


## Article

# Construction Cost Prediction Using Deep Learning with BIM Properties in the Schematic Design Phase

DoYoon Park and SeokHeon Yun \* 

Department of Architectural Engineering, Gyeongsang National University, Jinju 52828, Republic of Korea; hebelube@gnu.ac.kr

\* Correspondence: gfyun@gnu.ac.kr; Tel.: +82-55-772-1755

**Abstract:** In the planning and design stage, it is difficult to accurately predict construction costs only by estimating approximate cost. It is also very difficult to predict the change in construction costs whenever the design changes. However, using the BIM model's attribute information and machine learning techniques, accurate construction costs can be predicted faster than when using the existing approximate cost estimate. In this study, building information such as 'total area', 'floor water', 'usage', and BIM attribute information such as 'wall area', 'wall water', and 'floor circumference' were used together to predict construction costs in the schema design stage. As a result of applying the machine learning technique using both the building design information and the BIM model attribute information, it was found that the construction cost was improved compared to the result of individual predictions of the building information or BIM attribute information. While accurately predicting construction costs using BIM's attribute information has its limits, it is expected to provide more accuracy compared to predicting costs solely based on construction cost influencing factors.

**Keywords:** construction cost estimation; schematic design; deep learning; BIM



**Citation:** Park, D.; Yun, S.

Construction Cost Prediction Using Deep Learning with BIM Properties in the Schematic Design Phase. *Appl. Sci.* **2023**, *13*, 7207. <https://doi.org/10.3390/app13127207>

Academic Editors: Mauro Lo Brutto and Jürgen Reichardt

Received: 25 March 2023

Revised: 11 June 2023

Accepted: 14 June 2023

Published: 16 June 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Background

The architectural design process is generally divided into four stages: 'Pre-Design', 'Schematic Design', 'Design Development', and 'Working Drawing'. Accurately predicting construction costs during the 'Pre-Design' stage is crucial for evaluating project feasibility. While the 'Unit Cost of Construction', which is released annually for existing building types, is commonly used for construction cost estimation during the 'Pre-Design' or 'Schematic Design' stages, its accuracy is lower than that of 'Detailed Cost Estimate', with a margin of approximately 10%. The approximate cost estimate is used to predict construction costs for projects that have not yet been finalized. The estimated estimate is calculated using data from the completed project, and the error range is approximately 10–25% [1].

'Detailed cost estimation' is not available until 'Design Development', where the detailed components of all buildings are determined through drawings, because construction-related details are not provided for 'detailed cost calculation' at the 'Pre-Design' and 'Schematic Design' stages. Therefore, accurate cost prediction in the early stages of the architectural design process is a challenging task.

The 'unit area method' is the only way to predict construction costs in the current early stage of the architectural design process. Accordingly, various studies are being conducted to increase the accuracy of cost prediction using regression analysis, artificial neural networks, genetic algorithms, and surveys.

If the architectural design is changed, the "detailed construction cost estimate" should also be revised as a whole, so a simple and quick way to predict construction costs is needed instead of detailed construction cost estimates.

## 2. Literature Review

The objective of this paper is to review and synthesize prior research on construction cost prediction. The paper specifically focuses on identifying the direction and method of research related to predicting construction costs in response to changes in design content during the 'Schematic Design' stage of the architectural design process. Through the review of research cases, it is expected that the most effective and appropriate approach to predict construction costs according to changes in design contents in the 'Schematic Design' stage will be derived.

In order to improve the accuracy of the estimate, Koo et al. [2] analyzed the performance data of a high-rise office building project and presented a model for calculating construction closing costs based on the Object and Parameter based Schematic Estimation Model (OPSEM). In the study, necessary activities were defined based on the properties of the object, and construction costs were calculated by applying construction costs to these tasks.

Kim et al. [3] analyzed the efficiency of BIM estimation by comparing the working time in the 2D-based cost estimate and the BIM-based cost estimate for the same project.

Sung [4] developed a model for predicting asthma occurrence based on environmental data and frequency of asthma occurrence, and trained a deep learning model, showing higher accuracy than the existing model.

Pei-Ying Wang et al. [5] developed a deep learning model to predict housing prices, in which features were extracted from input data by combining convolutional neural network (CNN) and long short-term memory (LSTM) networks, and predictive performance was improved using a joint self-attention mechanism.

Jo and Yun [6] proposed an improved approximate cost prediction model for the 'Pre-Design' stage by identifying the optimal combination of influencing factors using regression analysis and the Pearson correlation coefficient. They found that the accuracy of the prediction model was significantly improved when using 'total area' and 'building area' as influencing factors. The study utilized the Pearson correlation coefficient to determine the degree of correlation between the influencing factors and construction costs.

Jung et al. [7] conducted a study on predicting construction costs for 'Smart Education Facilities' in the 'Schematic Design' stage, where the 'Unit cost of construction' technique was found to be inadequate due to the small construction samples involved. They used multiple regression analysis, DNN, and DBN models to predict construction costs, and compared and analyzed the error rates of each model. To ensure the reliability of the results, they employed the 'k-fold' technique. The study found that using the DBN model could reduce the problem of overfitting that could occur when less data are used. In the study, the authors used variables such as 'Total Area', 'Building Area', 'Land Area', 'Underground Floor', 'Underground Floor Area', 'Order Method', 'School Classification', 'Number of Classrooms', and 'Construction Period'. However, in this study, DNN was used as an artificial intelligence model to check the possibility of using BIM properties, and the 'construction period' variable was excluded from the analysis as it was difficult to specify in the 'Schematic Design'.

