ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0319
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

Data mining approach for identifying factors impacting construction accident costs: from indirect expenses perspectives

Ayesha Munira CHOWDHURY[1]*, Eun-Ju HA[2], Jae-ho CHOI[3]

[1] *ICT Integrated Safety Ocean Smart Cities Engineering Department, Dong-A University, Busan, Korea,* E-mail address: ayesha@dau.ac.kr
[2] *Department of Electronics, Dong-A University, Busan, Korea,* E-mail address: 19731jmj@donga.ac.kr
[3] *ICT Integrated Safety Ocean Smart Cities Engineering Department, Dong-A University, Busan, Korea,* E-mail address: jaehochoi@dau.ac.kr

Abstract: Construction projects account for a significant proportion of workplace hazards globally. While construction cost reports typically emphasize direct accident costs such as treatment expenses, nursing care costs, or disability benefits, indirect factors like work interruption loss costs or consolation costs are frequently overlooked, because it is relatively difficult to estimate those factors in advance. Recognizing and accurately estimating the indirect costs factors associated with construction accidents would not only shed light on the monetary impact these incidents have on overall project costs but also would enable to estimate the total accident cost in advance. The current study seeks to identify factors influencing indirect costs, which ultimately govern the total accident cost, through a data mining approach. A survey was conducted in domestic construction companies, resulting in a dataset of 1038 accident records collected from construction sites. First, statistical analysis was performed to uncover characteristics and patterns of factors affecting construction accident costs from both direct and indirect perspectives. Later, this study proposes four distinct machine learning (ML) models, comparing their performances in predicting the total accident cost (including indirect costs) in advance. Additionally, this research sheds light on an important issue in construction data analysis, which is the scarcity of data in a particular class, by applying random oversampling and random undersampling techniques. The suggested framework can assist practitioners and management in estimating construction accident costs and identifying the relevant attributes that impact accidents at the construction site for future practices.

Key words: construction accident costs, direct and indirect accident cost, data mining (DM), machine learning (ML), statistical analysis

1. INTRODUCTION

Occupational health and safety (OHS) is a major global issue. The International Labor Organization reported that one employee dies from an occupational accident or illness every 15 seconds, and approximately 2.34 million employees die annually due to occupation-related illnesses [1]. While occupational accidents are on a downward trend in the other industries, the construction industry is still experiencing a rising trend, amplifying social and economic losses [2].

Accidents at the construction site can be attributed to various factors; however, a lack of safety precautions and reluctance to adopt appropriate safety measures within the industry are predominant issues. According to Kim et al. [3], a study revealed that workers in their 40s and 50s, representing a significant portion of construction workers, exhibit notably higher safety insensitivity compared to those in their 20s and 60s. Any accidents occurring at the workplace accidents impose fiscal burdens not only to the employers but also on national economy. If companies and employers are aware of the costs associated with each accident, it will motivate them to implement a necessary safety management system.

Tang et al. [4] insisted that costs arising from the accidents are responsible for the majority of the fiscal losses construction industry faces every year. However, without properly identifying the factors attributing construction accidents or the costs that are resulting afterwards, appropriate safety management system cannot be created. Construction accident costs can be categorized into two groups: direct costs, which is known as the insurance costs and indirect or secondary costs. Indirect costs usually arise from lost production days, work interruption costs, legal fees, or new labor employment costs, etc. However, since indirect costs are not directly related to accidents, they can be highly variable. Another issue is that, depending on the nature of the cost, it might take years to evaluate the total indirect cost, and therefore it usually gets overlooked in the project cost ledgers.

