Prediction of Accident Types and Accident-Causing Objects Using Construction Project Data

Yoon S.[a], Chang T.[a,b], Chi S.[a]
[a] Department of Civil and Environmental Engineering, Seoul National University, South Korea
[b] Institute of Construction and Environmental Engineering, Seoul National University, South Korea
ynn8854@snu.ac.kr

Abstract. Accident prevention in the construction industry is more than important considering steady high accident rates. To prevent construction accidents, it is crucial to identify dangerous objects and possible accident types caused by such objects in advance. Thus, the authors developed machine-learning models that predict accident types (e.g., fell, crushed, tripped, struck by) and accident-causing objects (e.g., temporary facility, tool, machinery, materials) based on construction project data including activity types, work progress, weather conditions, and safety planning levels. The performances of four prediction algorithms were compared to determine the optimal algorithm for separate accident types and objects prediction models. As a result, XGBoost showed best performance for both models with the weighted average F1-score of 0.874 and 0.749, respectively. The results of this study can contribute to construction accident prediction by informing practitioners about dangerous and accident-inducing conditions in advance.

1. Introduction

Despite continuous efforts to improve safety, the construction industry is often considered more dangerous than other industries (Kim & Chi, 2019; Z. Zhang et al., 2023). Construction accounts for more than 30% of all fatal accidents across industrial sites, while the employment rate is approximately 7% worldwide. In addition, construction accident rates have been increasing steadily in recent years (KOSIS, 2022). Therefore, construction accidents are a burden on countries as they cause social anxiety and economic losses (Chen et al., 2022; Koc & Gurgun, 2022).

Accordingly, governments, including the Occupational Safety and Health Administration (OSHA) in the United States and the Korea Authority of Land and Infrastructure Safety (KALIS) in Korea, have collected data on construction projects and attempted to investigate the types and causes of recurring accidents using the data. Similarly, several studies have explored construction project data to identify the causes of accidents and prevent safety disasters. In particular, some researchers have developed models to predict potential accident types using the collected data (Cho et al., 2022; Lee et al., 2020; Tixier et al., 2016). However, existing studies are limited in that they consider detailed construction project data, such as activity types, work progress, weather conditions, and safety planning levels when predicting accidents, which are the important sources to explain accident-inducing conditions on site. In addition, because the previous studies focus mainly on predicting accident types, such as slips and falls, it is difficult to identify and respond to the accident-causing objects (e.g., heavy equipment and temporary facilities) because these are the key factors leading to accidents.

To overcome the limitations of previous approaches, this study aims to develop a machine learning (ML) model that uses construction project data to predict the types of construction accidents and the target objects. The model uses only basic information about a given construction site, and can predict the most likely accident objects and types. This allows construction site managers to respond to and prepare for the possible accidents in advance.

2. Literature Review

Research on preventing construction accidents by analyzing construction project data has been conducted in a variety of ways. Most early studies investigated statistical methods to identify the factors that cause construction accidents (Cheng et al., 2010; Chong & Low, 2014; Molenaar et al., 2009; Shapira et al., 2009; Wu et al., 2015). In studies that used simple statistical analysis, the accident frequency by factors such as worker age, a contractor size, and a day of the week was analyzed to determine the main factors that influence accidents. For example, López Arquillos et al. (2012) found that most accidents occurred on Monday. However, these simple statistics are often too general as common senses and do not contribute significantly to accident prevention. Other studies have conducted more detailed statistical analyses. Molenaar et al. (2009) employed structural equation modeling (SEM) to suggest guidelines for construction stakeholders to prevent accidents with the paths from sources to events. Cheng et al. (2012) used a classification and regression tree (CART) to analyze factors that are highly correlated with construction accidents, and succeeded in distinguishing accident situations in which falls and trips occur frequently. These studies are novel in that they identify the most important accident factors on construction sites; but, they are limited to representative construction-related accident factors, which makes it difficult for practitioners to use them on site.

The other researchers have conducted accident prevention studies by predicting accidents using machine learning algorithms (Cheng et al., 2020; Poh et al., 2018; Sadeghi et al., 2020; Zhang et al., 2019). For instance, Choi et al. (2020) developed a fatal accident prediction model using construction worker data; because the model only used data from workers, it did not reflect the characteristics of different sites. Kang and Ryu (2019) proposed a model to predict the accident types that are likely to occur based on construction site data. However, it was limited in its ability to provide safety managers with information about which objects might cause accidents.