Kwon et al. [8] developed a construction cost prediction model for public apartment buildings in the 'Schematic Design' stage. They used regression analysis and selected factors that influence construction costs, such as 'Total Area', 'Building Area', 'Land Area', 'Construction Area', 'Ground Floor Area', 'Underground Floor Area', and 'Ground floor'. They analyzed the 'Pearson correlation coefficient' for 'Total Construction Cost' and found that variables such as 'Building Area', 'Ground Floor Area', 'Number of Households per Floor', 'Total Household', 'Perimeter of the Building', 'Exterior Finish Area', 'Land Area', 'Parking Lot Area', 'Underground Parking lot Floor', and 'Number of Parking spaces' had an overall coefficient higher than 0.9. However, some items, such as 'Perimeter of the Building' or 'Exterior Finish Area', which may fluctuate significantly with design changes in the 'Schematic Design' stage, were also included. The study was focused on using influencing factors that could change frequently according to design changes, such as

‘Perimeter of the Building’ or ‘Exterior Finish Area’, rather than factors that do not change much in the ‘Schematic Design’ stage, such as ‘Total Area’ and ‘Building Area’.

Günaydın [9] proposed a cost estimation system for low-rise steel structures using ANN, and defined area, circumference length, height, and load as variables for estimating construction costs. When ANN was used compared to the result of using regression, the error rate was reduced by about 4%. Although ANN performs better than simple regression methods in cost estimation, its accuracy may vary depending on the configuration of the neural networks [10]. Mattel et al. [11] stated that the ANN method was inspired by the process in which the human brain operates. Polat [1] stated that ANN finds the correlation weight of the hidden layer in the relationship between the input layer and the output layer, and that it is a method of predicting the result value through the calculated weight.

Yang [12] emphasized the importance of economic feasibility and cost analysis in the early stages of architectural planning, highlighting their impact on project management and implementation. The study proposed a novel cost estimation model that combines Building Information Modeling (BIM) and parametric methodology. Unlike traditional cost estimation methods, this model takes into account the interactions between architectural elements, enabling more accurate cost prediction. The effectiveness of the proposed model has been demonstrated through its application in real construction projects.

Vitasek et al. [13] discussed the feasibility and adoption of Building Information Modeling (BIM) for cost estimation in the construction industry. The paper proposed methods for performing cost estimation using BIM and presented a process schema for this purpose. It also provided case studies illustrating how BIM is currently being utilized in the field of cost estimation. Furthermore, the document discussed measures necessary to maximize the utilization of BIM.

