Ensembled Machine Learning-Based Cost Prediction Model for Modular and Offsite Construction Projects in Canada

Qudrati Al Wasiew[1], Amirhossein Mehdipoor[2], Aryan Hojjati[2] and SangHyeok Han[1,3]

[1]Department of Building, Civil and Environmental Engineering, Concordia University, Montréal, QC, Canada
[2]Construction Research Centre, National Research Council Canada, Ottawa, ON, Canada
[3]Centre for Innovation in Construction and Infrastructure Engineering and Management (CICIEM), Gina Cody School of Engineering and Computer Science, Concordia University, Montréal, QC H3G 1M8, Canada
qudratial.wasiew@mail.concordia.ca, amirhossein.mehdipoor@nrc-cnrc.gc.ca, aryan.hojjati@nrc-cnrc.gc.ca, sanghyeok.han@concordia.ca

Abstract –

Recent studies have highlighted the necessity for a conceptual cost prediction model for Modular and Offsite Construction (MOC) projects. Barriers such as high costs and the absence of scientific methods for justifying cost comparisons between MOC and conventional construction methods impede the adoption of MOC. This study presents an ensemble machine learning (ML) technique that assesses the performance of three distinct algorithms: linear regression, support vector machine regression, and random forest. It illustrates the effectiveness of this approach for preliminary cost predictions in the conceptual stage. The analysis of predicted versus actual costs demonstrated strong correlation, with R^2 approximately 0.78, root mean square error (RMSE) around 13340 (0.053 when normalized), and mean absolute error (MAE) approximately 10331 (0.041 when normalized). Additionally, the implementation of the synthetic minority oversampling technique (SMOTE) eliminated data-related stochasticity, while the timeline-based data collection addressed time-related stochasticity. This ensemble machine learning technique serves as a foundational database for subsequent learning-based models in MOC projects.

Keywords –

Modular and Offsite Construction; Ensembled Machine Learning; Conceptual Cost Prediction

1 Introduction

Modular and Offsite Construction (MOC) is acknowledged as a superior approach for improving sustainability and productivity compared to conventional on-site construction. MOC is a procedure wherein diverse building elements, including walls, floors, roofs, and staircases, are manufactured in a controlled environment and subsequently transported to the construction site for assembly. In this study, MOC or modular refers to factory-produced housing units manufactured in controlled environments and transported to site for final assembly, typically consisting volumetric modules and panelized structural components. However, the execution of MOC has been obstructed by uncertainty concerning the costs and advantages of employing these novel approaches in the construction industry. MOC approaches, including prefabrication, panelization, and modularization, utilize lean manufacturing principles to tackle conventional in-situ building issues, such as productivity, logistics, safety, emissions, waste, quality, and the trade-offs between environmental limitations and prices [1]. Notwithstanding these advantages, there have been increasing obstacles to the adoption of MOC as an alternative in nations such as Canada, Hong Kong, Singapore, the United Kingdom, and the United States. The recent threats to Canada's construction industry resulting from post-pandemic inflation highlight the necessity of adopting MOC as a potential solution to mitigate the negative impacts of inflation in this vital sector. However, progress is impeded by the absence of a cost database that can effectively illustrate the feasibility of MOC compared to conventional methods [2]. The primary obstacle to MOC in Canada is its inadequate comprehension. In the United Kingdom, prefabricated houses have not achieved commercial success due to obstacles such as elevated costs and restricted opportunities for clients and designers to customize [3]. Researchers have observed that ambiguity surrounding initial expenses, capital expenditures, payback timeframes, and supply chain challenges has impeded the implementation of MOC in China [4]. The principal challenges related to insufficient information for adopting MOC include: (1) the conceptual ambiguity surrounding MOC costs, (2) the perception that offsite solutions incur higher costs than traditional alternatives, (3) the absence of cost data and information regarding MOC, and (4) the uncertainty regarding methods to reduce construction costs while enhancing efficiency.

