CTGAN-Based Model to Mitigate Data Scarcity for Cost Estimation in Green Building Projects

Eunbin Hong[1]; June-Seong Yi[2]; and Donghwan Lee[3]

Abstract: This study presents a method for estimating construction costs, even when dealing with limited and unreliable data, to enhance decision-making in the early project stages. Owners, particularly in green building projects, often face challenges due to the scarcity of usable data, making cost estimation a complex task. They struggle to differentiate between costs associated with existing buildings and green buildings. To address this issue, we introduce a novel approach that leverages conditional tabular generative adversarial networks (CTGANs) for data augmentation, overcoming the limitations of relying solely on historical data. This involves training an artificial neural network (ANN)–based model using synthetic data, effectively addressing the scarcity and imbalance present in the original small data set. Compared to models trained exclusively on the original data set, our approach yielded a remarkable reduction of approximately 66% in root-mean-square error (RMSE), while increasing the validity from 0% to 15.09%. This study not only improves construction cost estimation but also facilitates more informed decision-making for owners, even in cases with limited and unreliable data, ultimately contributing to the efficiency of the construction project planning process. **DOI: 10.1061/JMENEA.MEENG-5880.** *This work is made available under the terms of the Creative Commons Attribution 4.0 International license, https://creativecommons.org/licenses/by/4.0/.*

Author keywords: Cost estimation; Small data; Data augmentation; Conditional tabular generative adversarial networks (CTGANs); Green buildings.

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

Introduction

In every construction project, the ultimate decision maker is the owner, especially during the crucial preproject stage when decisions are made without the direct involvement of experts (Kim 2004). However, this phase is often confronted with challenges related to data limitations, particularly when estimating construction costs. Each building possesses unique characteristics, and construction projects exhibit inherent discontinuities (Segerstedt and Olofsson 2010; BuHamdan et al. 2019). Taking this into account, it is sometimes unavoidable to use small data to estimate the construction costs. Consequently, there are situations where the available data for cost estimation are both limited and unreliable, yet research adopting a data-driven cost estimation approach is being steadily conducted (Monghasemi and Abdallah 2021; Lee et al. 2024).

This challenge becomes even more pronounced when considering green building projects. Owners with limited construction knowledge and experience often find themselves lacking practical examples to inform their decisions in the realm of green construction. The scarcity of accumulated construction cost data, coupled with the difficulty in assessing its reliability, creates a formidable obstacle. These constraints underscore the need for innovative approaches that can address the dearth of usable data.

Feeley and Silman (2011) have raised similar concerns in the context of species distribution modeling, where "presence-only" data are often insufficient for accurate modeling. They emphasize the importance of improving the availability and quality of existing data to bridge this gap. Analogously, in the realm of construction cost estimation, particularly in green building projects that emphasize energy performance, the task of extracting meaningful training data from sparse historical cost data is formidable. This challenge has prompted research efforts in construction engineering and related fields (Matel et al. 2022; Song et al. 2023; Yang et al. 2023).

Matel et al. (2022) specifically address the problem of ineffective utilization of machine learning algorithms due to data limitations and information constraints. They propose an optimized artificial neural network (ANN)–based method for conceptual cost estimation. Furthermore, construction costs for green buildings differ significantly from those of traditional buildings due to variations in construction methods and energy performance considerations. Predicting the construction cost of a green building based solely on data from existing nongreen structures is inherently limited. With green buildings still in their nascent stage (Wan et al. 2022), research on their construction costs and relevant data remains scarce.

This study aims to develop a method that empowers owners to estimate the construction cost of green buildings even with limited information at the planning stage. Such an approach is vital to support owners' decision-making during the early phases of construction projects. Our primary objective is to develop an estimation model based on a data augmentation and machine learning that can predict construction costs within a reasonable margin of error using approximate information about buildings, such as location and energy performance grades, which can be determined during the planning phase of green building projects. To achieve this, we utilize cost data from certified zero-energy buildings (ZEBs) completed in Korea between 2020 and 2021.

[1]Ph.D. Student, Dept. of Architectural and Urban Engineering, Ewha Womans Univ., Seoul 03760, Korea. Email: heb@ewhain.net

[2]Professor, Dept. of Architectural and Urban Systems Engineering, Ewha Womans Univ., Seoul 03760, Korea (corresponding author). Email: jsyi@ewha.ac.kr

[3]Professor, Dept. of Statistics, Ewha Womans Univ., Seoul 03760, Korea. ORCID: https://orcid.org/0000-0002-3878-6876. Email: donghwan.lee@ewha.ac.kr

Note. This manuscript was submitted on August 26, 2023; approved on January 31, 2024; published online on April 27, 2024. Discussion period open until September 27, 2024; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Management in Engineering***, © ASCE, ISSN 0742-597X.

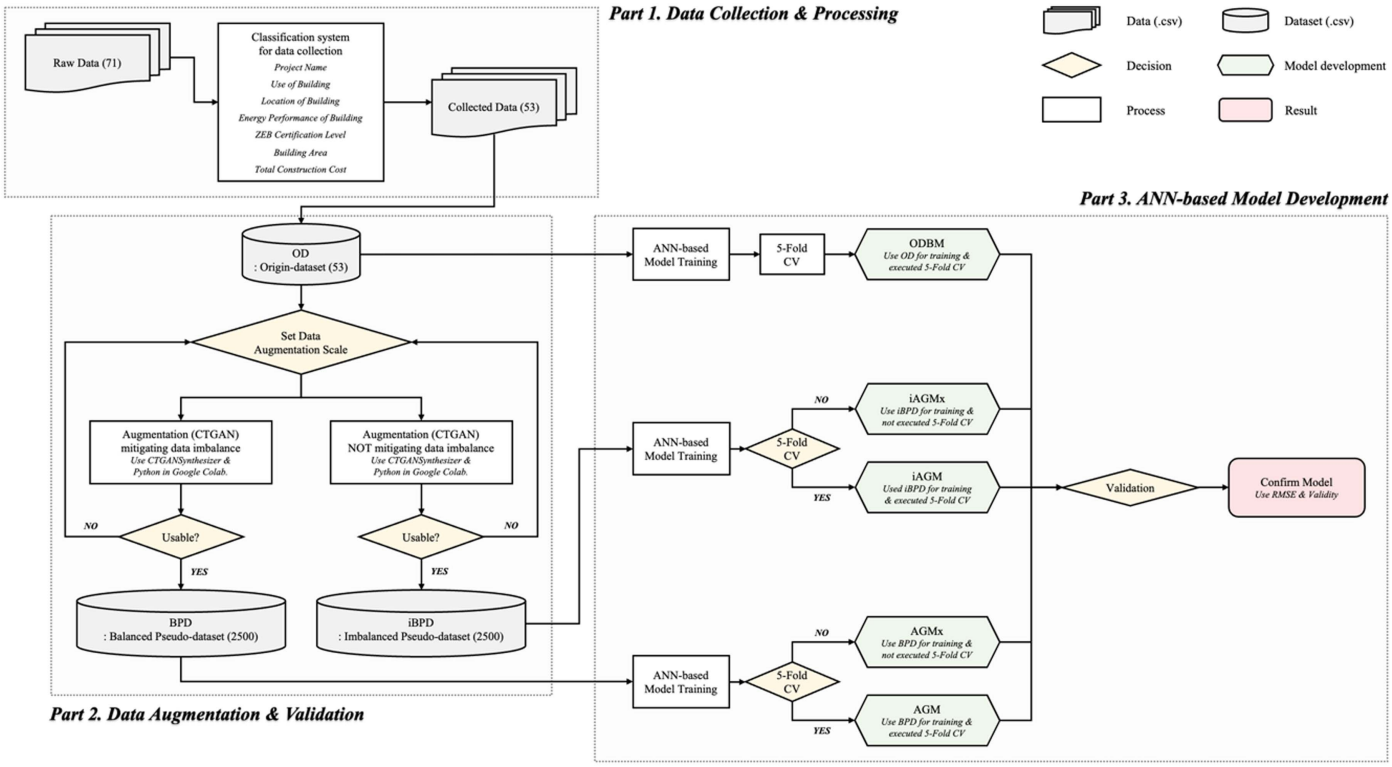

Fig. 1. Research workflow.

The construction cost is influenced by numerous complex factors, making machine learning a suitable tool to provide accurate estimations. However, the quality and quantity of real training data needed for optimal performance are often challenging to obtain. To address this issue, this study explores data augmentation techniques, which have gained prominence in recent research, to estimate construction costs and alleviate data inadequacy. The approach involves augmenting a limited pool of historical data and training a regression model on the augmented data set. Through this research, we evaluate the applicability of data augmentation to enhance construction cost estimation, ultimately paving the way for more effective use of available historical data (Fig. 1).

Literature Review

Supporting the Owners' Decision-Making in Early Stage

As the project progresses, key decisions are sequentially made. Accordingly, project information is incrementally delivered to each participant. By the time the construction stage is reached, the cost of construction is almost fixed. Thus, cost management is performed to ensure the project is completed within the target construction budget, which necessitates working in a timely and effective manner.

Conceptual design covers the period from when the owner identifies their need for the building to the completion of a basic design. At this stage, the requirements of the owner are materialized into project information. The potential for construction cost reductions is according to the progress of the construction project, and it is greatest in the planning stage and decreases as the project progresses (Zedan and Miller 2018). The cost of design change increases as the construction phase progresses. Therefore, calculating the appropriate construction cost at the initial stage is very important. In addition, alternatives could be considered to estimate the construction cost in the planning and design stages. In other words, calculating the construction cost in the predesign stage or in the alternative-analysis stage is essential (Cho 2015). This serves as a budgeting standard for the owner, enables rational design within the calculated amount for the designer, and allows the builder to calculate the execution budget or grasp the construction scale.

As described previously, the owner is the final decision-making entity in a construction project. They establish a budget and plan for project execution at the various stages, which greatly affects the success or failure of the project. Although construction cost estimation in the early stage determines the success or failure of a construction project, estimating it with high reliability and accuracy is difficult (Ahn et al. 2003). Accordingly, several studies have been conducted on providing decision-making support to the owner primarily before the construction project begins to ensure successful completion. In the preproject stage, especially when dealing with green building projects, owners often face challenges in making decisions without the guidance of experts, particularly during the planning and conceptual cost estimation. The process for this heavily relies on the expertise and knowledge of engineers (Holm and Schaufelberger 2021; Pan and Zhang 2021; Takefuji 2021). In the case of Korea, it is encouraged at the governmental level to seek advice from engineers in the field of building energy, especially during the preproject stages of green building projects. Just as the subject needs to understand the basics of design to make proper decisions, the estimator needs accurate cost information for decision-making, and it can be obtained by applying the basic principles of estimation (Carr 1989). Due to the nature of construction work, establishing a standardized construction method or design for accurately estimating the cost is essential. Simply using the arithmetic average of past performance data to make an estimate has marked limitations. To overcome these, regression analysis

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

(Lowe et al. 2006; Fragkakis et al. 2011; Kim 2011; Alshamrani 2017) and simulation (Lee 2004; Sonmez 2008; Kim et al. 2012; Jin et al. 2014) have been applied traditionally and have been developed, and machine learning-based construction cost estimation studies are currently being actively conducted.

Machine Learning for Construction Cost Estimation

Several algorithms for performing machine learning exist, and numerous studies have been conducted to estimate the cost of construction projects using them (Adeli and Wu 1998; Günaydın and Doğan 2004; Wilmot and Mei 2005; Cheng et al. 2010; Han et al. 2011; Lee et al. 2012; Vahdani et al. 2012; Feng and Li 2013; Juszczyk 2013, 2015; Cho 2015; Hyari et al. 2016; Jung et al. 2018; Chandanshive and Kambekar 2019; Kim et al. 2019; Mir et al. 2021), proving the superiority of machine learning in construction cost estimation. In particular, ANN is a deep learning algorithm capable of learning any nonlinear function (Rather et al. 2015) and has the capacity to learn weights that map any input to the output (Zhang et al. 1998). Due to these attributes, ANN-based models are known to exhibit strong estimation performance. Moreover, related studies have been actively conducted until recently. In particular, the ANN-based construction cost estimation model achieved excellent performance, and its superiority has been proven by the accuracy of its results, which exceeds those of expert systems, regression analysis, and simulation, as well as case-based reasoning (CBR), support vector machines (SVM), and random forest–based models (Kim et al. 2000; Günaydın and Doğan 2004; Kim 2003a; Han et al. 2011; Jung et al. 2018; Kim et al. 2022).