### 3. Purpose and Methodology

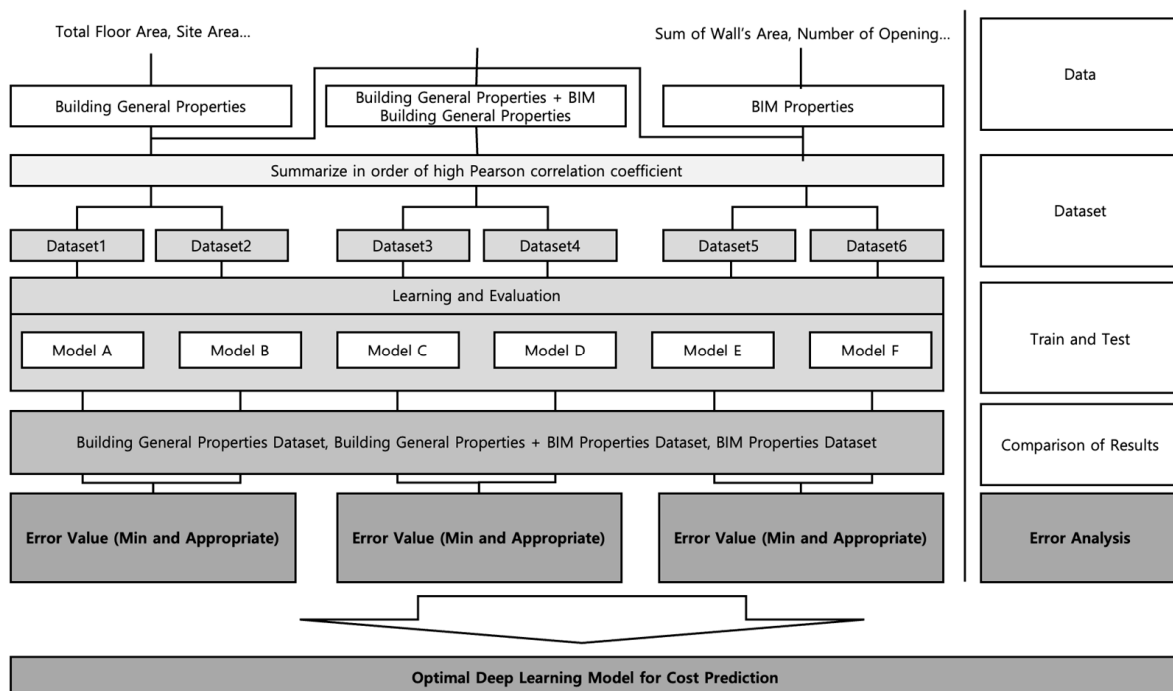
#### 3.1. Purpose of Research

Various factors influence the design of a building, such as “Area”, “Number of Floors”, and “Use of Building”, which can be easily extracted from the “Building General Properties” during the pre-design stage. Once determined, these factors usually remain relatively stable, even if the shape of the building undergoes changes during the schematic design stage. However, frequent changes may occur in the placement of space and building design during the schematic design stage. Therefore, the designer must have the ability to promptly and accurately predict the total construction cost based on the design elements to ensure the feasibility and success of the project.

This paper focuses on the use of Building Information Modeling (BIM) during the schematic design stage, which enables real-time extraction of data for each object of the building that may change with design modifications. The paper aims to identify the factors that can be extracted from BIM models and frequently undergo changes in response to design alterations. The objective of this study is to compare the accuracy of predicting total construction costs using the influencing factors extracted from BIM models with those obtained from the building general properties such as “total area” and “building heights”.

#### 3.2. Research Methodology

This study was structured around the following steps. First, the influencing factors that can affect the total construction cost were investigated and classified into two groups: those entered in the construction outline, and those obtainable from the BIM model. Second, data on the influencing factors and the total construction costs, which are dependent variables, were extracted and compiled from completed construction projects. Third, the influencing factors collected were summarized into several combination datasets to facilitate the machine learning process. Fourth, a machine learning model for predicting construction costs was prepared. Finally, datasets were applied to machine learning models to analyze error rates for each model. Details of these steps are shown in Figure 1.



**Figure 1.** Research flow.

### 3.3. Machine Learning Model for Construction Cost Prediction

In this study, the learning algorithm used is Deep Learning (DNN), with each deep learning model layer being composed of a fully connected layer. The concept behind this approach is as follows: DNN is a type of neural network model developed from Artificial Neural Networks (ANN), which features a deep structure comprising multiple hidden layers. DNN has the advantage of being able to model more complex data, and it represents an improvement over the overfitting problem encountered in ANN [14].

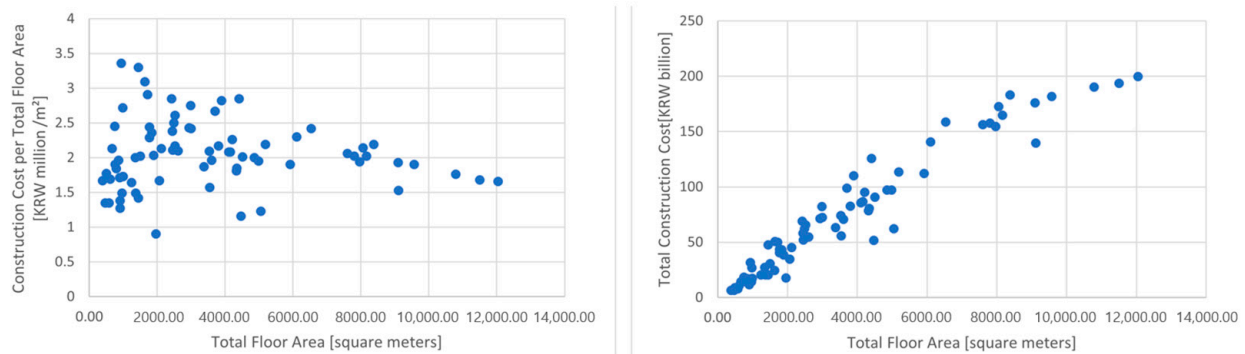
To identify the influencing factors suitable for use in the learning process, the Pearson correlation coefficient was employed. The Pearson correlation coefficient is a statistical measure used to quantify the strength and direction of the linear relationship between two variables,  $X$  and  $Y$ . The coefficient has a value between  $+1$  and  $-1$ , as dictated by the Kosh-Schwarz Inequality. A value of  $+1$  indicates a perfect positive linear correlation between the variables, while a value of  $0$  indicates no linear correlation. A value of  $-1$  indicates a perfect negative linear correlation between the variables. Jung et al. [15] used Pearson correlation coefficients to determine the strength of the correlation between variables, and found that the performance of the model can be improved by identifying and utilizing variables with stronger correlation.

Hashemie et al. [16] analyzed a variety of documents proposing cost estimation with machine learning techniques, finding that the size of the dataset used to train the model is critical to accuracy. Due to the limited number of data samples available, which is considered relatively insufficient for effective learning, the cross-validation technique was applied to ensure the reliability of the study results. Cross-validation is a widely used method for evaluating the performance of machine learning models. Its purpose is to assess how well a model will generalize to new, unseen data. One common reason for using cross-validation, specifically  $k$ -fold cross-validation, is to prevent overfitting of the model. Overfitting occurs when a model is trained too well on the training data, resulting in poor performance on new data. The  $k$ -fold verification can use the entire data for learning without separating verification and test data, which can be used to accurately determine whether training data are suitable for prediction. Shin [17] used  $k$ -fold cross-validation techniques to evaluate the accuracy of trained models on small datasets.