Without the indirect cost, it is not possible to estimate the true cost of the accident. Therefore, this study applied data mining (DM) techniques to a construction accident cost database, which was collected focusing on indirect cost information. The objectives of this study are twofold. First, exploratory data analyses (EDA) and statistical analyses are conducted to investigate the patterns among construction accident factors, direct costs, and indirect costs. Second, four machine learning (ML) algorithms (decision tree (DT), k-nearest neighbor (K-NN), random forest (RF), and extreme gradient boosting (XGBoost)) are employed to predict the total accident cost class following the guidelines of the Korea Construction Disaster Prevention Institute (KCDR), based on the relevant accident factors. Subsequently, this total accident cost class is used in turn to quantify the indirect cost, utilizing two ML regression models: RF and gradient boosting (GB). This study can help employers estimate the total cost in advance during accidents and thus become an important tool in decision-making in safety management. Additionally, a challenge was encountered in the ML application, namely, class imbalance in the output classes related to the total accident cost. Data scarcity is a prominent issue in the construction industry. Hence, this study can serve as a valuable resource for researchers and practitioners who encounter similar challenges.

2. LITERATURE REVIEW

Many studies on construction accidents primarily focus on analyzing accident reports and statistics for cause determination or evaluating risk magnitude [5]. Accidents at construction sites can lower production rates, profits, and limit future investments for companies. While previous studies acknowledge that the actual accident cost surpasses the direct or insurance cost, the extent of the indirect cost is often overlooked.

Heinrich [6] was the first to shed light on this indirect accident cost issue and suggested that indirect costs can be four times the direct cost. Since then, indirect construction accident costs are typically expressed as a ratio to the direct cost. Due to their variability, different literature portrays the direct-indirect accident cost relation in various ratios, ranging from 1:0.67 to 1:20 [7-11]. However, researchers often mention that providing a universally accepted ratio is not possible [12]. To date, there have been no studies on predicting accident costs, including the hidden indirect cost. Given these issues, this study focuses on the factors affecting indirect costs in construction accidents and the prediction of indirect costs to present more accurate industrial accident prevention activities and management measures.

DM methods have gained popularity for developing efficient predictive models. For instance, Shin [13] employed support vector machines (SVM) to predict safety plan costs in civil and architectural construction projects, utilizing 114 data samples. However, the research scope was limited to only two types of construction projects and did not consider other types, such as industrial or public facilities, etc. Ayhan and Tokdemir [14] used artificial neural network (ANN) models to forecast construction accident outcomes. In a two-step approach, Pham et al. [15] examined the effectiveness of thirteen regression-based machine learning models in optimizing building costs. First, costs and resources were predicted based on specific building features, and then potential building features were determined within a given budget. The authors concluded that ANN, GB, and XGBoost showed the best results. Therefore, DM facilitates the interpretation of a standard database and the generation of new knowledge, making it a key tool adopted in this study to achieve the research objectives.

3. RESEARCH METHODOLOGIES

To prepare the dataset for the DM application, a survey was conducted among domestic construction companies in Korea. Based on the survey, 1038 accident records were collected. The survey questionnaire included multiple criteria, such as information regarding construction projects, details about accidents and the victims, and finally, information about the relevant accident costs. Table 1 shows the relevant information received from the survey.

The research methodologies follow three steps. First, accident cost data were obtained through a questionnaire survey conducted on representative construction companies in Korea. Second, the patterns between the accident factors and the direct and indirect costs are investigated through EDA and statistical analyses. Third, based on the factors found in the statistical analysis, an accident cost prediction model is proposed using four machine learning algorithms: DT, K-NN, RF, and XGBoost. Figure 1 below illustrates the research flow.

Table 1. Description of the data received from the survey

No.	Attributes	Description
1	Death	Number of deaths during accident
2	Injury	Number of injured during accident
3	Construction type	Six types, such as buildings, industrial construction etc.
4	Accident type	Eighteen types (Falling, bumping, Slipping, etc.)
5	Specific work type	Thirty-nine work types such as painting, excavation works etc.
6	Injured area	Nineteen types, such as head, eyes, hands, etc.
7	Specific occupation of the worker	Sixty-two types, such as boring, bricks or masonry works etc.
8	Indirect accident cost	Cost such as consolation cost, work-interruption loss cost, accident investigation cost, labor replacement cost, material loss cost etc.
9	Direct accident cost	Insurance cost such as nursing care benefits, disability benefits, worker's compensation cost, medical treatment and medication expenses etc.
10	Total accident cost	Total cost = (Indirect accident cost + Direct accident cost)

