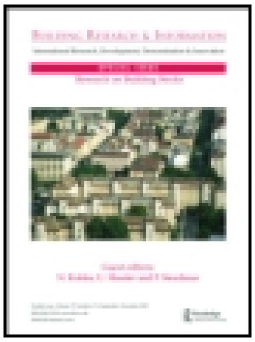

Building Research & Information

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rbri20

A construction cost estimation framework using DNN and validation unit

Salman Saeidlou & Nikdokht Ghadiminia

To cite this article: Salman Saeidlou & Nikdokht Ghadiminia (2023): A construction cost estimation framework using DNN and validation unit, Building Research & Information, DOI: 10.1080/09613218.2023.2196388

To link to this article: https://doi.org/10.1080/09613218.2023.2196388

Published online: 11 Apr 2023.

A construction cost estimation framework using DNN and validation unit

Salman Saeidlou[a] and Nikdokht Ghadiminia[b]

[a]School of Engineering, Technology and Design, Canterbury Christ Church University, Canterbury, United Kingdom; [b]School of Architecture, Computing, and Engineering, University of East London, London, United Kingdom

ABSTRACT

Accurate construction cost estimation is crucial to completing projects within the planned timeframe and expenditure. The estimation process depends on multiple variables maintaining complex relationships between themselves and the target cost. As a result, an in-depth analysis from an experienced construction consultant is required to estimate construction costs accurately. Machine learning (ML) technology can learn from previous data, which is equivalent to human experience. Many project-specific ML models estimate the construction cost, which misses the generalizability. This paper addresses the gap and designs, develops, implements, and analyzes a deep learning (DL) based novel framework that maps 94.67% of the independent variables with a mean average percentage error (MAPE) of 11.60%. The proposed framework is not limited to any specific project. It estimates the construction cost of similar projects, further validated by an innovative estimator validation unit.

ARTICLE HISTORY
Received 14 November 2022
Accepted 9 March 2023

KEYWORDS
Deep learning; construction cost estimation; framework; validator

Introduction

Constraining factors like costs, schedules, and quality all work together to ensure a project is completed successfully. The construction phase provides an opportunity to inspect and enhance a project's quality while ensuring that the schedule and cost remain within the bounds of its contracts. The contractors and the stakeholders are very involved in these estimates (Markiz & Jrade, 2022). When stakeholders and decision-makers have a realistic idea of how much a project will cost before it even begins, they can make informed decisions about feasibility studies, bidding, and cash flow management (Al-Nassafi, 2022). To the detriment of the project's stakeholders and contractors, cost overruns are a typical result of an underestimated project budget (Banks-Grasedyck et al., 2022). Several methods have been used in practice, and others have been proposed in the literature to accurately predict construction costs to limit losses and satisfy project profitability targets.

Costs can be estimated in a few different ways, but the two most common are qualitative and quantitative assessments. Qualitative methods relying on experts' opinions may be biased, resulting in erroneous estimates (Strömbäck & Tärnell, 2022). The proposed methodology estimates construction costs from quantitative assessment, and thus, the prediction by the proposed system is more accurate.

There has been a growing body of literature employing both classical statistical techniques (Akintoye & Fitzgerald, 2000; Chan & Park, 2005; Hitsanu, 2022) and machine learning (ML) models (Alshboul et al., 2022b; Kim & Cha, 2022; Matel et al., 2022; Shoar et al., 2022). The state-of-the-art machine learning approach is deep neural network (DNN)-based algorithms. This paper uses DNN to estimate construction costs. The DNNs are capable of establishing relations among complex, heterogenous, and multidimensional features, and so the proposed methodology generates better and more reliable estimations. Sometimes the data can be more crucial than the process itself (Amoore, 2022). However, developing and optimizing machine learning models for construction cost estimation in data-specific scenarios demonstrates promising results, which have been studied and presented in this paper.

Multiple construction variables impact the construction cost directly and indirectly. However, they demonstrate irregular patterns among themselves when the cost is considered the target variable and related to these influencing variables (Elhegazy et al., 2022). Moreover, these variables are further dependent on other indirect variables (Okonkwo et al., 2022). As a result,

CONTACT Salman Saeidlou salman.saeidlou@canterbury.ac.uk

the traditional simple summation of construction cost variables is not enough to accurately estimate the cost. The proposed framework overcomes this limitation. The complex distribution pattern, multiple internal, external, and hidden dependencies, and their temporal instability make construction cost estimation challenging (Dang-Trinh et al., 2022). This challenge has been beaten using the DNN-based construction cost estimation framework presented in this paper. Usually, rule-based approaches are not enough for complex relations between features and the target variable. The promising human-intelligence-like performance of deep learning (DL) technology is being used in this complicated field of studies, including: the medical imaging sector (Faruqui et al., 2021); stock market analysis (Kumbure et al., 2022); targeted marketing (Sun et al., 2022); autonomous vehicles (Hui et al., 2022); virtual assistants (Liao et al., 2022); robotics (Nguyen et al., 2022); and in many other fields. Inspired by DL's capability in human intelligence replication, this paper designs, studies, analyzes, and experiments, applying deep neural networks (DNN) in construction cost estimation.

Despite the successful application of artificial neural networks (ANN) in construction cost estimation (Baduge et al., 2022), the coherence and relevancy of the recent literature are not leaning towards a standardized solution in this regard (Zabin et al., 2022). It is a significant research gap in this sector, which has been observed and studied by the researcher of this paper, and thus, the framework presented in this paper leads the way to standardize the methodology of applying machine-learning solutions to estimate construction costs. This study proposes an innovative construction cost estimation framework using DNN to bridge the gap. Another research gap is the applied validation of the DNN-based cost estimator. It is a common practice in the deep learning domain to split the data set into training, test, and validation sets and evaluate the validity of the network (Kahloot & Ekler, 2021). However, construction cost estimation is an influential business factor that directly impacts the stakeholders' benefits (Doloi, 2013); thus, algorithmic and external validation is essential in order to rely on cost estimation by DNN practically. However, external validation is a significant research gap in the application of DNN in construction cost estimation. This research has evolved, encompassing these two research gaps, and contributes by carrying out the following:

- Design, implement, analyze, and evaluate a DNN-based framework to estimate construction costs using direct variables.
- Incorporate an estimation validator unit to address the credibility issue of practical uses of DNN in construction cost estimation.
- Analyze the limitation of the DNN-based construction cost estimator and the scope of mining opportunities from the limitations.