Machine learning is garnering attention as an excellent problem-solving method. Reliance on actual accumulated data only is a limitation in all research fields that utilize machine learning (Jordan and Mitchell 2015), including construction cost estimation. Data augmentation has been proposed as a new method for overcoming this limitation. Notably, few studies have utilized it to overcome data limitations (Frid-Adar et al. 2018; Mikołajczyk and Grochowski 2018; Park et al. 2018; Cho and Moon 2019; Tanaka and Aranha 2019; Jo et al. 2020; Kang and Shin 2021). Utilizing source data is essential to securing the performance of machine learning (Xiao et al. 2015; Bailly et al. 2022). Collecting large amounts of high-quality data that is fully representative of the real world has marked limits. Accordingly, insufficient amounts of training data, low-quality data, and underfitting and overfitting training data due to irrelevant features are major challenges facing machine learning presently (Sheng et al. 2008; Jordan and Mitchell 2015; Gudivada et al. 2017). Comparisons between the value predicted by the machine learning model and the actual value reveal that the model's performance is reliable if it simultaneously exhibits low bias and low variance. This study aims to examine data augmentation techniques to improve the performance of machine learning models through solving underfitting or overfitting by increasing the amount of data. This will be used in ANN-based model development for roughly estimating the construction cost. Accordingly, the feasibility of roughly estimating the construction cost at an effective grade by augmenting a small amount of data can be determined.

Data Augmentation for Machine Learning Performance

Data augmentation is a technique for increasing the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data (Cho 2020). It acts as a regularizer and helps reduce overfitting when training a machine learning model (Shorten and Khoshgoftaar 2019). Data can be augmented through these generative models. They learn from given data and generate data with similar distribution as the training data. A generative adversarial network (GAN) is a generative model proposed by Goodfellow et al. (2014). It is mainly applied in more advanced image generation and natural language processing fields. GAN uses two neural networks, unlike existing methods of learning that use one neural network. These two neural networks perform learning antagonistically to each other, namely, adversarial training (Jo et al. 2020). Unlike traditional data augmentation techniques, GAN generates synthetic data through synthesis.

GANs have exhibited superior performance to those of traditional data augmentation techniques and can improve the performance of machine learning models. GANs are differentiated from the traditional techniques in that they synthesize data through adversarial learning. Fig. 2 is the schematic diagram of it, and Goodfellow (2016) describes the framework how GANs work as follows: the basic idea of GANs is to set up a game between two players.

The first player, called the generator, produces samples that are designed to emulate the distribution of the training data. The second player, known as the discriminator, evaluates the samples to discern whether they are real or fake. The discriminator employs conventional supervised learning methods, classifying inputs into either real or fake classes (de Meer Pardo 2019). The generator is trained to deceive the discriminator. In order to emerge victorious in this game, the generator must learn to create samples that are identical to the genuine ones and must originate from the same distribution as the training data (Brownlee 2019). Because of their suitability for augmenting unstructured data such as images, their application to different types of data is being explored. For instance, conditional tabular GAN (CTGAN) is a generative model that can augment tabular structured data, and their ability to improve the performance of machine learning models has been proven in previous studies (Frid-Adar et al. 2018; Mikołajczyk and Grochowski 2018; Park et al. 2018; Cho and Moon 2019; Tanaka and Aranha 2019; Jo et al. 2020; Han et al. 2021; Kang and Shin 2021; Jia et al. 2022; Habibi et al. 2023).

Frid-Adar et al. (2018) used GAN to generate medical images by synthesizing computerized tomography (CT) images associated with liver lesions. (1) A model built from data before augmentation, (2) a model built using data augmented via a traditional method, and (3) a model built based on data augmented using GAN were compared, and the superiority of GAN was demonstrated. Park et al. (2018) proposed a method for improving the performance of machine learning models using training data augmented from a small amount of structured data using GAN. This method improved the performance of the machine learning model corresponding to the increase in structured data, and the feasibility of applying it to various types of data was demonstrated. Cho and Moon (2019) increased the size of data sets by hundreds of times using various methods based on existing image data sets. Using data augmentation to increase the accuracy between image objects by more than 5%–10%, the accuracy of the deep learning network was improved. Tanaka and Aranha (2019) used a GAN to generate artificial training data for machine learning. The benchmark data set was tested using different network architectures, such as imbalanced data sets and data containing sensitive information. Using the training data generated by the GAN, it was demonstrated that the trained decision tree (DT) classifier had the same or higher accuracy than the DT classifier trained on the original data set. Jo et al. (2020) solved the problem of imbalance in intrusion detection data using a CTGAN oversampling model based on GAN. High-quality synthetic data without duplicates could be generated. The superiority of the GAN-based model was demonstrated by comparing its

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

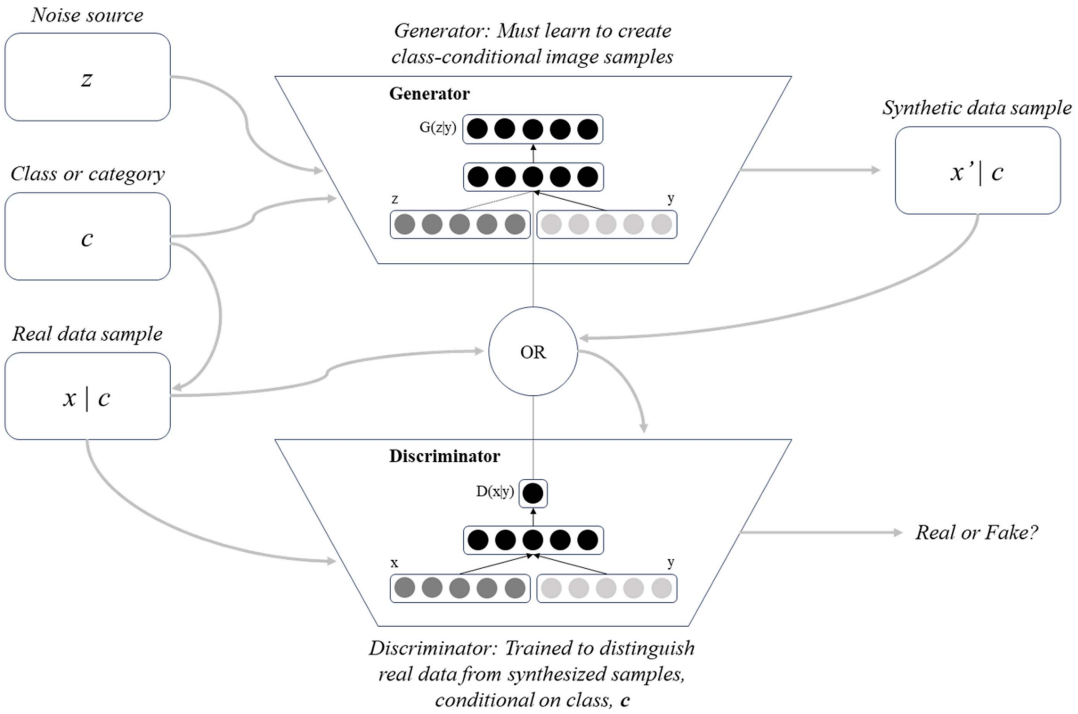

Fig. 2. Adversarial learning pipeline of GAN. (Data from Mizra and Osindero 2014; Creswell et al. 2018.)

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

performance with the traditional augmentation-based models, such as synthetic minority over-sampling technique (SMOTE). Kang and Shin (2021) improved the accuracy of the estimation model by augmenting highly biased data using CTGAN to mitigate its scarcity and imbalance.

While there exist prior studies that utilized traditional data augmentation techniques to predict construction project costs, research employing the CTGAN, which generates high-quality synthetic data through an adversarial learning pipeline, has been scarce to find. Delgado and Oyedele (2021) utilized undercomplete, sparse, deep, and variational autoencoders for predicting construction costs of plant projects. They augmented the data and built a cost prediction process based on deep neural network regressors, demonstrating its performance. Additionally, Takefuji (2021) employed the extra-trees algorithm and random-forest algorithm for data augmentation to establish a regression-based primary construction cost prediction process. They further proved the superior performance of the extra trees–based prediction process. As the validity of the cost prediction process constructed based on augmented construction cost data through prior research has been established, this study aims to prove the effectiveness of the cost prediction process using CTGAN, a generative model that learns from raw data and generates synthetic data resembling the distribution of the training data. This approach differs from conventional data augmentation techniques based on general data sampling, such as autoencoders. We aim to perform cost prediction using ANN, known for their excellence in approximate cost estimation with minimal parameters, as demonstrated by previous studies (Kim et al. 2000; Günaydın and Doğan 2004; Kim 2003a, b; Han et al. 2011; Jung et al. 2018) in the context of primary construction cost estimation. Furthermore, when generating augmented training data for deep learning–based predictions across various fields, we plan to leverage CTGAN, a generative model whose superiority has been demonstrated in comparison to other algorithms in recent research (Jo et al. 2020; Han et al. 2021; Jia et al. 2022; Habibi et al. 2023). In this regard, CTGAN is expected not only to augment the scale of ZEB construction cost data for improving the performance of the ANN-based cost prediction model but also to play a significant role in mitigating issues related to data scarcity and imbalance, thus making previously unusable data usable. The effectiveness and excellence of ANN-based construction cost estimation (Kim et al. 2000; Günaydın and Doğan 2004; Kim 2003a, b; Han et al. 2011; Jung et al. 2018) and data augmentation using CTGAN (Jo et al. 2020; Han et al. 2021; Jia et al. 2022; Habibi et al. 2023) were confirmed by previous studies. In addition, it is validated as a research methodology. In this study, data augmentation is performed using CTGAN to resolve the scarcity and imbalance of real cost data of ZEB. We examine whether unusable data become usable data through an augmentation technique such as CTGAN, and whether the construction cost of a ZEB can be predicted through an ANN-based machine learning model by learning the synthetic data.

Preparatory Analysis for Model Development

Deriving Factors Affecting the Construction Cost of ZEB

In this section, factors affecting the construction cost of new ZEBs are derived from related previous studies. First, we examine previous studies related to the construction cost of ZEBs. Studies deriving an optimal design or analyzing life-cycle cost (LCC) are excluded because the focus was on studies that derived influencing factors or calculated construction costs. Specifically, the primary goal of this study is to build a model for predicting the construction cost of a ZEB. Accordingly, three previous studies were identified (KICT et al. 2019; Hu 2019; Shim and Lee 2021).