To tackle with the cost related problems, researchers have followed traditional approaches, statistical methods, machine learning (ML) techniques and hybrid approaches. Traditional cost estimating methods employ detailed quantity take-offs, whereas comparative cost estimates depend on parameters including type, size, capacity, time, material, and other related factors. Traditional cost estimating relies on blueprints and specification breakdowns, which are time-consuming. Additionally, comparative cost estimates, such as case-based reasoning (CBR), assume a linear relationship between the final cost and fundamental design variables of the project. The assumption of a linear relationship is problematic and frequently leads to inaccurate cost estimates [5, 6]. Recent studies have increasingly focused on machine learning and deep learning algorithms with multiple layers, which are effective for comprehending the non-linear patterns inherent in the data [7, 8]. Stochastic parameters include temporal factors such as inflation rates and discount rates, as well as elements like missing data and outliers, which contribute to uncertainty in learning models. Researchers have employed methods such as Monte Carlo Simulation and Fuzzy Logic Theory to address these types of problems by completing the missing data [9, 10].

To address all these problems together, this study employs an ensemble machine learning-based cost prediction methodology, utilizing data from the US Census Bureau survey and RSMeans Square Foot Costs Book, to fulfill the following objectives: (1) to create a conceptual stage MOC project cost prediction model, (2) to establish a model for cost comparison between MOC and traditional alternatives, (3) to compile an initial cost database for MOC projects in Canada, and (4) to analyze the cost savings realized through the application of lean manufacturing techniques in previously completed projects across the U.S. The cross-country data were used because no complete dataset from the Canadian Modular Industry was available, and both the US Census Bureau and RSMeans have been widely adopted as international reference benchmarks in construction cost research. RSMeans provides normalized cost indexes that are transferable across regions and have been directly adopted in this study using two cross-country datasets. RSMeans Square Foot Cost data for Canada are already internally normalized using Canadian labor rates, material pricing, and provincial location factors. The objective of this study is not to establish absolute cost parity between the U.S. and Canada, but to examine whether the learned cost relationships from modular construction data remain valid when applied to region-specific Canadian baseline costs. Section 2 delineates the study approach employed to attain these aims. Section 3 delineates the findings and discourse of this investigation.

2 Methodology

As illustrated in Figure 1, the proposed methodology consists of following five steps: (1) data collection for modular buildings in U.S., data collection for conventional buildings in U.S., and data collection for conventional buildings in Canada; (2) data pre-processing stage combining preparation of data, exploratory data analysis, principal component analysis and mandatory pre-processors for a machine learning model such as one hot encoding and normalization; (3) developing the machine learning model using an ensembled approach with three algorithms; (4) data augmentation using the synthetic minority oversampling technique (SMOTE) and (5) model evaluation and validation. The following sections describe each step of the overview of the research framework.

2.1 Data Collection

The data collection step was mainly conducted through two major data sources which are the manufactured houses survey from US Census Bureau [11] and RSMeans Square Foot Costs Book [12]. 528 data points were collected from the survey results of U.S. Census Bureau ranging from houses 874 square feet gross area to 1879 square feet gross area. The survey was conducted across 44 states of U.S. from 2018 to 2023. Majority of the buildings were constructed using vinyl siding wood frame with concrete blocks and pressure treated wood as the foundation. Average buildings had 2 to 3 bedrooms as per the published information available. The price collected was the average sales price up to transportation to the client site excluding taxes and land acquisition fees. Depending on the available data from the survey results, required data were collected from the RSMeans Square Foot Costs Book for conventional houses in U.S. and conventional houses in Canada. 4483 data points were collected for conventional houses in U.S. ranging from 874 square feet gross area to 1879 square feet gross area to match the available manufactured housing data pattern, from 2018 to 2023. The data was collected for three different material types; wood siding wood frame, brick veneer wood frame, and stucco on wood frame. The data included three different number of floor representatives as in 1 floor, 1.5 floor and 2 floor as per RSMeans data set. The type selected was the single-family residential houses of average grade with continuous RCC footing and average mechanical and