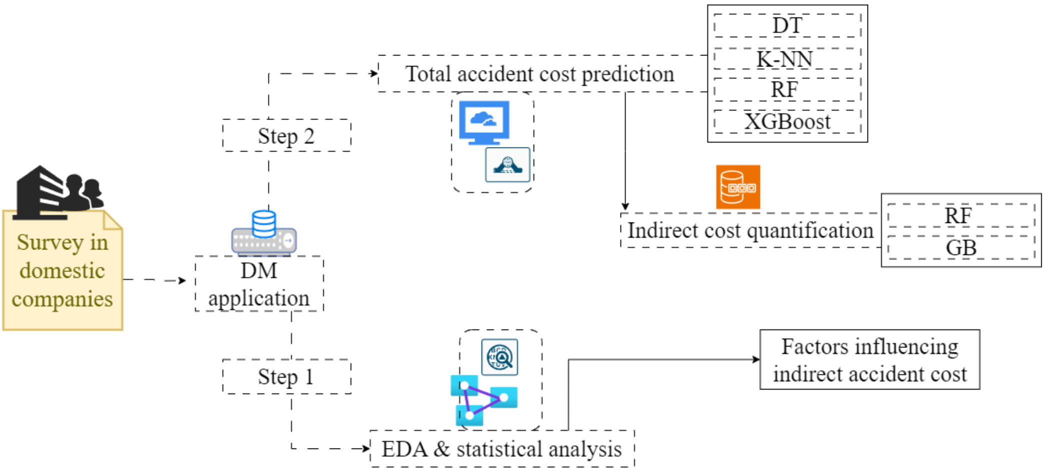

DT
K-NN
RF
XGBoost
Total accident cost prediction
Step 2
Survey in domestic companies
DM application
Indirect cost quantification
RF
GB
Step 1
Factors influencing indirect accident cost
EDA & statistical analysis

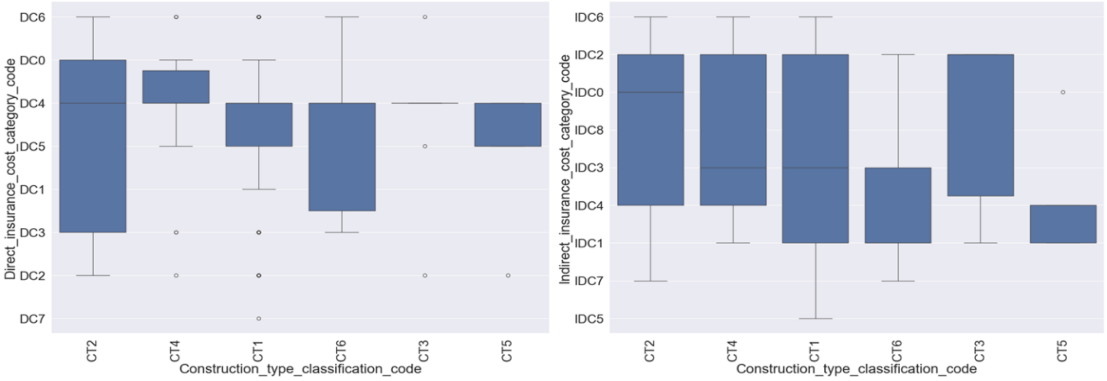
Figure 1: Proposed research methodologies