Literature review

The calculation of building costs has always been a task that places a premium on the knowledge and insight of industry professionals (Elhag et al., 2005). That means human intelligence with acquired experience can estimate construction costs. The DNN-based solution, which exhibits human-like intelligence, is thus a good fit in the field explored in this paper. Not every organization, especially small or newly established construction firms, can afford to allocate a budget to keep experienced consultants (Choudhry, 2016). Large organizations or companies may have the resources and experience to compile their in-house construction cost database. Apart from large construction firms, some individuals take construction responsibilities into their own hands, and those who aren't construction industry experts may rely on commercial vendors' published construction cost indexes (Zhang et al., 2017). Government construction cost statistics also assist people in estimating construction costs. However, these data are not always up to date (K'akumu, 2007). An automatic computerized system employed to estimate the construction cost, which can replace the necessity of using a consultant or government statistics, benefits many people. The research conducted in this paper is such an endeavour.

Like many different sectors, researchers in the construction sector have started using statistical and machine learning methods to improve the precision and timeliness of cost estimates (Makridakis et al., 2018). Statistical and machine learning methods improve the decision-making process by transforming data from the past into decision-support systems (Lee et al., 2016). This can overcome the lack of data for precise estimation at the beginning of a project. Based on this observation, the proposed framework focused on developing innovative network architecture to estimate construction costs precisely and validate the estimation with an external validation unit.

A study by Al-Momani (1996) builds an LR model for building cost prediction with three project features as explanatory variables. This research gives an idea of the variables to consider for construction cost estimation in the proposed methodology. The research focused on the effect of public procurement law on construction costs in Turkey and applied decision tree (DT), support vector machines (SVM), and artificial neural networks (ANN). Information related to projects, such as start and end dates, geographic scope, and discount percentages, were used as inputs (Erdis, 2013). Although this experiment aims to find the deviation from the estimated time and cost, it lays down an essential theoretical background to adopt in developing the construction cost estimator framework.

To better anticipate the costs of building in China, Shutian et al. (2017) created a fusion method that combines the Kalman filter with least-squares support vector machines (LS-SVM) and linear regression (LR). The output of the experiment is promising. However, the variable distribution of the construction cost is non-linear. This raises the question of using linear approaches. In the proposed methodology, the DNN has been used to address this issue. Sub-gradient SVM has been used to evaluate the network's performance along with LR. There are many variables to estimate the construction cost. Using those which are most influential in training, a DNN is essential. The construction area, application type, city hierarchy, and other project characteristics were used as inputs. The unit cost of concrete, the unit cost of formwork, the type of structural assembly, and the amount of superimposed load were all taken into account by Chakraborty et al. (2020) when estimating the total construction cost. A filtered version of these variables has been used in the proposed framework.

An attempt to increase the precision of BIM labour cost predictions was made by Huang and Hsieh (2020) by coupling random forest and linear regression. The apartment complex's total square footage and the total number of stories are part of the input variables. However, this approach is specific to a particular scenario. The proposed framework is not limited to a particular problem but instead has a more straightforward and lightweight network architecture. It aims to develop a uniform platform for estimating construction costs. An innovative construction cost estimation method based on statistical analysis, particularly regression analysis, was proposed by Li et al. (2022), and it studied the norm. A similar approach was published by Lowe et al. (2006) but with much more simplicity, clarity, and familiarity with the applied algorithms. Although these two methods are good construction cost estimators, they require a linear data set that maintains mathematical linearity. Compared to these approaches, the performance of the proposed methodology is invariant to data set linearity because of the use of a DNN. Otherwise, the performance tends to degrade.

One of the reasons behind choosing the machine learning (ML) method for the framework tested and developed in this paper is that it captures complicated correlations between input and output without requiring the specification of mathematical representations. A systematic review by Tayefeh Hashemi et al. (2020) discusses the common approaches to construction cost estimation using machine learning techniques, including the support vector machine (SVM), the dynamic tree (DT), and the random forest (RF). The SVM then separates the data along a hyperplane by nonlinearly mapping the raw data into a high-dimensional space. Using a recursive partitioning procedure, the DT can find a good tree structure inside a data set without requiring expert knowledge (Tayefeh Hashemi et al., 2020). However, the deep neural network approaches perform better than any other machine learning approach in construction cost estimation (Wang et al., 2022). Comparing the effectiveness, robustness, optimizable nature, and capability to map between target and complexly distributed dependent variables, the DNN has been used as the cost estimator in the proposed framework (Yanik et al., 2022).

Methodology

The framework consists of a data set manager, a deep neural network (DNN), and an estimation validator. These significant components have sub-components. The overview of the proposed framework is illustrated in Figure 1.

Data set preparation

The proposed framework uses a specific data-set variable pattern. It is essential to follow this pattern to get an optimized and accurate output from it. Table 1 lists and explains the variables. These variables are common in every building construction. Depending on the requirements, construction types, and specifications, additional variables come into consideration. However, the variables used in this paper are the fundamental features.

The core variables related to constructions are considered in this paper. Other variables influence the cost (Bernagros et al., 2021). However, these have been ignored for the simplicity of the framework.