## 4. Data for Machine Learning

### 4.1. Data for Machine Learning

This study analyzed 78 buildings completed by the Korea Public Procurement Service, with a ‘Total Construction Cost’ of less than 20 billion. The survey collected data on the total floor area, total construction costs, and construction costs per total floor area. The average total floor area of the surveyed buildings was 3536.10 m<sup>2</sup>, with a minimum of 387.25 m<sup>2</sup> and a maximum of 12,041.58 m<sup>2</sup>. The average total construction costs were KRW 71.31 billion, with a minimum of KRW 63.7 billion and a maximum of KRW 199.61 billion. The average construction costs per total floor area were KRW 2,040,000, with a minimum of KRW 900,000 and a maximum of KRW 3,360,000. The distribution of the collected data is shown in Figure 2.



**Figure 2.** Construction cost per total floor area according to total floor area (left), total construction cost according to total floor area (right).

### 4.2. Data Collection

The primary goal of this study was to develop a more accurate cost prediction model for building construction utilizing machine learning techniques in the ‘Schematic Design’ stage. To accomplish this, various data that can be extracted during the ‘Schematic Design’ phase, where architecture-related components are planned, were collected and analyzed, including total floor area, site area, building area, land area, landscape area, general floor height, total floor height, number of basement floors, number of ground floors, and number of parking spaces.

In addition to this data, LOD 200 level BIM models were newly built and utilized to collect additional BIM properties, such as the number of rooms, sum of room heights, sum of room perimeter, sum of room area, sum of wall length, sum of wall area, sum of floor area, number of steps, amount of floor space, sum of floor space, number of stairs, and sum of steps. Table 1 provides a comprehensive summary of the types of data collected through this study. The factors determined by “Building General Properties” include “Total area”, “Site area”, “Building area”, “Landscape area”, “Floor height”, “Total height”, “Ground floor level”, “Number of floors”, “Number of parking lots”, etc., and the correlation between these factors and construction costs is as shown in Table 1.

**Table 1.** Collected data for building cost prediction using machine learning.

Building General Properties	BIM Properties	
	Object	Data
Total Floor Area (m <sup>2</sup> )	Room	Number (ea)
Site Area (m <sup>2</sup> )		Sum of Height (m)
Building Area (m <sup>2</sup> )		Sum of Perimeter (m)
Landscape Area (m <sup>2</sup> )		Sum of Area (m <sup>2</sup> )

Table 1. Cont.

Building General Properties	BIM Properties	
	Object	Data
Typical Floor Height (m)	Wall	Number (ea)
Total Height (m)		Sum of Height (m)
Number of Basement Floors (ea)		Sum of Length (m)
Number of Floors (ea)		Sum of Area (m <sup>2</sup> )
Number of Parking lots (ea)	Floor	Sum of Perimeter (m)
		Sum of Area (m <sup>2</sup> )
	Column	Number (ea)
		Sum of Height (m)
	Stair	Number (ea)
		Sum of Steps (ea)
	Opening	Number (ea)
		Sum of Area (m <sup>2</sup> )

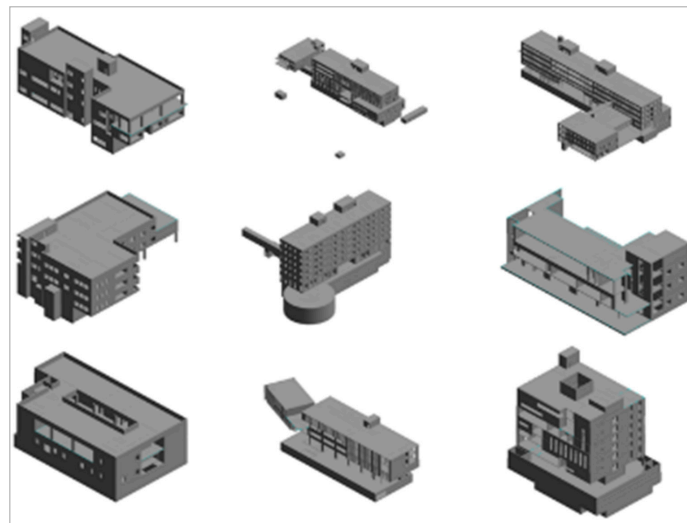
The relationship between each building general property and total construction cost is shown in Figure 3.



Figure 3. Construction cost distribution relation with building general properties.