4. ANALYSIS OF ACCIDENT COST FACTORS

4.1 Exploratory Data Analysis (EDA)

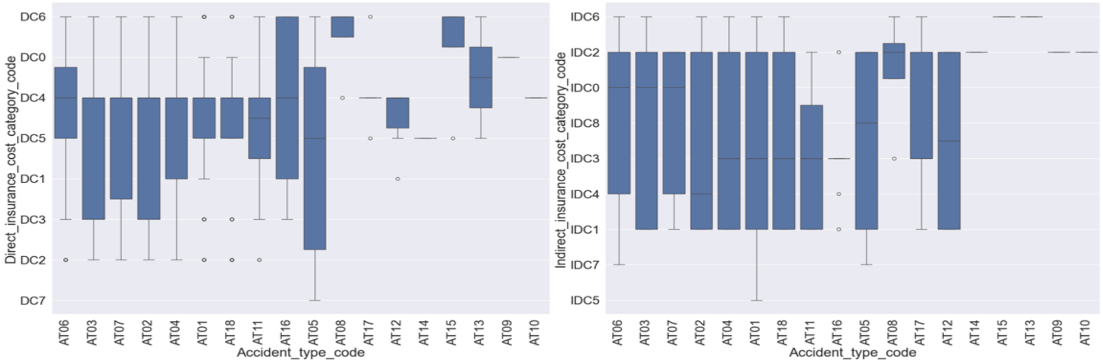
EDA is a DM technique used to investigate the characteristics of a particular dataset. EDA uncovers hidden patterns and visualizes the interrelationships between different variables. In the dataset, there are seven variables: death, injury, construction type, accident type, specific work type, worker's specific occupation, and injured area, in addition to the cost variables. Among the seven variables, death, injury construction type, accident type and injured area types are plotted against the cost categories.

For instance, Figure 2 illustrates the distribution of direct and indirect cost categories across different construction types during accidents. Notably, indirect costs are more prevalent in construction types CT1 (buildings) and CT2 (civil infrastructures), while direct costs are more common in CT2 and CT6 (miscellaneous structures). In Figure 3, the highest direct cost was associated with accident type AT05 (collapse), whereas the highest indirect costs stemmed from accident type AT01 (fall). Regarding injury types, IA01 (head injuries) occasionally resulted in indirect costs reaching the IDC7 (500,000~1mil$) range, whereas IA13 (complex injuries such as brain or nerve damage) led to direct costs in the DC7 (500,000~1mil$) range (Figure 4). Based on the EDA analysis, construction stakeholders may prioritize safety enhancements in CT1 and CT2 constructions, emphasizing the use of personal protective equipment to prevent fall or collapse accidents and mitigate head or complex injuries.



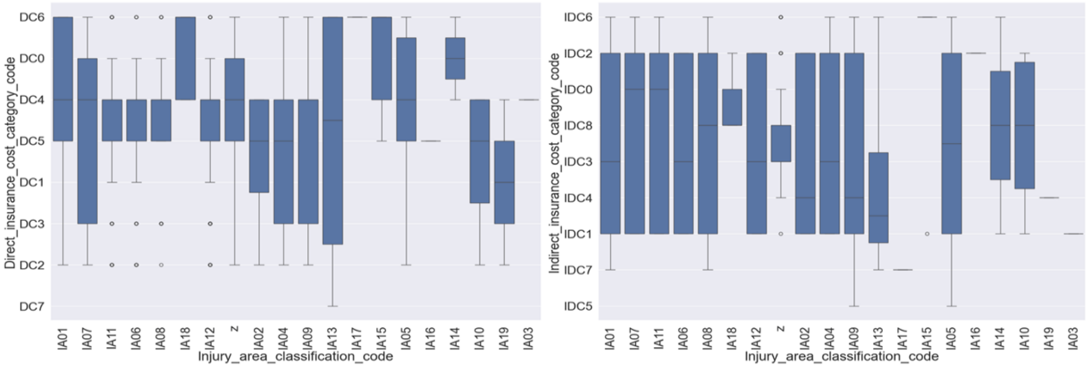
(a) (b)

Figure 2: Direct and indirect cost pattern in specific construction types



(a) (b)