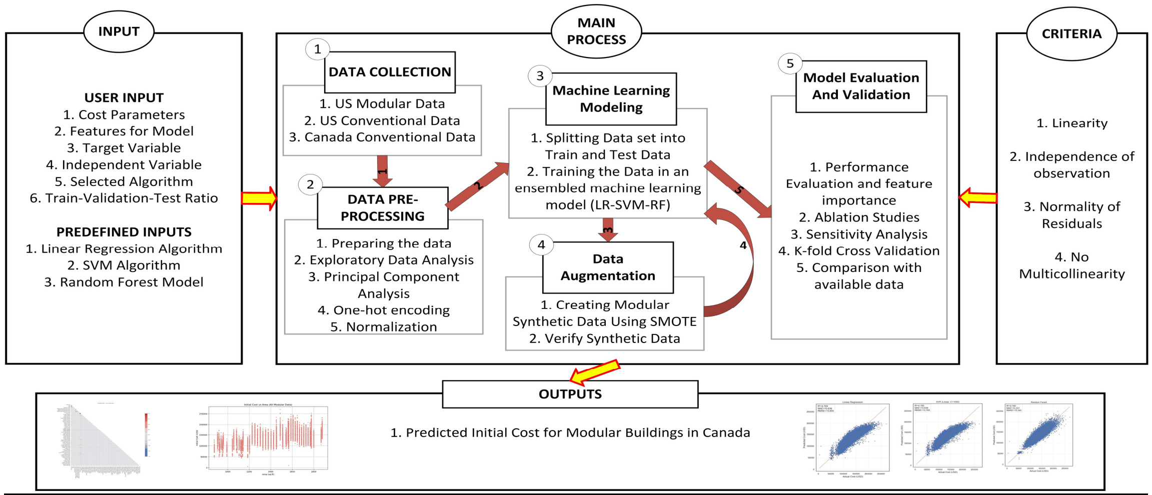

Figure 1: Overview of the research framework.

electrical equipment. The price collected was a sales price with general contractor overhead and profit without tax, which resembles a similar pattern from the conveyed survey results from U.S. Census Bureau. Similarly, conventional house price was collected for the 10 Provinces in Canada as well with a similar pattern of parameters such as area, floors, material and average building details. The total dataset was divided into 80:20 training and testing set ratio. The training dataset was used for model learning and hyperparameter tuning, while remaining 20% was used for independent model evaluation. Table 1 shows the statistical distribution of the collected data across different parameters.

Table 1: Summary of Dataset Distribution and Characteristics

Description	Details
Total Data Points	8738
Location	44 States/10 Provinces
Country	U.S. and Canada
Year	2018-2023
Material	4 Types
Number of Floors	1, 1.5, 2.0
Average Building	Single Family Single Section
Dataset Distribution	
Category	Number of Samples
Modular (U.S.)	528
Conventional (U.S.)	4483
Conventional (Canada)	3727
Total	8738
Model Training Configuration	
Parameter	Value
Train-Test Split	80:20

The Canadian dataset for conventional buildings collected using RSMeans was not used to train the final machine learning model, rather it was used up to data augmentation as a baseline cost for conventional buildings with Canadian parameter. This helped the dataset to learn cost pattern differences among U.S. and Canada regional barrier and the U.S. modular and conventional data helped to learn cost patterns between modular and conventional buildings.

2.2 Data Pre-processing

The second step included preparation of data according to the requirement for the model, exploratory data analysis to understand the existing relationship and patterns, principal component analysis to find out influential parameters connected to the target variable and finally one hot encoding along with normalization before feeding the data into the machine learning model.

2.3 Machine Learning Modeling

A three-algorithm ensemble machine learning model was selected to execute the core methodology. The three algorithms utilized were linear regression, support vector machine, and random forest. The objective of choosing these three algorithms was to furnish a list with diverse capabilities for capturing both linear and nonlinear patterns in the dataset. Linear regression forecasts cost by representing the linear increase or decrease of the target variable as a function of the independent variables. A support vector machine can be optimized by the selection of the kernel and hyperparameters to accommodate variations in the dataset's pattern. Furthermore, random forest is a tree-based algorithm that has recently demonstrated efficacy in identifying high-dimensional nonlinear patterns inside datasets [13]. Figure 2 illustrates the architecture of the generated machine learning model.

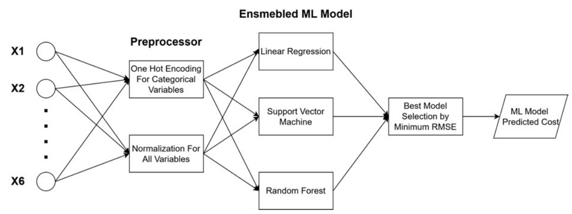

Figure 2: ML model architecture.