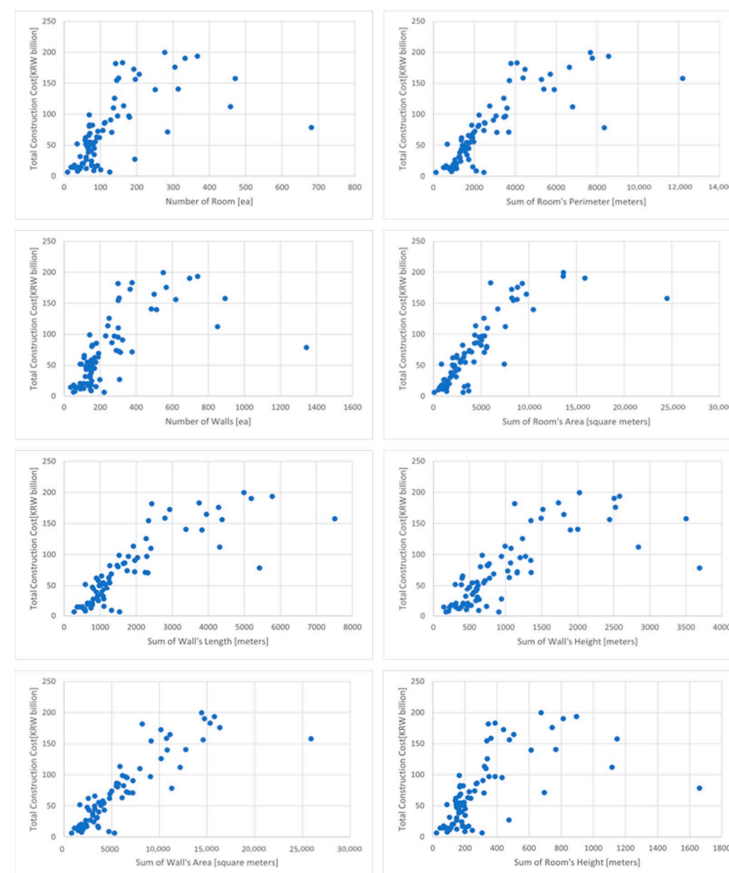
In order to extract BIM properties for predicting construction costs, 78 BIM models including objects such as room walls, floors, columns, stairs, windows, and doors were modeled using 2D drawings. The newly created BIM models are shown in Figure 4.





**Figure 4.** BIM model cases for cost prediction.

From the created BIM model, properties such as the number and height of walls and rooms, circumference, and area sum were extracted. These data determine the size and complexity of the building, and data such as length, area, and circumference are expected to affect construction costs such as structure construction and finishing work, and were extracted as BIM attribute values for predicting construction costs. Figure 5 shows the results of analyzing the correlation between these BIM attributes and construction costs. Figure 5 shows the correlation between these BIM attributes and construction costs, which shows that the extracted BIM attributes correlate with construction cost.



**Figure 5.** Total construction cost according to BIM properties (room, wall).

## 5. Machine Learning Analysis for Construction Cost Prediction with Properties

### 5.1. Dataset for Cost Prediction

In a general estimate, construction costs were predicted only by construction cost influencing factors such as floor area, but this paper intends to present a more accurate construction cost prediction method using BIM properties as well as these construction cost influencing factors.

Due to the limitation of the amount of data, we attempted to analyze the predictive performance of construction costs, focusing on the influencing factors that have the highest correlation with construction costs. “Total floor area” and “Building area” were derived as factors of building general properties, which are highly correlated with construction costs. Through the correlation analysis with construction cost, the General-1 model, which uses only the total area and building area as predictors of construction cost, and the General-2 case, which predicts construction cost by adding the number of floors, were defined. The details of the General-1 and General-2 cases for analyzing the predictive performance of construction costs are as shown in Table 2.

**Table 2.** Building general properties dataset.

Dataset	Influencing Factors	Pearson Correlation Coefficient
General-1	Total Floor Area	0.965048
	Building Area	0.625655
General-2	Total Floor Area	0.965048
	Building Area	0.625655
	Number of Floors	0.509851

Among the construction cost influencing factors extracted from BIM, the high correlation with construction cost was found for ‘sum of floors’ and ‘number of stairs’, using them to define a ‘BIM-1’ case that can predict construction cost, and then ‘sum of wall area’ with high construction cost impact was added to define ‘BIM-2’. These details are as shown in Table 3.

**Table 3.** BIM properties dataset.

Dataset	Influencing Factors	Pearson Correlation Coefficient
BIM-1	Sum of Floors’ Perimeter	0.895415
	Number of Stairs’ Steps	0.875447
BIM-2	Sum of Floors’ Perimeter	0.895415
	Number of Stairs’ Steps	0.875447
	Sum of Walls’ Area	0.870781

Finally, in the form of a combination of the two types, the Com-1 was defined using the sum of the total area and the length of the floor circumference as cost influencing factors, and the Com-2 model was defined by adding the number of stairs to it. Details of the Com-1 and Com2 datasets are shown in Table 4.

**Table 4.** Combination dataset of BIM properties and building general properties.

Dataset	Influencing Factors	Source
Com-1	Total Floor Area	Building General Properties
	Sum of Floors’ Perimeter	BIM Properties
Com-2	Total Floor Area	Building General Properties
	Sum of Floors’ Perimeter	BIM Properties
	Number of Stairs’ Steps	BIM Properties



### 5.2. Configurations of Test Cases

The prediction performance of the construction cost may vary depending on the activation function, the depth of the model, and the configuration of the node. In addition, the composition of the learning model may vary depending on the type of data, the amount of data, and the type of results to be predicted, so it is necessary to define and test various cases to find the optimal construction cost prediction model.

