in modeling complex nonlinear relationships. By selecting suitable models, researchers can achieve more accurate construction cost predictions, thereby providing robust support for informed project decision-making. Despite increasing attention to ML in construction cost estimation, current studies remain fragmented in scope. Most research either focuses on a narrow subset of algorithms or fails to apply consistent benchmarking across models. Moreover, few studies leverage both linear and ensemble methods within a unified evaluation pipeline using standardized industry datasets. Advanced algorithms such as Neural Networks [22] and XGBoost [23], though well-established in other domains, are rarely investigated in the context of construction cost forecasting. This gap highlights the need for a robust comparative framework to evaluate the predictive performance, generalization, and uncertainty estimation of multiple ML models under controlled experimental conditions.

To address this gap, this study formulates the hypothesis that ensemble learning models (e.g., XGBoost and CatBoost) can significantly outperform traditional linear and non-parametric models in predicting construction costs, particularly in modeling nonlinear interactions and managing categorical variables. This hypothesis is empirically tested using a standardized RSMeans dataset containing 4,477 samples, covering diverse structural assemblies and cost variables. The main contributions of this study are threefold:

1. The implementation of six machine learning models, i.e., Ridge Regression, Lasso Regression, Elastic Net, K-Nearest Neighbor (KNN) Regression, XGBoost and CatBoost, under a consistent training-validation-testing protocol.
2. The integration of confidence interval analysis to quantify predictive uncertainty, an aspect often neglected in prior studies.
3. The introduction of a process innovation in cost modeling by developing a scalable and robust framework that bridges predictive accuracy with practical deployment feasibility.

The organization of this paper is as follows: Section 2 reviews related work. Section 3 details the methodologies of the six ML models employed. Section 4 describes the data sources, the model training process, performance evaluation metrics, and discusses the results. Section 5 concludes with a summary of findings and potential avenues for future research.

II. Literature Review

As computer and information technologies continue to evolve rapidly, machine learning has emerged as a highly effective method for prediction and classification, enabling the discovery of underlying patterns in vast amounts of data. A growing number of studies are leveraging machine learning techniques to improve the accuracy of construction cost forecasting. Hai [22] employed Multivariate Linear Analysis to develop a regression model estimating 16 factors that impact construction costs. Using SPSS for weighted statistical analysis, four common factors were identified, revealing a maximum budget deviation rate of 4.80%, which falls within the acceptable threshold of 10%. In a study by Harrison et al. [23], XGBoost was applied to historical construction project data from Ghana, spanning the year of 2016 to 2018. The model demonstrated strong predictive capabilities, achieving RMSE, MSE, MAE, and Mean Absolute Percentage Error (MAPE) values of 0.202, 0.041, 0.069, and 0.306, respectively. Their innovative approach provided insights into key variables, enhancing the design of predictive models for cost overruns and improving cost estimation practices in the construction industry. Lowe et al. [24] developed Linear Regression (LR) models using 286 datasets from the UK to predict construction costs. Among the six models examined, a total of 41 candidate predictors were considered, with five factors—gross internal floor area (GIFA), building function, project duration, mechanical systems, and foundation work—repeatedly identified as primary cost influencers. The findings also indicated that conventional cost estimation techniques generally yield an MAPE close to 25%.

Jafarzadeh [25] collected real data from 183 confined masonry (CM) school buildings in Iran and employed Stepwise Regression to develop parametric models, with cross-validation used to test their predictive

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

performance. Four variables were identified as the best predictors of changes in retrofit net construction cost (RNCC), i.e., mortar quality, roof and floor diaphragm type, seismic weight index, and gross floor area. This study emphasized the importance of the glass floor area in RNCC prediction for CM buildings. It recommended the use of a double-log cost area model during the early design phase of seismic retrofits. Using comprehensive data from 93 construction projects in Australia, Skitmore and Ng [26] developed multiple predictive models aimed to estimate actual construction durations and costs. These models were built using Forward Cross-Validated Regression analysis, incorporating variables such as contract amount, duration, project type, contractual arrangements, client sector, and contractor selection. A range of regression approaches, including conventional regression and cross-validation methods, were employed, with the cross-validation model ultimately chosen due to the lowest deleted residual sum of squares. The sensitivity analysis indicated that estimation errors for construction duration declined with more extended contract periods, whereas the accuracy of cost predictions remained stable regardless of project size.

Drawing on data from approximately 300 construction projects, Emsley et al. [27] constructed cost estimation models based on Neural Networks (NN), using LR as a baseline for performance comparison. Their findings highlight the key strength of the NN method—its capability to capture complex nonlinear relationships within the dataset. The best NN model achieved an MAPE of 16.6%. This performance compares favorably with conventional estimates, which range from 20.8% to 27.9%. Shahandashti and Ashuri [28] conducted an extensive literature review to identify 16 potential predictors for the National Highway Construction Cost Index (NHCCI), including variables such as consumer price index, new housing starts and crude oil prices. Using Granger causality and unit root analyses, they found that the average hourly wage in the construction sector and crude oil prices served as significant leading indicators. A Vector Error Correction (VEC) model was then formulated based on cointegration test findings, providing a suitable multivariate framework for NHCCI forecasting. The VEC model, which incorporated crude oil price and NHCCI data, successfully passed diagnostic evaluations and outperformed a univariate model in prediction accuracy, as demonstrated by its lower MAPE and MSE in out-of-sample tests. Petruseva et al. [29] implemented both Support Vector Machine (SVM) and LR models, utilizing Bromilow's cost and time relationship model along with DTREG predictive modeling software. Their comparative analysis revealed that SVM significantly outperformed LR in prediction accuracy. Huang and Hsieh [30] proposed an innovative approach grounded in the Cross-Industry Standard Process for Data Mining (CRISP-DM), introducing a hybrid model that integrates Random Forest (RF) with LR to improve the precision of labor cost predictions in the BIM-based construction phase. Drawing on data from 19 finalized BIM projects, their comparative study showed that this combined methodology significantly reduce the uncertainty associated with estimating BIM labor costs.

Kim et al. [31] utilized a hybrid approach that integrates NN with Genetic Algorithm (GA) to predict early-stage construction costs for residential buildings developed in Seoul, South Korea, between 1997 and 2000. The research assessed three distinct models: the first tuned the backpropagation network's parameters via a trial-and-error process; the second optimized these parameters using a genetic algorithm; and the third applied a genetic algorithm specifically to train the NN weights. Findings revealed that the second model outperformed the others in accurately estimating the preliminary costs of residential construction projects. Cheng et al. [32] proposed a hybrid intelligent framework named ELSVM to capture variations in construction pricing as represented by the Construction Cost Index (CCI). This system combines the capabilities of Least Squares Support Vector Machine (LS-SVM) and Differential Evolution (DE). In the model, LS-SVM was responsible for establishing the functional relationship between the CCI and its influencing variables, while DE was employed to fine-tune the LS-SVM parameters. The model was trained using a dataset comprising 122 historical records. Results from the experiments demonstrated that ELSVM successfully modeled CCI dynamics, achieving a mean absolute percentage error (MAPE) of under 1%, indicating high predictive accuracy. Alshboul et al. [33] collected, preprocessed, analyzed, and evaluated the latest datasets from 3,578 green projects in North America, utilizing two advanced ML techniques (e.g., LightGBM and XGBoost) to identify key parameters influencing cost estimation for sustainable buildings. Their analyses suggest that public and private investments in sustainable buildings are likely to result in reduced costs.

Zhang et al. [34] sought to design a parameter-driven cost prediction method for state highway administrations (SHA) to predict project costs before execution, enabling preemptive measures against cost escalation. While ordinary least squares (OLS) regression is a prevalent method in cost estimation, it has notable limitations. This study expanded the variable set to include previously unused economic factors, that are generally considered influential in determining highway construction costs. The findings indicated that the criterion-based LASSO regression model surpassed OLS in terms of prediction accuracy. Shehadeh et al. [35] introduced a set of ML models, i.e., Modified Decision Tree (MDT), LightGBM, and XGBoost, to estimate the residual value of

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

construction equipment. The performance of these models was assessed using four key evaluation criteria: MAE, MSE, MAPE, and R^2. This study highlighted the capacity of ML to enhance automation within the construction industry. Simić et al. [36] constructed cost prediction methods utilizing Neural Networks, XGBoost and Multiple Regression Analysis. Their findings suggest that satisfactory prediction accuracy can be attained using a limited number of cost drivers, specifically three from the owner's viewpoint and five from the contractor's. Expanding the number of input variables does not always enhance model precision. The survey results also revealed differing priorities: owners have more concern over environmental impacts, whereas contractors focused on fluctuations in resource prices, particularly in light of recent increases driven by the Russia-Ukraine conflict and the COVID-19 pandemic.

Alshboul et al. [37] utilized FR, Deep NN (DNN) and XGboost to estimate the costs of green buildings by considering both hard and soft cost-related factors. The performance of these models was evaluated using standard metrics, including MAE, MSE, MAPE, and R^2. Among the models, XGBoost achieved the highest predictive accuracy with an R^2 of 0.96, followed by DNN at 0.91 and RF at 0.87. In a separate study, Kim [38] assessed the effectiveness of three cost estimation methods, i.e., SVM, NN, and Regression Analysis (RA), with historical construction cost data. The results demonstrated that the NN model outperformed others in accuracy, making it the most appropriate choice for predicting school construction costs. Cheng and Hoang [39] proposed an innovative cost forecasting framework for construction projects, known as EAC-LSPIM, which combines LS-SVM, DE, and Machine Learning-based Interval Estimation (MLIE). This EAC-LSPIM generates both point estimates and prediction bounds, thus offering a confidence level that addresses inherent uncertainties in project costing. Yun [40] advanced construction cost prediction by employing a neural network with a multi-output regression model to estimate seven sub-construction costs, instead of a single aggregate figure. This approach enables the detailed prediction of individual cost factors simultaneously, thereby facilitating the estimation of various construction types or partial costs.

Reviewing the literature, there is a noticeable trend towards utilizing ML methods in construction engineering, particularly for predicting construction costs due to their superior predictive capabilities. However, the application of ML in this domain is still in its early stages, presenting significant opportunities for further research. It is observed that similar ML models (e.g., LR, SVM, NN, RF, XGBoost) are frequently applied across various contexts. In contrast, the potential of other ML models in construction cost prediction remains underexplored. Moreover, most previous studies have limited their performance comparisons to a small selection of models, typically three to four. In response to these identified gaps, this research undertakes an in-depth comparative analysis by applying six diverse ML models, namely Ridge Regression, Lasso Regression, Elastic Net, KNN Regression, XGBoost, and CatBoost, on a unified construction cost dataset. Notably, most of these selected ML models have not been previously applied to construction cost prediction and are evaluated using metrics such as R^2, RMSE, and MBE.