KICT et al. (2019) classified factors affecting construction costs into individual, general, and contingent factors. The certification grade, scale, and use of ZEBs were identified as affecting the

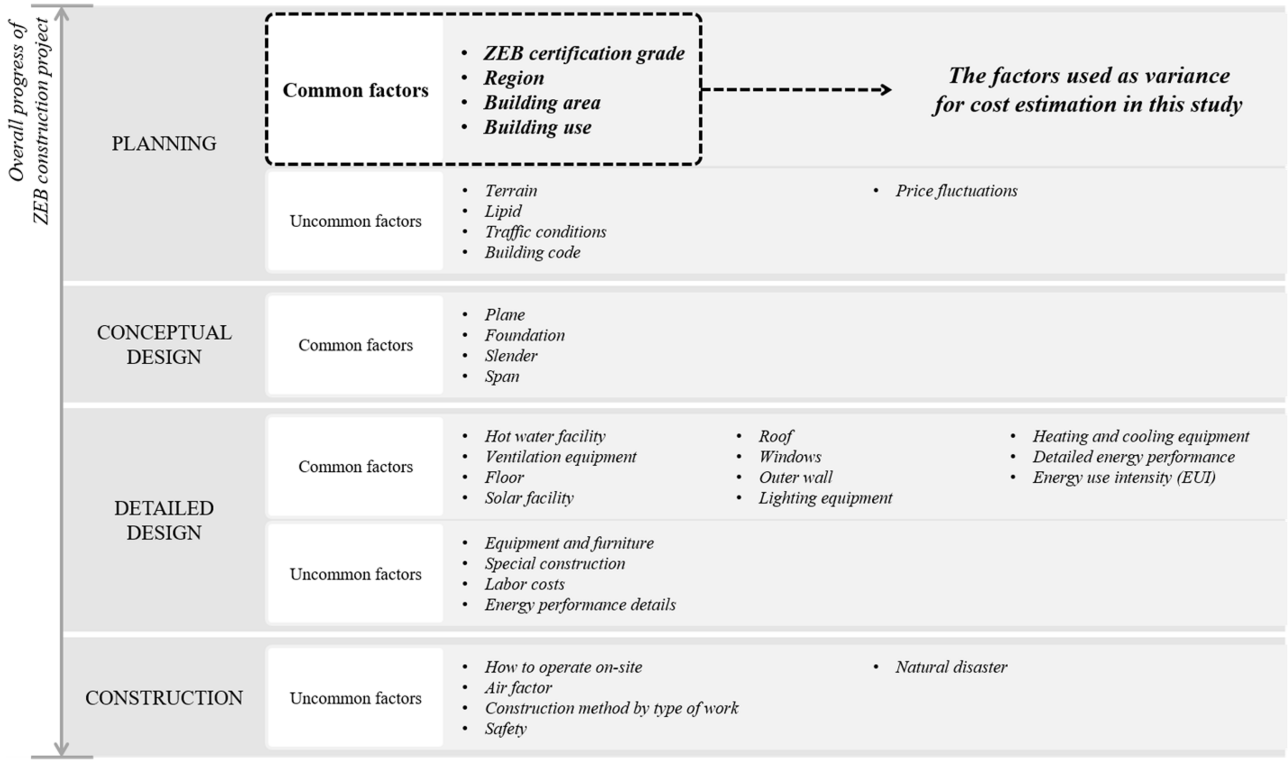

Fig. 3. Cost factors considered in the ZEB construction project in previous studies.

construction cost. Hu (2019) determined the factors affecting construction costs based on the *National Building Cost Manual* and RSMeans' *Square Foot Costs* book and the standard for calculating the construction cost of an existing building. The data collected from the established standards were categorized, and the difference between the construction cost of a ZEB and that of an existing building was analyzed. Location of city/state, use of building, building area, energy use intensity, and year were included in the criteria. Detailed factors affecting the construction cost of ZEBs were also identified. Shim and Lee (2021) analyzed the factors that increase the construction cost of ZEBs. Accordingly, the criteria for collecting construction cost data were presented. Based on a detailed statement of construction cost collected according to the suggested standards, factors affecting the increase in the construction cost of a ZEB were derived. Through expert interviews, construction methods by region, city center, and construction type were derived as factors affecting the construction cost of a ZEB. The target audience was a general contractor with experience in the construction of ZEBs and multiple experts from construction manufacturing and construction companies.

Comparing the factors affecting the construction cost derived from previous studies (KICT et al. 2019; Hu 2019; Shim and Lee 2021) and expert interviews (Hong 2022) confirmed that the derived factors for analyzing ZEB construction cost data in each study were practicable; therefore, factors shown in Fig. 3 as commonly derived are herein utilized. In this process, we categorized the factors that influence the cost variation of energy-efficient buildings, extracted considering both the construction project phases and building energy performance indicators, into two groups. The factors that were commonly utilized in all previous studies were classified as "common factors," while those used in some studies but not universally were designated as "uncommon factors." However, different energy performance indicators of a building were used in each study, and we determined the representative indicator as a ZEB certification grade being used in Korea. It is a system that assigns ratings based on the energy performance of buildings, such as Leadership in Energy and Environmental Design (LEED) a green building certification program used worldwide. Considering the energy harvesting and energy conservation per unit area per year of a building, grades 1 to 5 are given in the order of highest energy performance.

These were classified according to the stage of the construction project in which each factor was determined (Fig. 3). In keeping with the purpose of this study, building aspects that (1) the owner cannot decide at the planning stage or (2) cannot be utilized by the owner to calculate the construction cost should be excluded. Moreover, the factors determined after the basic design stage were excluded. Cost estimation exerts a substantial influence on the project life cycle, particularly during its initial stage. Hence, the primary rationale for dedicating resources to conduct a comprehensive analysis and address uncertainties in cost data at the early stages of process design is that this stage carries the greatest significance in shaping the overall project's economic viability and feasibility, accounting for the customary project life cycle (Cheali et al. 2015). As such, (1) certification grade, (2) region, (3) size, and (4) use were derived as factors affecting the ZEB construction cost in this study.

Setting the Accuracy Standard for Cost Estimation

The standard to be used for evaluating the estimation models is set in this section, and it is based on the suggested accuracy range according to the preproject stage (CII 1986; Bredehoeft et al. 2020; Verzuh 2021). The allowable accuracy range is set conservatively, and an error rate range of ±20% is used as standard in this study (Fig. 4).

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

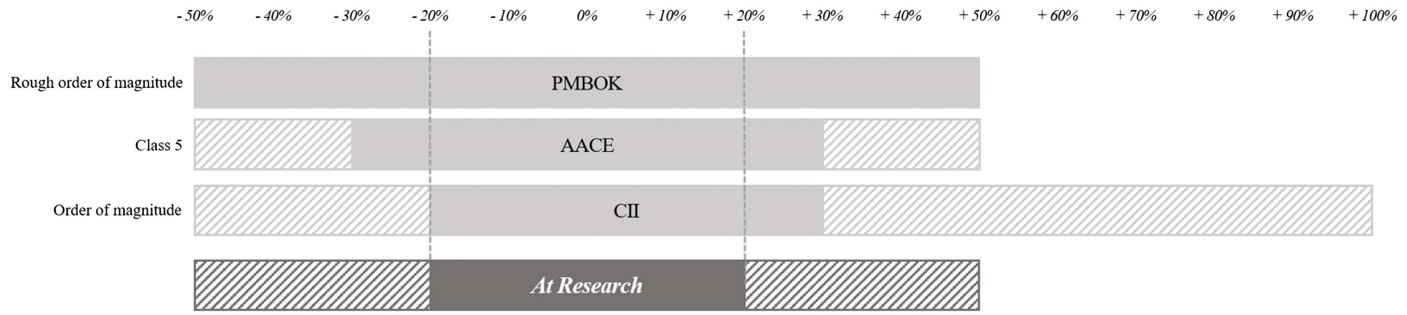

Fig. 4. Setting the accuracy standard for evaluating the estimation performance: PMBOK = project management body of knowledge; AACE = Association for the Advancement of Cost Engineering; and CII = Construction Industry Institute.

Establishing the Classification System to Collect Cost Data

The types and details of information used in construction cost management vary. The information used differs according to the stage of the construction. The design document is not prepared during the planning stage. In this case, historical data are used to estimate the construction cost. Performance data should include construction cost information classified by facility type (office facility, educational facility, etc.), space (room), and part. Therefore, a classification system for different types of construction cost information is needed. The classification system should consider variable factors relevant to the calculation of the construction cost.

Therefore, a classification system for collecting construction cost data was established based on a derivation of factors affecting ZEB construction cost, and it is shown in Fig. 5. Using the classification system, a reliable and accurate model for estimating the ZEB construction cost was established.

CTGAN-Based Estimation Model Development

Data Collection and Preprocessing

We obtained a data set comprising 71 ZEB construction cost breakdowns of certified ZEBs completed in Korea between 2020 and 2021 from the Korean Energy Agency. To ensure data quality, we used a classification system (Fig. 5) during data collection, excluding breakdowns that lacked essential building information or total construction cost. After this filtering process, we were left with 53 high-quality breakdowns, which formed the basis of our data set. We then linked this data set with information regarding the building and its energy performance from the Korean ZEB certification system's webpage (Korea Energy Agency 2023). Consequently, we created a data set of 53 data points, which we subsequently augmented.

For the purpose of our study, we carefully selected predictor variables, including region, building use, and ZEB certification grade. The target variable was set as the construction cost per unit area, often referred to as unit construction cost, calculated by dividing the total construction cost by the total building area. These three predictor variables are known to influence construction costs and are information that building owners can typically obtain during the planning stage (Table 1).

Data Attributes

Our analysis of the raw data set revealed important insights into its characteristics. Figs. 6–8 illustrate the distribution of unit construction costs with respect to the predictor variables. Initially, we observed that establishing a linear relationship between the predictor variables and the target variable was challenging. However, we did observe a weak correlation: as the ZEB certification grade increased, the cost per unit also tended to increase.

Furthermore, the data exhibited an imbalance, with certain grades, building uses, and regions being overrepresented. Notably,

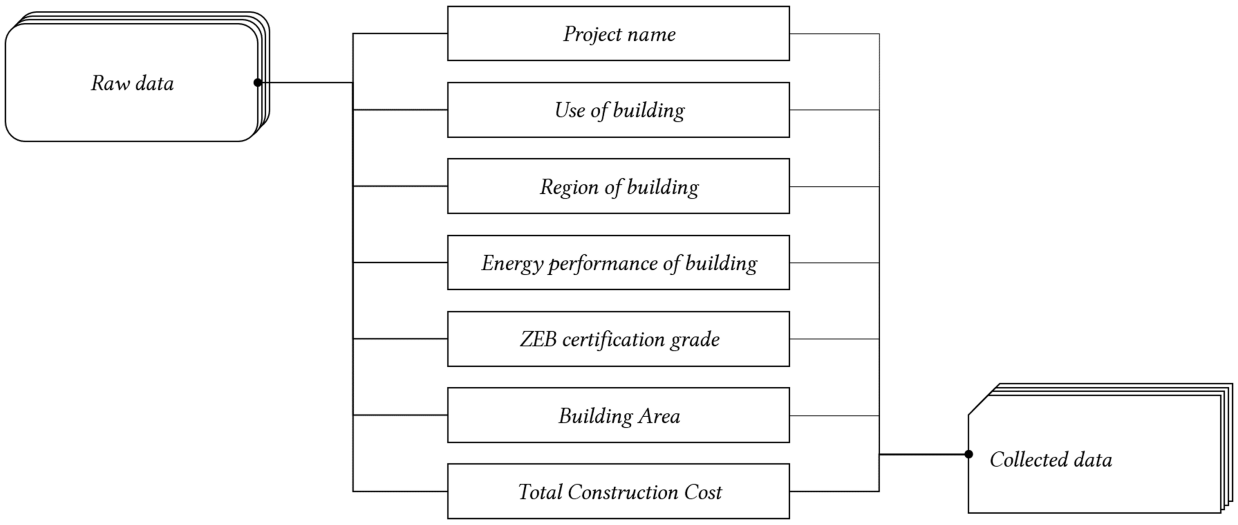

Fig. 5. Classification system for collecting cost data.