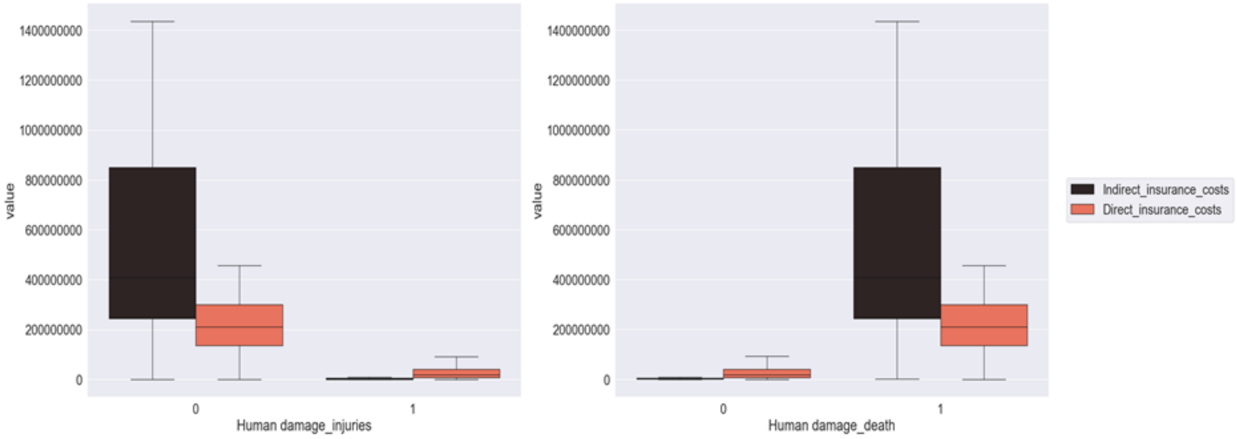
Figure 3: Direct and indirect cost pattern in specific accident types



(a) (b)

Figure 4: Direct and indirect cost pattern in specific injury types

Figures 5(a) and 5(b) below show the indirect and direct costs incurred by accidents at the construction site for both injury and death cases. It can be observed that the indirect cost is higher in death cases, while the direct cost is higher in injury cases. The 0 and 1 in both the figure indicates the number of death or injury cases in the dataset



(a) (b)

Figure 5: Direct and indirect cost occurring from injury vs death cases

4.2 Statistical Analysis

Following the EDA, specific patterns are found between the accident factors and the direct and indirect accident costs. Next step is the accident cost prediction model with ML algorithms. However, it is important to find the relevance of the factors to the target output, i.e., total accident cost. Therefore, statistical analysis is performed to assess the significance of the relevant factors and to remove any outlier variables in the accident cost data This study adopted the Kruskal-Wallis test for categorical variables and the Pearson correlation matrix test for numerical variables. The results of these statistical tests are presented in Table 2.

The statistical analysis indicates that all categorical values are significant, with a p-value less than 0.05. Additionally, numerical values exhibit a strong correlation, with coefficients exceeding 0.5.

