A Graph Neural Network Approach to Conceptual Cost Estimation

Hao Liu[1], Jack C.P. Cheng[1] and Chimay J. Anumba[2]

[1]Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong
[2]College of Design, Construction and Planning, University of Florida, United States of America
hliuci@connect.ust.hk, cejcheng@ust.hk, anumba@ufl.edu

Abstract –

Conceptual cost estimation plays an important role in construction projects since it is the basis for stakeholders to produce financial plans (e.g., establishing project budgets). The current practice, heavily dependent on cost engineers' subjective judgment and manual work, tends to be error-prone and labor-intensive. In response, this paper introduces a Graph Neural Network (GNN) approach to accurate and efficient conceptual cost estimation. Firstly, cost factors impacting construction costs, as well as their relationships, are identified based on literature review to form a graph representation. Afterwards, a GNN model is deployed to predict the construction cost. A real-world dataset from school projects is used for validation. The results show that the proposed approach achieved high accuracy, demonstrating the potential of graph neural networks in conceptual cost estimation.

Keywords –

Conceptual cost estimation; Deep learning; Graph neural network

1 Introduction

Conceptual cost estimation predicts the construction cost at the early stages of the project (e.g., conceptual design, budget setup) [1]. The estimation result is vital for the success of the project since stakeholders rely on it to set project budgets and make cost management plans before and during construction [2]. However, design information at project early stages is limited and full of uncertainties, which leads to low level of confidence on the estimation [3]. Traditional methods are heavily dependent on the experience and subjective assessments of cost engineers [4]. Such subjective evaluations, however, can vary and be unreliable, often resulting in inaccurate estimations and potential financial losses in the project [5]. Moreover, the dependence on the expertise of cost engineers makes the estimation process laborious and time-consuming, which is problematic for construction projects that typically operate on tight timelines [6].

To this end, deep learning, as a data-driven approach, presents an appealing alternative. It refines its accuracy autonomously by learning from historical data, and has demonstrated its reliability in data analysis and prediction within the construction sector [7]. Yet, standard deep learning models fall short in expressing the nuances of construction cost estimation. Specifically, construction projects exhibit intricate interrelations among cost factors (e.g., influence of contract type on project duration), highlighting their complex dependencies [8]. Traditional deep learning models, which typically use isolated factors to form tabular inputs, fail to account for these real-world characteristics of construction cost factors [9], missing out on capturing the interactive effects crucial for accurate construction cost estimation.

Graph deep learning, a branch of deep learning adept at representing intricate interrelations among input variables, offers a solution for encapsulating the complexity of construction cost factors. In this method, data is represented in a graph format where the edges in the graph provide linkage between the nodes, thereby depicting the impact relationships between them [9]. A number of studies have recognized the effectiveness of graph deep learning in representing input data in an expressive way for accurate predictions in the construction domain [10]. Nevertheless, its application in construction cost estimation is still in its infancy.

Thus, this research targets at developing a graph deep learning-based approach for accurate and efficient conceptual cost estimation for construction projects. A graph representation that integrates various construction cost factors and their relationships will be established, enabling expressive input data modeling. Then, a Graph Neural Network (GNN) will be developed and deployed using the graph representation to conduct data-driven conceptual cost estimation.

2 Related Work

Numerous studies have been conducted for accurate and efficient conceptual cost estimation. Early research leveraged statistical models to fit historical project data. For instance, Williams employed a univariate linear regression model to predict construction costs of highway projects [11]. In contrast, Stoy et al. used regression analysis with various factors (e.g., floor height, project type) to estimate construction costs for residential buildings [12]. The traditional statistical models have highlighted the dependency between the construction costs and their influential factors. However, they can be complex to implement and may lack robustness to intricate nonlinearity [13].

As a result, researchers began exploring the use of machine learning, known for its automatic pattern recognition and good prediction capabilities, to facilitate conceptual cost estimation. Fang et al. developed a Support Vector Machine (SVM) model with design information as the inputs to predict construction costs of building projects [14]. Based on the gradient boosting model, Chakraborty et al. used information structure and material design for construction cost estimation at the value engineering stage [15]. Although the majority of studies using conventional machine learning models for conceptual cost estimation yielded credible outcomes, they encounter limitations associated with shallow learning [16]. This limits their ability to discern more intricate patterns that could enhance estimation accuracy.