Data processing

The ranges of the variables are not uniform, but it is essential to transform them into a uniform scale. In this experiment, the mean normalization has been

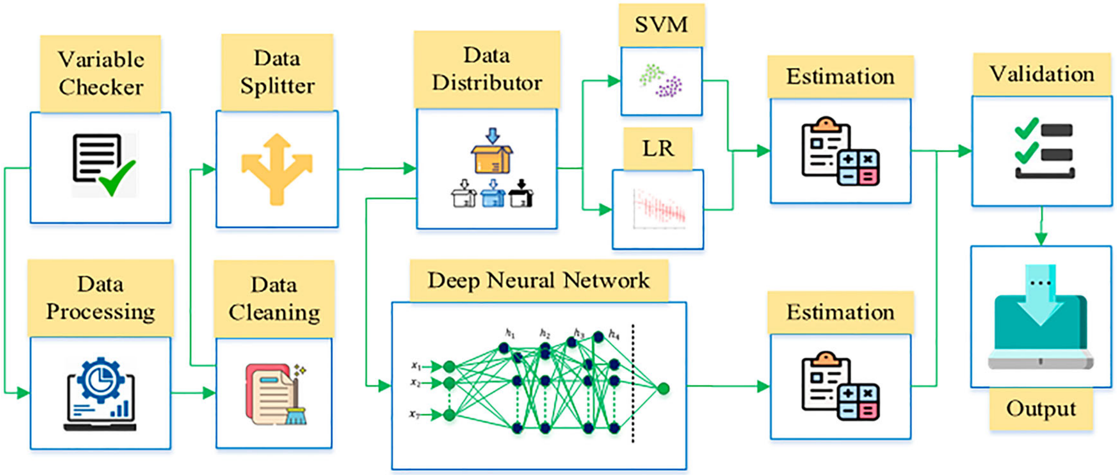

Figure 1. Overview of the proposed framework.

Table 1. Variable list, description, and pattern.

Variable	Role	Description
O	Cost	This is the final estimated cost.
x_1	Building type	Different types of buildings, categorical type. Here 0 = residential, 1 = 0 commercial
x_2	Number of rooms per floor	Numerical number of floors, including the ground floor
x_3	Number of special facilities	Three facilities have been considered in this framework Categorial type. Here, 0 = no, 1 = lift, 2 = garage, 3 = both 1 and 2
x_4	Total floor area	Total area, including all floors; the numerical value measured in square metres
x_5	Number of levels	Number of floors, including the ground floor
x_6	Floor area per level	Area of individual floors
x_7	Construction worker cost	Average construction worker cost of a project

used to scale the variable values between 0 and 1. Equation (1) has been used for mean normalization (Saranya & Manikandan, 2013).

$$o_i = \frac{x_i - \mu}{\max(x_i) - \min(x)} \tag{1}$$

Here, o_i is the normalized value of the ith feature. The μ is the simple mean. The mean of the categorical values has not been used. Equation (1) applies to numerical variables only.

Data cleaning and splitting

Irrelevant observations, structural errors, outliers, and missing values severely impact the overall performance of machine-learning models. The proposed framework requires a clean data set. The experimenting data set has been manually cleaned in this research. The proposed framework splits the data set into training, testing, and validation. The literature review on the state-of-art machine learning approaches suggests that the training and test ratio of 70:30 is a standard data splitting ratio (Presnell & Alper, 2019). Thus, the same ratio has been used in this paper.

Subset of the data set

A subset of the data set after processing using Equation (1) is listed in Table 2. Except for building type and the number of special facilities, the rest of the features are scaled between 0 and 1. The x_1 and x_3 are encoded using the one hot encoding method. The rest of the values are directly used during the training and testing period.

The subset of the data set presented in Table 2 has been selected randomly, which gives a general idea about the overall characteristics of the complete data set.

Deep neural network (DNN) architecture

This experiment uses a deep neural network with four hidden layers to estimate the construction cost. A fully connected network architecture with 28 hidden nodes has been used (Bird et al., 2019). It is illustrated in Figure 2.

The network has seven input nodes, one output node, and 28 nodes in each hidden layer. The output vector of

Table 2. Subset of the data set.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	O
1	0.50	0	0.90	0.75	0.81	0.65	0.72
1	0.78	0	0.91	0.79	0.85	0.77	0.85
0	0.85	0	0.85	0.64	0.49	0.92	0.89
1	0.90	3	0.81	0.81	0.55	0.75	0.91
0	0.80	1	0.79	0.67	0.61	0.75	0.90

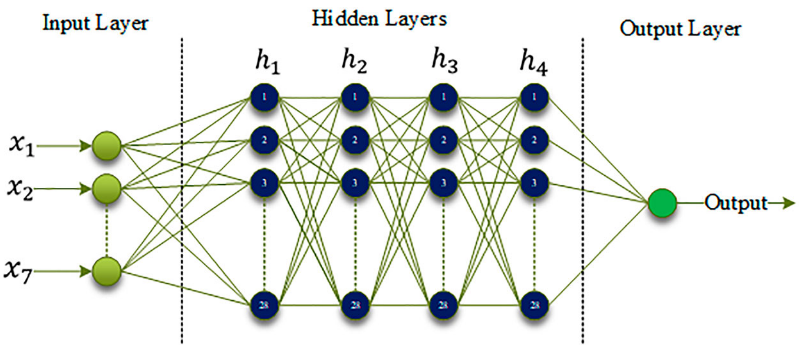

Figure 2. Deep neural network architecture.

the network is defined using Equation (2).

$$K^l = \sigma^l(B^l + W^l K^{l-1}) \tag{2}$$

Here the K^l is the output vector; B^l is the bias vector; W^l is the weight matrix; and σ^l is the activation function. The tanh, ReLU (Agarap, 2018), and sigmoid functions have been used as the activation functions of the input layer, hidden layers, and output layers, respectively, and are defined by Equations (3), (4), and (5) sequentially.

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3}$$

$$k(x) = \max(0,\ x) \tag{4}$$

$$k(x) = x \tag{5}$$

The bias for active neurones is set to 1, and dropped-off neurones are 0 (Mianjy et al., 2018).