2.4 Data Augmentation

The data augmentation process was executed to address some gaps in the U.S. Census Bureau survey data. Data for gross areas ranging from 1202 to 1452 square feet and from 1576 to 1768 square feet were absent. The initial model exhibited subpar performance due to an inadequate variation in the distribution of the chosen independent variable. As some ranges of modular housing sizes were missing in the U.S. Census Bureau data, to ensure a balanced dataset and avoid bias toward certain building sizes, the Synthetic Minority Oversampling Technique (SMOTE) was used to generate additional samples for the underrepresented ranges. At first more data were collected from RSMeans for the underrepresented ranges for conventional U.S. and conventional Canada buildings. Finally, data augmentation was applied to generate synthetic data for the missing ranges in the U.S. modular data as well as some data for Canada modular so that the model is not overfitting to any independent variables. Various data augmentation approaches were examined, including the SMOTE, random oversampling technique, adaptive synthetic oversampling technique, TOMEK under-sampling technique, and k-means and k-medians imputation. Given the availability of real-world datasets pertaining to conventional buildings in RSMeans as a primary data source, the SMOTE was selected for the data augmentation phase. This method generates synthetic data by establishing neighbors within the existing minority data, while a custom relationship logic can elucidate the observed relationships between the majority and minority data throughout the entire dataset. The generated synthetic data was assessed by distribution skewness, correlation preservation tests (such as Pearson and Spearman), and the Kolmogorov-Smirnov (KS) test of the cumulative distribution function [14]. Table 2 shows the data type ratio before and after augmentation process.

Table 2: Data type ratio for collected and augmented dataset

Dataset	Data Type	Ratio
Collected Data (8738)	Real Modular: Real Conventional	6:94
First Augmented Data (36042)	Real Modular: Synthetic Modular: Real Conventional Including Missing Area Ranges	1.5:48.5:50
Final Model Dataset (18153)	Real Modular: Synthetic Modular	2.9:97.1

2.5 Model Evaluation and Validation

The model evaluation and validation phase commenced with an assessment of the created models' performance regarding correctness and errors, with the resultant feature importance in relation to the original principal component analysis. Upon attaining the requisite performance level, multiple ablation tests were undertaken, including the elimination of existing outliers in the dataset, to ascertain whether performance could be enhanced further or if the outliers were merely noise. The building type was concurrently utilized as a variable across all accessible data, including conventional data, to evaluate potential enhancements in model performance and to ascertain the presence of overfitting. A sensitivity analysis was conducted to assess variations in the target variable and to identify any excessively sensitive elements within the model. K-fold cross-validation was ultimately conducted to confirm that the final model was neither overfitting nor underfitting. To make sure the synthetic data behaved like real data, statistical tests such as Pearson, Spearman, and Kolmogorov-Smirnov were used to compare distributions. The results showed no significant difference between real and synthetic samples, meaning the added data followed the same patterns as the original dataset.

3 Results and Conclusions

To assess the proposed machine learning-based cost prediction model, various evaluation methodologies were employed at this stage such as the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) to yield a thorough comprehension of the model and identify the optimal algorithm as shown in Figure 3. Subsequently, feature importance, ablation experiments, sensitivity analysis, and k-fold cross-validation were conducted.

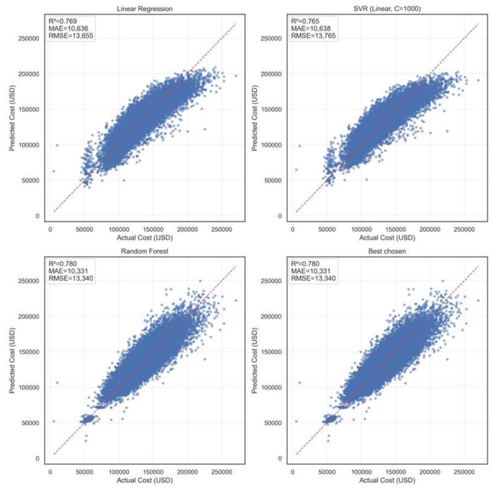

Figure 3: Performance of the machine learning model.

3.1 Performance Evaluation and Feature Importance

As per the ensembled model architecture, random forest is chosen as the optimized algorithm for the dataset. Random forest feature importance was conducted to analyse the influential features that influence the final model. Table 3 shows the performance comparison against existing conceptual cost studies in conventional construction as there is no comparable study in modular construction till date.