To address the limitations of traditional machine learning models, deep learning, a branch of machine learning renowned for its advanced capability to autonomously extract features and provide state-of-the-art accuracy, has been investigated in recent years. Saeidloua and Ghadiminia implemented a Deep Neural Network (DNN) to estimate construction costs of buildings and found that DNN outperformed traditional machine learning models, such as SVM, in terms of estimation accuracy [17]. Similarly, Kim and Cha applied a DNN model to predict construction costs of data-scarce renovation projects, with the consideration of probability distributions of cost factors [18]. The deep learning methods have advanced conceptual cost estimation models regarding accuracy and efficiency. However, they rely on structured tabular inputs and tend to overlook the intricate relationships between cost factors in construction projects, which can lead to unreliability in real-world applications [9]. Therefore, to bridge the research gap, this paper introduces a graph deep learning approach, including a graph representation for modeling interrelated cost factors, and a GNN model for accurate conceptual cost estimation in a data-driven manner.

3 Proposed Method

Figure 1 provides an overview of the proposed graph neural network-based approach. Firstly, a thorough literature review is undertaken to identify factors influencing construction costs and their relationships. The identified results are validated by professionals specializing in construction cost estimation. Subsequently, a graph representation is established to effectively express the cost factors and relationships. A GNN model is then developed to conduct model training based on the formulated graph data representation. With the trained GNN model, the construction cost is predicted in an end-to-end manner. The main focus of this research is building projects, given their applicability to a diverse range of stakeholders, including public/private developers and contractors of various sizes. Detailed explanations of the proposed method are presented in subsequent subsections.

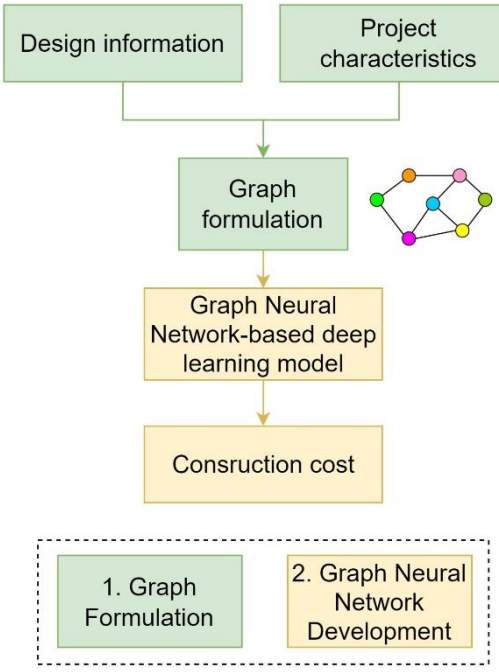

Figure 1. Overview of the proposed GNN-based approach

3.1 Graph Formulation

A comprehensive review of existing literature on construction cost estimation was carried out to pinpoint various factors that influence the construction costs, along with how these factors interrelate. Professionals in construction cost estimation then validated the findings. For inclusion in the graph deep learning model, these factors should meet specific criteria relevant to the focus

and approach of this research, as follows:

- The factors should be those influencing the overall construction cost, rather than cost indices or unit rates.
- The factors should be relevant and easily obtainable in conceptual cost estimation (i.e., project early stages).
- The factors should be quantifiable for inputting into deep learning models.

Table 1 Description of the cost factors

Cost factor		Description
Design information	Construction floor area	Construction floor area. A numerical variable in m^2.
	Building height	Height of the building. A numerical variable in m.
	Soil condition	Geology condition type based on the geological map. A categorical variable.
Project characteristics	Project type	Project type. A categorical variable.
	Project duration	Planned duration of the construction works. A numerical variable in $month$.
	Project location	Location of the project by region. A categorical variable.
	Contract type	Contract type. A categorical variable.

Figure 2 displays the identified cost factors and their relationships, with Table 1 providing the descriptions of the factors. As shown in Figure 2, there are three levels of the identified hierarchy of cost factors. The first level is the target of the estimation, i.e., the construction cost. The second level represents the influential aspects. The third level consists of specific cost factors in each aspect. In general, design information and project characteristics are two major aspects that impact construction costs. The design information aspect refers to construction floor area [19], building height [20], and soil condition [21]. As for the project characteristics aspect, it covers project type [1], project duration [1], project location [22], and contract type [23].

Moreover, there are relationships between these cost factors. The dotted lines among the specific cost factors represent their impact relationships. In terms of design information, the soil condition plays a pivotal role in determining the building height [24]. Besides, there is a notable relationship between the height of a building and its construction floor area, since taller buildings may have larger floor areas, and vice versa. Regarding project characteristics, the project type has impacts on the choice of contract type [8]. Furthermore, both the contract type and project location can influence the duration of the project [25]. In addition, the project location often decides the soil condition according to geological maps.

