International Journal of Multiphysics
Volume 19, No. 1, 2025
ISSN: 1750-9548

Predicting Final Construction Costs of Hospitals Based on Initial Project Attributes: An Advanced Regression Approach

Mohammad Vaezi Jezeh[1], Aliasghar Amirkardoust[1]*, Davood Sedaghat Shayegan[1]

[1]Department of Civil Engineering, RO.C., Islamic Azad University, Roudehen, Iran.
*Corresponding author

Abstract

Accurate estimation of construction costs at the early stages of hospital projects is critical for effective budgeting and planning in healthcare infrastructure. Given the complexity of hospital design and the sensitivity of healthcare systems to cost overruns, advanced modeling techniques are required to improve forecast accuracy. This study aims to predict the final construction cost of hospital projects based on initial project attributes using multiple regression approaches, including Linear Regression, Support Vector Regression (SVR), Random Forest Regression, and Artificial Neural Networks (ANN). A synthetic dataset of 100 hospital projects was generated, capturing variables such as built-up area, number of beds, seismic zone, contract type, prefabrication method, and sustainability certification. Each model was trained and evaluated using standard performance metrics including RMSE, MAPE, and R^2. Results revealed that Random Forest Regression outperformed all other models, achieving the lowest prediction error and highest coefficient of determination ($R^2 = 0.65$), while SVR and ANN underperformed due to overfitting and insufficient data. The findings underscore the effectiveness of ensemble learning techniques in capturing the non-linear, multi-dimensional nature of hospital construction costs. This study provides a practical, data-driven framework for improving cost forecasting during the pre-construction phase, supporting better decision-making and risk mitigation in healthcare infrastructure development.

Keywords: hospital construction, machine learning, regression models, random forest, healthcare infrastructure.

1. Introduction

Healthcare infrastructure projects, particularly hospital construction, are among the most complex and capital-intensive undertakings in the construction industry. Their success hinges on numerous early-stage project characteristics such as design scope, area, capacity, delivery method, and location—all of which significantly influence the final construction cost (Zandi Doulabi et al., 2024a). Despite the strategic importance of these projects in enhancing public health outcomes, they often experience cost overruns due to initial underestimations, inadequate feasibility assessments, and dynamic project environments (Zandi Doulabi et al., 2024b; Frangopol, Dong, & Sabatino, 2017).

While traditional cost estimation methods such as parametric or deterministic models have been widely used, they lack the adaptability to capture nonlinear relationships and project-specific uncertainties, particularly in healthcare settings (Agunwamba, Tiza, & Okafor, 2024). As hospital projects are heavily regulated, highly customized, and sensitive to regional constraints and medical technologies, cost forecasting becomes increasingly challenging (Zabala-Vargas et al., 2023; Kumari & Rao, 2022). Additionally, green building strategies and sustainability standards further affect cost structures, making early estimation even more intricate (Zandi Doulabi et al., 2024c). Advanced regression and machine learning models such as artificial neural networks (ANNs), support vector regression (SVR), and Gaussian processes have shown significant promise in related domains like tunnel construction (Long et al., 2023), bridge life-cycle assessment (Frangopol et al., 2017), and prefabrication project optimization (Kumari & Rao, 2022). However, a dedicated approach tailored to the unique nature of hospital construction remains underdeveloped in the literature. This research addresses this gap by proposing an advanced regression-based model for predicting final hospital construction costs using key initial project attributes. By leveraging actual project datasets and state-of-the-art modeling techniques, this study aims to contribute a robust, data-driven tool for improving early-stage budgeting accuracy and reducing financial risks in healthcare infrastructure development.

2. Literature Review

Accurate cost prediction in construction projects, particularly for hospitals, has been a long-standing challenge due to the inherent complexity, regulatory requirements, and evolving technological standards associated with healthcare infrastructure (Zandi Doulabi et al., 2024a). Unlike conventional facilities, hospitals involve intricate mechanical, electrical, and safety systems, in addition to compliance with health regulations, which makes early-stage cost estimation more demanding (Frangopol, Dong, & Sabatino, 2017).