Training and optimization

The proposed framework has experimented with the Residential Building Data Set publicly available at the Machine Learning Repository of the University of California, Irvine (UCI) (Asuncion & Newman, 2007). The initial weight plays a role in learning optimization, which has been done using the normalized Xavier weight initialization (Datta, 2020) defined by Equation (6).

$$W_i = P_D\left[-\sqrt{\frac{6}{n+m}},\ \sqrt{\frac{6}{n+m}}\right] \tag{6}$$

Here, W_i is the initial weight; P_D a uniform probability distribution between the range $-\sqrt{\frac{6}{n+m}}$ and $\sqrt{\frac{6}{n+m}}$. The n and m are the numbers of input and output nodes, respectively, in this range. The adaptive moment estimation (ADAM) optimizer (Kingma & Ba, 2014) has been used in this paper to optimize the learning process.

Estimation validation unit

An innovative estimation validation unit introduces uniqueness to the proposed framework. This unit has been created using four different machine-learning algorithms to validate the estimation done by the DNN. These two machine learning models are linear regression (LR) (Weisberg, 2005) and support vector machines (SVM) (Hearst et al., 1998) with a sub-gradient descent algorithm (Shalev-Shwartz et al., 2011).

The LR model consists of eight variables (x_n) with eight different learning parameters. The LR mode is defined by Equation (7). It discovers the linear relation between the price estimation and construction cost estimating variables.

$$y = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_7 x_7 + \varepsilon \tag{7}$$

Here, the y is the estimated price; $\theta_0 = 1;\ x_0 = 1;\ \theta_1$ to θ_7 are the learning parameters of x_1 to x_7 respectively. The ε is the error term.

The variable distribution for building cost estimation demonstrates a non-linear pattern in different segments. To address this issue, the SVM with a sub-gradient descent algorithm has been used in this experiment to validate the estimation by the DNN. The soft-margin-based SVM (Hu et al., 2010) used in this paper tries to minimize the expression of Equation (8).

$$f(W, b) = \left[\frac{1}{n}\sum_{i=1}^{n}\max(0,\ 1 - y_i(W^T x_i - b))\right] + \lambda W^2 \tag{8}$$

Here, the $f(W, b)$ is the convex function of weight matrix W and bias b. The construction variable distribution does not follow a specific scale. Equation (8) does not scale with n in iterations, which makes it a good fit for construction cost estimation.

Result and experimental evaluation

The proposed framework estimates the construction cost using a deep neural network and validates the result using SVM and LR. Each of these three algorithms performs regression in this context. The literature review shows that the state-of-art evaluation matrices for regression are the coefficient of determination (R^2) (Ozer, 1985) defined by Equation (9); the root mean square error (RMSE) (Chai & Draxler, 2014) expressed in Equation (10); the mean of absolute error (MAE) (Chai & Draxler, 2014) formulated in Equation (11); and the mean average percentage error (MAPE) (Goodwin & Lawton, 1999) demonstrated in Equation (12).

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (a_i - p_i)^2}{\sum_{i=1}^{m} (a_i - mean(a))^2} \tag{9}$$

$$RMSE = \sqrt{\frac{1}{m} \times \sum_{i=1}^{m} (p_i - a_i)^2} \tag{10}$$

$$MAE = \frac{1}{m} \times \sum_{i=1}^{m} |p_i - a_i| \tag{11}$$

$$MAPE = \frac{1}{m} \times \sum_{i=1}^{m} \left| \frac{p_i - a_i}{a_i} \right| \tag{12}$$

Here, the a_i is the target variable provided in the data set; p_i is the corresponding predicted target variable, and m is the number of instances in the data set.

Experimental result

The proposed framework has experimented with two different data sets and one augmented data set. These data sets are the Residential Building Data Set developed by the University of California, Irvine (UCI) (Asuncion & Newman, 2007), the Reinforced Concrete Building Data Set prepared by M. Y. Cheng and Hoang (2018), and an augmented combination of these two data sets. Table 3 lists the performance of the proposed framework on the UCI data set.

Table 3. Performance on UCI data set.

Model	R^2	RMSE	MAE	MAPE
DNN	0.96	27.94	17.22	11.60%
LR	0.82	34.10	25.09	16.99%
SVM	0.91	31.44	21.38	13.92%

The performance of the framework on the Reinforced Concrete Building (RCB) Data Set is listed in Table 4. The data set's relevant variables that align with the proposed framework have been used in this experiment.

Table 4. Performance on reinforced concrete building data set.

Model	R^2	RMSE	MAE	MAPE
DNN	0.969	23.42	14.50	9.81%
LR	0.843	31.65	21.72	14.20%
SVM	0.930	27.11	19.76	11.52%

Multiple attempts have been made during the experiment to compare the results with other similar approaches. However, the attempts failed because of the unavailability of the experimenting data set and different evaluation matrices of measurement. As a result, the experimenting data sets have been merged and augmented by introducing random noise to the normalized values. The range of the noise is between 0.01 and 0.2. The augmented data set's experimental results have been listed in Table 5.

The experimental results demonstrate that the performance of the framework is satisfactory. The R^2 for the UCI, RCB, and augmented data sets are 0.96, 0.969, and 0.91, respectively. That means the model properly fits 94.67% data.

Performance evaluation

The coefficient of determination (R^2) value for the DNN for the three experimenting data sets is illustrated in Figure 3. It shows a minor variation for the three data sets.

The R^2 value for DNN and SVM are similar. The LR works best for linear relations. The data set is not

Table 5. Performance on augmented data set.