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

Table 1. Encoding variable for model training

Classification	Factor	Variable name	Variable specification	Encoding	Data type
Predictor variable	Region	Region	Narrow central region	0	Nominal
			Wide central region	1	
			Southern region	2	
			Island region	3	
	Building use	Use	Educational research facilities	0	Nominal
			Child and geriatric welfare institution	1	
			Detached house	2	
			Power plant facilities	3	
			Business facilities	4	
			Residential neighborhood facilities: Class 1	5	
			Residential neighborhood facilities: Class 2	6	
			Cultural and assembly facilities	7	
	Building area	Grade	Grade 1	1	Ordinal
			Grade 2	2	
			Grade 3	3	
			Grade 4	4	
			Grade 5	5	
Target variable	Total construction cost per unit area (KRW/m^2)	Unit cost	—	—	Continuous

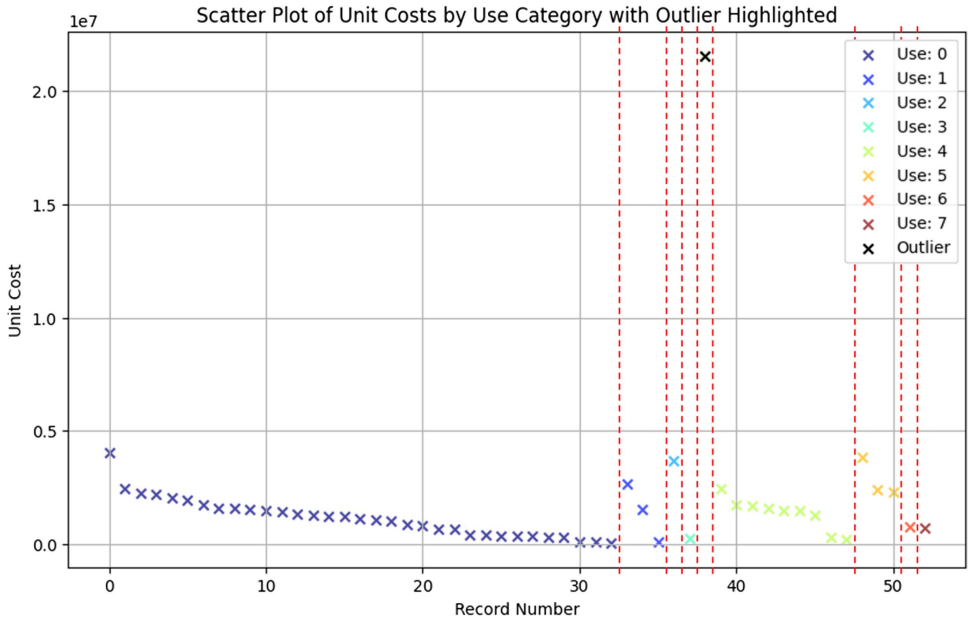

Fig. 6. Data distribution according to use.

we identified an outlier, referred to as "Building I," which stood out due to its exceptionally high cost per unit compared to other samples. Although this outlier could be considered for removal due to its statistical attributes, we decided to retain it because it provided valuable cost information based on actual construction cost breakdowns.

To summarize the data set attributes: there is no clear linear relationship, the data set exhibits class imbalance, and there is an outlier. Considering these attributes, we identified key considerations for developing our construction cost estimation model. We recognized the need for a model based on a nonlinear function, addressing class imbalance, and evaluating the impact of the potential outlier. Given the limited data, class imbalance, and the presence of the outlier, we adopted two strategies: (1) using artificial neural networks for its superior estimation capabilities and (2) employing CTGAN to augment our data set based on the energy performance of ZEBs.

Data Augmentation Process

Our data set construction comprised three main steps. First, we created the original raw data set (OD) following the outlined process. Second, we generated a synthetic data set (iBPD) by augmenting the data without considering the raw data set's attributes, maintaining the same class weights as the original data. Third, we constructed an augmented synthetic data set (BPD) by considering the raw data set's attributes and maintaining class weight parity. These three data sets were created:

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

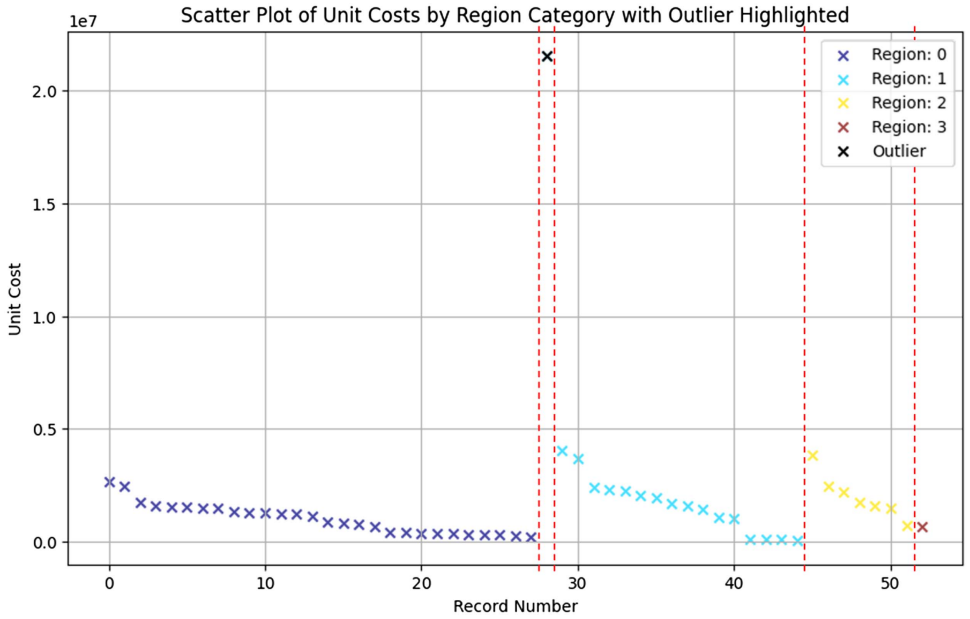

Fig. 7. Data distribution according to region.

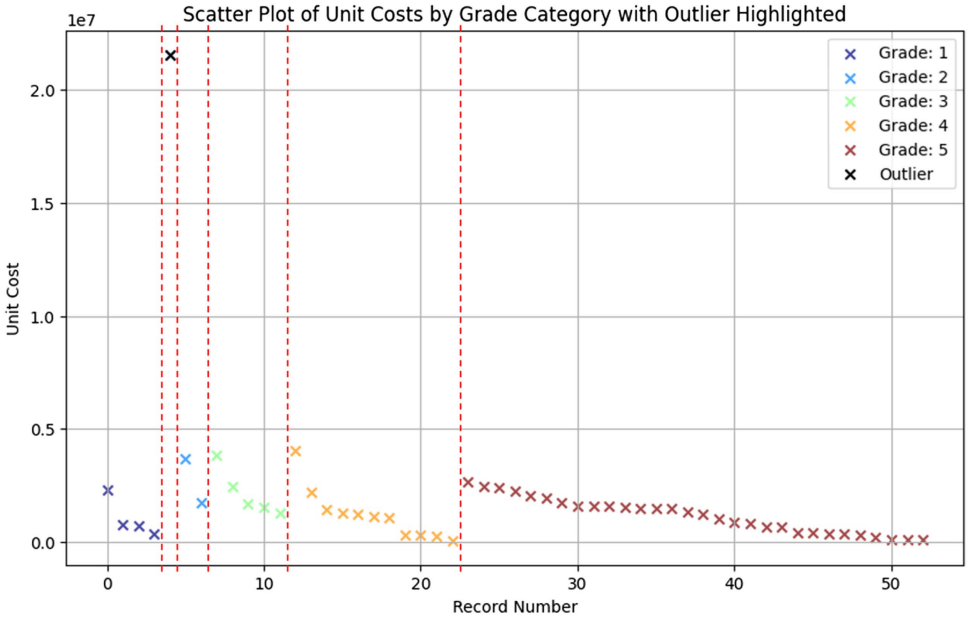

Fig. 8. Data distribution according to grade.

- OD (original data set): the preprocessed data set
- iBPD (imbalanced pseudo–data set): a synthetic data set that does not address class imbalance
- BPD (balanced pseudo–data set): a synthetic data set that rectifies class imbalance

Class imbalance can negatively impact machine learning model performance (Lee 2020). Thus, when augmenting data, it is essential to address this imbalance. We conducted a detailed analysis to compare the performance of estimation models trained on synthetic data sets with and without consideration of data attributes.

Our data augmentation was carried out using Python with CTGAN, an open-source library provided by the Synthetic Data Vault (SDV) project at the Massachusetts Institute of Technology's Data to AI Lab (Xu et al. 2019). After generating synthetic data sets by augmenting the raw data, we conducted a thorough examination of their statistical attributes to determine if the synthetic data sets could replace the original data set. To facilitate this verification, we utilized the "evaluate" function, an open-source tool provided by SDV (Figs. 9 and 10).

When determining the appropriate augmentation scale, it is crucial to consider that machine learning, particularly with ANNs, typically requires thousands of data points for effective training (Brownlee 2017; Mitsa 2019). Model performance tends to improve with larger data sets (Jordan and Mitchell 2015; Xiao et al. 2015; Bailly et al. 2022). We initiated data augmentation by expanding the 53 raw data points to 1,000 data points. We evaluated

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

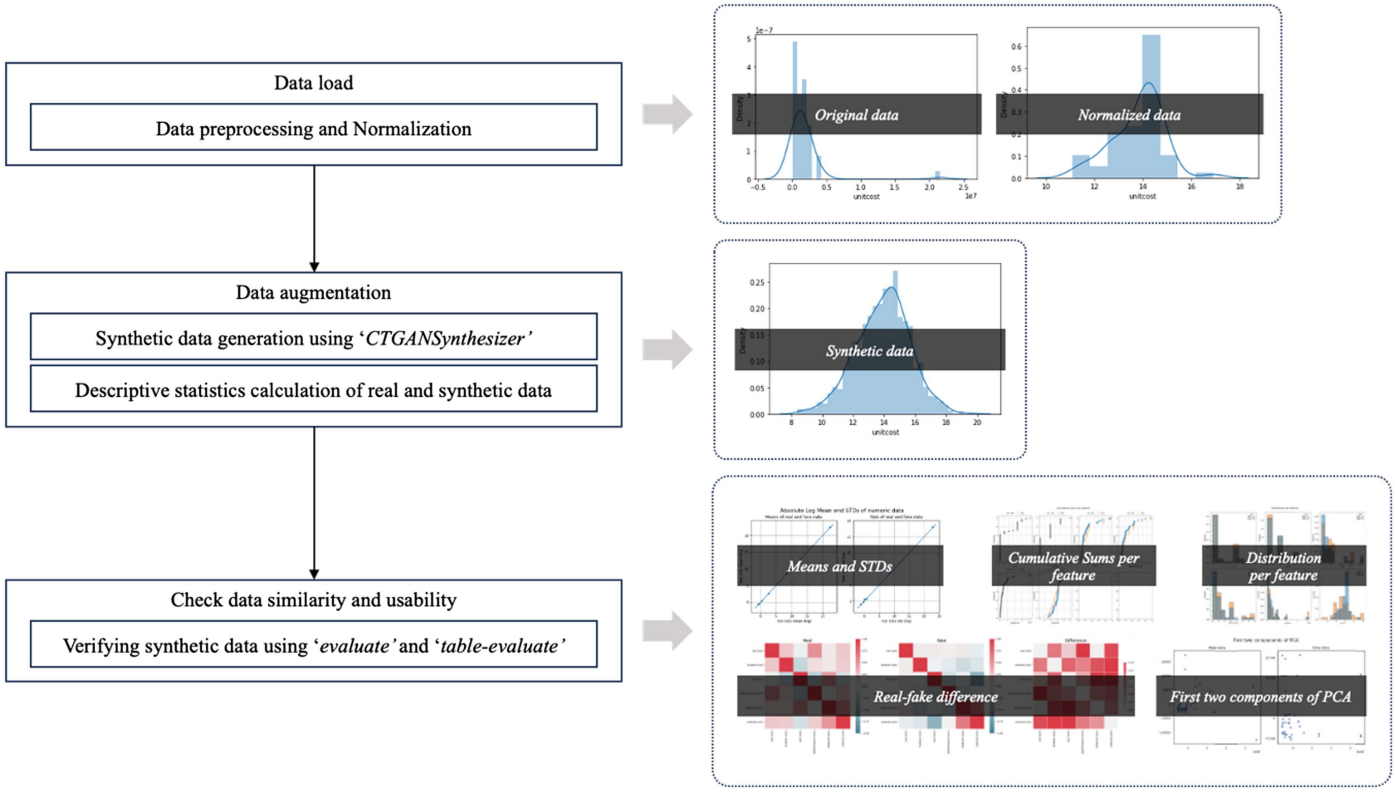

Fig. 9. Data augmentation and verifying process using CTGAN.