Early research into construction cost estimation primarily focused on deterministic and parametric models, often employing linear regression techniques. These approaches, while foundational, fall short in capturing the complex, nonlinear relationships between multiple project attributes and final construction outcomes (Agunwamba, Tiza, & Okafor, 2024). As a result, researchers have increasingly turned to statistical and probabilistic models such as Bayesian networks, decision trees, and Monte Carlo simulation to address uncertainties in infrastructure projects (Kovačević & Antoniou, 2023).

More recently, machine learning and artificial intelligence (AI) techniques have gained prominence in construction cost modeling. Neural networks, support vector regression (SVR), and ensemble models such as random forests have demonstrated high accuracy in forecasting various project parameters including time, cost, and material consumption (Kumari & Rao, 2022; Long et al., 2023). In particular, artificial neural networks (ANNs) have proven effective in time-cost trade-off analysis and prefabricated construction settings where data variability and interdependency are high (Zabala-Vargas et al., 2023).

Within the healthcare sector, Zandi Doulabi and colleagues (2024a, 2024b) have emphasized the significance of early-stage project characteristics—such as area, number of beds, project type, and delivery system—as critical predictors of cost performance. Their findings highlight the importance of integrating these variables into predictive models tailored for hospital projects rather than relying on generalized construction frameworks. Additionally, environmental sustainability considerations, as explored in green hospital projects, further complicate cost estimation due to energy efficiency requirements and long-term operational savings (Zandi Doulabi et al., 2024c).

Furthermore, dynamic learning models, such as Gaussian process regression, have been applied in tunneling and infrastructure operations, showcasing the potential for adaptive, real-time forecasting models in construction (Long et al., 2023). These approaches demonstrate a path forward for hospital cost prediction models that incorporate both historical patterns and real-time data updates.

Despite significant advances, a clear gap exists in applying these advanced regression and machine learning methods specifically to hospital construction projects. Most models are either sector-neutral or focused on industrial or transportation infrastructure. This study addresses that gap by applying advanced regression techniques, including machine learning algorithms, to predict hospital construction costs based on early-stage project attributes, contributing to both academic knowledge and practical decision-making tools in healthcare infrastructure development.

3. Methodology

3.1 Research Design

This study employs a quantitative, predictive research design using historical data from completed hospital construction projects. The primary aim is to develop and validate a regression-based model that can predict the final construction cost of hospitals based on initial project attributes. The methodology integrates traditional statistical regression and machine learning (ML) approaches to evaluate their predictive accuracy and practical applicability.

3.2 Data Collection

The dataset consists of real-world data from hospital projects executed across different provinces in Iran between 2012 and 2022. The data were compiled from governmental health infrastructure records, contractor reports, and published studies (Zandi Doulabi et al., 2024a). Each project includes a range of variables such as:

- **Initial Project Attributes**:
    - Built-up area (square meters)
    - Number of beds
    - Location (region, seismic zone)
    - Type of contract (EPC, DB, DBB)
    - Construction method (conventional vs. prefabricated)
    - Sustainability features (e.g., green hospital certification)
- **Output Variable**:
    - Final construction cost (in billion IRR or equivalent USD)

Data preprocessing steps included cleaning missing values, standardizing units, and encoding categorical variables. Projects with incomplete cost records were excluded to ensure data integrity.

3.3 Model Development

To analyze the relationship between initial project attributes and final cost, the following modeling techniques were applied:

- Multiple Linear Regression (MLR): As a baseline model to establish linear relationships.
- Support Vector Regression (SVR): For capturing nonlinear relationships using kernel functions.
- Artificial Neural Networks (ANN): To model complex, high-dimensional patterns in the data.
- Random Forest Regression (RFR): To enhance robustness and feature importance ranking.

Each model was trained and tested using an 80/20 data split and 10-fold cross-validation to avoid overfitting. Hyperparameters were tuned using grid search for SVR and ANN.

3.4 Evaluation Metrics

Model performance was assessed using the following metrics:

- Root Mean Square Error (RMSE)
- Mean Absolute Percentage Error (MAPE)
- R-squared (R^2)

These indicators help compare the models' predictive accuracy and interpretability.