Model	R^2	RMSE	MAE	MAPE
DNN	0.91	29.44	20.63	13.54%
LR	0.78	36.83	28.87	18.01%
SVM	0.88	33.58	23.42	15.26%

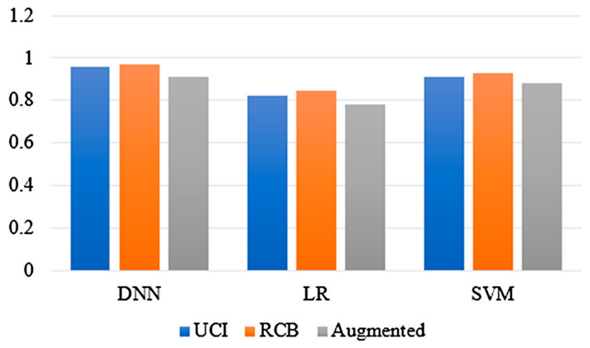

Figure 3. Coefficient of determination (R^2) evaluation.

linear. As a result, the R^2 for the LR is lower than for the DNN and SVM. However, the average of R^2 is 94.67%, which indicates models properly fit the data distribution. The mean average percentage error (MAPE) has been illustrated for the three experimenting data sets in Figure 4.

The MAPE is a little higher for the LR, as expected. However, the DNN and SVM exhibit similar MAPE values. The average MAPE for the DNN is 11.65%, which demonstrates the correctness of the cost estimation. The average MAPE of the SVM is 13.56%. There is only a 1.91% difference between the MAPE of the DNN and the SVM. It indicates the estimation from the DNN is valid. The experimental values show a similar nature to the MAE values, which have been illustrated in Figure 5.

The performance evaluation of the proposed framework highlights the validity of the estimation of the DNN. The comparison of the three ML models with three different data sets leaves no scope for questioning the correct estimation made by the framework.

Performance comparison

The performance of the proposed system has been compared to four similar methodologies. Each of these approaches used different data sets and features. Because of the similarity in the methodology, these four papers have been compared with the proposed framework and listed in Table 6.

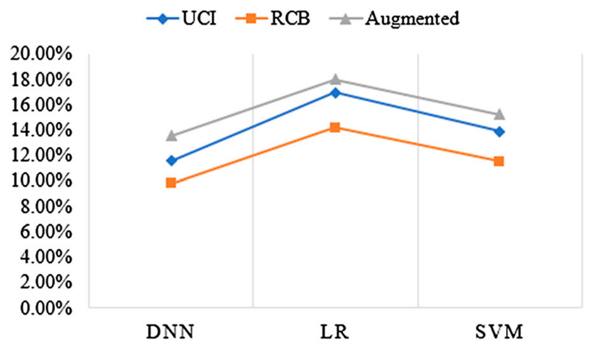

Figure 4. Mean average percentage error (MAPE) comparison.

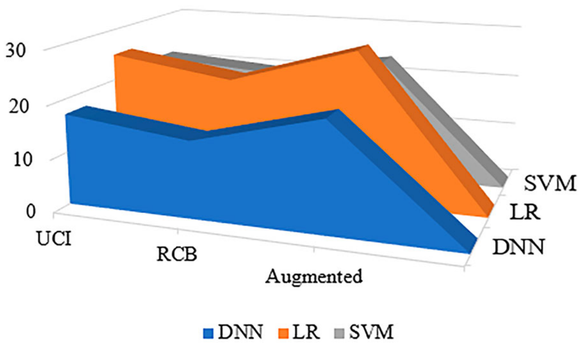

Figure 5. The MAE values of DNN, SVM, and LR.

The accuracy of the proposed construction cost estimation framework using DNN and the validation unit is 94.67% which is higher than three of the comparison papers. However, the result obtained by Hashemi et al. (2019) is 0.04% higher than the proposed framework, which is a marginal difference. The methodology of Alex et al. (2010) is much more complex than the proposed framework. The proposed method still holds the superiority in terms of architectural simplicity, even if the accuracy is 0.04% lower than the work of Alex et al. (2010).

Limitations and future scope

Any computerized system has limitations. The proposed framework is not an exception. There are four main limitations of this framework, which have been discussed here.

Variable limit

The first limitation of the proposed framework is the variable limit. It has been designed with the most common and influential seven building construction cost estimators. It cannot handle more than seven variables. One of the key contributions of this research is obtaining accurate results with a limited number of variables. Moreover, a limited internal variation of the variables has been used in the proposed methodology. A modern residential building may have multiple unique features. These features vary from building to building. As a result, it is not possible to incorporate every unique feature a building may possess (Juszczyk, 2017). This is a major limitation of the proposed framework and any framework. At the same time, there are no hard and fast rules limiting the number of construction variables, which imposes another challenge that this paper has not solved. It is another drawback related to the variable limit. However, these variable limitations pave the way for more research on this topic: to find the optimum number of variables and systematically handle the unique features of buildings. These opportunities will be explored in subsequent papers.

Table 6. Performance comparison.

Paper	Method	Accuracy
Alshemosi and Alsaad (2017)	Multifactor Linear Regression	92%
Alex et al. (2010)	Artificial Neural Network (ANN)	80%
Hashemi et al. (2019)	ANN & Genetics Algorithm (GA) Hybridization	**94.71%**
Rafiei and Adeli (2018)	Machine Learning (ML)	89.90%
Proposed	**Deep Learning (DL)**	94.67%

Data set structure heterogeneity

The heterogeneity of the data set structure imposes a challenge on an informed approach to estimating building costs through a machine learning algorithm (Scheres, 2016). The methodology followed by the researchers while developing the data sets varies based on the perspective, core focus, building type, time, location, and many other factors. Different data sets come up with different structures depending on the point of interest, the emphasis of certain factors, or the goal of the data collection project. As a result, applying a uniform framework becomes challenging. Designing, implementing, and training a DNN based on a particular data set leaves no scope for complexity. However, comparing the trained network with similar approaches becomes difficult because there is no available similar enough published research on the same data set structure. This is a major limitation of the proposed framework for any suitable methods. The data set deformation method, where any construction cost-related data sets are dissolved into the stream of fragmented data to reconstruct it into a uniform structure by maintaining a standard proportional range of values, is a potential solution to this problem. However, the level of complexity of such an approach requires a separate study, which leaves scope for another field of research in this domain.