Drawing from the identified cost factors and the relationships between them, Figure 3 illustrated the formulated graph representation. The nodes in the graph denote the identified cost factors, while the arrows describe the impact relationships between them (i.e., the dotted lines shown in Figure 2). The unidirectional arrows represent the certain impact from one factor to the other, while the bidirectional arrows denote the fuzzy interactions between two factors. Consequently, the graph representation encapsulates the cost factors and their correlations. This forms the foundational input for the GNN model in the next step.

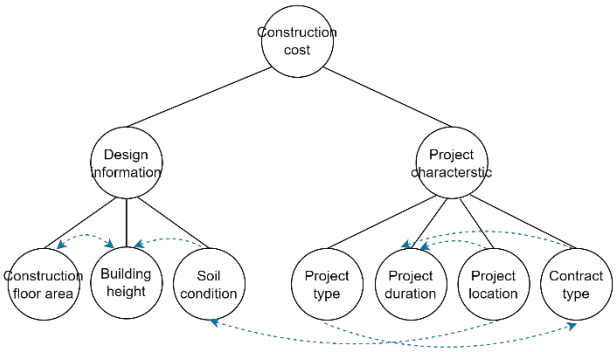

Figure 2. The identified cost factors and their relationships

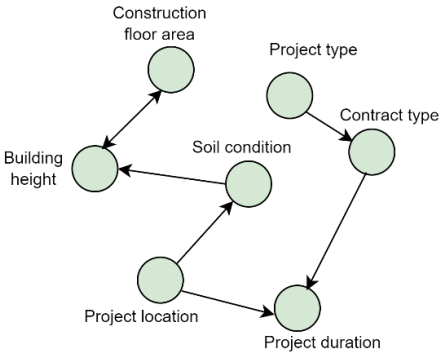

Figure 3. The formulated graph representation

3.2 Graph Neural Network Development

After establishing the graph representation of cost factors, a GNN model is developed to take the formulated graph as the input, utilize the graph structure in the model learning, and predict the construction cost. Figure 4 describes the architecture of the GNN model.

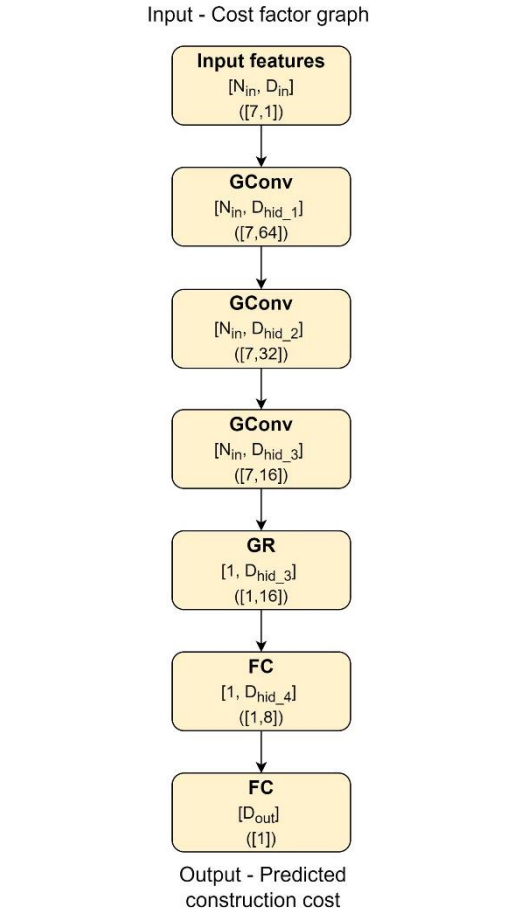

Figure 4. The developed GNN model

As shown in Figure 4, the developed model features graph convolutional layers to take advantage of the input graph structure in the neural network model training. The graph convolutional layer is a core component of Graph Convolutional Network (GCN), which is a mainstream GNN model and has demonstrated state-of-the-art accuracy performance in various graph-based prediction applications [26]. The graph convolution mechanism utilizes the graph structure to capture both node features and their topological relationships [26]. Equation (1) describes the operation in a graph convolutional layer as follows:

$$H^{(l+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad (1)$$

Where:

- $H^{(l)}$ stands for the node feature matrix at the l^{th} layer.
- $\hat{A} = A + I$ denotes the adjacency matrix of the graph, A, with added self-connections represented by the identity matrix I.
- $\hat{D}$ is the diagonal degree matrix of $\hat{A}$
- $W^{(l)}$ means the trainable weight matrix at the l^{th} layer.
- σ is the non-linear activation function

Firstly, each node aggregates features from its neighbors, which learns contextual information in the graph. This is achieved by the multiplication of the adjacency matrix with the node feature matrix. The symmetric normalization, applied via $\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$, considers different node degrees to ensure balanced influences from each node's neighbors. Then, a linear transformation is performed for the aggregated features through the trainable weight matrix. By doing so, the features are projected to a higher-level feature space, which facilitates the learning of complex patterns in the graph. Afterwards, a non-linear activation function, such as Rectified Linear Unit (ReLU), is used to introduce non-linearity in the neural network model for capturing more complex data patterns. Multiple graph convolutional layers are stacked as the essential parts of the developed GNN model. An average graph readout operation is conducted to compile the features of all the nodes into a single feature vector as a graph-level representation. The operation is defined in Equation (2), where N is the number of nodes in the graph, h_i stands for the feature vector of node i, and h_G denotes the graph readout feature vector to represent the entire graph. After producing the graph-level representation, fully connected layers are integrated to output the predicted construction costs.

$$h_G = \frac{1}{N}\sum_{i=1}^{N} h_i \quad (2)$$

Upon training the developed GNN model, a prediction model is established for conceptual cost estimation. The trained model can autonomously generate construction cost predictions for new projects that have the same graph input information. Such an end-to-end process helps to reduce the need for manual labor.

4 Experiments and Results

4.1 Experimental Design

The proposed method was implemented on a personal computer equipped with the Windows 10 operating system, an Intel(R) Core(TM) i7-11700KF @ 3.60GHz processor, an NVIDIA GTX 3060Ti GPU, and 32GB of RAM. The experiments were conducted with the Python 3.9.7 [27] programming language. PyTorch 1.10.2 [28] and scikit-learn 1.0.2 [29] libraries are adopted as the development platforms.

This study employs a dataset from school construction projects for validation. The dataset is from the Development Bureau of the Government of the Hong Kong Special Administrative Region, an authoritative organization overseeing building and infrastructure projects in Hong Kong. The dataset encompasses 50 school projects in Hong Kong, including their actual construction costs and the cost factor values required for the graph representation described in Section 3.1. The project scope includes various school types: primary, secondary, secondary-cum-primary, and special schools. Secondary-cum-primary schools offer continuous education from primary through secondary levels, while special schools cater to students with physical or intellectual disabilities.

The dataset undergoes min-max normalization as part of its preprocessing, a technique aimed at enhancing model performance and expediting convergence speed [30]. It is randomly split into training and testing sets in an 8:2 ratio. The training set is used to develop and train the GNN model as detailed in Section 3.2. For the training process, Adam optimizer, which is recognized for its superior performance over other common optimizers such as SGDNesterov and RMSprop [31], is used for model optimization. After training, the model is evaluated using the testing set to assess the model's prediction accuracy.

4.2 Model Prediction Results

Regarding the model evaluation, the widely used metric for regression problems, Mean Absolute Percentage Error (MAPE), is adopted to assess the performance of the model in predicting construction costs. MAPE is a straightforward and scale-invariant metric that measures the discrepancy between predicted and actual values. The calculation is defined in Equation (3), where n represents the total number of projects being tested, A_i and P_i denote the actual and predicted construction costs, respectively. The MAPE indicates the percentage variance between predicted and actual costs, with a range from 0% to 100%. A low MAPE value means that the prediction is close to the actual cost.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - P_i}{A_i}\right| \quad (3)$$

Table 2 summarizes the prediction accuracy results of the developed GNN model and the comparison with common statistical, machine learning, and deep learning models in previous literature on conceptual cost estimation. The 5-fold cross-validation is conducted to thoroughly evaluate the performance of the GNN and other baseline models. The dataset is split into 5 folds, each of them is used for validation iteratively. The final performance is obtained by averaging the 5-fold cross-validation results to reduce the risk of overfitting in a relatively small dataset and provide a more reliable performance comparison.