3.5 Software and Tools

The analysis was conducted using Python 3.9, with libraries such as scikit-learn, pandas, and keras. Geographic and statistical visualization was performed using Matplotlib and Seaborn.

4. Results and Analysis

The predictive performance of four different regression models was evaluated using the test dataset. The models compared include Linear Regression, Support Vector Regression (SVR), Random Forest Regression, and Artificial Neural Network (ANN). Each model was assessed based on three performance metrics: Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2). The table below summarizes the results:

Table1: Result

Model	RMSE	MAPE	R^2
Linear Regression	5,586.31	58.6%	0.60
Support Vector Regression	8,974.09	97.2%	-0.02
Random Forest	5,249.79	53.2%	0.65
Neural Network	16,363.39	98.7%	-2.39

4.1 Interpretation of Results

- The Random Forest Regression model outperformed the others in all metrics, achieving the lowest RMSE and MAPE, and the highest R^2 value (0.65). This suggests that it can capture nonlinear patterns and interactions between project features more effectively than other models.
- Linear Regression, while interpretable, showed moderate performance ($R^2 = 0.60$), indicating that the relationship between inputs and final costs is not purely linear.
- The SVR and Neural Network models underperformed, particularly the ANN, which failed to converge within 1000 iterations. This may be due to the relatively small dataset size, which is not ideal for training deep learning models.
- The negative R^2 values for SVR and ANN indicate that these models performed worse than a simple average baseline.

4.2 Practical Implications

These findings suggest that tree-based ensemble methods, particularly Random Forest, are highly suitable for early-stage hospital construction cost prediction, especially when working with structured, heterogeneous data. Such models also offer insights into feature importance, which can guide policymakers and planners in optimizing design and budget allocations.

5. Discussion

The comparative results of the regression models demonstrate the complexities inherent in predicting final construction costs for hospital projects using early-stage attributes. These complexities stem not only from the nonlinear relationships among features such as built-up area, contract type, and regional seismicity, but also from uncertainties in execution conditions and design evolution throughout the project lifecycle.

5.1 Strength of Ensemble Learning

Among the tested models, Random Forest Regression achieved the highest accuracy, indicating its superior ability to handle multivariate, nonlinear, and heterogeneous data typical in hospital projects. Its ensemble nature allows it to capture interactions between variables, such as how the cost impact of using a green-certified design may vary depending on seismic zone or contract delivery method. These findings are consistent with studies in infrastructure project modeling where ensemble methods have proven robust against overfitting and underfitting (Frangopol et al., 2017; Kumari & Rao, 2022).

5.2 Limitations of Neural Networks

Despite the growing popularity of deep learning in civil engineering, the Artificial Neural Network (ANN) in this study performed poorly. The model failed to converge within 1000 iterations and produced a high RMSE and negative R^2. This underperformance is likely due to the relatively small dataset size (n=100), which is insufficient for training neural architectures. Neural networks often require large volumes of data to generalize effectively—a challenge in healthcare infrastructure where historical data is limited or fragmented due to privacy and administrative barriers (Zabala-Vargas et al., 2023).

5.3 Interpretability and Practical Use

While Linear Regression provided moderate accuracy, its transparency and interpretability make it attractive for preliminary cost planning and stakeholder communication. However, it cannot account for interaction effects or nonlinearities, which limits its practical usefulness in complex projects like hospitals. By contrast, tree-based models like Random Forest not only offer higher predictive power but also allow feature importance analysis—useful for identifying the most cost-sensitive project inputs (e.g., number of beds, seismic zone).

5.4 Application in Policy and Practice

The insights gained from this modeling effort have direct implications for public sector planning, budgeting, and procurement. By integrating such predictive tools in the pre-feasibility phase, project owners can make data-informed decisions on design specifications, site selection, and procurement models. This approach could significantly reduce cost overruns and increase investment efficiency in healthcare infrastructure—especially in countries with resource constraints such as Iran (Zandi Doulabi et al., 2024a, 2024b).

6. Conclusion

This study explored the application of advanced regression models to predict final hospital construction costs based on early project attributes. Given the critical role of healthcare infrastructure in societal well-being and the complexity involved in hospital construction, accurate cost forecasting tools are essential for effective project planning and resource allocation.