Table 3: Performance Comparison

Article	Algorithm	R^2	Dataset
[15]	LR and ANN	0.69 – 0.98	Historical Residential Buildings from Germany
[16]	LR	0.95	Historical Care Retirement Facility
[17]	ANN	0.89	Water Rehabilitation Projects
[8]	GB	0.99	RSMeans Structural Assemblies
Proposed Study	RF	0.78	Combined US Census Bureau and RSMeans

The ensemble architecture was initially developed to evaluate three candidate algorithms' (LR, SVM and RF) performance on the final dataset. As shown in Figure 3, RF was the best performer as per developed ensembled architecture, so final feature importance was conducted for RF. Depending on internal pattern of the dataset, and hyperparameter tuning, the best performer can change in case of any other dataset or this dataset when additional data from Canada modular industry is available. Testing the model's performance on any other dataset is considered outside the scope of this study. Figure 4 shows the output for the random forest feature importance test. Area and year appear as dominant predictors because construction cost estimation is inherently dependent on building size and temporal cost escalation. However, other variables such as location, material type, and number of floors also contributed to model predictions. The scope of this study is within the prediction for a conceptual stage cost for modular buildings, and with the availability of more granular data regarding phases and components future work can investigate more into the dominant predictors such as specific modular building parameters like module sizes, panel sizes and joints.

3.2 Ablation Studies

Two varieties of ablation investigations were performed subsequent to the selection of the final machine learning model. The final model dataset had 18,153 data points, encompassing both actual and synthetic modular building data. The dataset had outliers across various model features, including area, location, material, and year. Subsequent to the final model selection, these outliers were eliminated from the final dataset to ascertain if the model's performance would improve. The outliers were identified using interquartile range (IQR) method, where observations outside 1.5xIQR from the first and third quartiles were considered potential outliers. The R^2 value decreased from 0.78 to 0.75, suggesting that the remaining outliers in the sample represented significant distributional variance rather than mere noise, and their removal led to underfitting. A dataset with typical RSMeans data included 36,042 data points. The inclusion of “building type” as an additional training feature was examined, yielding an R^2 value of 0.95 and the maximum importance for this feature suggesting overfitting, which resulted in the model's exclusion. The inclusion of “building type” resulted in a significant increase of R^2 value because the model captured dataset-specific correlations rather than generalizable patterns. This is expected since RSMeans data encodes strong cost differences directly through building type (e.g., economy, average etc.). A comparison between initial principal component analysis and this model’s random forest feature importance revealed that “building type” did not consistently appear as dominant contributor on principal component analysis explaining overall data variance, however in this model it emerged as an overly dominant predictor, suggesting potential bias. Therefore, instead of applying explicit regularization, a feature exclusion strategy was adopted to improve model generalizability.

3.3 Sensitivity Analysis

A sensitivity study was conducted to evaluate the model's performance across all chosen features. The primary objective of this test was to identify any oversensitive characteristics within the designed model and to ascertain whether the model remained susceptible to uncertain parameters, such as absent outliers, incomplete data, temporal variables like inflation, and other deficiencies.

Figure 5 illustrates the outcomes of the sensitivity analysis. The final model demonstrated a balanced sensitivity across all selected features. The previously observed unclear behavior, stemming from ambiguous parameters, was largely mitigated in the final model.

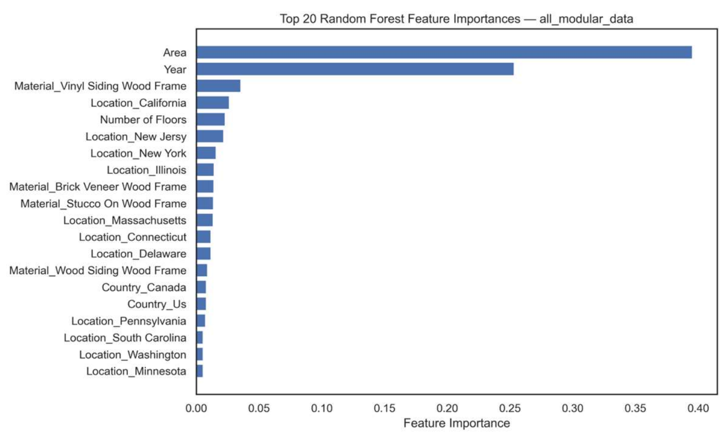

Figure 4: Random Forest feature importance of the final tested model.