This study attempted to derive an optimal learning model by defining the six types of ‘Case A’, ‘Case B’, ‘Case C’, ‘Case D’, ‘Case E’, and ‘Case F’, which consist of various nodes, learning rates, and epochs. All cases (Case A–Case F) used ‘Adam’ as a loss function, and the ratio of test data to training data was fixed at 50:50. Details of each case are as shown in Table 5.

**Table 5.** Configurations of test cases for machine learning.

Case	Hidden Layer Configurations	Learning Rate	Epochs	Test Data Ratio	Loss Function
Case A	Softmax(100)/ Elu(64)/ Elu(32)	0.0005	1500	0.5	Adam
Case B	Tanh(100)/ Relu(64)/ Relu(32)	0.001	1500	0.5	Adam
Case C	Elu(100)/ Elu(50)	0.001	1500	0.5	Adam
Case D	Softmax(100)/ Relu(64)/ Relu(32)	0.0005	1500	0.5	Adam
Case E	Tanh(100)/ Tanh(64)/ Tanh(32)	0.01	1500	0.5	Adam
Case F	Relu(100)/ Relu(64)/ Elu(32)	0.0001	2500	0.5	Adam

To ensure reliable results, the k-fold technique was employed in this study, given the limited number of samples available for artificial intelligence learning, which was only 78. Additionally, the ‘DNN’ algorithm and the ‘fully connected layer’ were utilized. To assess potential issues of underfitting or overfitting, the difference between the error rate of the test data and that of the learning data was quantified as the ‘relative error between error rates’.

### 5.3. Error Rates Analysis of Test Cases

This study classified data that can impact construction costs into three categories: the building general properties dataset, the BIM properties dataset, and the combination dataset of BIM properties and building general properties. The purpose of this study was to evaluate how much construction cost prediction errors can be relatively reduced when estimating construction costs using BIM properties compared to using only general building properties.

In construction cost prediction, there are two types of error rates: the “training error rate” derived during the training phase and the “testing error rate” obtained after the completion of training. If there is a significant difference between the values of the training error rate and the testing error rate, the reliability of the predicted values can be questionable. Therefore, in addition to the training error rate and the testing error rate, a measure

indicating the difference between these two values, called the “relative error rate”, was calculated. The formula for calculating the relative error rate is as Equation (1).

$$\text{Relative Error Rate} = \frac{\text{Training Error Rate}}{\text{Testing Error Rate}} \times 100 \quad (1)$$

In other words, a larger magnitude of the relative error rate indicates lower confidence in the predicted values, while a smaller magnitude of the relative error rate suggests higher confidence in the predicted values. This allows us to assess the reliability of the predicted values based on the size of the relative error rate.

Considering that the optimal artificial intelligence model varies for each type of data (building general properties dataset, BIM properties dataset, building general + BIM properties dataset), a uniform artificial intelligence model was not applied to all data types. Instead, diverse combinations (Case A–Case F) were applied to each individual dataset, the building general properties dataset, BIM properties dataset, and building general + BIM properties dataset. As a result, the training error rate, testing error rate, and relative error rate were calculated for all cases.

Finally, by comparing the average values of the training error rate and testing error rate when the relative error rate was the lowest for each case of building general properties dataset (General-1, General-2), BIM properties dataset (BIM-1, BIM-2), and building general + BIM properties dataset (Com-1, Com-2), the “error rate” was determined. By comparing the derived error rates, it was possible to determine which data combination had the lowest error rate.

The building general properties dataset, when the General-2 dataset was trained as Case A, yielded the minimum relative error rate value. The average of the training error rate and testing error rate at this point, which was 33.04%, was determined as the error rate. This information can be found in Table 6.

**Table 6.** Error rates from building general properties.

Dataset	Case	Training Error Rate	Testing Error Rate	Relative Error Rate
General-1	Case A	34.34%	37.22%	8.38%
	Case B	21.33%	8.24%	61.37%
	Case C	18.77%	20.02%	6.67%
	Case D	47.79%	76.64%	60.39%
	Case E	24.75%	4.19%	83.06%
	Case F	31.67%	37.01%	16.85%
General-2	Case A	33.62%	32.46%	3.45%
	Case B	24.13%	11.12%	53.90%
	Case C	21.18%	17.45%	17.60%
	Case D	36.24%	39.34%	8.55%
	Case E	25.81%	7.48%	71.02%
	Case F	32.14%	33.96%	5.65%

The BIM properties dataset, when the BIM-1 dataset was trained as Case B, yielded the minimum relative error rate value. The average of the training error rate and testing error rate at this point, which was 37.33%, was determined as the error rate. This information can be found in Table 7.

The building general + BIM properties dataset, when the Com-2 dataset was trained as Case C, yielded the minimum relative error rate value. The average of the training error rate and testing error rate at this point, which was 20.82%, was determined as the error rate. This information can be found in Table 8.