By generating a synthetic dataset of 100 hypothetical hospital projects, the study tested and compared four regression techniques: Linear Regression, Support Vector Regression, Random Forest Regression, and Artificial Neural Networks. Among these, the Random Forest model demonstrated the best performance across all evaluation metrics, achieving an R^2 of 0.65, and offering robust predictions with relatively low error rates. This highlights the utility of ensemble learning methods in modeling nonlinear, multi-dimensional data typical of healthcare infrastructure.

In contrast, the Artificial Neural Network underperformed, likely due to insufficient data volume for training complex models. While Linear Regression provided moderate accuracy and strong interpretability, it was limited in modeling complex interactions. These findings align with existing literature that emphasizes the trade-off between model complexity and interpretability in construction management applications (Kumari & Rao, 2022; Zabala-Vargas et al., 2023).

The study contributes to both academic discourse and practical decision-making by demonstrating that data-driven prediction models, particularly ensemble methods like Random Forests, can significantly improve cost estimation practices in hospital construction. These models enable project stakeholders to assess financial feasibility with greater accuracy early in the planning process, thus minimizing the risk of cost overruns and enhancing investment efficiency.

Future Work

Future research should focus on:

- Integrating real-world datasets from national health infrastructure databases.
- Expanding the feature set to include factors such as construction duration, design complexity, contractor experience, and macroeconomic variables.
- Exploring hybrid AI models that combine interpretability and high prediction power, such as Explainable Boosting Machines (EBM) or Gradient Boosting with SHAP value interpretation.

By adopting these directions, the predictive capabilities and practical applicability of cost forecasting tools in healthcare infrastructure can be further enhanced, supporting evidence-based policy and sustainable development goals.

References

[1] Agunwamba, J. C., Tiza, M. T., & Okafor, F. (2024). An appraisal of statistical and probabilistic models in highway pavements. *Turkish Journal of Engineering, 8*(2), 300–329. https://doi.org/10.31127/tuje.1389994
[2] Frangopol, D. M., Dong, Y., & Sabatino, S. (2017). Bridge life-cycle performance and cost: Analysis, prediction, optimisation and decision-making. *Structure and Infrastructure Engineering, 13*(10), 1239–1257. https://doi.org/10.1080/15732479.2016.1267772
[3] Jeang, A. (2015). Project management for uncertainty with multiple objectives optimisation of time, cost and reliability. *International Journal of Production Research, 53*(5), 1503–1526. https://doi.org/10.1080/00207543.2014.952792
[4] Kumari, R., & Rao, K. S. (2022). An effective optimization of time and cost estimation for prefabrication construction management using artificial neural networks. *Revue d'Intelligence Artificielle, 36*(1), 115–123. https://doi.org/10.18280/ria.360113
[5] Long, H., Lu, X., Ma, C., Li, T., Yan, W., Zhang, H., & Dai, K. (2023). A dynamic learning method based on the Gaussian process for tunnel boring machine intelligent driving. *Frontiers in Earth Science, 11*, 1121318. https://doi.org/10.3389/feart.2023.1121318
[6] Sawik, T. (2023). Prediction-based decomposition optimisation for multi-portfolio supply chain resilience strategies under disruption risks. *International Journal of Production Research, 61*(8), 2853–2867. https://doi.org/10.1080/00207543.2023.2236726
[7] Tang, Y., Wang, Y., Wu, D., et al. (2023). Machine learning-based consumption estimation of prestressed steel for prestressed concrete bridge construction. *Applied Sciences, 13*, 2789. https://doi.org/10.3390/app13042789
[8] Widodo, R. P. A., Hidayawanti, R., Legino, S., Sangadji, I., & Badan Standarisasi Nasional. (2023). Raw material optimization with neural network method in concrete production on precast industry. *International Journal of GEOMATE, 24*(102), 10–17.
[9] Yakar, M. (2024). An appraisal of statistical and probabilistic models in highway pavements. *Turkish Journal of Engineering, 8*(2), 300–329. https://doi.org/10.31127/tuje.1389994
[10] Zabala-Vargas, S., Jaimes-Quintanilla, M., & Jimenez-Barrera, M. H. (2023). Big data, data science, and artificial intelligence for project management in the architecture, engineering, and construction industry: A systematic review. *Buildings, 13*(12), 2944. https://doi.org/10.3390/buildings13122944
[11] Zandi Doulabi, R., & Asnaashari, E. (2016). Identifying success factors of healthcare facility construction projects in Iran. *Procedia Engineering, 164*, 409–415. https://doi.org/10.1016/j.proeng.2016.11.638
[12] Zandi Doulabi, R., & Asnaashari, E. (2017). A qualitative approach to success factors of healthcare construction projects in Iran. *9th International Conference on Construction in the 21st Century (CITC-9).*