Table 2. Siginificance/Correlation of the variables

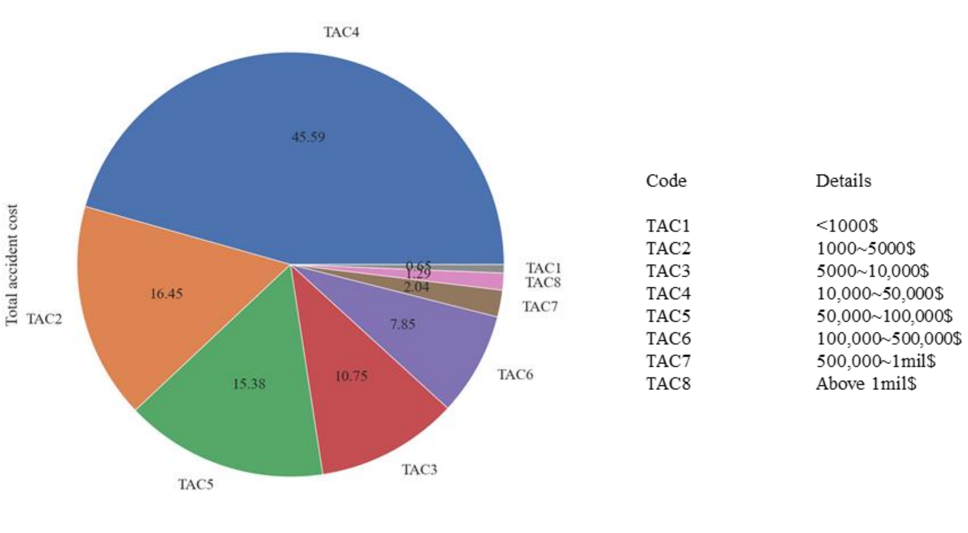
No.	Variable Name	Significance/ Correlation	Remarks
1	Construction type	0.00165	Less than 0.05, therefore siginificant
2	Accident type	1.0083e-14	Less than 0.05, therefore siginificant
3	Specific work type	0.0002589	Less than 0.05, therefore siginificant
4	Injured area	2.352e-22	Less than 0.05, therefore siginificant
5	Specific occupation type	2.530e-15	Less than 0.05, therefore siginificant
6	Direct cost type	0.00081	Less than 0.05, therefore siginificant
7	Death (N)	0.68	Positively related
8	Injury (N)	-0.68	Negatively related

5. ACCIDENT COST PREDICTION MODEL DEVELOPMENT

In this study, four ML classifiers—namely DT, K-NN, RF, and XGBoost, are employed to predict the total accident cost type following an accident event. Subsequently, two ML regressors (RF and GB) are implemented to quantify the indirect cost, utilizing the predicted total accident cost type obtained from the classifier. The input variables include death, injury, construction type, accident type, specific occupation, specific work, injured area, and direct cost type for the total accident cost type. The inclusion of direct cost type is essential as it is known and originates from the insurance cost. Conversely, the regressors use death, injury, direct cost type, and the total accident cost type. The ultimate objective is to predict the indirect cost without the need to wait for extended periods or follow-up years.

5.1 Data Resampling Strategies

It is found that the output variable, or the total accident cost type, exhibits a class imbalance issue, a phenomenon where one of the output classes possesses the majority portion compared to other classes. This hampers the generalization of the developed model and, therefore, needs resolution. To address this class imbalance problem, two different techniques, such as random oversampling (ROS) and random undersampling (RUS), are employed. The former duplicates the minority classes, whereas the latter removes the majority classes. Therefore, this study applied ML classifiers in three different approaches: a regular approach, an ROS approach, and an RUS approach. Figure 6 below illustrates the distribution of output classes, and it is found that TAC4 holds the majority percentage of the output classes.



Code	Details
TAC1	<1000$
TAC2	1000~5000$
TAC3	5000~10,000$
TAC4	10,000~50,000$
TAC5	50,000~100,000$
TAC6	100,000~500,000$
TAC7	500,000~1mil$
TAC8	Above 1mil$

Figure 6: Distribution of the total accident cost

5.2 ML Model Results & Discussion

The datasets were divided into 80%-20% ratios for training and testing subsets for both classifiers and regressors. Table 3 below presents the training results of ML classifiers in three strategies. As it can be seen, data resampling with ROS significantly improves the results for each of the models. Based on the F-1 score, DT performs the best in both regular and RUS training strategies, whereas K-NN performs the best in the ROS training strategy, as well as among all three strategies. Except for K-NN in the regular and RUS strategy, the precision, recall, and F-1 scores for all the models are close to 0.65, showing significant promise in the utilization of ML models in evaluating possible total accident costs.