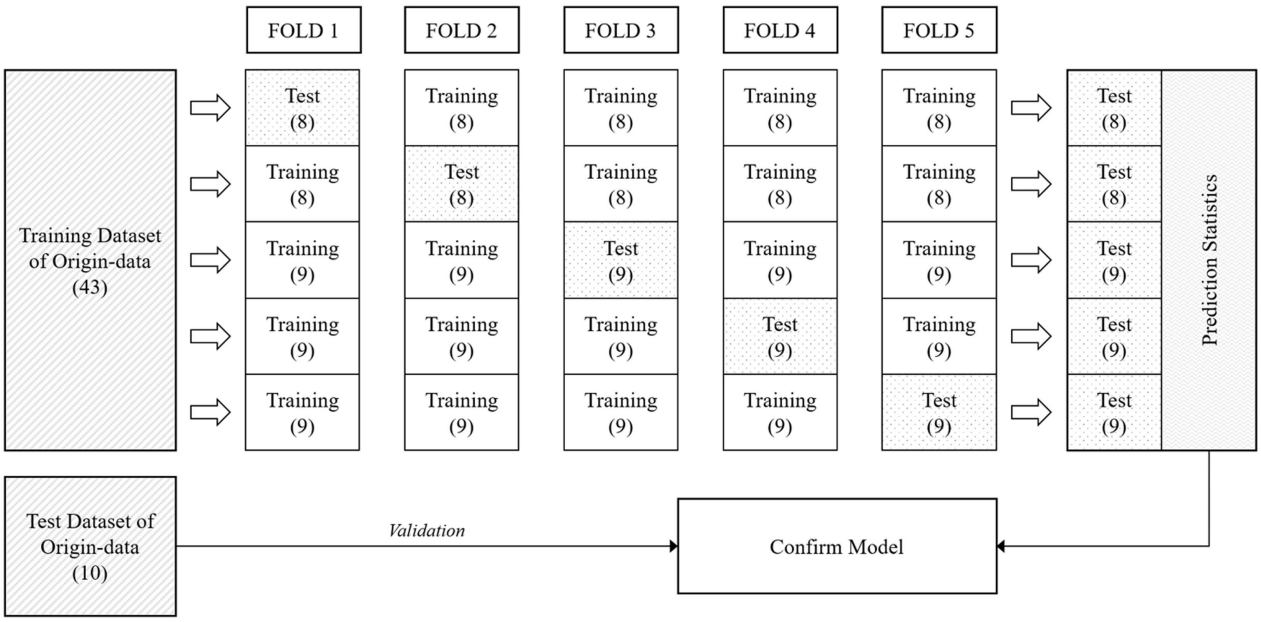

Fig. 10. Fivefold cross-validation for OD training.

the quality of this synthetic data and then gradually increased the augmentation scale, verifying the generated synthetic data at each step (Fig. 11).

Verification of Synthetic Data

To ascertain the adequacy of the synthetic data augmentation, we conducted two main examinations. First, we checked whether the synthetic data accurately captured the attributes of the raw data, and second, we evaluated their suitability for use as training data. We employed the normalized unit cost value, taking into account the data set's distribution and attributes. The statistical indicators used for verifying the synthetic data include the following:

- Data distribution plot
- Means and standard deviation (STD) plot
- Cumulative sums per feature

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

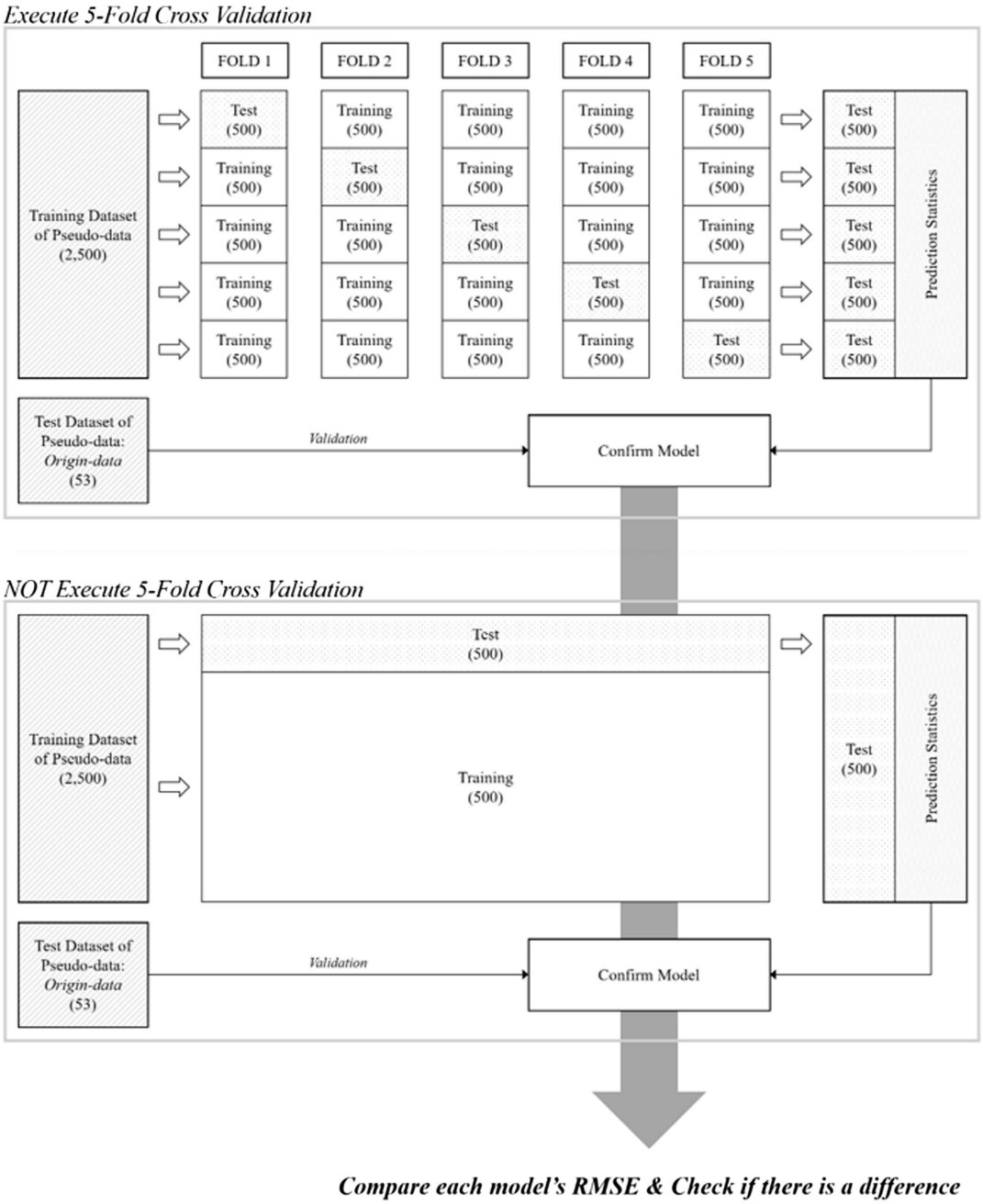

Fig. 11. Fivefold cross-validation for iBPD and BPD training.

- Distribution per feature
- Box plot
- Descriptive statistics (size, mean, STD, minimum, quartile, maximum)
- Similarity

These indicators were utilized to determine whether the synthetic data sufficiently represented the attributes of the raw data. Additionally, they were employed to assess the suitability and usability of the synthetic data for training purposes. We examined two key aspects: (1) the descriptive statistical attributes and (2) the similarity between the raw data and synthetic data. These metrics were calculated using "evaluate," an open-source library provided by SDV for synthetic data verification. The following indicators were used to verify the feasibility of substituting raw data with synthetic data and to determine whether the synthetic data effectively captured the attributes of the raw data, ensuring their suitability for use as training data. It is worth noting that we also employed a linear regression coefficient to assess whether the trends observed in the real data were preserved in the synthetic data. The assessment of synthetic data utility typically relies on analysis-specific criteria, which involve comparing data summaries and/or model coefficients fitted to synthetic data with those obtained from the original data. When the results from both the original and synthetic data align, the synthetic data are considered to exhibit high utility (Snoke et al. 2018).

Data Augmentation Results

A total of 53 raw data points were augmented to generate synthetic data with similar attributes, including mean, deviation, and correlation. We compared the descriptive statistics of the synthetic data set and the similarity with the original data set for augmenting the

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

Table 2. Descriptive statistics of normalized unit cost value for each case

Classification	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Data size	500	1,000	1,500	2,000	2,500	3,000
Mean (difference with OD)	13.97 (−0.18)	13.90 (−0.10)	13.77 (0.03)	13.83 (−0.03)	13.92 (−0.12)	13.85 (−0.06)
STD (difference with OD)	1.69 (−0.62)	1.70 (−0.63)	1.72 (−0.65)	1.70 (−0.63)	1.74 (−0.67)	1.71 (−0.64)
Minimum (difference with OD)	8.59 (2.50)	8.42 (2.67)	8.55 (2.54)	8.42 (2.67)	8.38 (2.71)	8.38 (2.71)
Lower quartile (difference with OD)	12.85 (0.08)	12.85 (0.08)	12.68 (0.26)	12.70 (0.23)	12.79 (0.14)	12.73 (0.20)
Median (difference with OD)	14.07 (−0.02)	14.02 (0.04)	13.82 (0.24)	13.97 (0.09)	14.04 (0.02)	13.96 (0.10)
Upper quartile (difference with OD)	15.11 (−0.73)	14.99 (−0.61)	14.95 (−0.58)	14.98 (−0.61)	15.06 (−0.68)	14.97 (−0.60)
Maximum (difference with OD)	19.76 (−2.87)	19.76 (−2.87)	19.71 (−2.82)	19.00 (−2.12)	19.76 (−2.87)	19.76 (−2.87)
Similarity, %	69.16	68.89	71.08	71.87	71.38	70.53
Multiple linear regression coefficient (difference with OD)	0.488 (0.081)	0.471 (0.098)	0.500 (0.069)	0.473 (0.096)	0.480 (0.089)	0.465 (0.104)

raw data to 500, 1,000, 1,500, 2,000, and 2,500 scales and examined whether it could be usable. The descriptive statistics examined for verifying the generated synthetic data in each case, the similarities, and the linear regression coefficient representing the explanatory power of the data set are listed in Table 2.

Our goal was to obtain synthetic data that closely resembled the raw data in terms of attributes. Ultimately, we determined that the synthetic data corresponding to a scale of 2,500 were the most suitable for building a construction cost estimation model.

We successfully confirmed that data augmentation was performed effectively and produced meaningful results. The codes used to create iBPD and BPD were identical (as listed in Table 1). iBPD exhibited an 89.4% similarity to the original data, and a comparison of descriptive statistics between the two data sets confirmed minimal differences (Table 3). BPD was created by uniformly increasing the data scale for each class to 500, resolving class imbalance. It had a 76.3% similarity to the raw data, and comparing the descriptive statistics of the two data sets also showed minimal differences (Table 3). In conclusion, both iBPD and BPD accurately reflected the attributes of the raw data set, confirming their suitability for training machine learning models.

Develop Estimation Models

We developed five estimation models for predicting the construction costs of ZEBs using the Machine Learning and Statistics Toolbox in MATLAB R2021b. These models were created by varying three conditions: (1) whether to augment data, (2) whether to address class imbalance, and (3) whether to perform cross-validation (CV). Each model was named according to its specific conditions:

- ODBM (OD-based model): trained using the original data set
- iAGMx (iBPD-based model): trained using iBPD without fivefold CV
- iAGM (iBPD-based model): trained using iBPD with fivefold CV
- AGMx (BPD-based model): trained using BPD without fivefold CV
- AGM (BPD-based model): trained using BPD with fivefold CV

Table 3. Descriptive statistics of OD, iBPD, and BPD

Descriptive statistics	OD	iBPD	BPD
Amount	53	2,500	2,500
Mean	13.797733	13.8809	14.1476
STD	1.069416	1.6702	1.3741
Minimum	11.086522	8.4151	10.9857
Lower quartile	12.931074	12.8037	13.1223
Median	14.058629	13.9757	14.1614
Upper quartile	14.377147	15.0397	15.1053
Maximum	16.86025	19.7589	16.7191

Our initial analysis revealed a low correlation between variables within the OD data set, necessitating fivefold CV to evaluate model performance effectively. As such, we performed fivefold CV, dividing OD into five subsets. Each model was trained on 80% of the data in each fold and tested on the remaining 20% for five iterations.