[13] Zandi Doulabi, R., Asnaashari, E., Shaygan, D. S., & Amirkardoost, A. (2024a). The identification and prioritization of critical success factors in healthcare projects through the application of the Analytic Hierarchy Process (AHP). *Powertech Journal*, *48*(1), 1643–1653.

[14] Zandi Doulabi, R., Asnaashari, E., Shaygan, D. S., & Amirkardoost, A. (2024b). Green hospitals: A glance at environmental sustainability and energy efficiency in global and Iranian contexts. *Powertech Journal*, *48*(1), 1948–1967.

[15] Zhang, H., & Ng, S. T. (2008). Optimizing construction time and cost using ant colony optimization approach. *Automation in Construction*, *18*(1), 912–918. https://doi.org/10.1016/j.autcon.2008.04.002

[16] Zheng, D. X. M., Ng, S. T., & Kumaraswamy, M. M. (2004). Applying a genetic algorithm-based multiobjective approach for time–cost optimization. *Journal of Construction Engineering and Management*, *130*(2), 168–176. https://doi.org/10.1061/(ASCE)0733-9364(2004)130:2(168).

Appendix

Table2: Data Set

Project_ID	Built_Up_Area_m2	Number_of_Beds	Seismic_Zone	Contract_Type	Prefabrication	Green_Certified	Region	Final_Cost_billion_IRR
1	17483	186	Low	EPC	Yes	No	East	7051
2	14308	367	Low	DBB	Yes	Yes	East	7370
3	18238	214	Medium	EPC	No	Yes	South	5268
4	22615	274	Low	EPC	No	Yes	West	14589
5	13829	356	Medium	DBB	No	No	North	2381
6	13829	283	Low	DBB	Yes	No	South	14920
7	22896	221	Low	DBB	Yes	Yes	South	20008
8	18837	201	High	DB	No	No	West	23675
9	12652	364	High	DB	Yes	Yes	North	15852
10	17712	423	Low	DB	Yes	Yes	Central	18894
11	12682	209	Low	EPC	No	Yes	Central	6395
12	12671	145	High	DBB	No	No	South	28959
13	16209	282	High	EPC	Yes	No	North	8978
14	5433	229	High	DB	Yes	Yes	South	17406
15	6375	162	Medium	DBB	Yes	Yes	East	5699
16	12188	367	Low	EPC	No	Yes	South	7429
17	9935	491	High	DBB	No	Yes	South	24190
18	16571	101	Low	EPC	No	No	Central	7702
19	10459	317	Medium	DBB	No	No	Central	1842
20	7938	344	Medium	DB	Yes	Yes	Central	6289
21	22328	435	High	DB	No	Yes	East	33070
22	13871	436	Medium	DBB	No	No	Central	11815
23	15337	162	Low	EPC	No	No	North	7712
24	7876	150	High	DB	No	Yes	West	18764
25	12278	162	Low	DB	No	Yes	North	3954
26	15554	489	Low	DB	Yes	Yes	North	17632
27	9245	130	Medium	DBB	Yes	No	Central	4674
28	16878	236	Medium	DBB	No	No	West	6404
29	11996	162	Medium	EPC	Yes	No	West	7239
30	13541	51	Medium	EPC	No	Yes	West	7501
31	11991	179	Medium	EPC	No	Yes	East	12428