III. Method

The methodology adopted in this study follows a structured machine learning (ML) workflow, illustrated in Fig.1, aimed at systematically evaluating and comparing the predictive performance of six ML models for construction cost estimation. These models include Ridge Regression, Lasso Regression, Elastic Net, K-Nearest Neighbors (KNN) Regression, Extreme Gradient Boosting (XGBoost), and CatBoost. This diverse selection of models encompasses linear, non-parametric, and ensemble-based learners, enabling a comprehensive assessment of algorithmic effectiveness across a range of data complexities and non-linear relationships commonly encountered in construction datasets.

The methodological pipeline begins with data acquisition and preprocessing, followed by partitioning the dataset into training and testing subsets. To ensure a robust evaluation and reduce the risk of overfitting, a stratified 5-fold cross-validation was integrated into the hyperparameter tuning process. GridSearchCV was employed to systematically explore hyperparameter spaces for each model, optimizing for the highest R^2 score on validation folds. All models were implemented using Python (version 3.13) with packages from scikit-learn (version 1.6.9), XGBoost (version 2.1.4), and CatBoost (version 1.2.7).

The model performance is assessed using three key metrics: the coefficient of determination (R^2), root mean square error (RMSE), and mean bias error (MBE). To enhance the interpretability and reliability of performance estimates, 95% confidence intervals were computed using bootstrapped resampling (n = 1,000 iterations) on the test set predictions.

A. Description of the RSMeans Dataset

The construction cost prediction models in this study were developed and evaluated using a dataset compiled from the RSMeans Assemblies Books published between 1998 and 2018. This dataset provides comprehensive historical records of construction component costs in the United States

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

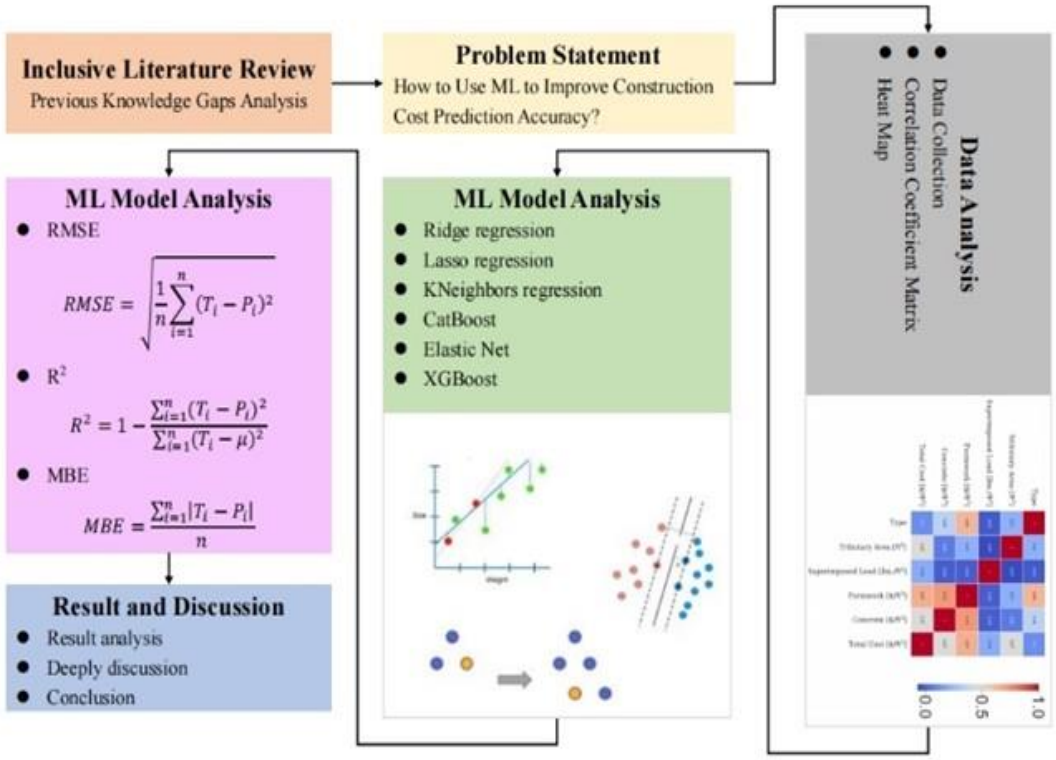


Fig.1. Workflow of the proposed methodology for construction cost prediction.

and is widely recognized as a benchmark in the industry. The target variable used is the total construction cost of structural assemblies, measured in USD per square foot [$/ft²]. The feature set includes five independent variables: structural assembly type (categorical), tributary area [ft²], superimposed load [lbs./ft²], unit cost of formwork [$/ft²], and unit cost of concrete [$/ft³].

The dataset contains 4,477 samples, each representing a distinct floor assembly in a medium- or high-rise structure. Assemblies include one-way slabs, two-way slabs, flat slabs with and without drop panels, multi-span joist slabs, and waffle slabs. These are encoded as an integer-valued categorical variable (1 to 6), representing the structural type. To retain ordinal information while minimizing dimensionality, label encoding was used instead of one-hot encoding. This approach was selected due to the low cardinality of the categorical variable and its implicit rank structure, as described in the RSMeans documentation.

Data integrity was thoroughly assessed prior to modeling. All features were complete with no missing values or null entries, thereby eliminating the need for imputation. Independent variables, such as unit costs and tributary area, were inspected for skewness and outliers using boxplots and kernel density estimation (KDE). Values beyond three standard deviations were capped using a mild winsorization strategy to reduce distortion while preserving data integrity. The descriptive statistics (mean, standard deviation, min, max) for each variable are summarized in Table 1.

RSMeans, now maintained by Gordian, undergoes continuous verification by a dedicated team of cost engineers. Its extensive coverage across regions and project types ensures that the dataset used is both authoritative and representative of real-world cost conditions. Despite its breadth, the dataset primarily reflects North American construction practices, which is acknowledged as a limitation in the generalizability of our models.

Table 1. Descriptive statistics of RSMeans dataset

Variables	Avg.	Std. Dev.	Min.	Max.
Structural Assembly Types	-	-	1	6
Triburaty Area [ft²]	763.23	421.86	225	1800
Superimposed Load [lbs./ft²]	107.04	58.8	40	200
Formwork [$/ft²]	7.43	2.28	4.19	13.75
Concrete [$/ft²]	3.55	0.96	1.73	5.23
Total Cost [$/ft²]	16.81	4.32	7.7	29.75

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

B. Data Preprocessing and Data Splitting

The raw dataset underwent a structured three-stage preprocessing pipeline: feature analysis, normalization, and data partitioning. First, Pearson correlation analysis was conducted to assess the linear relationships between input features and the target variable (Total Cost in $/ft²). As shown in the heatmap in Fig.2., the unit cost of formwork and the unit cost of concrete exhibited strong positive correlations with the target variable (r = 0.82 and 0.79, respectively), confirming their importance in cost estimation. Conversely, superimposed load and tributary area had weak correlations (r < 0.15), suggesting a lower predictive contribution in linear space.

Second, all continuous numerical variables were normalized using MinMaxScaler to a [0, 1] range. This transformation addressed scale disparities among features and improved convergence stability for gradient-based models such as Ridge, Lasso, and XGBoost. Normalization was consistently applied to both training and test data using the same scaling parameters derived from the training set.

Third, the dataset was partitioned into training and testing subsets, with 70% of the data allocated for training and 30% reserved for testing. A randomized shuffle split (*random_state* = 2021) was employed to ensure reproducibility and preserve the natural distribution of data. As the dataset targets a continuous variable, simple random sampling was used rather than stratified sampling, which is more appropriate for classification tasks. The split aimed to facilitate an unbiased evaluation of model generalization on unseen data.

Throughout preprocessing, the dataset was verified to contain no missing values, null entries, or inconsistent formatting. As the structural assembly type (categorical feature) had already been label-encoded in the dataset preparation phase (described in Section III.A), no further encoding was required at this stage. Although the training-test split is essential for performance evaluation, additional validation procedures, such as 5-fold cross-validation with shuffled splits, were subsequently applied during model training and hyperparameter optimization to ensure robust assessment across different subsets of the data.

C. ML Models for Construction Costs Prediction

This section presents the six machine learning models evaluated in this study: Ridge Regression, Lasso Regression, Elastic Net, KNN Regression, XGBoost, and CatBoost. All models were implemented using the Anaconda Python distribution, which includes essential scientific computing libraries such as Scikit-learn, NumPy, and SciPy. XGBoost and CatBoost were integrated via their latest official releases from GitHub. This setup ensures reproducibility and compatibility across all training, tuning, and evaluation tasks.

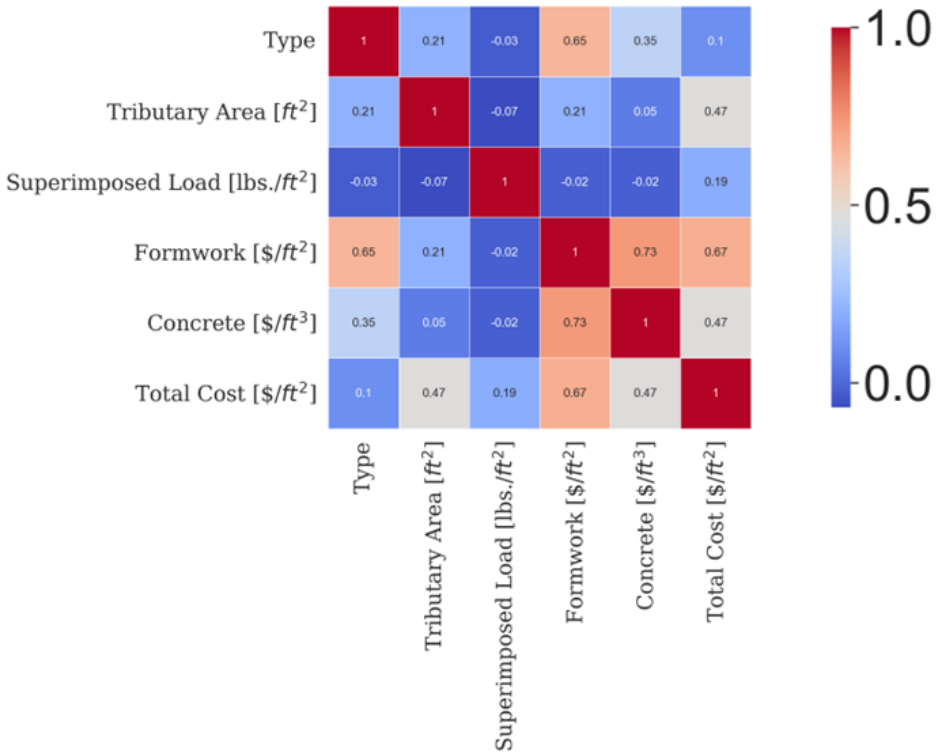


Fig.2. Correlation heatmap showing the relationships between features and total construction cost ($/ft²).