Limited data repository

The data are like fuel to machine-learning-based approaches. Machine-learning models learn from previous data to estimate or predict the target variable. It has been observed during this experiment that the data repository for building construction cost estimation is not enriched enough (Scheres, 2016). The literature reviews suggest that most published literature uses UCI data sets or privately collected data that are not publicly available. Moreover, the data collection methodologies are yet to be standardized. These challenges limit the capability of machine-learning algorithms to estimate building construction costs. However, the future scope of the research, to develop a standard methodology for creating construction-cost-related data sets to build a shared and rich data repository, which would accelerate the machine-learning research in this sector is revealed.

Economic variables

The economic variables are indirectly valuable but directly impact the construction cost (Tas & Yaman, 2005). The proposed framework's limitation is that the economic variables have not been considered. However, the economic variables are not ignored either. It has been observed that the framework becomes complicated when both economic and non-economic variables are considered. For the sake of simplicity, the economic variables are ignored in this study. However, research is ongoing by the author of this paper to develop another framework to estimate the impact of economic variables on construction costs. In the future, these two frameworks will be merged to prepare a complete construction cost estimation, including economic and non-economic variables.

A study conducted by Rafiei and Adeli (2018) included both economic and non-economic variables in construction cost estimation using machine learning. The proposed methodology loses its superiority to the research published by Rafiei and Adeli (2018) from the variable diversity perspective. Green building construction is a recent and eco-friendly trend that is turning into modern construction standards. The proposed research does not take the green building construction parameters into consideration; however, this was done by Alshboul et al. (2022b). Considering the heavy construction equipment as one of the features, which was done by Alshboul et al. (2022a), would make the proposed methodology more reliable. However, these limitations have not been overcome in the current state of the research.

Discussion and conclusion

Accurate cost estimation is a vital step in project success. Usually, consultants with years of experience enjoy handsome consultation fees to estimate construction costs accurately. Today's advanced machine-learning algorithms, especially deep neural networks, learn from data and demonstrate human-like estimation capability. However, applying deep learning and machine learning in construction cost estimation is still nominal. This research aimed to utilize the potential of DNN to prepare a practical framework to estimate construction costs. The implementation principle used in this research mapped 94.67% of independent variables into target variables. Furthermore, it estimated the price with a MAPE of 11.60%. The estimator validation unit validates the estimation from the DNN to increase the reliability of the estimation. Whereas most of the research focuses on project-specific cost estimation, this paper took an innovative, generalized approach to estimate the cost of any similar project.

A machine-learning model designed to predict the construction cost of a particular construction project is made redundant after the project is completed unless a similar project is launched. Therefore, the generalization addressed in this paper is significant in this research field. However, the nature of the problem imposes multiple challenges in developing an all-in-one framework; thus, the research scope of this paper was limited to residential buildings with seven cost estimators. However, this limitation opens the door to further research. This paper is a milestone in applied machine learning with an outstanding demonstration of accurately estimating construction costs. It has been observed in the literature review that the data processing approaches are different in multiple papers. The data set used is also exclusive to unique construction projects. The methodology presented in this paper generalizes the data processing approaches into a standard structure, potentially a state-of-the-art method to process data for construction cost estimation. Making an important business decision, such as large-scale construction, requires cross-validation. The existing construction cost estimation methodologies estimate the costs. However, those approaches' external cross-validation of the predicted cost is absent. The proposed construction cost estimation framework not only predicts the target variables with 94.67% accuracy but also validates the prediction to enhance the acceptability of the projection.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

Akintoye, A., & Fitzgerald, E. (2000). A survey of current cost estimating practices in the UK. *Construction Management and Economics*, *18*(2), 161–172. https://doi.org/10.1080/014461900370799

Al-Momani, A. H. (1996). Construction cost prediction for public school buildings in Jordan. *Construction Management and Economics*, *14*(4), 311–317. https://doi.org/10.1080/014461996373386

Al-Nassafi, N. M. (2022). The effect of cash flow variation on project performance: An empirical study from Kuwait. *The Journal of Asian Finance Economics and Business*, *9*(3), 53–63. https://doi.org/10.13106/jafeb.2022.vol9.no3.0053

Alex, D. P., Al Hussein, M., Bouferguene, A., & Fernando, S. (2010). Artificial neural network model for cost estimation: City of Edmonton's water and sewer installation services. *Journal of Construction Engineering and Management*, *136*(7), 745–756. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000184

Alshboul, O., Shehadeh, A., Al-Kasasbeh, M., Al Mamlook, R. E., Halalsheh, N., & Alkasasbeh, M. (2022a). Deep and machine learning approaches for forecasting the residual value of heavy construction equipment: A management decision support model. *Engineering, Construction and Architectural Management*, *29*(10), 4153–4176. https://doi.org/10.1108/ECAM-08-2020-0614

Alshboul, O., Shehadeh, A., Almasabha, G., & Almuflih, A. S. (2022b). Extreme gradient boosting-based machine learning approach for green building cost prediction. *Sustainability*, *14*(11), 6651. https://doi.org/10.3390/su14116651

Alshemosi, A. M. B., & Alsaad, H. S. H. (2017). Cost estimation process for construction residential projects by using multifactor linear regression technique. *Criterion*, *6*(6), 7. https://doi.org/10.21275/ART20174128

Amoore, L. (2022). Machine learning political orders. *Review of International Studies*, *49*(1), 1–17. https://doi.org/10.1017/S0260210522000031

Asuncion, A., & Newman, D. (2007). UCI machine learning repository.