The developed GNN model demonstrated superior accuracy in predicting construction costs when measured against other typical baseline models. More specifically, it achieved better accuracy results than the typical statistical model, Least Absolute Shrinkage and Selection Operator (LASSO), highlighting its enhanced performance over traditional statistical techniques. A comparison between the GNN model and a conventional machine learning model, Support Vector Regression (SVR), was conducted, and the results showed the advantage of the GNN model in improving the prediction accuracy. Both the DNN and GNN models belong to deep learning methods. The DNN refers to an artificial neural network with an input layer (i.e., the features), hidden layers, and an output layer (i.e., predicted cost). For a fair comparison, the DNN adopts the same number of hidden layers (i.e., 5) as the GNN. The grid search is used to decide other optimal hyperparameters (e.g., learning rate, batch size) for the DNN and GNN, respectively. The performance comparison between the formulated graph representation incorporating interrelationships in the GNN and normal flat tabular inputs in the DNN verifies whether the graph representation is more effective or not. The higher accuracy provided by the GNN model highlights the superiority of the proposed approach compared with typical deep learning models that adopt tabular input formats. This indicates the importance of considering the relationships between cost factors within the deep learning model.

Table 2. Prediction performance of the developed GNN model and its comparison with other baseline models

	LASSO	SVR	DNN	The GNN model
MAPE	25.5%	21.0%	22.4%	**15.2%**

5 Conclusions

Conceptual cost estimation is pivotal in establishing

the financial scale for ensuring the success of the project. This study introduces a graph neural network approach to predict construction costs accurately and efficiently. The contributions are twofold. Firstly, an expressive graph representation is established for cost data modeling based on identification of cost factors and their complex relationships. Secondly, a novel GNN-based deep learning model is developed to predict construction costs, which demonstrates the effectiveness and potential of graph deep learning in conceptual cost estimation.

Of note is that while the proposed GNN-based method is more complex than simple techniques (e.g., regression, decision tress), the complexity is a response to the intricate nature of construction cost estimation where multiple cost factors are interrelated. Simple models fail to capture these interactions, which can lead to significant errors, as shown in the results. We argue that the proposed GNN-based approach provides a structured and systematic method for encapsulating the complex relationships while maintaining operational simplicity. The model's input requirements and the process flow remain straightforward in an end-to-end manner. In addition, the ability to handle non-structured data allows the proposed method to handle irregularities and maintain performance where simpler models might falter.

Although the outcomes of this study are promising, several limitations exist. The formulated graph representation reflects the impact relationships among the specific cost factors, but does not comprehensively capture the hierarchical relationships in the identified two-level cost factor hierarchy shown in Figure 2. In the future, such relationships can be incorporated for more expressive modeling of cost factors. Besides, the black-box nature of the neural networks raises concerns on the explainability of the cost estimation results. Future research is suggested to investigate and integrate advanced explainable artificial intelligence (XAI) techniques to reveal the hidden cost patterns learnt by the model to facilitate a trustable and reliable decision-making process.

Acknowledgement

The authors would like to thank The Government of the Hong Kong Special Administrative Region - Development Bureau for providing partial support to this research, including dataset preparation. Any opinions and findings are those of the authors and do not necessarily reflect the views of The Government of the Hong Kong Special Administrative Region - Development Bureau.

References

[1] He X., Liu R. and Anumba C.J. Data-Driven Insights on the Knowledge Gaps of Conceptual Cost Estimation Modeling. *Journal of Construction Engineering and Management*, 147 (2): 04020165, 2021.

[2] Zhang S., Migliaccio G.C., Zandbergen P.A. and Guindani M. Empirical Assessment of Geographically Based Surface Interpolation Methods for Adjusting Construction Cost Estimates by Project Location. *Journal of Construction Engineering and Management*. 140 (6): 04014015, 2014.

[3] Elmousalami H.H. Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review. *Journal of Construction Engineering and Management*, 146 (1): 03119008, 2020.

[4] Liu H., Cheng J.C.P., Gan V.J.L. and Zhou S. A novel Data-Driven framework based on BIM and knowledge graph for automatic model auditing and Quantity Take-off. *Advanced Engineering Informatics*, 54: 101757, 2022.

[5] Akintoye A. Analysis of factors influencing project cost estimating practice. *Construction Management and Economics*, 18 (1): 77–89, 2000.

[6] Zhang Y., Minchin R.E., Flood I. and Ries R.J. Preliminary Cost Estimation of Highway Projects Using Statistical Learning Methods. *Journal of Construction Engineering and Management*, 149 (5): 04023026, 2023

[7] Akinosho T.D., Oyedele L.O., Bilal M., Ajayi A.O., Delgado M.D., Akinade O.O. and Ahmed A.A. Deep learning in the construction industry: A review of present status and future innovations. *Journal of Building Engineering*, 32: 101827, 2020.