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

D. ML Models for Construction Costs Prediction

1. Ridge Regression

Ridge Regression is a regularization technique particularly valuable for addressing multicollinearity, an issue often encountered in construction cost prediction where input features such as labor cost, material cost, and project scale are strongly correlated. Multicollinearity can cause instability in coefficient estimation under ordinary least squares (OLS) regression, where minor changes in the input data may result in significant fluctuations in predicted outcomes. Ridge Regression alleviates this by adding a penalty term to the OLS cost function, thereby shrinking the magnitudes of the coefficients and reducing model variance.

In the context of construction cost prediction, Ridge Regression stabilizes the model by incorporating a regularization parameter α, a non-negative scalar that controls the strength of the penalty. The Ridge cost function is formulated in Eq. (1) [41] as follows:

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \right\} \quad (1)$$

where *n* denotes the total number of data points; *p* is the total number of features used for construction cost prediction; $i = 1, \ldots, n$ and $j = 1, \ldots, p$ refer to the indices of the data point and feature, respectively; y_i represents the actual construction costs; x_{ij} represents the features (e.g., square footage, number of floors, labor hours, etc.); β_j are the coefficients; α serves as the tuning parameter that determines the strength of the regularization applied. The closed-form solution to this minimization problem is calculated using Eq. (2) [41]:

$$\hat{\beta}_{ridge} = (X^T X + \alpha I)^{-1} X^T y \quad (2)$$

where X represents the matrix of input variables, with rows denoting individual observations and columns representing specific features; y is the vector of dependent variables, representing the construction costs of reach data point; I is the identity matrix, ensuring that the regularization is applied uniformly across all coefficients.

To ensure numerical stability and mitigate feature dominance due to scale differences (e.g., Tributary Area ranging from 225 to 1800 ft² versus Concrete Cost ranging from \$1.73 to \$5.23/ft³), all features were normalized to the [0, 1] range using MinMaxScaler. This ensure the regularization terms affects all features equitably.

Hyperparameter tuning was conducted using GridSearchCV with 5-fold cross-validation (*random seed* = 2021). Six candidate values of α, logarithmically spaced between 0.0001 and 10, were evaluated using R^2 as the scoring metric. The optimal configuration, $\alpha = 1$, achieved the highest validation score (R^2 = 0.836). When assessed on the test dataset, the model attainted strong performance with an R^2 of 0.827, confirming its ability to generalize to unseen data.

2. Lasso Regression

Lasso Regression enhances traditional LR by incorporating an L1 regularization term into the objective function, which penalizes the absolute values of the coefficients. This penalty not only controls model complexity to mitigate overfitting, but also enforces sparsity, shrinking less significant coefficients to precisely zero. Such automatic feature selection is especially advantageous in construction cost prediction, where datasets may contain numerous variables with varying levels of relevance.

Mathematically, Lasso Regression modifies the OLS loss function by using Eq. (3) [42]:

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (3)$$

where λ is the regularization parameter that determines the penalty strength, and the second term promotes sparsity in β.

In construction cost modeling, this capability is vital, as it eliminates redundant or weak predictors (e.g., overly correlated economic factors), thereby yielding simpler, more interpretable models. After applying MinMaxScaler normalization to scale all features to the [0,1] range, a GridSearchCV procedure was used to identify the optimal value of λ from the set [0.0001, 0.001, 0.01, 0.1, 1, 10], using 5-fold cross-validation (*random seed* = 2021).

The optimal model was obtained with $\lambda = 0.001$, corresponding to a relatively mild penalty. This weaker regularization was necessary to retain informative features while still encouraging sparsity. The selected model achieved an R^2 of 0.836 on the training set. Significantly, on the test set, the model maintained robust performance with an R^2 of 0.827, indicating good generalization despite the reduced complexity.

The final trained Lasso model retained a subset of the original features, suggesting that only the most informative predictors, such as formwork cost and concrete cost, were utilized. This sparsity supports streamlined cost forecasting pipelines and enhances explainability, a critical consideration in the construction industry.

3. Elastic Net

Elastic Net is a regularized regression method that effectively combines the strengths of both Ridge and Lasso Regression, instrumental in scenarios with high-

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

dimensional feature spaces or strongly correlated predictors, i.e., conditions often found in construction cost prediction tasks. While Ridge is known for coefficient shrinkage and Lasso for automatic feature elimination, Elastic Net offers a flexible compromise by applying both L1 and L2 regularizations.

The Elastic Net loss function is defined using Eq. (4) [43]:

$$\hat{\beta}_{elastic_{net}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \alpha_{overall} \left[\rho \sum_{j=1}^{p} |\beta_j| + \frac{(1+\rho)}{2} \sum_{j=1}^{p} \beta_j^2 \right] \right\} \quad (4)$$

where $\alpha_{overall}$ is the overall regularization strength; $\rho \in [0,1]$ is the mixing parameter that control the trade-off between L1 (sparsity-inducing) and L2 (shrinkage-inducing) penalties. When $\rho = 1$, Elastic Net becomes equivalent to Lasso, and when $\rho = 0$, it reduces to Ridge Regression. In construction cost prediction, many cost-related features, such as unit prices for formwork and concrete, often exhibit strong correlations. Elastic Net retains these grouped features (via L2) while eliminating irrelevant predictors (via L1), ensuring that no critical variables are inadvertently discarded.

For this study, all features were scaled to the [0, 1] range using MinMaxScaler to maintain uniform regularization. A grid search over five values of $\alpha_{overall} \in \{0.001, 0.01, 0.1, 1, 10\}$ was conducted using 5-fold cross-validation (*random seed* = 2021), with a fixed $\rho = 0.5$. This configuration was selected to leverage both L1 and L2 penalties evenly, allowing the model to strike a balance between sparsity and stability. The best-performing Elastic Net model was found at $\alpha_{overall} = 0.001$, achieving an R^2 of 0.836 on the training set and 0.827 on the test set, matching Ridge and Lasso in generalization, but offering better feature handling. The selected model retained all major predictors while shrinking minor ones, thereby producing a robust yet interpretable model. In practice, this balance is essential for the construction industry, where model explainability is critical for stakeholder trust. Elastic Net not only avoids the instability of Ridge in the presence of collinearity, but also overcomes the aggressive zeroing tendencies of Lasso.

4. KNN Regression

K-Nearest Neighbors (KNN) Regression is a non-parametric machine learning technique widely used to estimate continuous outcomes by leveraging the proximity of similar instances. Unlike parametric methods that assume a functional relationship between input variables and the target, KNN relies purely on data-driven similarities, making it well-suited for problems with complex or unknown interactions, such as construction cost estimation.

To estimate the construction cost for a new input x, KNN identifies the *k* most similar instances from the training data, determined using distance measures such as Euclidean metric, and computes the averages of their corresponding target values. The predicted value y_{pred} for the new input x is then calculated as shown in Eq. (5) [44]:

$$y_{pred} = \frac{1}{k} \sum_{a \in N_k(x)} y_a \quad (5)$$

where y_a is the target value (e.g., construction cost) of the *a*-th nearest neighbor in the training set; $N_k(x)$ is the group of *k* closest data points to x, determined using the selected distance measurement method.

In this study, the KNN Regressor was implemented using the Scikit-learn library and optimized via 5-fold cross-validation (*random seed* = 2022). A grid search explored different values of $k \in \{2,4,6,8,10,12,14,16\}$ and two distance metrics, i.e., Euclidean (L2) and Manhattan (L1). The best performance was achieved with $k = 8$ and Euclidean distance, which yielded the highest validation R^2 score. Although Manhattan distance was considered, it did not improve predictive performance and was therefore excluded from the final configuration. Prior to training, all features were normalized using MinMaxScaler to mitigate the influence of feature scale on distance calculations, a critical step, as KNN is highly sensitive to feature magnitude. The model's performance on the test set confirmed moderate predictive ability, with signs of overfitting due to its instance-based learning nature and lack of internal regularization mechanisms. KNN proved most effective on training data, but its generalization performance was inferior compared to ensemble methods such as XGBoost and CatBoost. This can be attributed to its local learning strategy, which fails to model broader structural patterns in the data, especially in high-dimensional spaces. Additionally, its computational complexity grows linearly with the dataset size, making it less suitable for large-scale real-time applications in construction management.

5. XGBoost

XGBoost is a powerful ML algorithm known for its ability to make accurate predictions by combining the outputs of multiple weak models, typically decision trees, into a strong predictive model. As an enhancement of the traditional Gradient Boosting Decision Tree (GBDT) algorithm, XGBoost introduces several improvements that make it particularly effective for large-scale datasets and complex models, such as those used in construction cost prediction. Unlike traditional GBDT, XGBoost incorporates a regularization term in its objective function, which helps prevent overfitting and enhances the model's generalization ability. This feature is crucial in construction cost prediction, where the model must perform reliably across a variety of

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

projects with differing characteristics. Additionally, XGBoost improves efficiency and accuracy by using a second-order Taylor expansion of the loss function, allowing the algorithm to handle more complex loss functions that may be difficult to differentiate directly.