**Table 7.** Error rates from BIM properties.

Dataset	Case	Training Error Rate	Test Error Rate	Relative Error Rate
BIM-1	Case A	25.66%	59.89%	133.41%
	Case B	35.47%	39.19%	10.48%
	Case C	28.64%	65.84%	129.88%
	Case D	38.90%	91.40%	134.95%
	Case E	48.34%	21.83%	54.84%
	Case F	23.90%	64.03%	167.92%
BIM-2	Case A	23.39%	58.86%	151.61%
	Case B	34.45%	27.34%	20.64%
	Case C	24.33%	63.22%	159.82%
	Case D	29.89%	70.97%	137.46%
	Case E	40.26%	14.39%	64.25%
	Case F	23.07%	63.54%	175.42%

**Table 8.** Error rates from ‘building general + BIM properties dataset’.

Dataset	Case	Training Error Rate	Test Error Rate	Relative Error Rate
Com-1	Case A	19.73%	34.24%	73.58%
	Case B	22.27%	11.27%	49.38%
	Case C	20.94%	20.24%	3.33%
	Case D	28.56%	55.88%	95.69%
	Case E	22.85%	4.99%	78.16%
	Case F	22.24%	39.33%	76.80%
Com-2	Case A	18.04%	36.14%	100.32%
	Case B	23.40%	9.11%	61.09%
	Case C	20.54%	21.10%	2.73%
	Case D	22.46%	41.61%	85.26%
	Case E	26.69%	5.72%	78.59%
	Case F	17.55%	39.04%	122.48%

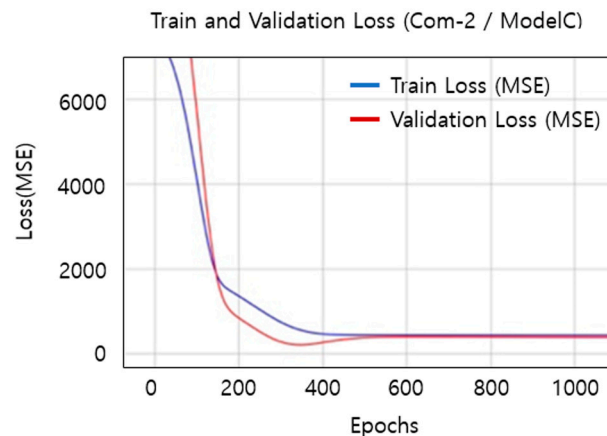
In terms of the minimum error rate of test data, Com-2 (20.82%), using a combination of building general properties and BIM properties had the smallest error rate, followed by the building general properties dataset (33.04%), and BIM properties (37.33%).

Upon investigation of the causes, among the factors within the building general properties dataset, the ‘total floor area’ showed a very high Pearson correlation coefficient of 0.9 or above with the total construction cost. However, the remaining factors exhibited low correlation coefficients, at 0.63 or below. On the other hand, within the BIM properties dataset, there was no factor with a correlation coefficient as high as the floor area of the building general properties dataset (0.965048). However, all the factors within the BIM properties dataset showed correlation coefficients of 0.87 or higher. Therefore, it can be inferred that the combination of three factors based on the highest Pearson correlation coefficients, as observed in the Com-2 dataset, resulted in the lowest error rate when trained.

As a result, compared to using general building properties or BIM properties independently to predict construction costs, it is expected that predicting construction costs by combining them can increase accuracy. The results are shown in Table 9. Additionally, in the case of Com-2, which had the lowest error rate, the training results of Model C are shown in Figure 6.

**Table 9.** Error rate comparison.

	Building General Properties	BIM Properties	Building General + BIM Properties
Mean Error Rate ((Training + Test)/2)	33.04%	37.33%	20.82%

**Figure 6.** Training result of model C (Com-2).

## 6. Conclusions and Limitation

### 6.1. Conclusions

An accurate estimate of the total construction cost of a building is possible only after the “Detailed Design” phase. In addition, estimating construction costs requires calculating the quantity of all components in the drawing, which takes considerable time. If the cost exceeds the budget, a comprehensive plan revision may be required, and in some cases, a complete overhaul of the initial plan may be required.

Although the building components are not yet accurately determined, it is necessary to accurately predict the total construction cost in the schematic design phase in which the main structural system is determined. This helps to prevent changes to the overall plan due to budget shortages. However, so far, low-accuracy unit area-based cost estimation methods are mainly used. BIM has major property values that determine the characteristics of buildings, and these values are highly correlated with construction costs. In addition, it is expected that accurate prediction of construction costs will be possible in the early stages by utilizing factors that affect construction costs.

This study attempted to verify whether the predictive performance of total construction costs can be improved when BIM object characteristics are used as an influencing factor compared to general characteristics of buildings.

For this study, a total of 78 buildings with construction costs of less than 20 billion were analyzed, and data extracted from each building’s ‘Building Register’ or ‘Building Overview Data’ were defined as ‘Building General Properties’, and data extracted from BIM objects as ‘BIM Properties’.

As a result of testing construction cost prediction performance on more than 30 AI learning models, six of the lowest error rates, Case A, Case B, Case C, Case D, Case E, and Case F, were defined as final analysis targets, and relative construction cost prediction performance was analyzed in these six cases. All datasets were trained and tested using all six models. As a result, the building general properties dataset showed an error rate of 33.04% and the BIM properties showed an error rate of 37.33%. When using the building general properties and BIM properties datasets together, the error rate was reduced to 20.82%. Therefore, it is expected that accurate construction cost prediction is possible by combining the building general properties and BIM properties datasets.