Baduge, S. K., Thilakarathna, S., Perera, J. S., Arashpour, M., Sharafi, P., Teodosio, B., & Mendis, P. (2022). Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications. *Automation in Construction*, *141*, 104440. https://doi.org/10.1016/j.autcon.2022.104440

Banks-Grasedyck, D., Lippke, E., Oelfin, H., Schwaiger, R., & Seemann, V. (2022). *The underestimated success factor: People, in: Successfully managing S/4HANA projects* (pp. 125–176). Springer.

Bernagros, J. T., Pankani, D., Struck, S. D., & Deerhake, M. E. (2021). Estimating regionalized planning costs of green infrastructure and low-impact development stormwater management practices: Updates to the US environmental protection agency's national stormwater calculator. *Journal of Sustainable Water in the Built Environment*, *7* (2), 2. https://doi.org/10.1061/JSWBAY.0000934

Bird, J. J., Ekárt, A., Buckingham, C. D., & Faria, D. R. (2019, July). Evolutionary optimisation of fully connected artificial neural network topology. In *Intelligent Computing-Proceedings of the Computing Conference* (pp. 751–762). Springer.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chakraborty, D., Elhegazy, H., Elzarka, H., & Gutierrez, L. (2020). A novel construction cost prediction model using hybrid natural and light gradient boosting. *Advanced Engineering Informatics*, *46*, 101201. https://doi.org/10.1016/j.aei.2020.101201

Chan, S. L., & Park, M. (2005). Project cost estimation using principal component regression. *Construction Management and Economics*, *23*(3), 295–304. https://doi.org/10.1080/01446190500039812

Cheng, M. Y., & Hoang, N. D. (2018). Estimating construction duration of diaphragm wall using firefly-tuned least squares support vector machine. *Neural Computing and Applications*, *30*(8), 2489–2497. https://doi.org/10.1007/s00521-017-2840-z

Choudhry, R. M. (2016). Appointing the design consultant as supervision consultant on construction projects. *Journal of Legal Affaires and Dispute Resolution in Engineering Construction*, *8*(4), 04516005. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000195

Dang-Trinh, N., Duc-Thang, P., Nguyen-Ngoc Cuong, T., & Duc-Hoc, T. (2022). Machine learning models for estimating preliminary factory construction cost: Case study in southern Vietnam. *International Journal of Construction Management*, 1–9. https://doi.org/10.1080/15623599.2022.2106043

Datta, L. (2020). *A survey on activation functions and their relation with xavier and he normal initialization*. arXiv preprint arXiv:2004.06632.

Doloi, H. (2013). Cost overruns and failure in project management: Understanding the roles of key stakeholders in construction projects. *Journal of Construction Engineering and Management*, *139*(3), 267–279. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000621

Elhag, T. M. S., Boussabaine, A. H., & Ballal, T. M. A. (2005). Critical determinants of construction tendering costs: Quantity surveyors' standpoint. *International Journal of Project Management*, *23*(7), 538–545. https://doi.org/10.1016/j.ijproman.2005.04.002

Elhegazy, H., Chakraborty, D., Elzarka, H., Ebid, A. M., Mahdi, I. M., Aboul Haggag, S. Y., & Abdel Rashid, I. (2022). Artificial intelligence for developing accurate preliminary cost estimates for composite flooring systems of multi-storey buildings. *Journal of Asian Architecture and Building Engineering*, *21*(1), 120–132. https://doi.org/10.1080/13467581.2020.1838288

Erdis, E. (2013). The effect of current public procurement law on duration and cost of construction projects in Turkey. *Journal of Civil Engineering and Management*, *19*(1), 121–135. https://doi.org/10.3846/13923730.2012.746238

Faruqui, N., Yousuf, M. A., Whaiduzzaman, M., Azad, A. K. M., Barros, A., & Moni, M. A. (2021). Lungnet: A hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data. *Computers in Biology and Medicine*, *139*, 104961. https://doi.org/10.1016/j.compbiomed.2021.104961

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*(4), 405–408. https://doi.org/10.1016/S0169-2070(99)00007-2

Hashemi, S. T., Ebadati E, O. M., & Kaur, H. (2019). A hybrid conceptual cost estimating model using ANN and GA for power plant projects. *Neural Computing and Applications*, *31*(7), 2143–2154. https://doi.org/10.1007/s00521-017-3175-5

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28. https://doi.org/10.1109/5254.708428

Hitsanu, M. S. (2022). *Conceptual Cost Estimation of Highway Earthwork Construction in Iowa Using Spatial Statistical Modeling*. [Doctoral dissertation, North Dakota State University].

Hu, Q., Che, X., Zhang, L., & Yu, D. (2010). Feature evaluation and selection based on neighborhood soft margin. *Neurocomputing*, *73*(10-12), 2114–2124. https://doi.org/10.1016/j.neucom.2010.02.007

Huang, C. H., & Hsieh, S. H. (2020). Predicting BIM labor cost with random forest and simple linear regression. *Automation in Construction*, *118*, 103280. https://doi.org/10.1016/j.autcon.2020.103280

Hui, F., Wei, C., ShangGuan, W., Ando, R., & Fang, S. (2022). Deep encoder–decoder-NN: A deep learning-based autonomous vehicle trajectory prediction and correction model. *Physica A: Statistical Mechanics and its Applications*, *593*, 126869. https://doi.org/10.1016/j.physa.2022.126869

Juszczyk, M. (2017). The challenges of nonparametric cost estimation of construction works with the use of artificial intelligence tools. *Procedia Engineering*, *196*, 415–422. https://doi.org/10.1016/j.proeng.2017.07.218

K'akumu, O. A. (2007). Construction statistics review for Kenya. *Construction Management and Economics*, *25*(3), 315–326. https://doi.org/10.1080/01446190601139883

Kahloot, K. M., & Ekler, P. (2021). Algorithmic splitting: A method for dataset preparation. *IEEE Access*, *9*, 125229–125237. https://doi.org/10.1109/ACCESS.2021.3110745