3.4 K-fold Cross Validation and Comparison with Available Data

The predictions from the machine learning model were ultimately confirmed by 5-fold cross-validation and by comparing them with the cost data available from the U.S. Census Bureau survey results and the RSMeans database. Figure 6 illustrates that there were no significant alterations in the R^2 value during the 5-fold cross-validation, demonstrating that the model was not overfitted to any subset of the dataset and performs consistently on unseen data. The model achieved a mean R^2 value of 0.779 with a standard deviation of 0.008 and standard error of 0.0036. The 95% confidence interval of the mean R^2 was estimated as [0.77, 0.79], indicating low variability and stable performance across different data partitions.

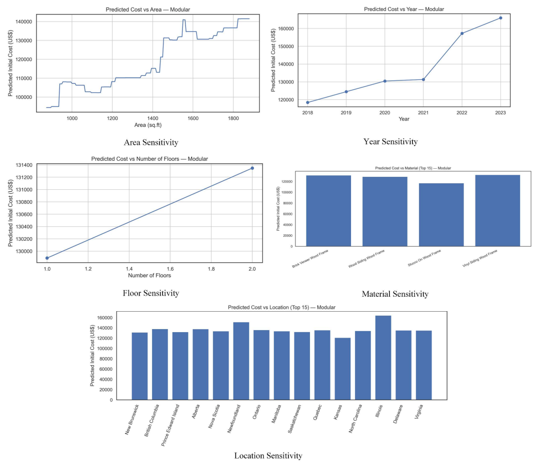

Figure 5: Sensitivity analysis results from the final model.

Subsequently, the projected modular construction cost data was compared with the actual data. The comparison involved analyzing the cost data disparity between modular and conventional buildings from the U.S. Census Bureau survey and contrasting it with the conventional costs from RSMeans in Canada and the projected modular costs derived from the model.

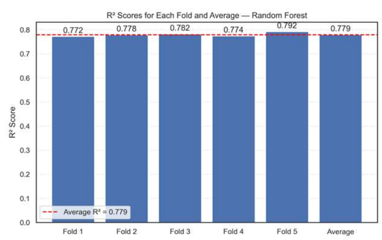

Figure 6: 5-fold cross validation results.

The comparison of analogous features among the selected independent variables in the U.S. and Canada revealed a mere 0.9% discrepancy between the two. This further demonstrates that the model has discerned the relational patterns present in the empirical data. The objective of the comparison in Table 4 is not statistical inference but illustrative comparison of cost patterns between modular and conventional construction. A detailed statistical hypothesis testing framework is considered outside the scope of this conceptual stage study and will be explored in future research. The purpose of this cross-country comparison is also not to establish exact cost parity but to evaluate whether learned cost relationships remain consistent across datasets. Full economic normalization between U.S. and Canadian markets is considered outside the scope of this conceptual model.

Table 4: Comparison with real world data for similar features

Building Type and Country	Area (Sq. ft.)	Cost (US $)	Average % Difference
Conventional, U.S.	1777	239138	-
Modular, U.S.	1777	141200	40%
Conventional, Canada	1777	299350	-
Predicted Modular, Canada	1777	176892	40.9%

3.5 Conclusion

This paper presents an ensemble machine learning strategy to tackle current cost-related issues in conceptual cost prediction for MOC. The proposed framework establishes a thorough methodology to create a verified learning-based cost prediction tool for the MOC sector in Canada.

This study's contributions are as follows. A complete strategy for creating a conceptual-stage cost estimation tool for MOC projects is provided. This framework tackles the dual aims of forecasting modular construction expenses and establishing a preliminary database for MOC projects, which may be utilized for subsequent cost prediction models. This study determines the most significant cost parameters for a conceptual cost prediction model by the integration of comprehensive feature selection and feature importance analysis. This model does not differentiate between specific cost components rather, it illustrates the cost reductions of a MOC project relative to a conventional project. Nevertheless, it can acquire and convey information and expertise from one regional research to another by employing conceptual cost factors. This may prove efficient for future research in sectors lacking data, but notable advancements have occurred in other regional areas. Knowledge and information produced in one region can be effectively assimilated and disseminated using a comparable machine learning model to enhance the designated industry.

Moreover, the established machine learning model can be calibrated to function as a cost-comparison tool between MOC and traditional projects. The ambiguity regarding the advantages of MOC projects can be alleviated. Despite the initial overhead expenses associated with establishing factories and machines potentially surpassing those of traditional projects, the elimination of waste, enhancement of quality, and acceleration of timelines can markedly decrease costs, hence boosting sustainable construction practices.