Let $\hat{y}_i$ be the predicted construction cost for each *i*-th data point (i.e., x_i) in XGBoost and it is calculated using Eq. (6) [45] as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F, \quad i = 1, \dots, n \tag{6}$$

where K is the total number of trees in the XGBoost model; $f_k(x_i)$ represents the output of the *k*-th decision tree for the input x_i; $F = \{f(x) = \omega_{q(x)}\}$ is the set of all possible trees, where $q(x)$ maps each input to a corresponding leaf in the tree, and ω is the weight associated with that leaf. XGBoost seeks to minimize an objective function that combines the loss function and regularization terms, formulated in Eq. (7) [45] as:

$$\hat{f} = \underset{f_1,\dots,f_K}{\operatorname{argmin}} \{\sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)\} \tag{7}$$

where $L(y_i, \hat{y}_i)$ is the loss function measuring the differences between the actual cost y_i and the predicted cost $\hat{y}_i$; $\Omega(f_k)$ is the regularization term for the complexity of the *k*-th tree, typically defined in Eq. (8) [45]:

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_{kj}^2 \tag{8}$$

where T_k is the number of leaves in the *k*-th tree; ω_{kj} is the weight of the *j*-th leaf in *k*-th tree; γ and λ are regularization parameters that control the trade-off between model complexity and performance.

In this study, XGBoost was implemented using the xgboost library in Python. Hyperparameter tuning was conducted using grid search with 5-fold cross-validation (*random seed* = 2022). The following hyperparameters and their ranges were explored: (a) number of estimators $n\_estimators \in \{100,200,\dots,2000\}$, with a step size of 100, (b) maximum depth of the trees, $max\_depth \in \{1,2,\dots,10\}$, (c) learning rate, $learning\_rate \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$, (d) regularization on tree complexity, $\gamma \in \{0, 0.1, 0.5\}$ and (e) L2 penalty, $\lambda \in \{0.5, 1, 2\}$.

The best-performing hyperparameter configuration was $n_{estimators} = 1400$, $max\_depth = 6$, $learning\_rate = 0.01$, $\gamma = 0.1$ and $\lambda = 1.0$. This configuration yielded a training R^2 of 0.994 and a test R^2 of 0.988, with an RMSE of 0.43 and near-zero MBE, indicating excellent predictive accuracy and generalization. The model's ability to suppress overfitting is primarily attributed to the regularization parameters and its iterative additive training structure.

XGBoost also supports feature importance analysis. In our study, formwork cost and concrete cost emerged as the most influential predictors, i.e., consistent with findings from the correlation heatmap. This interpretability facilitates decision-making in budget planning and resource allocation. Given the high dimensionality and nonlinearity of construction cost data, XGBoost's scalability, ability to model complex interactions, and resilience to multicollinearity make it a compelling choice for practical deployment in cost estimation workflows.

6. CatBoost

CatBoost is another gradient boosting algorithm that builds upon the foundations of GBDT, following in the advancement of algorithms like XGBoost and LightGBM. Developed and open-sourced by Yandex in 2018, CatBoost offers unique features that make it suitable for construction cost prediction, where categorical variables often play a significant role. One of CatBoost's key advantages is its native handling of categorical features, eliminating the need for extensive preprocessing, such as one-hot encoding. This capability reduces the risk of overfitting and enhances prediction accuracy, which is crucial in construction cost prediction, where features or variables such as project type, location, and material grade are commonly categorical. Additionally, CatBoost employs Ordered Boosting, a method that improves model robustness and efficiency by sorting the training data and selecting a subset of relevant samples to train each decision tree iteration. This approach helps to reduce overfitting and ensure better generalization.

Denote $F_m(x_i)$ as the predicted construction cost for each *i*-th data point (i.e., x_i) in CatBoost after adding the *m*-th decision tree, and it is calculated using Eq. (9) [46] as:

$$F_m(x_i) = F_{m-1}(x_i) + h_m(x_i) \tag{9}$$

where $F_{m-1}(x_i)$ is the cumulative prediction from the first $(m-1)$ decision trees; $h_m(x_i)$ represents the output of the *m*-th decision tree, which fits the residuals from the previous trees. Similar to XGBoost, CatBoost minimizes an objective function, as shown in Eq. (7) [45] and Eq. (8) [45], that combines the loss function with a regularization term to control model complexity and prevent overfitting. The key difference lies in substituting the predicted value $\hat{y}_i$ in Eq. (7) [45]with $F_m(x_i)$ as computed through Eq. (9) [46].

In this study, CatBoost Regressor was applied to predict total construction costs using the RSMeans dataset. The model was implemented using the catboost library in Python. Hyperparameter tuning was conducted using 5-fold cross-validation with randomized search, optimizing the following key parameters: (a) $n_{estimators} = 1000$ that determines the total number of boosting iterations, (b) $max\_depth = 4$ that controls tree complexity, aiming to balance the expressiveness and overfitting risk, (c) $learning\_rate = 0.1$ that determines the contribution of each tree to the ensemble, (d) $subsample = 0.9$ that specifies the fraction of samples used per boosting round to

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

introduce variance and mitigate overfitting and (e) $olsample\_bylevel = 0.9$ that denotes the proportion of features used at each level of tree construction.

After hyperparameter tuning, the model was trained on the training set and evaluated on the test set using RMSE, MAE, MAPE, and maximum error as the key performance indicators. The CatBoost model achieved a test R^2 of 0.987 and RMSE of 0.47, comparable to XGBoost and significantly outperforming linear models. Additionally, CatBoost provides built-in feature importance metrics, which were visualized via bar charts. These insights help identify the most influential cost drivers, such as formwork and concrete unit costs, thereby enhancing interpretability and enabling informed decision-making by practitioners.

Overall, CatBoost's native handling of categorical variables, robust boosting technique, and efficient optimization process make it particularly suitable for construction cost modeling, where data complexity, heterogeneity, and inter-variable dependencies are common.

E. Performance Metrics Used for Model Evaluations

The performances of the six ML models for construction cost prediction are evaluated and compared using three performance metrics: coefficient of determination (R^2), root mean square error (RMSE), and mean bias error (MBE). The detailed definitions of these performance metrics are as follows.

The R^2 metrics assess how well the regression model fits the data and it is defined in Eq. (10) [47] as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \mu)} \quad (10)$$

where y_i is the *i*-th actual data point; $\hat{y}_i$ is the *i*-th predicted data point; *n* is the total number of data points; *i* is the index of data point; μ is the mean of the actual data. Higher R^2 values indicate a better fit of the model to the data, as they represent a higher proportion of variance in the dependent variable that the model explains.

The RMSE metric is a commonly used evaluation metric that measures the average magnitude of the error between predicted and actual data points. It is mathematically defined in Eq. (11) [47] as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (11)$$

Smaller RMSE values are preferable because they indicate lower prediction errors, reflecting higher prediction accuracy and model performance.

The MBE metric measures the average difference between the actual and predicted values of data points, reflecting the overall bias in the model's predictions. MBE provides insight into whether the model tends to overestimate or underestimate the actual values on average. It is mathematically defined in Eq. (12) [47] as:

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (12)$$

A smaller absolute value of MBE indicates less bias and a better overall prediction accuracy. The closer the MBE metric is to zero, the better the model's performance, as it signifies minimal average deviation between predicted and actual values.

IV. Result

A. Performance Comparisons of ML Models

The performance of six selected ML models (Ridge Regression, Lasso Regression, Elastic Net, KNN Regression, XGBoost, and CatBoost) for solving construction cost prediction problems is compared in Table 2 and Fig. 3. These comparisons are based on the R2, RMSE, and MBE metrics, evaluated on both the training dataset (70% of the entire dataset) and testing dataset (30% of the whole dataset).

The training dataset results provide insights into how well each model captures the underlying patterns in the dataset. The first three models (Ridge Regression, Lasso Regression, and Elastic Net) yield identical performance metrics, with R^2 values of 0.836 and RMSE values of 1.750. Their MBE values are zero, indicating no significant bias in the training predictions. The moderate R^2 and relatively high RMSE suggest that while these models capture the data patterns to a reasonable extent, they fall short in accuracy compared to more complex models. KNN Regression, with an R^2 value of 0.986 and a significantly lower RMSE of 0.503, demonstrates a much better fit to the training data than the linear models. However, the slightly positive MBE of 0.029 indicates a minor tendency to overestimate costs on average. XGBoost and CatBoost show exceptional performance on the training dataset, with R^2 values close to 1.000 (0.995 and 0.996, respectively). Their RMSE values (i.e., 0.306 for XGBoost and 0.275 for CatBoost) are the lowest among all models, indicating high precision in capturing the training data patterns. The testing dataset results are crucial as they reveal the models' generalization capabilities on unseen data. Similar to the training dataset, Ridge Regression, Lasso Regression, and Elastic Net perform consistently on the testing dataset, with R^2 values of 0.827 and RMSE values around 1.792. Their positive MBE values (0.102) indicate a slight overestimation in predictions. The similarity in results across these models suggests limited flexibility, impacting their ability to generalize effectively to new data. KNN Regression maintains a high R^2 value of 0.951 on the testing dataset, although the RMSE increases to 0.951, and the MBE reduces to 0.042.

These results indicate that while KNN generalizes relatively well, it exhibits some overfitting compared to

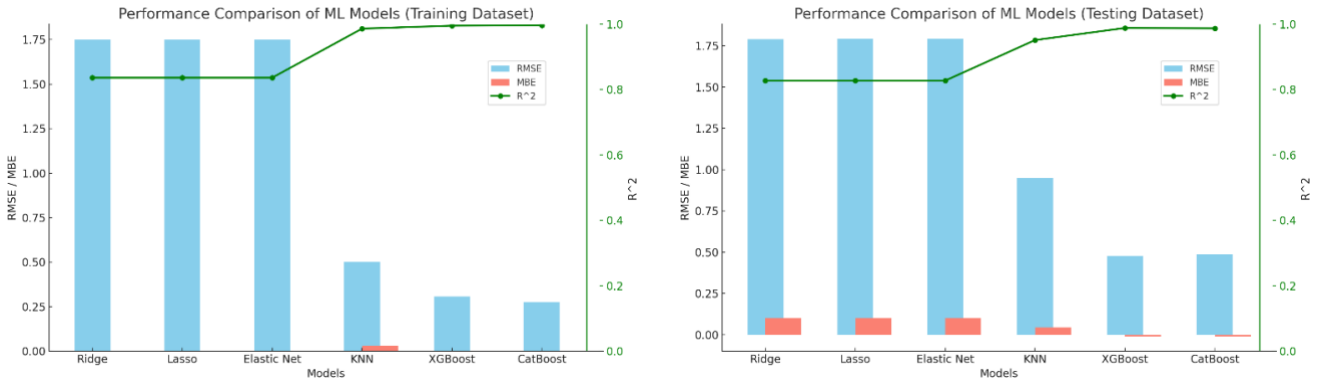
Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

the training dataset, as reflected in the increased RMSE. XGBoost and CatBoost continue to outperform the other models on the testing dataset, with R^2 values of 0.988 and 0.987, respectively. Although their RMSE values increase slightly to 0.478 for XGBoost and 0.485 for CatBoost, they remain the lowest among all models, indicating excellent predictive accuracy. Additionally, the near-zero and slightly negative MBE values suggest that these models have minimal bias and are highly reliable for predicting construction costs.