The model built based on OD was named ODBM. The amount of synthetic data augmented with CTGAN was not small, namely, 2,500. Specifically, the difference in the model performance depended on whether CV was investigated. Therefore, second, we built a model based on iBPD that did not perform fivefold CV. This model was trained on 80% of iBPD and tested on 20% of them (Fig. 12), and it was named iAGMx.

Third, we built a model that performed fivefold CV based on iBPD. Accordingly, the iBPD was divided fivefold: this model was trained on 80% of data in each fold and tested on 20% of it for five times (Fig. 12), and it was named iAGM. Fourth, we built a model that did not perform fivefold CV based on BPD, and this model was trained on 80% of BPD and tested on 20% of them, and it was named AGMx. Fifth, we built a model that performed fivefold CV based on BPD: the BPD was divided fivefold, and this model was trained on 80% of each fold and tested on 20% of it for five times. The corresponding model was named AGM.

Validation of Estimation Model

Two indicators were utilized to validate the model. The first indicator is the residual. The evaluation index for measuring the performance of the model depends on the algorithm used for learning and the goal to be achieved through it (Kang and Shin 2021). ANNs involve learning in the direction of minimizing the mean square error (MSE) that occurs in the hidden input and output layers. Therefore, to evaluate the performance of the ANN model, the MSE should be evaluated. RMSE is generally used to evaluate the performance of a regression model because this gives a large penalty when a large error value occurs, the influence of outliers that cannot be discriminated through mean absolute error (MAE) and MSE can be removed. Herein, the performance of the ANN-based regression learning model in this study is evaluated through RMSE and MSE.

The second indicator is "validity," which is defined as the ratio of valid predicted values among all values, and it is operationally used as the accuracy of an estimation model to verify the usability of the model. Previously, developing an estimation model required separating training and test data to verify that the model trained with the training data achieves adequate estimation performance on new data. The predicted values obtained by inputting the test data were compared with the actual values. An error rate between

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

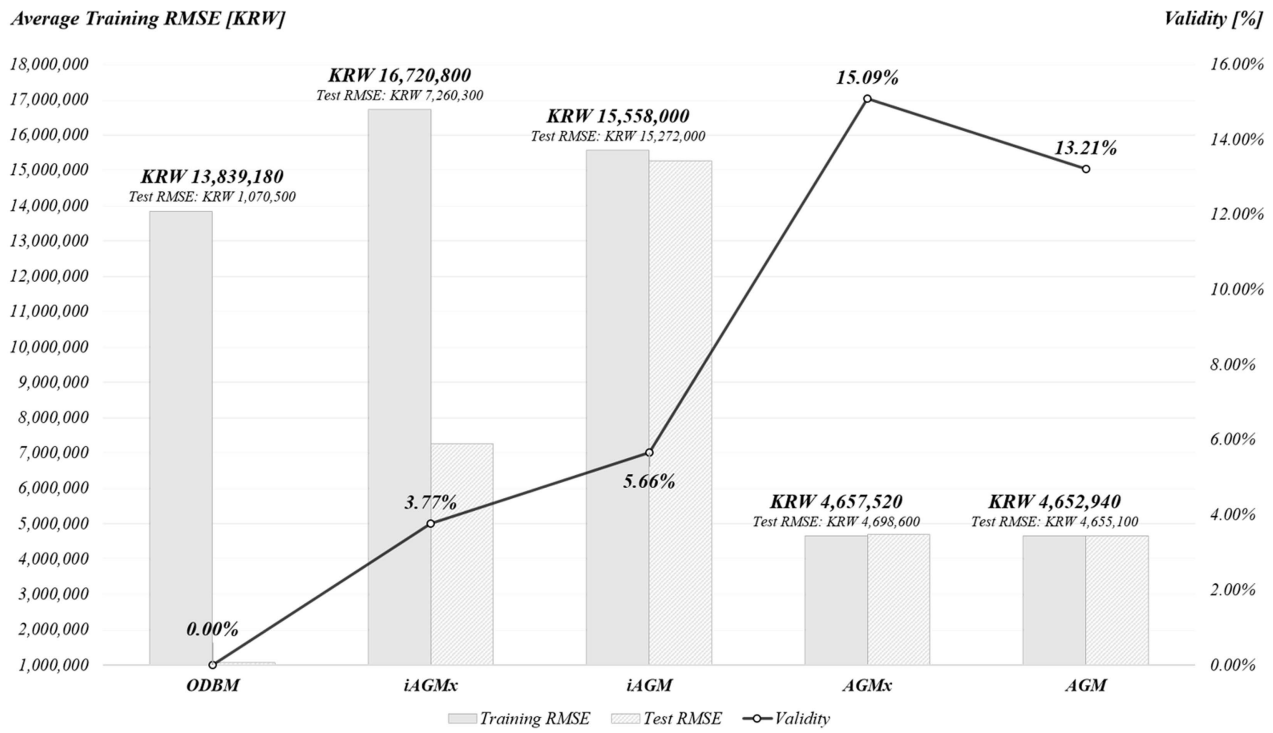

Fig. 12. Estimation performance and validity of each model.

them of [−20%, 20%] was defined as valid, and outside this range were defined as not valid.

Model Estimation Results

Fig. 12 summarizes the estimation performance and validity of the five models: ODBM, iAGMx, iAGM, AGMx, and AGM. Table 4 provides the average error rates for valid estimation results produced by each model. The RMSE for ODBM was 13,839,180 (13.84 million KRW), indicating the approximate difference between actual and predicted construction costs. However, the validity for ODBM was 0.00%, with 0 out of 10 test data points achieving valid estimation results.

For iAGMx, the RMSE was 16,720,800 (16.72 million KRW), and the validity improved to 3.77%, with two out of 53 test data points yielding valid results. iAGM, with an RMSE of 15,558,000

Table 4. Valid estimation results of each model

Model	Variable				Error rate of predicted value (%)
	Name	Use	Region	Grade	
ODBM			—		
iAGMx	AP	Detached house	Wide central region	2	7
	AH	Residential neighborhood facilities: Class 1	Wide central region	5	−10
iAGM	AG	Residential neighborhood facilities: Class 1	Southern region	3	2
	AP	Detached house	Wide central region	2	18
	AH	Residential neighborhood facilities: Class 1	Wide central region	5	−7
AGMx	D	Child and geriatric welfare institution	Narrow central region	5	14
	AH	Residential neighborhood facilities: Class 1	Wide central region	5	14
	J	Educational research facilities	Wide central region	5	10
	AV	Educational research facilities	Wide central region	5	1
	AC	Educational research facilities	Wide central region	5	−4
	T	Educational research facilities	Southern region	5	−11
	AD	Business facilities	Southern region	5	−15
	AQ	Educational research facilities	Wide central region	4	−12
AGM	AG	Residential neighborhood facilities: Class 1	Southern region	3	−10
	AP	Detached house	Wide central region	2	−12
	D	Child and geriatric welfare institution	Narrow central region	5	2
	BB	Educational research facilities	Narrow central region	3	−13
	AH	Residential neighborhood facilities: Class 1	Wide central region	5	−2
	J	Educational research facilities	Wide central region	5	−6
	AV	Educational research facilities	Wide central region	5	−17

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

(15.56 million KRW), achieved a 5.66% validity rate, producing valid results for three out of 53 test data points.

In the case of AGMx, the RMSE was 4,657,520 (4.66 million KRW), and the model achieved a 15.09% validity rate, with 8 out of 53 test data points delivering valid results. Lastly, AGM exhibited an RMSE of 4,652,940 (4.65 million KRW) and a validity rate of 13.21%, with seven out of 53 test data points being valid.

Across all models, the test RMSE was lower than the training RMSE, suggesting that the models produced appropriate estimations. The presented results offer a comprehensive overview of the performance and validity of each model.

Discussion

Estimating Construction Costs with Limited Data

This study aimed to propose a method for estimating construction costs, even when faced with the challenge of limited and unreliable data. We recognize that the construction industry encompasses a wide range of building types, construction methods, and materials. This diversity often results in small and unreliable data sets, making accurate cost estimation a complex task, particularly for nonexperts. To address this issue, we chose ZEBs as a representative example due to the inherent challenges posed by their data. Our approach involved not only overcoming the limitations of a small data set but also addressing the uncertainty associated with data reliability. We achieved this by leveraging CTGAN, a generative adversarial network-based data augmentation technique. Through our experimental results, we demonstrate the effectiveness of our approach.

Model Performance Evaluation

We initiate our discussion by evaluating the performance of various estimation models developed in this study. These models were constructed under different conditions, including data augmentation, class imbalance resolution, and cross-validation. Here, we summarize the key findings.

ODBM

Our initial model, OD-Based Model (ODBM), produced valid estimation results for none (0%) of the 10 test data points. Notably, the lowest error rate was observed for Building F, which belongs to the class with the largest amount of data. Conversely, buildings with specific uses, such as power plant facilities (3), residential neighborhood facilities: Class 1 (5), and cultural and assembly facilities (7), exhibited significant estimation errors, ranging from −1,861% to −3,386%. This pattern was also observed for buildings located in the southern region (2) and those belonging to certification grades 1, 2, and 3. In essence, ODBM struggled to provide valid estimates, primarily due to the scarcity of the training data and the presence of an outlier (Building I).

iAGM and iAGMx (iBPD-Based Models)

We proceeded to explore the impact of data augmentation on model performance through iAGM and iAGMx, which were trained on iBPD generated by augmenting OD. The RMSE of iAGM was 15,558,000 KRW, higher than that of ODBM (13,839,180 KRW), and it achieved valid estimation results for three (5.66%) of the 53 test data points. Meanwhile, iAGMx exhibited an RMSE of 16,720,800 KRW, displaying similar performance to ODBM (13,839,180 KRW), and achieving valid estimates for two (3.77%) out of 53 test data points. Three important conclusions can be drawn from these findings. Firstly, cross-validation improves model performance when the data are inadequate for training. Secondly, data augmentation without resolving data imbalance can exacerbate performance issues. Finally, neither iAGM nor iAGMx provides satisfactory results for practical use in cost estimation at the planning stage.

AGM and AGMx (BPD-Based Models)

In our pursuit to overcome data scarcity and imbalance, we examined the performance of AGM and AGMx, trained on BPD, which resolved both issues. AGM achieved an RMSE of 4,652,940 KRW, significantly lower than ODBM (13,839,180 KRW), iAGM (15,558,000 KRW), and iAGMx (16,720,800 KRW). Impressively, it delivered valid estimates for seven (13.21%) of the 53 test data points. AGMx similarly exhibited an RMSE of 4,657,520 KRW, on par with AGM (4,652,940 KRW), and provided valid estimates for eight (15.09%) out of 53 test data points. These outcomes yield three notable conclusions. Firstly, when augmenting data, addressing data imbalance is crucial. Secondly, with high-quality data suitable for model training, satisfactory performance can be achieved without cross-validation. Lastly, the presence of an outlier, such as Building I, is mitigated during the data augmentation process using CTGAN, resulting in improved model performance.

Interpretation of Key Findings and Implications for Practice

In conclusion, both AGM and AGMx outperformed other models in terms of estimation performance (Fig. 12). While they demonstrated relatively low validity in estimation results, their significance lies in their contribution to proposing a novel construction cost estimation method, particularly when dealing with limited data. It is important to note that the accuracy of these models can be further improved with the accumulation of diverse raw cost data for ZEBs, facilitating future data augmentation efforts. This study underscores the potential of data augmentation techniques, such as CTGAN, in addressing data limitations and improving estimation models' performance. Additionally, it highlights the importance of data quality, class imbalance resolution, and the management of outliers in construction cost estimation. Looking ahead, future research directions may include the collection of more comprehensive and diverse data sets, the exploration of advanced data augmentation methods, and the refinement of machine learning models to enhance accuracy and applicability in the construction cost estimation domain.