Applying cross-country normalization is considered outside the scope of this conceptual-stage model. Future work will investigate explicit cross-country normalization once sufficient real modular cost data become available in Canada. Moreover, although synthetic data were used to fill missing ranges, the model was carefully validated to ensure these data did not distort results. Future work will include more real modular housing data from Canadian projects to further verify model robustness. As the objective of this study is conceptual-stage cost prediction rather than detailed cost breakdown modeling, component-level decomposition such as production, transportation and overhead cost will be addressed in future work when more granular modular factory data become available.

This research facilitates the creation of a viable machine learning-based cost predictor by the implementation of methodologies including timeline data collection, data augmentation, and synthetic data synthesis, which mitigate uncertainties stemming from inflation rates, absent data, and outliers. While the model is proficient in making certain predictions, enhancements can be implemented to evaluate its capacity to address uncertainties. The model could be expanded to function as a more detailed cost breakdown structure for MOC projects when analogous data is accessible from factories. Ultimately, the model may be employed to formulate a multi-objective optimization problem whereby cost serves as either a constraint or the objective function of the favored methodology.

Acknowledgement

The authors gratefully acknowledge the National Research Council Canada (NRC) for their collaboration and funding for this project (Grant ID: CSDP-004-01).

References

[1] C. Mao, et al. Cost analysis for sustainable off-site construction based on a multiple-case study in China. *Habitat International*, 57:215-222, 2016.

[2] A. Mehdipoor, I. Iordanova and M. Al-Hussein. Off-Site Construction Resilience: A Strategic Response to Inflation Challenges in Construction in Postpandemic Canada. *Journal of architectural engineering*, 31(1):04024045, 2025.

[3] E. Vakaj, et al. An ontology-based cost estimation for offsite construction. *Journal of information technology in construction (ITcon)*, 28:220-245, 2023.

[4] C. Mao, et al. Major Barriers to Off-Site Construction: The Developer's Perspective in China. *Journal of Management in Engineering*, 31(3):04014043, 2015.

[5] S. Z. Doğan, D. Arditi and H. M. Günaydın. Determining attribute weights in a CBR model for early cost prediction of structural systems. *Journal of Construction Engineering and Management*, 132(10):1092-1098, 2006.

[6] R. Sonmez. Range estimation of construction costs using neural networks with bootstrap prediction intervals. *Expert Systems with Applications*, 38(8):9913-9917, 2011.

[7] M.-Y. Cheng, H.-C. Tsai and E. Sudjono. Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. *Expert Systems with Applications*, 37(6):4224-4231, 2010.

[8] D. Chakraborty, et al. A novel construction cost prediction model using hybrid natural and light gradient boosting. *Advanced Engineering Informatics*, 46:101201, 2020.

[9] M. Ammar, T. Zayed and O. Moselhi. Fuzzy-based life-cycle cost model for decision making under subjectivity. *Journal of Construction Engineering and Management*, 139(5):556-563, 2013.

[10] O. S. Alshamrani and A. Alshibani. Automated decision support system for selecting the envelope and structural systems for educational facilities. *Building and Environment*, 181:106993, 2020.

[11] U. S. C. Bureau. Manufactured Housing Survey Annual Data On-line: https://www.census.gov/data/tables/time-series/econ/mhs/annual-data.html, Accessed: 11/07/2025.

[12] GORDIAN. Square Foot Costs with RSMeans Data. On-line: https://www.rsmeans.com/products/books/2025-cost-data-books/2025-square-foot-costs-book?, Accessed: 11/07/2025.

[13] H. A. Salman, A. Kalakech and A. Steiti. Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024:69-79, 2024.

[14] V. Buggineni, C. Chen and J. Camelio. Enhancing manufacturing operations with synthetic data: a systematic framework for data generation, accuracy, and utility. *Frontiers in Manufacturing Technology*, 4:1320166, 2024.

[15] O. Dursun and C. Stoy. Conceptual estimation of construction costs using the multistep ahead approach. *Journal of Construction Engineering and Management*, 142(9):04016038, 2016.

[16] R. Sonmez. Conceptual cost estimation of building projects with regression analysis and neural networks. *Canadian Journal of Civil Engineering*, 31(4):677-683, 2004.

[17] T. Shehab and M. Farooq. Neural network cost estimating model for utility rehabilitation projects. Engineering, *Construction and Architectural Management*, 20(2):118-126, 2013.