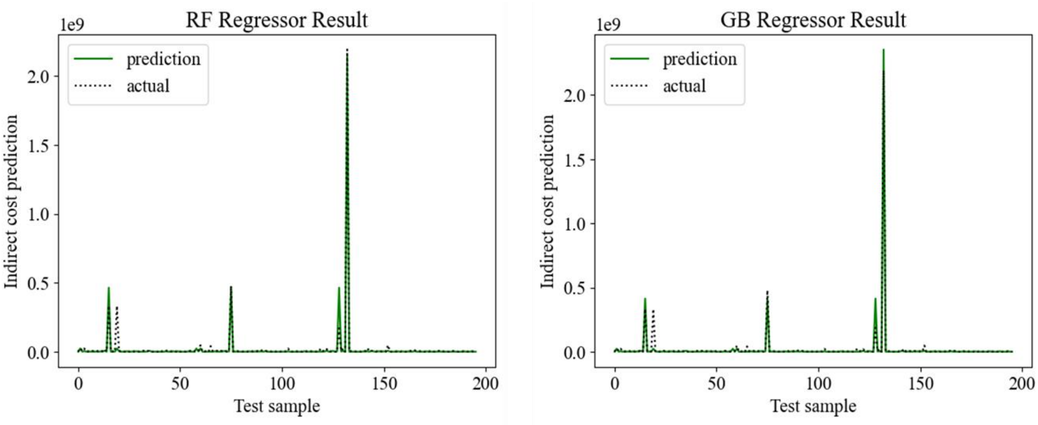
Table 3. Performance comparison of the ML classifiers

Model	Accuracy			Precision			Recall			F1-score		
	Reg.	ROS	RUS	Reg.	ROS	RUS	Reg.	ROS	RUS	Reg.	ROS	RUS
DT	0.84	0.82	0.79	0.79	0.84	0.79	0.78	0.83	0.79	0.78	0.83	0.77
RF	0.83	0.79	0.79	0.73	0.80	0.73	0.76	0.78	0.72	0.74	0.76	0.69
K-NN	0.99	0.97	0.98	0.53	0.93	0.49	0.55	0.93	0.52	0.5	0.93	0.48
XGBoost	0.78	0.90	0.76	0.77	0.90	0.67	0.78	0.90	0.69	0.76	0.90	0.63

On the other hand, Table 4 represents the result of the ML regressor, and Figure 7 illustrates the predictive performance of the regression models. Both regressors show similar performance; however, in terms of MAE (mean absolute error) and MSE (mean squared error), RF shows better performance, although GB had a slightly higher correlation coefficient.

Table 4. Performance comparison of the ML regressors

No.	Model	R^2-score	MAE	MSE
1	RF	0.9618123	0.0262	0.1248
2	GB	0.963	0.03	0.12



(a) (b)

Figure 7: Indirect cost prediction result by the ML regressors

This approach enables construction managers and stakeholders to estimate the total accident cost that may occur immediately after an accident and eradicates the need for following up on hidden and unpredictable indirect costs, saving time and effort. In contrast to previous studies, this two-step prediction technique provides a more direct quantification approach rather than expressing the indirect cost as a ratio to the direct cost. Additionally, the proposed approach uses information readily available after an accident; therefore, employers and stakeholders can immediately use this information during budget management or record the cost in the accounting system. Doing so would allow for future budget allocation and resource management.

6. CONCLUSION

Accidents at construction sites are a frequent phenomenon, and every year the construction industry faces financial losses due to costs incurred from construction accidents. Existing studies on construction accidents mostly focus on causes or risk severity, often neglecting associated cost factors. This study analyzes construction accidents from a cost perspective in an attempt to identify patterns in accident factors that involve indirect costs through DM applications.

A construction accident cost database is created from a survey among domestic construction companies. EDA is performed to characterize the patterns between accident factors and cost variables. Four ML classifiers are used to predict the total accident cost category after an accident. Subsequently, two ML regressors are employed, and using the total accident cost information from the classifier, the regressors predict the amount of indirect cost that may occur after an accident at the construction site. To the best of the authors' knowledge, this is the first attempt to predict the indirect cost incurred by construction accidents. Estimating the accident cost in real time would make decision-makers realize the true financial impact these accidents bring to the project budget and, consequently, inspire them to enhance safety precautions. Additionally, the outcomes of the study would be helpful for stakeholders and practitioners during budget planning and future safety management planning.