[8] Chen Q., Xia B., Jin Z., Wu P. and Hu Y. Choosing Appropriate Contract Methods for Design-Build Projects. *Journal of Management in Engineering*, 32 (1): 04015029, 2016.

[9] Mostofi F., Toğan V., Başağa H.B., Çıtıpıtıoğlu A. and Tokdemir O.B. Multiedge Graph Convolutional Network for House Price Prediction. *Journal of Construction Engineering and Management*, 149 (11): 04023112, 2023.

[10] Li M., Liu Y., Wong B.C.L., Gan V.J.L. and Cheng J.C.P. Automated structural design optimization of steel reinforcement using graph neural network and exploratory genetic algorithms. *Automation in Construction*, 146: 104677, 2023.

[11] Williams T.P. Predicting final cost for competitively bid construction projects using regression models. *International Journal of Project Management*, 21 (8): 593–599, 2003.

[12] Stoy C., Pollalis S. and Schalcher H.-R. Drivers for Cost Estimating in Early Design: Case Study of Residential Construction. *Journal of Construction*

Engineering and Management, 134 (1): 32–39, 2008.

[13] Wang R., Asghari V., Cheung C.M., Hsu S.-C. and Lee C.-J. Assessing effects of economic factors on construction cost estimation using deep neural networks. *Automation in Construction*, 134: 104080, 2022.

[14] Fang S., Zhao T. and Zhang Y. Prediction of construction projects' costs based on fusion method. *Engineering Computations*, 34 (7): 2396–2408, 2017.

[15] Chakraborty D., Elhegazy H., Elzarka H. and Gutierrez L. A novel construction cost prediction model using hybrid natural and light gradient boosting. *Advanced Engineering Informatics*, 46: 101201, 2020.

[16] Pan Y., Fu X. and Zhang L. Data-driven multi-output prediction for TBM performance during tunnel excavation: An attention-based graph convolutional network approach. *Automation in Construction*, 141: 104386, 2022.

[17] Saeidlou S. and Ghadiminia N. A construction cost estimation framework using DNN and validation unit. *Building Research & Information*, 1–11, 2023.

[18] Kim J. and Cha H.S. Expediting the Cost Estimation Process for Aged-Housing Renovation Projects Using a Probabilistic Deep Learning Approach. *Sustainability*, 14 (1): 564, 2022.

[19] Jafarzadeh R., Ingham J.M., Wilkinson S., González V. and Aghakouchak A.A. Application of Artificial Neural Network Methodology for Predicting Seismic Retrofit Construction Costs. *Journal of Construction Engineering and Management*, 140 (2): 04013044, 2014.

[20] Elmousalami H.H. Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction: A Case Study and Comparative Analysis. *IEEE Transactions on Engineering Management*, 68 (1): 183–196, 2021.

[21] Cheng M.-Y., Tsai H.-C. and Sudjono E. Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. *Expert Systems with Applications*, 37 (6): 4224–4231, 2020.

[22] Lowe D.J., Emsley M.W. and Harding A. Predicting Construction Cost Using Multiple Regression Techniques. *Journal of Construction Engineering and Management*, 132 (7): 750–758, 2006.

[23] Matel E., Vahdatikhaki F., Hosseinyalamdary S., Evers T. and Voordijk H. An artificial neural network approach for cost estimation of engineering services. *International Journal of Construction Management*, 22 (7): 1274–1287, 2019.

[24] Coduto D., Yeung M. and Kitch W. Geotechnical Engineering: Principles & Practices 2nd Edition. Pearson, London, United Kingdom, 2010.

[25] Kumaraswamy M.M. and Chan D.W.M. Determinants of construction duration. *Construction Management and Economics*, 13 (3): 209–217, 1995.

[26] Wu Z., Pan S., Chen F., Long G., Zhang C. and Yu P.S. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32 (1): 4–24, 2021.

[27] Python Software Foundation. Python. On-line: https://www.python.org/, Accessed: 22/11/2022.

[28] Linux Foundation and Meta AI PyTorch. On-line: https://pytorch.org/, Accessed: 28/12/2022.

[29] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. and Duchesnay É. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12 (85): 2825–2830, 2011.

[30] Goodfellow I., Bengio Y. and Courville A. Deep Learning, Illustrated Edition. The MIT Press, Cambridge, Massachusetts, United States, 2016.

[31] Kingma D.P. and Ba J. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA, 2015.