The scatter plots in Fig. 4 illustrate the relationship between predicted and actual construction costs for each ML model, providing a visual validation of the performance metrics presented in Table II. The scatter plots for Ridge Regression, Lasso Regression, and Elastic Net on both training and testing datasets show a moderate spread around the diagonal line of perfect prediction, indicating that while the models' predictions are somewhat aligned with actual costs, there is room for improvement. The consistent spread across both datasets suggests these models generalize reasonably well but lack the flexibility needed to capture more complex patterns. The KNN Regression model shows a tight clustering of points around the diagonal in the training dataset, reflecting its high R^2 value of 0.986. However, in the testing dataset, while the clustering remains relatively tight, the increased spread indicates a slight reduction in accuracy, consistent with its R2 value of 0.951. This suggests some overfitting, where the model performs exceptionally well on known data but less so on new data. Finally, both XGBoost and CatBoost demonstrate near-perfect alignment of predicted and actual costs in both training and testing datasets, as evidenced by the close clustering of points along the diagonal line. The minimal deviation from the line of perfect prediction, along with high R2 values, confirms their robustness and generalization capability. The consistency in their performance across both datasets suggests that these models are well-tuned and highly effective for construction cost prediction.

Table 2. Performance comparison of six ML models

ML Models	Training Dataset			Testing Dataset		
	R^2	*RMSE*	*MBE*	R^2	*RMSE*	*MBE*
Ridge Regression	0.836	1.750	0.000	0.827	1.790	0.102
Lasso Regression	0.836	1.750	0.000	0.827	1.792	0.102
Elastic Net	0.836	1.750	0.000	0.827	1.792	0.102
KNN Regression	0.986	0.503	0.029	0.951	0.951	0.042
XGBoost	0.995	0.306	0.000	0.988	0.478	- 0.010
CatBoost	0.996	0.275	0.000	0.987	0.485	- 0.010



(a) (b)

Fig. 3. Performance comparison of six ML models using: (a) training dataset and (b) testing dataset.

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

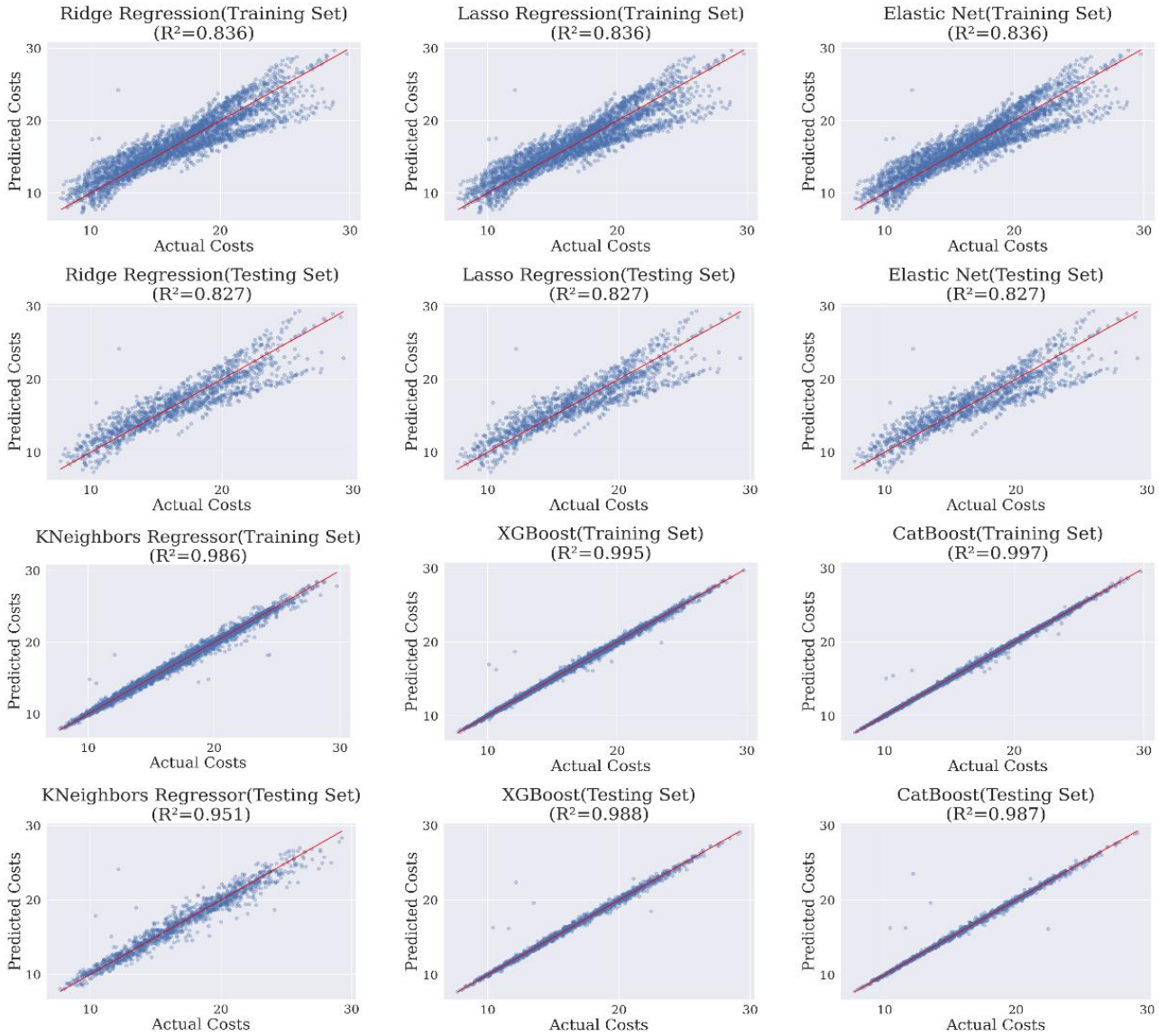


Fig.4. Scatter plots of all six ML models in training and testing datasets.

B. Confidence Intervals of ML Predictions

To enhance the interpretability and reliability of model performance, we included 95% confidence intervals (CI) for all predicted costs across both training and testing sets, as shown in Fig. 5. These intervals were computed using bootstrapped residuals over 1,000 iterations, capturing the range of variability in the prediction estimates. The blue vertical lines indicate the prediction uncertainty for each individual sample, where narrower intervals denote higher model confidence, and wider intervals suggest less stable forecast.

From the visualization, it is evident that the linear models (Ridge, Lasso, and Elastic Net) exhibit relatively wide confidence intervals across both training and testing sets. This suggests these models have higher variance in their prediction errors, likely due to their limited ability to model nonlinearities inherent in construction cost data. Moreover, their average RMSE values remain above 1.75, confirming their weaker generalization performance.

In contrast, ensemble methods such as XGBoost and CatBoost consistently show narrow confidence intervals, especially in the testing set. This reflects their superior robustness and lower prediction variance. For instance, the CatBoost model achieves an RMSE of 0.488 on the test set, accompanied by relatively tight confidence bounds, indicating a strong fit to unseen data while maintaining consistent predictive accuracy.

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

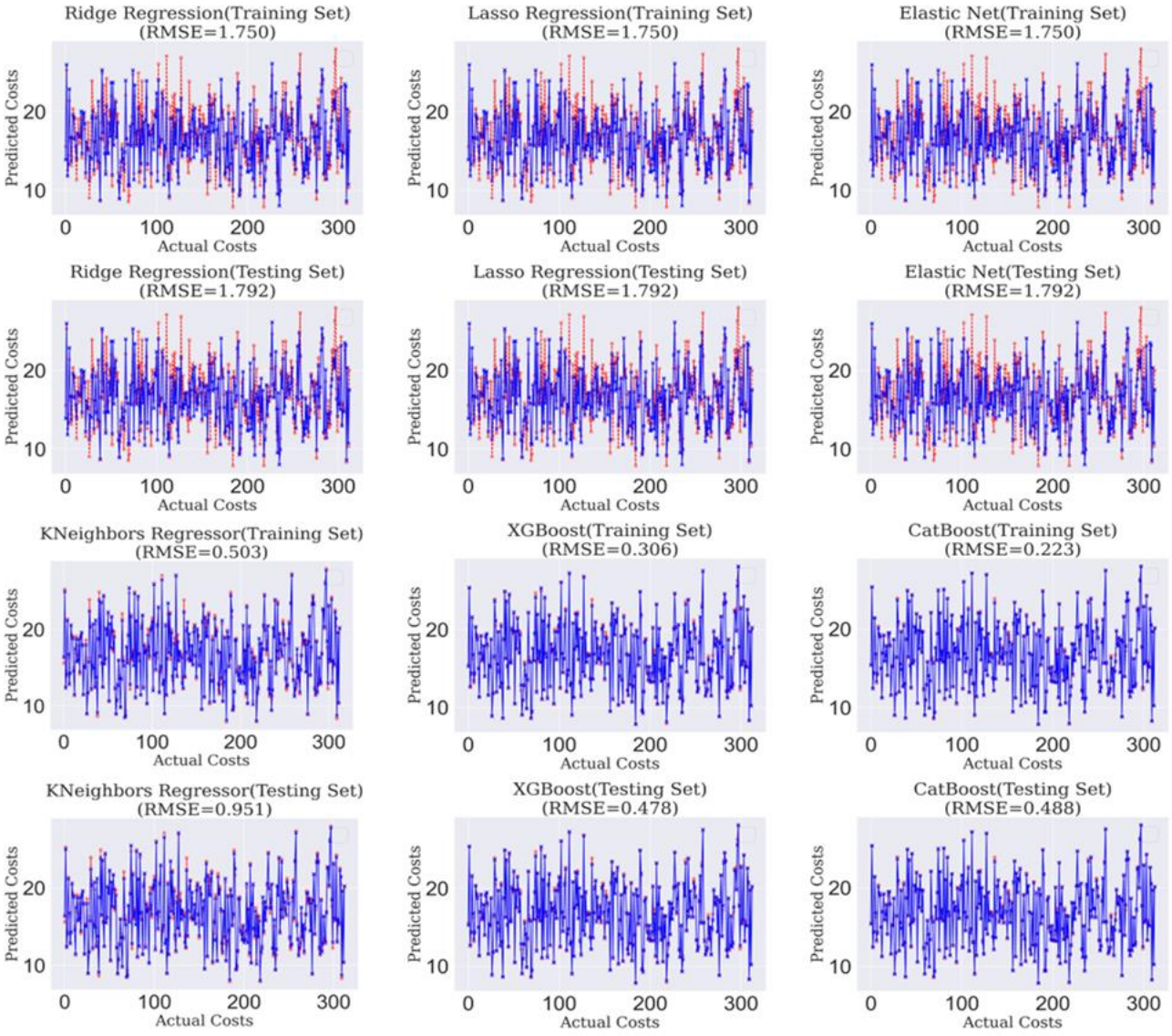


Fig.5. Confidence Intervals of all six ML models in training and testing datasets.