Kim, J., & Cha, H. S. (2022). Expediting the cost estimation process for aged-housing renovation projects using a probabilistic deep learning approach. *Sustainability*, *14*(1), 564. https://doi.org/10.3390/su14010564

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, *197*, 116659. https://doi.org/10.1016/j.eswa.2022.116659

Lee, H., Chung, S. H., & Choi, E. J. (2016). A case study on machine learning applications and performance improvement in learning algorithm. *Journal of Digital Convergence*, *14*(2), 245–258. https://doi.org/10.14400/JDC.2016.14.2.245

Li, Q., Guo, L., & Zhou, H. (2022). Construction quality evaluation of large-scale concrete canal lining based on statistical analysis. *FAHM, and Cloud Model. Sustainability*, *14*(13), 7663. https://doi.org/10.3390/su14137663

Liao, S. W., Hsu, C. H., Lin, J. W., Wu, Y. T., & Leu, F. Y. (2022). A deep learning-based Chinese semantic parser for the almond virtual assistant. *Sensors*, *22*(5), 1891. https://doi.org/10.3390/s22051891

Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting construction cost using multiple regression techniques. *Journal of Construction Engineering and Management*, *132*(7), 750–758. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(750)

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS One*, *13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Markiz, N., & Jrade, A. (2022). Integrating an expert system with BrIMS, cost estimation, and linear scheduling at conceptual design stage of bridge projects. *International Journal of Construction Management*, *22*(5), 913–928. https://doi.org/10.1080/15623599.2019.1661572

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

Mianjy, P., Arora, R., & Vidal, R. (2018, July). On the implicit bias of dropout. In Jennifer Dy & Andreas Krause (Eds.), *International conference on machine learning* (pp. 3540–3548). PMLR.

Nguyen, H. T., Cheah, C. C., & Toh, K. A. (2022). An analytic layer-wise deep learning framework with applications to robotics. *Automatica*, *135*, 110007. https://doi.org/10.1016/j.automatica.2021.110007

Okonkwo, C., Evans, U. F., & Ekung, S. (2022). Unearthing direct and indirect material waste-related factors underpinning cost overruns in construction projects. *International Journal of Construction Management*, 1–7. https://doi.org/10.1080/15623599.2022.2052431

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin*, *97*(2), 307. https://doi.org/10.1037/0033-2909.97.2.307

Presnell, K. V., & Alper, H. S. (2019). Systems metabolic engineering meets machine learning: A new era for data-driven metabolic engineering. *Biotechnology Journal*, *14*(9), 1800416. https://doi.org/10.1002/biot.201800416

Rafiei, M. H., & Adeli, H. (2018). Novel machine-learning model for estimating construction costs considering economic variables and indexes. *Journal of Construction Engineering and Management*, *144*(12), 04018106. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001570

Saranya, C., & Manikandan, G. (2013). A study on normalization techniques for privacy preserving data mining. *International Journal of Engineering Technology (IJET)*, *5*(3), 2701–2704.

Scheres, S. H. (2016). Processing of structurally heterogeneous cryo-EM data in RELION. *Methods in Enzymology*, *579*, 125–157. https://doi.org/10.1016/bs.mie.2016.04.012

Shalev-Shwartz, S., Singer, Y., Srebro, N., & Cotter, A. (2011). Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical Programming*, *127*(1), 3–30. https://doi.org/10.1007/s10107-010-0420-4

Shoar, S., Chileshe, N., & Edwards, J. D. (2022). Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: Application of random forest regression. *Journal of Building Engineering*, *50*, 104102. https://doi.org/10.1016/j.jobe.2022.104102

Shutian, F., Tianyi, Z., & Ying, Z. (2017). Prediction of construction projects' costs based on fusion method. *Engineering Computations*, *34*(7), 2396–2408. https://doi.org/10.1108/EC-02-2017-0065

Strömbäck, A., & Tärnell, E. (2022). *Evaluation and Learning about Social Sustainability in the Real Estate Industry: A Qualitative and Quantitative Study of how Real Estate Companies can Contribute to Society and Profitability*.

Sun, C., Adamopoulos, P., Ghose, A., & Luo, X. (2022). Predicting stages in omnichannel path to purchase: A deep learning model. *Information Systems Research*, *33*(2), 429–445. https://doi.org/10.1287/isre.2021.1071

Tas, E., & Yaman, H. (2005). A building cost estimation model based on cost significant work packages. *Engineering Construction and Architechtural Management*, *12*(3), 251–263. https://doi.org/10.1108/09699980510600116

Tayefeh Hashemi, S., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. *SN Applied Sciences*, *2*(10), 1–27. https://doi.org/10.1007/s42452-020-03497-1

Wang, R., Asghari, V., Cheung, C. M., Hsu, S. C., & Lee, C. J. (2022). Assessing effects of economic factors on construction cost estimation using deep neural networks. *Automation in Construction*, *134*, 104080. https://doi.org/10.1016/j.autcon.2021.104080

Weisberg, S. (2005). *Applied linear regression (Vol. 528)*. John Wiley & Sons.

Yanik, E., Intes, X., Kruger, U., Yan, P., Diller, D., Van Voorst, B., … De, S. (2022). Deep neural networks for the assessment of surgical skills: A systematic review. *The Journal of Defense Modeling and Simulation*, *19*(2), 159–171. https://doi.org/10.1177/15485129211034586

Zabin, A., González, V. A., Zou, Y., & Amor, R. (2022). Applications of machine learning to BIM: A systematic literature review. *Advanced Engineering Informatics*, *51*, 101474. https://doi.org/10.1016/j.aei.2021.101474

Zhang, S., Bogus, S. M., Lippitt, C. D., & Migliaccio, G. C. (2017). Estimating location-adjustment factors for conceptual cost estimating based on nighttime light satellite imagery. *Journal of Construction Engineering and Management*, *143*(1), 04016087. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001216