This study's findings make a substantial contribution to the body of knowledge in construction cost estimation, particularly in the context of green buildings. Our research demonstrated that the construction cost of green buildings could be effectively estimated using data augmentation techniques and machine learning models, addressing the challenges posed by data scarcity and imbalance. The use of CTGAN for data augmentation proved to be a vital tool in overcoming the limitations of small data sets. By augmenting 53 original data points to create a synthetic data set, we were able to provide a more robust and comprehensive foundation for training our machine learning models. The fact that the augmented data set, at a scale of 2,500, closely resembled the raw data in terms of key attributes like mean, deviation, and correlation, underscores the efficacy of our data augmentation approach. Moreover, the development of different estimation models, each catering to various conditions such as class imbalance and cross-validation, showcased the flexibility and adaptability of our approach. The varying levels of RMSE and validity across these models highlighted the nuanced understanding required in model selection and optimization for accurate cost estimation in green buildings. For practitioners in the field of construction management and green

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

building projects, our findings offer several practical implications. Firstly, the demonstrated ability of ANNs to predict construction costs with considerable accuracy, even with limited data, is a significant advancement. This implies that project managers and decision makers can rely on more accurate cost predictions in the early stages of green building projects, facilitating better budget planning and resource allocation. Secondly, our research highlights the importance of considering data attributes such as class imbalance and the scale of data augmentation. This insight is crucial for practitioners who are looking to apply machine learning techniques in their construction cost estimation processes.

Conclusions

This study underscores the potential of data augmentation techniques, specifically CTGAN, in overcoming data limitations to estimate construction costs, particularly in cases where data availability is limited or in the early stages of adoption, as is often the case with ZEBs. By augmenting a small data set with CTGAN, we successfully built construction cost estimation models based on both raw and synthetic data. Two significant findings emerge from our research. Firstly, we confirmed the effectiveness of data augmentation by comparing models trained on nonuniformly augmented raw data with those trained on uniformly augmented raw data. The superior performance of models learning from synthetic data demonstrates the viability of CTGAN as an effective data augmentation method. Secondly, the efficacy of CV in enhancing model performance was highlighted through comparisons between models with and without CV. Our research indicates that models trained on synthetic data achieved lower error rates in cost estimation, reinforcing the value of CTGAN for improving construction cost forecasting accuracy. In essence, our study demonstrates the feasibility of augmenting construction cost data to enhance the accuracy and validity of estimation results. This approach proves valuable for predicting construction costs in scenarios where data availability is constrained. Furthermore, the application of CTGAN to augment construction cost data contributes to the field of construction cost estimation and offers a novel method to address data limitations, reducing the reliance on historically accumulated data. This approach is poised to benefit future research endeavors, particularly in determining factors influencing green buildings' construction costs during the planning stage, drawing from previous research and expert insights. This study is not without its limitations. The reliance on data from a specific geographic region (Korea) and a limited time frame (2020–2021) may affect the generalizability of the findings. Future research could address this by incorporating a more diverse data set spanning different regions and periods.

Despite the constraints imposed by the limited data set, this study lays the foundation for future improvements and investigations. Several avenues for future research are proposed:

- Addressing data imbalance: Observed across all classes in this study, data imbalance was primarily addressed for certification grade classes. Future research should focus on developing data augmentation techniques to achieve a balanced distribution across all classes. This approach aims to provide a more comprehensive and equitable data set, enhancing the overall model performance.
- Incorporating continuous variables: While this study concentrated on ordinal variables related to a building's energy performance, there is an opportunity to expand the range of input features. Future studies could explore the integration of additional variables that might affect the construction cost of green buildings. The research plans to incorporate continuous variables such as material costs, labor rates, and economic factors. This expansion will allow for a more versatile and effective cost estimation model, particularly beneficial for small data scenarios.
- Data quality enhancement: Emphasizing the crucial role of data quality in model performance, future efforts will prioritize data quality assurance, validation, and the elimination of outliers. These steps are essential for improving the robustness and reliability of the constructed models.

In summary, this study has demonstrated the efficacy of data augmentation using CTGAN for improving construction cost estimation models in scenarios with limited and unreliable data. Future endeavors will focus on addressing data imbalance, incorporating continuous variables, and enhancing data quality to advance the accuracy and applicability of construction cost estimation models, ultimately aiding in the establishment of initial plans for buildings, especially those with unique attributes such as green buildings.

The result of this study presents a significant advancement in the field of management in engineering, particularly in managing the financial aspects and predictive modeling of construction costs for green buildings. Through meticulous data collection, preprocessing, augmentation, and the development of estimation models using ANNs and CTGAN, this study addresses critical challenges in cost estimation, a fundamental aspect of project management in civil engineering. The primary contribution of this study lies in its innovative approach to overcoming data scarcity and imbalance in the context of green buildings construction cost estimation. The application of CTGAN for data augmentation, coupled with the use of ANNs, has demonstrated a promising path for achieving more accurate and reliable cost predictions, essential for effective financial management in construction projects. This approach not only enhances the predictive accuracy but also offers a practical solution to the often-limited data set issues in construction project management. From a management perspective, the findings of this study provide actionable insights for civil engineers, project managers, and decision makers in the construction industry. The ability to predict construction costs accurately at early stages empowers owners and managers with critical information for decision-making, budgeting, strategic planning, and resource allocation, thereby improving the overall efficiency and viability of construction projects.

In conclusion, this study significantly contributes to the field of management in engineering by providing a robust methodology for the accurate estimation of construction costs, particularly in the context of sustainable building projects like green building projects. It exemplifies the integration of technical expertise and management acumen, essential for contemporary civil engineering practice. The methods and findings presented here extend its scope by introducing innovative approaches to managing and predicting essential project metrics. The study also underscores the importance of integrating advanced data-driven techniques in contemporary engineering practice. It highlights the evolving role of civil engineers as not only technical experts but also managers who must adeptly handle complex financial and technological aspects of projects.

Data Availability Statement

Due to confidentiality agreements, supporting data can only be made available to bona fide researchers subject to a nondisclosure agreement. Models or code that support the findings of this study

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

are available from the corresponding author upon reasonable request.

Acknowledgments

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure, and Transport (Grant No. RS-2020-KA158109).

References

Adeli, H., and M. Wu. 1998. "Regularization neural network for construction cost estimation." *J. Constr. Eng. Manage.* 124 (1): 18–24. https://doi.org/10.1061/(ASCE)0733-9364(1998)124:1(18).

Ahn, Y. S., K. R. Song, and J. M. Heo. 2003. "Improving the accuracy of screening of cost estimating in early construction project phase." *J. Archit. Inst. Korea* 19 (11): 133–140.

Alshamrani, O. S. 2017. "Construction cost prediction model for conventional and sustainable college buildings in North America." *J. Taibah Univ. Sci.* 11 (2): 315–323. https://doi.org/10.1016/j.jtusci.2016.01.004.

Bailly, A., C. Blanc, É. Francis, T. Guillotin, F. Jamal, B. Wakim, and P. Roy. 2022. "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models." *Comput. Methods Programs Biomed.* 213 (Jan): 106504. https://doi.org/10.1016/j.cmpb.2021.106504.

Bredehoeft, P. R., L. R. Dysert, J. K. Hollmann, and T. W. Pickett. 2020. *AACE International Recommended Practice No. 18R-97 Cost Estimate Classification System–As Applied in Engineering, Procurement, and Construction for the Process Industries (TCM Framework: 7.3–Cost Estimating and Budgeting)*. Morgantown, WV: AACE.

Brownlee, J. 2017. *How much training data is required for machine learning?* Vermont, VIC, Australia: Machine Learning Mastery.

Brownlee, J. 2019. *A gentle introduction to generative adversarial networks (GANs)*, 17. Vermont, VIC, Australia: Machine Learning Mastery.

BuHamdan, S., A. Alwisy, A. Bouferguene, and M. Al-Hussein. 2019. "Novel approach to overcoming discontinuity in knowledge: Application in value-adding frameworks in construction industry." *J. Constr. Eng. Manage.* 145 (8): 04019045. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001670.

Carr, R. I. 1989. "Cost-estimating principles." *J. Constr. Eng. Manage.* 115 (4): 545–551. https://doi.org/10.1061/(ASCE)0733-9364(1989)115:4(545).

Chandanshive, V., and A. R. Kambekar. 2019. "Estimation of building construction cost using artificial neural networks." *J. Soft Comput. Civ. Eng.* 3 (1): 91–107.

Cheali, P., K. V. Gernaey, and G. Sin. 2015. "Uncertainties in early-stage capital cost estimation of process design—A case study on biorefinery design." *J. Front. Energy Res.* 3 (Feb): 3.

Cheng, M. Y., H. C. Tsai, and E. Sudjono. 2010. "Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry." *Expert Syst. Appl.* 37 (6): 4224–4231. https://doi.org/10.1016/j.eswa.2009.11.080.

Cho, H. C., and J. S. Moon. 2019. "A layered-wise data augmenting algorithm for small sampling data." *J. Internet Comput. Serv.* 20 (6): 65–72.

Cho, J.-H. 2020. "Data augmentations to improve deep learning network model accuracy." Ph.D. dissertation, Dept. of Electronics and Communications Engineering, Kwangwoon Univ.

Cho, N. 2015. "A study on estimating construction cost using machine learning techniques: Focusing on temporary retaining wall method." Master's thesis, Graduate School of Business Administration, Hanyang Cyber Univ.

CII (Construction Industry Institute). 1986. *SD-6: Control of construction project scope*. Austin, TX: CII.

Creswell, A., T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath. 2018. "Generative adversarial networks: An overview." *IEEE Signal Process Mag.* 35 (1): 53–65. https://doi.org/10.1109/MSP.2017.2765202.

Delgado, J. M. D., and L. Oyedele. 2021. "Deep learning with small datasets: Using autoencoders to address limited datasets in construction management." *Appl. Soft Comput.* 112 (Nov): 107836. https://doi.org/10.1016/j.asoc.2021.107836.

de Meer Pardo, F. 2019. "Enriching financial datasets with generative adversarial networks." Ph.D. dissertation, Institute of Applied Mathematics, Delft Univ. of Technology.

Feeley, K. J., and M. R. Silman. 2011. "The data void in modeling current and future distributions of tropical species." *Global Change Biol.* 17 (1): 626–630. https://doi.org/10.1111/j.1365-2486.2010.02239.x.

Feng, G. L., and L. Li. 2013. "Application of genetic algorithm and neural network in construction cost estimate." In Vol. 756 of *Advanced materials research*, 3194–3198. Bäch SZ, Switzerland: Trans Tech Publications.

Fragkakis, N., S. Lambropoulos, and G. Tsiambaos. 2011. "Parametric model for conceptual cost estimation of concrete bridge foundations." *J. Infrastruct. Syst.* 17 (2): 66–74. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000044.

Frid-Adar, M., I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. 2018. "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification." *Neurocomputing* 321 (Dec): 321–331. https://doi.org/10.1016/j.neucom.2018.09.013.

Goodfellow, I. 2016. "Nips 2016 tutorial: Generative adversarial networks." Preprint, submitted December 31, 2016. http://arxiv.org/abs/1701.00160.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative adversarial nets" In *Advances in neural information processing systems*, 27. Montreal: Neural Information Processing Systems 2014.

Gudivada, V., A. Apon, and J. Ding. 2017. "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations." *Int. J. Adv. Software* 10 (1): 1–20.

Günaydın, H. M., and S. Z. Doğan. 2004. "A neural network approach for early cost estimation of structural systems of buildings." *Int. J. Proj. Manage.* 22 (7): 595–602.

Habibi, O., M. Chemmakha, and M. Lazaar. 2023. "Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT botnet attacks detection." *Eng. Appl. Artif. Intell.* 118 (Feb): 105669. https://doi.org/10.1016/j.engappai.2022.105669.

Han, G., S. Liu, K. Chen, N. Yu, Z. Feng, and M. Song. 2021. "Imbalanced sample generation and evaluation for power system transient stability using CTGAN." In *Proc., Int. Conf. on Intelligent Computing & Optimization*, 555–565. Cham, Switzerland: Springer.