These confidence interval plots provide a more comprehensive understanding of each model's predictive reliability and highlight the practical advantage of ensemble approaches in capturing uncertainty with greater precision, an important consideration for real-world construction cost estimation tasks.

V. Discussion

A. Machine Learning for Construction Cost Estimation

This study provides a comprehensive evaluation of six machine learning models Ridge Regression, Lasso Regression, Elastic Net, KNN Regression, XGBoost, and CatBoost for predicting construction costs using the standardized RSMeans dataset The results strongly indicate that ensemble-based models, particularly XGBoost and CatBoost, significantly outperform traditional linear models and KNN in prediction accuracy, generalization, and robustness across both training and testing datasets.

Specifically, XGBoost and CatBoost consistently achieved R^2 values above 0.98, RMSE values below 0.5, and near-zero MBE, demonstrating their exceptional capability to model nonlinear relationships and capture complex feature interactions often present in construction datasets. These findings confirm that ensemble methods are well-suited for high-dimensional data with multicollinearity and nonlinear dependencies, i.e., common characteristics in cost estimation tasks involving structural, economic, and categorical variables. By comparison, Ridge Regression, Lasso Regression,

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

and Elastic Net, although known for their stability and interpretability, yielded only moderate predictive performance, with R^2 values around 0.83 and RSME values ranging from 1.75 to 1.79. These models' inherent assumption of linearity limits their ability to account for intricate variable interactions, making them less effective for capturing the complex cost dynamics in construction data. KNN Regression, while demonstrating strong accuracy on the training set, showed a significant increase in error on the testing set, indicating overfitting. This suggests that KNN is highly sensitive to local data structures and noise, which may hinder its generalization performance in real-world applications with varying project profiles.

This evaluation confirms that ensemble models, particularly XGBoost and CatBoost, are highly capable of delivering accurate, stable, and generalizable predictions in the context of construction cost estimation. Their superior performance can be attributed to their capacity for iterative learning, regularization, and handling of both continuous and categorical variables without extensive preprocessing. These strengths make them highly effective tools for practitioners seeking data-driven approaches to improve cost estimation accuracy and reliability.

B. Comparative Analysis with Existing Studies

Numerous studies have employed machine learning techniques for predicting construction costs, each contributing uniquely to the field. For instance, Yun [40] explored the use of a multi-output regression model based on artificial neural networks (ANN) to predict seven sub-itemized construction costs simultaneously. By comparing the multi-output model (error rate = 16.80%) with a traditional single-output model (error rate = 17.67%), the study demonstrated that itemized prediction slightly improved the accuracy of total cost estimation. This method provided more granular insights into the influence of individual construction activities on overall cost. However, the study also highlighted limitations related to varying error rates across project types, the lack of optimization strategies for specific items, and high data annotation costs, all of which could hinder model generalizability in practice. Expanding on this, Simić et al. [36] incorporated XGBoost into a cost prediction framework for highway projects, integrating multiple regression analysis and neural networks. The study emphasized different stakeholder perspectives, such as owners and contractors, by identifying key cost drivers through surveys. Interestingly, their findings indicated that including too many input variables did not necessarily improve model accuracy, underscoring the importance of feature selection and model simplification. Similarly, Harrison et al. [23] applied XGBoost to forecast cost overruns in Ghanaian construction projects. The model demonstrated strong predictive performance, as indicated by RMSE, MSE, MAE, and MAPE metrics, validating the algorithm's ability to effectively model cost deviations using real-world data. In another study, Alshboul et al. [33] employed a hybrid approach that combines mathematical modeling and machine learning to estimate green building costs. Utilizing data from 3,578 green projects in the northern United States, the authors developed a supply-demand equilibrium model that incorporated macroeconomic factors, including inflationary cycles and external investments. Their results revealed that both public and private investments significantly lowered green building costs, especially during deflationary periods. However, the study's regional scope limits its generalizability to non-green or international construction projects.

As summarized in Table 3, most prior studies focused on either single-output prediction, limited machine learning techniques, or narrow cost variables, and often lacked a unified evaluation framework. Few studies comprehensively benchmarked multiple machine learning models on a standardized dataset using consistent evaluation metrics. This study addresses these gaps by applying six different ML algorithms to a unified RSMeans dataset and evaluating them using RMSE, MBE, and R^2 metrics. Compared to prior research, the integration of XGBoost and CatBoost in this study offers superior accuracy, generalizability, and modeling efficiency for multidimensional and nonlinear construction datasets.

Table 3. Comparison with previous studies

Authors	Methodology	Metrics
Yun [40]	Multiple-Regression, ANN	R^2
Simić et al. [36]	Multiple-Regression, ANN, XGBoost	R^2, MAPE
Harrison et al. [23]	XGBoost	RMSE, MSE, MAE, MAPE
Alshboul et al. [33]	LightGBM, XGBoost	MAE, RMAE, MAPE, R^2
This study	Ridge Regression, Lasso Regression, Elastic Net, KNN Regression, XGBoost, CatBoost	RMSE, MBE, R^2

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

C. Justifying the Superior Performance of Ensemble Learning Models in Cost Prediction

The superior performance of ensemble-based models, such as XGBoost and CatBoost, observed in this study can be attributed to their ability to capture complex, nonlinear relationships and automatically manage diverse data characteristics, which are critical challenges in construction cost prediction tasks. Construction datasets often include intricate interactions between features such as material type, labor costs, project scale, structural complexity, and economic variables. These variables rarely exhibit linear dependencies, rendering traditional linear models, such as Ridge Regression, Lasso, and Elastic Net, suboptimal due to their inherent assumption of linearity.

Unlike linear regressors, XGBoost and CatBoost are built upon decision tree ensembles that iteratively correct prediction errors through boosting mechanisms. Specifically, XGBoost uses a second-order Taylor expansion of the loss function and incorporates both L1 and L2 regularization, which not only accelerates convergence but also improves generalization by penalizing model complexity. Its ability to handle missing values, incorporate sparsity-aware learning, and parallelize tree construction further enhances its suitability for large, noisy, and high-dimensional construction datasets.

CatBoost, on the other hand, introduces additional innovations tailored to handling categorical features, a crucial characteristic in construction datasets that include variables such as project location, structural type, and material categories. Its use of Ordered Boosting avoids target leakage and overfitting by ensuring that each data point is used for both training and evaluation without introducing artificial bias. Moreover, CatBoost implements gradient bias correction during tree construction, further improving robustness and accuracy even on smaller datasets or when feature distributions are imbalanced.

These strengths explain the high predictive performance observed in this study, where both models achieved R^2 values above 0.98 and RMSE values below 0.5 across the test data. Their architecture enables the capture of subtle dependencies and nonlinear effects that are often overlooked or misrepresented by simpler models. Furthermore, their embedded feature selection and handling of multicollinearity ensure that redundant or weakly correlated inputs do not degrade model performance, a crucial advantage in cost modeling, where correlated features are standard.

In essence, XGBoost and CatBoost are not merely predictive models but comprehensive learning systems capable of adaptive learning, robust generalization, and scalable deployment, making them highly effective for the dynamic and multidimensional nature of construction cost prediction. Their demonstrated superiority in this study highlights the strategic importance of ensemble learning in modern construction informatics and supports their integration into future intelligent cost estimation frameworks.

D. Overfitting Concerns and Practical Deployment Considerations

Although KNN Regression demonstrated promising accuracy on the training dataset, its performance deteriorated substantially on the test dataset, highlighting a classic case of overfitting. This limitation is especially problematic in real-world deployment scenarios where unseen data frequently deviates from the training distribution. KNN's reliance on memorizing training instances and its sensitivity to noise and irrelevant features make it prone to overfitting, especially in high-dimensional spaces. This is evident in the inflated RMSE and reduced R^2 scores observed during testing.

To mitigate such issues, future implementations of KNN or similar non-parametric models should incorporate cross-validation, feature selection, and dimensionality reduction techniques such as Principal Component Analysis (PCA). Additionally, applying distance-weighted variants of KNN or ensemble versions (e.g., bagged KNN) could enhance robustness. However, such improvements often come with increased complexity and diminishing returns, particularly when more sophisticated models are available.

In contrast, while XGBoost and CatBoost exhibit outstanding performance in both training and testing scenarios, with near-zero bias and minimal variance, they introduce a different set of challenges. Their ensemble nature and iterative training process inherently demand greater computational resources and longer training times. This complexity can impede real-time applications in construction project environments, where decisions must often be made rapidly and based on evolving data. Furthermore, model interpretability becomes more difficult as the number of trees and depth increases. While both XGBoost and CatBoost offer feature importance metrics and compatibility with SHAP (SHapley Additive exPlanations) for explainable AI, these tools may still fall short of the intuitive transparency desired by construction project managers or stakeholders with limited technical expertise.

To bridge this gap, future research should focus on streamlining these models for operational use. Techniques such as model pruning, quantization, or the development of surrogate models that approximate ensemble behavior using simpler structures could provide viable solutions. Additionally, integrating these models into cloud-based decision support systems can

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

help alleviate the burden of local computational demands, enabling scalable deployment in industry practice. While ensemble models like XGBoost and CatBoost deliver state-of-the-art accuracy in construction cost prediction, addressing their deployment barriers —namely, computational overhead and interpretability —will be essential for widespread adoption in the industry. Striking a balance between predictive performance and operational feasibility remains a critical area for future exploration.