Additionally, the EDA indicated that most of the direct and indirect costs are more frequent in buildings, civil infrastructures, and miscellaneous structures construction type. Collapse and fall-type accidents are more common than others, with head and complex types of injuries having high indirect and direct costs, respectively. Hence, workers in these types of construction should be closely monitored and encouraged to maintain PPE to avoid any fatal accidents and, consequently, the resulting costs.

ACKNOWLEGEMENTS

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (National Research for Smart Construction Technology: Grant 225MIP-A156381-03), and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1A6A1A03012812).

REFERENCES

[1] Y. Wang, H. Chen, B. Liu, M. Yang, and Q Long, "A Systematic Review on the Research Progress and Evolving Trends of Occupational Health and Safety Management: A Bibliometric Analysis of Mapping Knowledge Domains", Frontiers in Public Health, vol. 8, pp. 81, 2020.

[2] Y. Shim, J. Jeong, J. Jeong, J. Lee, Y. Kim, "Comparative Analysis of the National Fatality Rate in Construction Industry Using Time-Series Approach and Equivalent Evaluation Conditions", Int. J. Environ. Res. Public Health, vol. 19, pp. 2312, 2022.

[3] S. Kim, S. Cha, Y. Cha, & S. Han, "A Study on the Characteristics of Safety Insensitivity in Construction Workers", Korean Journal of Construction Engineering and Management, vol. 22, no. 2, pp. 88–96, 2021.

[4] S.L. Tang, K.C. Ying, W.Y. Chan, and Y.L. Chan, "Impact of social safety investments on social costs of construction accidents", Construction Management and Economics, vol. 22, no. 9, pp. 937-946, 2004.

[5] O. N. Aneziris, I. A. Papazoglou, D. Kallianiotis, "Occupational risk of tunneling construction," Safety Science, vol. 48, no. 8, pp. 964–972, 2010.

[6] H.W. Heinrich, Industrial Accident Prevention: A Scientific Approach, 1st ed., New York, NY: McGraw Hill, 1931, pp. 2.

[7] J. Hinze and L.L. Appelgate, "Costs of construction injuries", Journal of Construction Engineering and Management, vol. 117, no. 3, pp. 537-550, 1991

[8] E. Leopold and S. Leonard, "Costs of construction accidents to employers," Occupational Accidents, 8, 1987.

[9] J.G. Everret and B.P. Frank, "Costs of accidents and injuries to the construction industry," Journal of Construction Engineering and Management, pp. 158-164, 1996.

[10] S.D. Choi, "A survey of the safety roles and costs of injuries in the roofing contracting industry," Journal of Safety, Health and Environmental Research, vol. 3, no. 1, pp. 1-20, 2006.

[11] N.N.K.N.M. Azman, A.C. Ahmad, M.M. Derus, and I.F.M. Kamar, "Determination of direct to indirect accident cost Ratio for railway construction project," in MATEC Web of Conferences, vol. 266, p. 03009, EDP Sciences, 2019.

[12] E.A.L. Teo and Y. Feng, "Costs of construction accidents to Singapore contractors," International Journal of Construction Management, vol. 11, no. 3, pp. 79-92, 2011.

[13] S. W. Shin, “Construction safety and health management cost prediction model using support vector machine”, Journal of the Korean Society of Safety, vol. 32, no. 1, pp.115-120.

[14] B.U. Ayhan and O.B. Tokdemir, "Predicting the outcome of construction incidents," Safety Science, vol. 113, pp. 91-104, 2019.

[15] T.Q.D. Pham, T. Le-Hong, X.V. Tran, "Efficient estimation and optimization of building costs using machine learning," International Journal of Construction Management, vol. 23, no. 5, pp. 909-921, 2023.