Han, H. D., J. H. Kim, J. H. Yoon, and J. W. Seo. 2011. "Road construction cost estimation model in the planning phase using artificial neural network." *KSCE J. Civ. Environ. Eng. Res.* 31 (6D): 829–837.

Holm, L., and J. E. Schaufelberger. 2021. *Construction cost estimating*. Abingdon-on-Thames, UK: Routledge.

Hong, E. 2022. "An ANN-based conceptual estimating of zero-energy building using CTGAN." Master's thesis, Dept. of Architecture and Urban Systems Engineering, Ewha Womans Univ.

Hu, M. 2019. "Does zero energy building cost more? An empirical comparison of the construction costs for zero energy education building in United States." *Sustainable Cities Society* 45 (Feb): 324–334. https://doi.org/10.1016/j.scs.2018.11.026.

Hyari, K. H., A. Al-Daraiseh, and M. El-Mashaleh. 2016. "Conceptual cost estimation model for engineering services in public construction projects." *J. Manage. Eng.* 32 (1): 04015021. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000381.

Jia, J., P. Wu, K. Zhang, and J. Zhong. 2022. Imbalanced disk failure data processing method based on CTGAN." In *Proc., Int. Conf. on Intelligent Computing*, 638–649. Cham, Switzerland: Springer.

Jin, Z.-X., R.-Z. Jin, S.-W. Han, and C. T. Hyun. 2014. "Stochastic cost estimation process using case-based reasoning and Monte Carlo simulation." In Vol. 34 of *Proc., Journal of the Architectural Institute of Korea Autumn Annual Conf.*, 439–440. Seoul: Architectural Institute of Korea.

Jo, Y. J., K. M. Bae, and J. Y. Park. 2020. "Research trends of generative adversarial networks and image generation and translation." *Electron.*

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

Telecommun. Trends 35 (4): 91–102. https://doi.org/10.22648/ETRI.2020.J.350409.

Jordan, M. I., and T. M. Mitchell. 2015. "Machine learning: Trends, perspectives, and prospects." *Science* 349 (6245): 255–260. https://doi.org/10.1126/science.aaa8415.

Jung, S.-S., M.-S. Park, H.-S. Lee, J.-G. Lee, and I.-S. Yoon. 2018. "Forecasting construction cost of public construction project using machine learning." In *Proc., Korean Journal of Construction Engineering and Management Annual Conf.*, 55–56. Seoul: Korea Institute of Construction Engineering and Management.

Juszczyk, M. 2013. "The use of artificial neural networks for residential buildings conceptual cost estimation." In Vol. 1558 of *Proc., AIP Conf.*, 1302–1306. College Park, MD: American Institute of Physics.

Juszczyk, M. 2015. "Application of committees of neural networks for conceptual cost estimation of residential buildings." In Vol. 1648 of *Proc., AIP Conf.*, 600008. College Park, MD: American Institute of Physics.

Kang, S., and K. S. Shin. 2021. "Conditional generative adversarial network based collaborative filtering recommendation system." *J. Intell. Inf. Syste.* 27 (3): 157–173.

KICT (Korea Institute of Civil Engineering and Building Technology), SK Ecoplant, HAEAHN Architecture, and IPODIUM. 2019. *Development of technology and cost optimization simulator to promote the implementation of low-cost zero-energy buildings*. Gyeong-gi, Korea: Korea Agency for Infrastructure Technology Advancement.

Kim, H. J., Y. C. Seo, and C. T. Hyun. 2012. "A hybrid conceptual cost estimating model for large building projects." *Autom. Constr.* 25 (Aug): 72–81. https://doi.org/10.1016/j.autcon.2012.04.006.

Kim, J.-H. 2003a. "An approach to facilitate knowledge streams of occasional individual building industry clients at the pre-project stage." Doctoral dissertation, School of Construction Management and Engineering, Univ. of Reading.

Kim, J.-H. 2004. "Theoretical reviews of client briefing and suggestions for conducting it in Korea." *Korean J. Constr. Eng. Manage.* 5 (3): 79–87.

Kim, S., C. Y. Choi, M. Shahandashti, and K. R. Ryu. 2022. "Improving accuracy in predicting city-level construction cost indices by combining linear ARIMA and nonlinear ANNs." *J. Manage. Eng.* 38 (2): 04021093. https://doi.org/10.1061/(ASCE)ME.1943-5479.0001008.

Kim, S. G. 2003b. "A study on the cost prediction and the analysis of decisive cost factors in multi-family housing." Master's thesis, Graduate School of Engineering, Yonsei Univ.

Kim, S.-H. 2011. "An empirical analysis on the presumption of public apartment's construction cost in housing land development project." *Korean J. Constr. Eng. Manage.* 12 (2): 81–88. https://doi.org/10.6106/KJCEM.2011.12.2.81.

Kim, S. K., J. S. Shin, I. H. Koo, and Y. K. Kim. 2000. "A statistical cost model for road construction project at the planning stage." *J. Civ. Eng.* 20 (2-D): 171–180.

Kim, Y., Y. Kim, I. Lee, and H. J. Lee. 2019. "Increasing accuracy of stock price pattern prediction through data augmentation for deep learning." *J. Bigdata* 4 (2): 1–12.

Korea Energy Agency. 2023. "Zero-energy building certification system." Accessed February 23, 2023. http://zeb.energy.or.kr.

Lee, B. 2020. "Impact of data imbalance on machine learning model performance." In *Proc., Journal of Computing Science and Engineering Conf.*, 697–699. Seoul: Journal of Computing Science and Engineering.

Lee, D. J. 2004. "A method of forecasting EAC (estimate at completion) using probability concept simulation." Master's thesis, Dept. of Architecture, Seoul National Univ.

Lee, G., T. Chang, and S. Chi. 2024. "Data-driven bridge maintenance cost estimation framework for annual expenditure planning." *J. Manage. Eng.* 40 (2): 04023068. https://doi.org/10.1061/JMENEA.MEENG-5706.

Lee, H.-S., H.-K. Lee, M.-S. Park, S.-Y. Kim, and J.-S. Ahn. 2012. "Conceptual cost estimating system development for public apartment projects." *Korean J. Constr. Eng. Manage.* 13 (4): 152–163. https://doi.org/10.6106/KJCEM.2012.13.4.152.

Lowe, D. J., M. W. Emsley, and A. Harding. 2006. "Predicting construction cost using multiple regression techniques." *J. Constr. Eng. Manage.* 132 (7): 750–758. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(750).

Matel, E., F. Vahdatikhaki, S. Hosseinyalamdary, T. Evers, and H. Voordijk. 2022. "An artificial neural network approach for cost estimation of engineering services." *Int. J. Constr. Manage.* 22 (7): 1274–1287. https://doi.org/10.1080/15623599.2019.1692400.

Mikołajczyk, A., and M. Grochowski. 2018. "Data augmentation for improving deep learning in image classification problem." In *Proc., 2018 Int. Interdisciplinary PhD Workshop (IIPhDW)*, 117–122. New York: IEEE.

Mir, M., H. D. Kabir, F. Nasirzadeh, and A. Khosravi. 2021. "Neural network-based interval forecasting of construction material prices." *J. Build. Eng.* 39 (Jul): 102288. https://doi.org/10.1016/j.jobe.2021.102288.

Mirza, M., and S. Osindero. 2014. "Conditional generative adversarial nets." Preprint, submitted November 6, 2014. http://arxiv.org/abs/1411.1784.

Mitsa, T. 2019. *How do you know you have enough training data*. Toronto: Towards Data Science.

Monghasemi, S., and M. Abdallah. 2021. "Linear optimization model to minimize total cost of repetitive construction projects and identify order of units." *J. Manage. Eng.* 37 (4): 04021036. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000936.

Pan, Y., and L. Zhang. 2021. "Roles of artificial intelligence in construction engineering and management: A critical review and future trends." *Autom. Constr.* 122 (Feb): 103517.

Park, S., C. Jeong, S. Seo, and J. Kim. 2018. "A study on structured data augmentation using generative adversarial nets." In *Proc., Journal of Computing Science and Engineering Conf.*, 947–949. Seoul: Journal of Computing Science and Engineering.

Rather, A. M., A. Agarwal, and V. N. Sastry. 2015. "Recurrent neural network and a hybrid model for prediction of stock returns." *Expert Syst. Appl.* 42 (6): 3234–3241.

Segerstedt, A., and T. Olofsson. 2010. "Supply chains in the construction industry." *Supply Chain Manage. Int. J.* 15 (5): 347–353. https://doi.org/10.1108/13598541011068260.

Sheng, V. S., F. Provost, and P. G. Ipeirotis. 2008. "Get another label? Improving data quality and data mining using multiple, noisy labelers." In *Proc., 14th ACM SIGKDD Int. Conf. on Knowledge disc*. New York: ACM Digital Library.

Shim, H. S., and S. Lee. 2021. "A study on the increase in construction cost for zero energy building." *J. Korea Acad.-Ind. Cooperation Society* 22 (1): 603–613.

Shorten, C., and T. M. Khoshgoftaar. 2019. "A survey on image data augmentation for deep learning." *J. Big Data* 6 (1): 1–48. https://doi.org/10.1186/s40537-019-0197-0.

Snoke, J., G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic. 2018. "General and specific utility measures for synthetic data." *J. R. Stat. Soc. Ser. A: Stat. Society* 181 (3): 663–688. https://doi.org/10.1111/rssa.12358.

Song, Z., H. Shi, X. Bai, and G. Li. 2023. "Digital twin-assisted fault diagnosis system for robot joints with insufficient data." *J. Field Rob.* 40 (2): 258–271. https://doi.org/10.1002/rob.22127.

Sonmez, R. 2008. "Parametric range estimating of building costs using regression models and bootstrap." *J. Constr. Eng. Manage.* 134 (12): 1011–1016. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:12(1011).

Takefuji, Y. 2021. "Converting detailed estimates to primary estimates with data augmentation." *Adv. Eng. Inf.* 49 (Feb): 101354.

Tanaka, F. H., and C. Aranha. 2019. "Data augmentation using GANs." Preprint, submitted April 19, 2019. http://arxiv.org/abs/1904.09135.

Vahdani, B., S. M. Mousavi, M. Mousakhani, M. Sharifi, and H. Hashemi. 2012. *A neural network model based on support vector machine for conceptual cost estimation in construction projects*. Qazvin, Iran: Journal of Optimization in Industrial Engineering.

Verzuh, E., and American Psychological Association. 2021. *A guide to the project management body of knowledge: PMBOK Guide*. Newtown Square, PA: Project Management Institute.

Wan, Y., Y. Zhai, X. Wang, and C. Cui. 2022. "Evaluation of indoor energy-saving optimization design of green buildings based on the intelligent GANN-BIM model." *Math. Probl. Eng.* https://doi.org/10.1155/2022/3130512.

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.

Wilmot, C. G., and B. Mei. 2005. "Neural network modeling of highway construction costs." *J. Constr. Eng. Manage.* 131 (7): 765–771. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:7(765).

Xiao, X., S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li. 2015. "A learning-based approach to direction of arrival estimation in noisy and reverberant environments." In *Proc., 2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2814–2818. New York: IEEE.

Xu, L., M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. 2019. "Modeling tabular data using conditional gan." Preprint, submitted July 1, 2019. http://arxiv.org/abs/1907.00503v2.

Yang, H., H. Tian, Y. Zhang, P. Hao, B. Wang, and Q. Gao. 2023. "Novel bootstrap-based ellipsoidal convex model for non-probabilistic reliability-based design optimization with insufficient input data." *Comput. Methods Appl. Mech. Eng.* 415 (Oct): 116231. https://doi.org/10.1016/j.cma.2023.116231.

Zedan, S., and W. Miller. 2018. "Quantifying stakeholders' influence on energy efficiency of housing: Development and application of a four-step methodology." *Constr. Manage. Econ.* 36 (7): 375–393. https://doi.org/10.1080/01446193.2017.1411599.

Zhang, G., B. E. Patuwo, and M. Y. Hu. 1998. "Forecasting with artificial neural networks: The state of the art." *Int. J. Forecasting* 14 (1): 35–62.

This work is made available under the terms of the Creative Commons Attribution 4.0 International license.