VI. Conclusion

This study presents a comprehensive evaluation of six machine learning (ML) models, namely Ridge Regression, Lasso Regression, Elastic Net, K-Nearest Neighbors (KNN) Regression, XGBoost, and CatBoost, for predicting construction costs using the standardized RSMeans dataset, which comprises 4,477 data points. Models were evaluated on both training and testing sets using three key metrics: R^2, RMSE, and MBE. Simulation results reveal that ensemble-based models, particularly XGBoost and CatBoost, significantly outperform traditional linear and non-parametric models. With R^2 values exceeding 0.98 and RMSE below 0.5, these models demonstrated exceptional predictive accuracy and generalization, reinforced by their near-zero MBE values. This strong performance highlights their capability to effectively capture complex, nonlinear relationships that simpler models typically overlook. In contrast, Ridge, Lasso, and Elastic Net, although interpretable and computationally efficient, underperformed due to their inability to handle such complex data. KNN Regression, though accurate on the training data, suffered from overfitting and poor generalization.

However, this study acknowledges several limitations. First, the RSMeans dataset is primarily focused on the North American construction context, which may limit the global applicability of the findings. Variations in labor cost structures, material pricing, building codes, and regulatory factors across regions could reduce the accuracy of these models in international contexts. Future research should validate the model on more diverse, region-specific datasets and assess its robustness across economic environments and construction practices. Second, while the computational demands and scalability of XGBoost and CatBoost are not prohibitive, their training time and resource requirements require consideration. Further studies could explore model compression techniques, parallel processing, or hybrid methods that strike a balance between performance and computational efficiency. Moreover, the "black-box" nature of ensemble models poses concerns about transparency and decision traceability in industry applications. To bridge this gap, future work should incorporate explainable AI (XAI) techniques, such as SHAP (SHapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations), to enhance stakeholder trust and model accountability.

Beyond cost prediction, XGBoost and CatBoost also show promise in broader construction management domains. These include resource allocation, project risk assessment, schedule optimization, and sustainability forecasting. Integrating diverse datasets (e.g., sensor data, supply chain dynamics, economic indicators) could improve the robustness and adaptability of these models. In summary, this study contributes to process innovation in construction cost estimation by demonstrating the superiority of ensemble ML models in handling complex, multi-dimensional datasets. To translate this academic insight into industrial impact, future research should focus on interpretability, model scalability, and real-time integration, enabling these models to support intelligent decision-making across the construction lifecycle.

Acknowledgment

This study was funded by the Malaysian Ministry of Higher Education through the Fundamental Research Grant Scheme (FRGS/1/2024/ICT02/UCSI/02/1).

References

[1] Y. Wang, et al., "Cost prediction of building projects using the novel hybrid RA-ANN model," *Eng. Constr. Archit. Manag.*, Jan. 2023.

[2] C. S. Chan, J. Lu, and B. Zhang, "Attaining cost efficiency in constructing sports facilities for Beijing 2008 Olympic Games by use of operations simulation," in *Proc. Winter Simulation Conf.*, Dec. 2006.

[3] W. Jennings, "Why costs overrun: risk, optimism and uncertainty in budgeting for the London 2012 Olympic Games," *Constr. Manag. Econ.*, vol. 30, no. 6, pp. 455–462, Jun. 2012.

[4] D. Blomberg, P. Cotellesso, W. Sitzabee, and A. E. Thal, "Discovery of internal and external factors causing military construction cost premiums," *J. Constr. Eng. Manag.*, vol. 140, no. 3, pp. 04013060, Mar. 2014.

[5] S. Ahn, S. Shokri, S. Lee, C. T. Haas, and R. C. G. Haas, "Effectiveness of interface-management practices in large-scale construction projects," *J. Manag. Eng.*, vol. 33, no. 2, pp. 04016039, Mar. 2017.

[6] O. Swei, J. Gregory, and R. Kirchain, "Construction cost estimation: a parametric approach for better estimates of expected cost and variation," *Transp. Res. Part B Methodol.*,

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

vol. 101, pp. 295–305, Jul. 2017.

[7] H. H. Elmousalami, "Artificial intelligence and parametric construction cost estimate modeling: State-of-the-art review," *J. Constr. Eng. Manag.*, vol. 146, no. 1, pp. 03119008, Jan. 2020.

[8] S.-W. Yang, S.-W. Moon, H. Jang, S. Choo, and S.-A. Kim, "Parametric method and building information modeling-based cost estimation model for construction cost prediction in architectural planning," *Appl. Sci.*, vol. 12, no. 19, pp. 9553, Sep. 2022.

[9] L. H., C. L, and Z. R, "Research on project cost management under the mode of bill of quantities valuation," *Int. J. Front. Eng. Technol.*, vol. 4, no. 2, 2022.

[10] H. Al‐Tabtabai, N. Kartam, I. Flood, and A. P. Alex, "Expert judgment in forecasting construction project completion," *Eng. Constr. Archit. Manag.*, vol. 4, no. 4, pp. 271–293, Apr. 1997.

[11] S. M. AbouRizk, G. M. Babey, and G. Karumanasseri, "Estimating the cost of capital projects: an empirical study of accuracy levels for municipal government projects," *Can. J. Civ. Eng.*, vol. 29, no. 5, pp. 653–661, Oct. 2002.

[12] S. S. Khanal, P. W. C. Prasad, A. Alsadoon, and A. Maag, "A systematic review: machine learning based recommendation systems for e-learning," *Educ. Inf. Technol.*, Dec. 2019.

[13] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, "Explainable, trustworthy, and ethical machine learning for healthcare: A survey," *Comput. Biol. Med.*, vol. 149, no. 106043, p. 106043, Oct. 2022.

[14] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Comput. Electron. Agric.*, vol. 151, pp. 61–69, Aug. 2018.

[15] M. J. Esfandiari and G. S. Urgessa, "Progressive collapse design of reinforced concrete frames using structural optimization and machine learning," *Structures*, vol. 28, pp. 1252–1264, Dec. 2020.

[16] M. Flah, I. Nunez, W. Ben Chaabene, and M. L. Nehdi, "Machine learning algorithms in civil structural health monitoring: A systematic review," *Arch. Comput. Methods Eng.*, vol. 28, no. 4, pp. 2621–2643, Jul. 2020.

[17] H. G. Melhem and Y. Cheng, "Prediction of remaining service life of bridge decks using machine learning," *J. Comput. Civ. Eng.*, vol. 17, no. 1, pp. 1–9, Jan. 2003.

[18] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, and S. P. Ong, "A critical review of machine learning of energy materials," *Adv. Energy Mater.*, vol. 10, no. 8, pp. 1903242, Jan. 2020.

[19] P. Davis, F. Aziz, M. T. Newaz, W. Sher, and L. Simon, "The classification of construction waste material using a deep convolutional neural network," *Autom. Constr.*, vol. 122, pp. 103481, Feb. 2021.

[20] C.-H. Huang and S.-H. Hsieh, "Predicting BIM labor cost with random forest and simple linear regression," *Autom. Constr.*, vol. 118, pp. 103280, Oct. 2020.

[21] G.-H. Kim, J.-E. Yoon, S.-H. An, H.-H. Cho, and K.-I. Kang, "Neural network model incorporating a genetic algorithm in estimating construction costs," *Build. Environ.*, vol. 39, no. 11, pp. 1333–1340, Nov. 2004.

[22] C. Hai, "Construction and application of multiple linear regression model for construction project cost," in *Int. Conf. Advanc. Enterp. Inf. Syst.*, Jun. 2021.

[23] George Harrison Coffie and F. Cudjoe, "Using extreme gradient boosting (XGBoost) machine learning to predict construction cost overruns," *Int. J. Constr. Manag.*, pp. 1–9, Dec. 2023.

[24] D. J. Lowe, M. W. Emsley, and A. Harding, "Predicting construction cost using multiple regression techniques," *J. Constr. Eng. Manag.*, vol. 132, no. 7, pp. 750–758, Jul. 2006.

[25] R. Jafarzadeh, J. M. Ingham, K. Q. Walsh, N. Hassani, and G. R. Ghodrati Amiri, "Using statistical regression analysis to establish construction cost models for seismic retrofit of confined masonry buildings," *J. Constr. Eng. Manag.*, vol. 141, no. 5, pp. 04014098, May 2015.

[26] R. Martin Skitmore and S. Thomas Ng, "Forecast models for actual construction time and cost," *Build. Environ.*, vol. 38, no. 8, pp. 1075–1083, Aug. 2003.

[27] M. W. Emsley, D. J. Lowe, A. R. Duff, A. Harding, and A. Hickson, "Data modelling and the application of a neural network approach to the prediction of total construction costs," *Constr. Manag. Econ.*, vol. 20, no. 6, pp. 465–472, Sep. 2002.

[28] S. M. Shahandashti and B. Ashuri, "Highway Construction Cost Forecasting Using Vector Error Correction Models," *J. Manag. Eng.*, vol. 32, no. 2, p. 04015040, Mar. 2016.

[29] S. Petruseva, V. Z. Pancovska, V. Zujo and A. Brkan-Vejzovic, "Construction costs forecasting: comparison of the accuracy of linear regression and support vector machine models," *Tech. Vjesn.*, vol. 24, no. 5, Oct. 2017.

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

[30] C.-H. Huang and S.-H. Hsieh, "Predicting BIM labor cost with random forest and simple linear regression," *Autom. Constr.,* vol. 118, p. 103280, Oct. 2020.

[31] G. H. Kim, D. Seo, and K.-I. Kang, "Hybrid models of neural networks and genetic algorithms for predicting preliminary cost estimates," *J. Comput. Civ. Eng.*, vol. 19, no. 2, pp. 208–211, Apr. 2005.

[32] M.-Y. Cheng, N.-D. Hoang, and Y.-W. Wu, "Hybrid intelligence approach based on LS-SVM and differential evolution for construction cost index estimation: A Taiwan case study," *Autom. Constr.*, vol. 35, pp. 306–313, Nov. 2013.

[33] O. Alshboul, A. Shehadeh, G. Almasabha, R. E. A. Mamlook, and A. S. Almuflih, "Evaluating the impact of external support on green building construction cost: A hybrid mathematical and machine learning prediction approach," *Buildings*, vol. 12, no. 8, p. 1256, Aug. 2022.

[34] C. Zhang, J. Zhu, T. Shi, and X. Li, "Influence line estimation of bridge based on elastic net and vehicle-induced response," *Meas.*, vol. 202, pp. 111883–111883, Oct. 2022.

[35] A. Shehadeh, O. Alshboul, R. E. Al Mamlook, and O. Hamedat, "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression," *Autom. Constr.*, vol. 129, p. 103827, Sep. 2021.