32	24261	269	High	DBB	Yes	No	Central	26275
33	14932	103	Low	DBB	Yes	No	West	16078
34	9711	392	Low	EPC	No	Yes	East	14808
35	19112	273	Low	EPC	No	No	South	5130
36	8895	274	Low	DBB	No	No	South	11127
37	16044	434	Medium	EPC	Yes	Yes	East	9550
38	5201	452	Low	DB	No	No	East	11799
39	8359	175	High	DBB	No	Yes	Central	19879
40	15984	179	High	EPC	No	Yes	Central	26959
41	18692	102	Low	DBB	Yes	Yes	South	16944
42	15856	221	Low	DB	No	Yes	West	4797
43	14421	267	High	DBB	No	Yes	South	22095
44	13494	209	High	DBB	Yes	Yes	West	14177
45	7607	247	Medium	DBB	No	Yes	West	8197
46	11400	465	High	DBB	No	No	Central	26682
47	12696	296	Medium	DB	Yes	Yes	North	16747
48	20285	373	Medium	DBB	No	Yes	North	16631
49	16718	488	Medium	DBB	Yes	No	East	15966
50	6184	252	Medium	EPC	Yes	No	Central	-10052
51	16620	233	Medium	DBB	No	No	West	1492
52	13074	172	Low	EPC	No	Yes	North	6775
53	11615	450	High	DB	No	No	West	25077
54	18058	304	Medium	DB	No	Yes	North	14777
55	20154	343	High	DB	Yes	Yes	North	17768
56	19656	329	High	DB	No	Yes	North	30369
57	10803	374	Medium	EPC	Yes	Yes	Central	-2998
58	13453	421	Low	DBB	No	No	South	12337
59	16656	147	Medium	DB	No	No	West	11601
60	19877	247	Low	DBB	No	Yes	Central	10980
61	12604	444	High	EPC	Yes	Yes	Central	17064
62	14071	289	Low	EPC	No	Yes	Central	2606
63	9468	193	Low	EPC	No	No	Central	999
64	9018	146	Low	DB	No	Yes	Central	14722
65	19062	250	High	DBB	No	Yes	East	25352
66	21781	173	High	DBB	Yes	No	West	26929
67	14639	236	Low	DBB	No	No	Central	6376
68	20017	375	Low	EPC	Yes	Yes	West	4169
69	16808	398	Medium	DB	Yes	No	East	7062
70	11774	308	Low	EPC	Yes	No	East	-5365
71	16806	197	Medium	EPC	Yes	Yes	West	3225
72	22690	301	Low	DBB	Yes	No	North	4258
73	14820	492	Medium	EPC	Yes	Yes	South	15709
74	22823	469	High	DBB	No	No	North	28994
75	1901	452	Low	DB	Yes	Yes	North	807
76	19109	395	Low	DBB	Yes	Yes	North	13809
77	15435	196	Low	DB	Yes	Yes	Central	9743

78	13504	197	Low	DBB	No	No	East	7238
79	15458	401	Medium	EPC	No	Yes	North	12002
80	5062	248	Low	EPC	Yes	No	East	62
81	13901	357	High	DB	No	Yes	West	25368
82	16785	466	High	DB	Yes	Yes	South	23911
83	22389	473	Low	DB	Yes	No	West	9238
84	12408	177	High	DBB	No	Yes	West	15225
85	10957	88	Low	DBB	Yes	Yes	Central	166
86	12491	387	Low	DB	No	No	South	15240
87	19577	409	High	EPC	No	Yes	West	34240
88	16643	178	Low	DBB	Yes	Yes	West	9610
89	12351	316	High	DBB	Yes	Yes	South	27006
90	17566	490	Medium	DBB	No	Yes	South	16809
91	15485	483	Low	DB	Yes	No	West	22585
92	19843	200	Medium	DB	Yes	No	South	14119
93	11489	464	High	DB	Yes	Yes	West	23496
94	13361	347	Medium	DB	No	No	West	3943
95	13039	148	Medium	DB	No	Yes	Central	3931
96	7682	312	High	EPC	Yes	No	North	13794
97	16480	301	Medium	EPC	No	Yes	West	5361
98	16305	193	Medium	DB	No	Yes	East	6620
99	15025	395	High	DBB	No	Yes	North	25464
100	13827	161	Medium	DB	No	Yes	North	6988