As a result of analyzing the prediction performance of construction costs through Deep learning, using BIM properties together can increase the accuracy of predicting total construction costs rather than using general building properties alone. Using this, it is expected that accurate construction costs due to design changes can be predicted at the basic design stage.

## 6.2. Limitations

Among the various factors considered in this study, the ‘total floor area’ showed an overwhelmingly high Pearson correlation coefficient value. This is attributed to the fact that the budget was set based on the unit price per total area in the budgeting process at the “pre-design” stage, and the design and construction process was conducted based on the budget afterward. Although the total floor area is the biggest factor influencing the construction cost, there are limitations to predicting various construction cost changes, and it is still difficult to define accurate construction cost influencing factors.

In addition, compared to other data, construction cost data that can be collected for one year are insufficient, and there are difficulties such as requiring a large amount of effort to investigate the factors influencing construction costs. In the future, if a plan is prepared to systematically collect and organize construction costs, it is expected that more accurate construction cost predictions will be possible.

**Author Contributions:** D.P. conceived the experiments, analyzed the data and wrote the paper; S.Y. supervised the research. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2021-KA163269).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare that they have no known competing financial interest or personal relationships that could have influenced the work reported in this study.

## References

1. Polat, G. ANN approach to determine cost contingency in international construction project. *J. Appl. Manag. Investig.* **2012**, *1*, 195–201.
2. Koo, K.; Park, S.; Park, S.C.; Song, J.K. Object & Parameter based Schematic Estimation Model for Predicting Cost of Building Interior finishing. *Korean J. Constr. Eng.* **2008**, *9*, 175–183.
3. Kim, S.; Park, G.; Song, B.; Choi, C.; Chin, S. A study on setting up work conditions for improving productivity of BIM-based Cost Estimation. *Korean J. Constr. Eng.* **2016**, *17*, 56–64. [\[CrossRef\]](#)
4. Sung, T. A Study on Asthmatic Occurrence Using Deep Learning Algorithm. *J. Korea Inst. Build. Const.* **2020**, *20*, 674–681.
5. Wang, P.; Chen, C.; Su, J.; Wang, T.; Huang, S. Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism. *IEEE Access.* **2021**, *9*, 50095–50107. [\[CrossRef\]](#)
6. Jo, Y.; Yun, S. Analysis of Impact Factors for the Improvement of Conceptual Cost Estimation Accuracy for Public Office Building. *J. Korea Inst. Build. Const.* **2021**, *21*, 495–506.
7. Jung, S.; Gwon, O.; Son, J. A Study on the Analysis and Estimation of the Construction Cost by Using Deep learning in the SMART Educational Facilities-Focused on Planning and Design Stage. *J. Korean Inst. Edu. Facil.* **2018**, *25*, 35–44.
8. Kwon, H.; Moon, H.; Lee, S.; Hong, T.; Koo, K.; Hyun, C. Cost prediction model of Public Multi-housing Projects in Schematic Design Phase. *Korean J. Constr. Eng. Manag.* **2008**, *9*, 65–74.
9. Günaydin, H.M.; Dogan, S.Z. A neural network approach for early cost estimation of structural systems of buildings. *Int. J. Proj. Manag.* **2022**, *22*, 595–602. [\[CrossRef\]](#)
10. Fachrurrazi; Saiful, H.; Mubarak, T. Neural network for the standard unit price of the building area. *Sustain. Civ. Eng. Struct. Constr. Mater.* **2016**, *171*, 282–293.
11. Matel, E.; Vahdatikhaki, F.; Hosseinyalamdary, S.; Evers, T.; Voordijk, H. An Artificial Neural Network approach for cost estimation of engineering services. *Int. J. Constr. Manag.* **2019**, *22*, 1274–1287. [\[CrossRef\]](#)
12. Yang, S.-W.; Moon, S.-W.; Jang, H.; Choo, S.; Kim, S.-A. Parametric Method and Building Information Modeling-Based Cost Estimation Model for Construction Cost Prediction in Architectural Planning. *Appl. Sci.* **2022**, *12*, 9553. [\[CrossRef\]](#)

13. Vitasek, S.; Zak, J. Cost estimation and building information modeling. In Proceedings of the 3rd International Conference on Engineering, Science and Technology, North Bangkok, Thailand, 19–22 April 2018; pp. 1–10.
14. Bharath, R.; Reza, B.Z. *Tensorflow for Deep Learning: From Linear Regression to Reinforcement Learning*, 1st ed.; O'Reilly Media: Sebastopol, CA, USA, 2018; p. 126.
15. Jung, J.; Park, S.; Lee, Y.; Gim, J. The Development of Infrared Thermal Imaging Safety Diagnosis System Using Pearson's Correlation Coefficient. *Korean Sol. Energy Soc.* **2019**, *39*, 55–65. [[CrossRef](#)]
16. Hashemi, S.T.; Ebadati, O.M.; Kaur, H. Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. *SN Appl. Sci.* **2020**, *2*, 1–25.
17. Shin, H.J. Automated Anomaly Diagnosis of Plant Drawings Using Machine Learning. Master's Thesis, Chung-Ang University, Seoul, Republic of Korea, 2021.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.