[36] N. Simić, N. Ivanišević, Đ. Nedeljković, A. Senić, Z. Stojadinović, and M. Ivanović, "Early highway construction cost estimation: Selection of key cost drivers," *Sustainability*, vol. 15, no. 6, p. 5584, Mar. 2023.

[37] 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, May 2022.

[38] G.-H. Kim, J.-M. Shin, S. Kim, and Y. Shin, "Comparison of school building construction costs estimation methods using regression analysis, neural network, and support vector machine," *J. Build. Constr. Plan. Res.*, vol. 01, no. 01, pp. 1–7, Mar. 2013.

[39] M.-Y. Cheng and N.-D. Hoang, "Interval Estimation of Construction Cost at Completion Using Least Squares Support Vector Machine," ," *J. Civ. Eng. Manag.*, vol. 20, no. 2, pp. 223–236, Mar. 2014.

[40] S. Yun, "Performance Analysis of Construction Cost Prediction Using Neural Network for Multioutput Regression," *Appl. Sci.*, vol. 12, no. 19, p. 9592, Sep. 2022.

[41] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, vol. 12, no. 1, pp. 55–67, Feb. 1970, doi: https://doi.org/10.1080/00401706.1970.10488634

[42] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[43] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x

[44] F. VALAFAR, "Pattern Recognition Techniques in Microarray Data Analysis," Annals of the New York Academy of Sciences, vol. 980, no. 1, pp. 41–64, Dec. 2002, doi: https://doi.org/10.1111/j.1749-6632.2002.tb04888.x

[45] 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 - KDD '16, vol. 1, no. 1, pp. 785–794, Aug. 2016, doi: https://doi.org/10.1145/2939672.2939785

[46] Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, "CatBoost: unbiased boosting with categorical features," arXiv (Cornell University), Jun. 2017, doi: https://doi.org/10.48550/arxiv.1706.09516

[47] 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, p. 101201, Oct. 2020, doi: https://doi.org/10.1016/j.aei.2020.101201

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

Author Biography

Lifei Chen is currently a PhD candidate in Engineering at UCSI University, with a research focus on construction cost control and resource optimization. She obtained her Master's degree in Engineering in 2022. During her postgraduate studies, she served as the project leader for a research initiative titled "Construction of the Whole Life Cycle Cost Optimization Model of Construction." Her work provided innovative solutions for enhancing the precision of construction cost control throughout the project lifecycle. Since January 2024, she has been pursuing her doctoral research at the intersection of artificial intelligence and construction cost management. Her current work actively explores the application of AI techniques in construction cost prediction and resource optimization, aiming to improve decision-making and operational efficiency in the construction industry.

Sew Sun Tiang received the Bachelor of Engineering (Hons.) Electronics majoring in Telecommunications from Multimedia University (MMU) in 2008 and PhD degree in Electrical and Electronics Engineering from Universiti Sains Malaysia. She was the Senior Lecturer in the School of Engineering, Asia Pacific University from 2015 to 2027. Dr. Tiang is currently working as an Assistant Professor in the Faculty of Engineering, Technology, and Built Environment, UCSI University. She has published over 90 research articles in the research areas related to antenna design, wireless communication, metaheuristic optimization, machine learning, and deep learning. Dr Tiang is also actively involved in various professional bodies. To date, she has been awarded the qualifications of Chartered Engineer (CEng) from the UK Engineering Council, Professional Engineer (PEng) qualification from the Board of Engineer Malaysia, and Professional Technologist (PTech) qualification from the Malaysia Board of Technologist (MBOT).

Kim Soon Chong has been an Assistant Professor and Head of Department in the Faculty of Engineering at UCSI University in Malaysia since 2023. He received his PhD in Electrical, Electronics & Systems Engineering from Universiti Kebangsaan Malaysia in 2022. Dr Chong is actively involved in various industrial collaborations, and he is currently working with several industrial grants. His main research interests are biomedical and healthcare technologies, machine learning, and deep learning. Dr Chong is also actively involved in various professional bodies. To date, she has been awarded the qualifications of Professional Engineer (PEng) qualification from the Board of Engineer Malaysia and Professional Technologist (PTech) qualification from the Malaysia Board of Technologists (MBOT).

Abhishek Sharma received the bachelor's degree in electronics and communication engineering from ITM-Gwalior, India, in 2012, and the master's degree in robotics engineering from the University of Petroleum and Energy Studies (UPES), Dehradun, India, in 2014. He was a Senior Research Fellow in a DST funded project under the Technology Systems Development Scheme and worked as an Assistant Professor with the Department of Electronics and Instrumentation, UPES. He also worked as a research associate in Ariel university (Israel) and received Emerging Scientist award in 2021. Currently he is working as an Associate Professor in computer science and engineering department (Graphic era deemed to be university, India) and as a guest lecturer in UCSI university, Malaysia. His research interests include machine learning, optimization theory, swarm intelligence, embedded system, control, and robotics.

Tarek Berghout is a distinguished researcher specializing in industrial informatics and manufacturing. He earned both his Master's and Ph.D. degrees from the University of Batna 2, Algeria, completing his doctoral studies in 2021. Currently, Dr. Berghout serves as the Chief Laboratory Technician at the University of Batna 2, where he focuses on developing machine learning algorithms for condition monitoring, predictive maintenance, and cybersecurity. Throughout his academic career, Dr. Berghout has made significant contributions to the fields of machine learning and deep learning, particularly in their applications to industrial processes. His research interests encompass condition monitoring, cybersecurity, and the use of MATLAB for developing predictive models. He has authored numerous publications that delve into these areas, reflecting his commitment to advancing knowledge in industrial engineering.

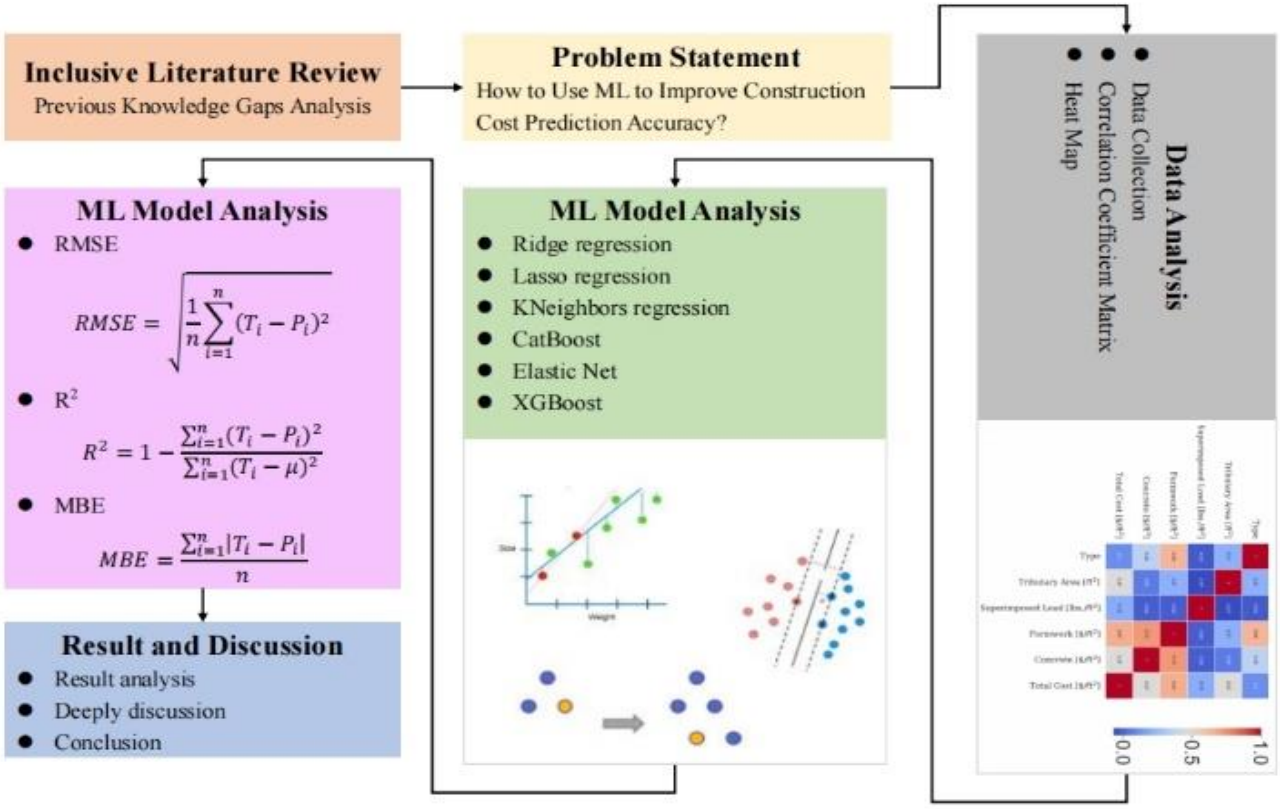
Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

Wei Hong Lim is currently an Associate Professor and a researcher at the Faculty of Engineering, Technology and Built Environment in UCSI University. He obtained his BEng (Hons) in Mechatronic Engineering and Ph.D. in Computational Intelligence from Universiti Sains Malaysia, Penang, Malaysia in the year 2011 and 2014, respectively. Dr. Lim was affiliated with the Intelligent Control Laboratory at National Taipei University of Technology, Taiwan as a Postdoctoral Researcher from 2015 to 2017 and as a Visiting Researcher in 2019. He has published more than sixty research articles in research areas related to computational intelligence, metaheuristic search optimization algorithms, deep learning, machine learning, energy management, digital image processing, among others. He is an active academic editor and reviewer for various reputable journals. Dr. Lim is also involved in multipleprofessional bodies. To date, he has been awarded with the qualifications of Chartered Engineer (CEng) and International Professional Engineer in UK Section (IntPE(UK)) from UK Engineering Council, European Engineer qualification from European Federation of National Engineering Associations (FEANI), Professional Engineer (PEng) qualification from Board of Engineer Malaysia, Professional Technologist (PTech) qualification from Malaysia Board of Technologist (MBOT) and Senior Membership (SMIEEE) from IEEE.

Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799

Graphical Abstract

This study systematically compares six ML models (Ridge, Lasso, Elastic Net, KNN, XGBoost, CatBoost) for construction cost prediction using RSMeans data. Results demonstrate XGBoost and CatBoost's superiority (R^2>0.98, RMSE<0.5) in modeling nonlinear relationships, outperforming linear regressors and KNN in accuracy and robustness.



Manuscript received March 8, 2025; Accepted May 10, 2025; date of publication May 20, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.799