To prevent construction site's accidents, it is necessary to prepare not only for the types of accidents but also for the objects that cause them. Construction-site safety managers still have difficulty in determining the objects that cause accidents. Therefore, in this study, the authors developed a model to predict the types of accident and a model that predicts the object causing the accidents using construction project data.

3. Methodology

In this study, an accident prediction model was developed based on Construction Safety Management Integrated Information (CSI) data managed by KALIS. Figure 1 illustrates the

research process, which consists of three steps. In the data preparation step, the CSI data were cleaned, the key variables were determined, and training and test datasets were prepared by data balancing. In the second step of model optimization, four machine learning models (Random Forest; RF, Light Gradient Boosting Model; LGBM, extreme Gradient Boosting; XGBoost, Categorical Boosting; CatBoost) were learned using the training dataset for predicting accident types and target objects. Then, cross-validation was performed to compare the model performances and select the optimal model. In the final step of model evaluation, the optimal model was evaluated using four metrics: accuracy, F1-score, precision, and recall. The overall process was performed using Python version 3.6.8.

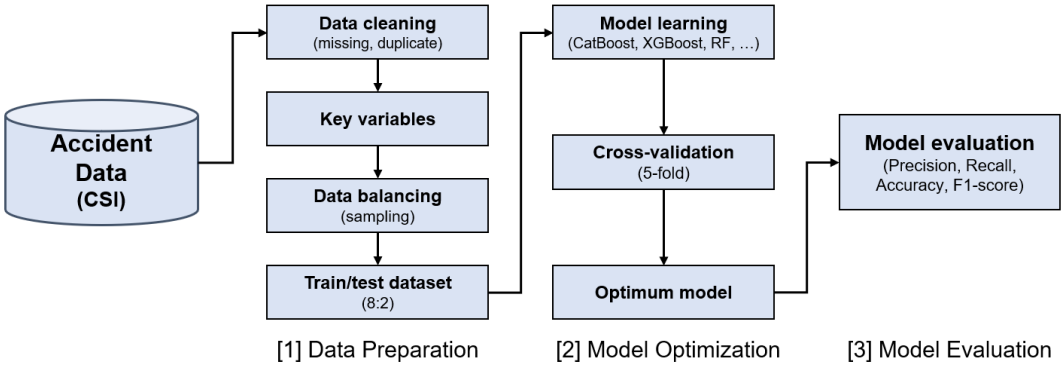

Figure 1: Research Process

3.1 Data Preparation

The data collected from the CSI includes 13,045 construction accident records from 2019 to 2022. The dataset from CSI consists of scene, worker, and accident information. The scene information has 21 variables (e.g., a process type, a process rate, and the number of workers), and the worker information includes 20 variables (e.g., gender, age, a level of injuries). Accident details are explained by the accident information, which consists of 30 variables (e.g., an accident type, an accident object, and an accident location).

Data cleaning was performed to generate input data for the model. The data with 15% or more missing values and the variables having 50 or more categories were eliminated first. Second, the variables that were highly skewed toward a particular category were removed because they were deemed meaningless. For example, male workers, who made up 98% of the entire dataset, were excluded. Numeric variables (e.g., construction cost, a process rate, and a bid rate) were converted into categorical variables. Finally, using correlation analysis, the authors removed the variables with low importance from the combinations of two variables with Cramer's V coefficients of 0.5 or higher to avoid a multicollinearity problem. Eventually, the authors obtained 9,671 data of 18 variables out of a total of 13,045 data of 71 variables.

The pre-processed data were divided into training (80%) and test (20%) data. The training data were then divided into classes, which resulted in a data imbalance problem. Oversampling techniques are normally used to prevent algorithm degradation (Chang et al., 2023). In this study,

the authors used the synthetic minority oversampling technique (SMOTE), a common oversampling technique used to solve overfitting problems (Douzas & Bacao, 2019; Feng et al., 2021). This sampling process was performed using the scikit-learn 0.24.2 and imbalanced-learn 0.7.0, Python libraries. The dataset information for the accident-object and accident-type prediction models is summarized in Table 1, and Table 2.

Table 1: Variables of the final dataset

Category	Variable	Value type
Scene information (13)	Ordering organization	Categorical
	Weather	Categorical
	Temperature	Numerical
	Humidity	Numerical
	Facility type	Categorical
	Construction type (Major)	Categorical
	Construction type (Minor)	Categorical
	Construction cost	Categorical
	Process rate	Categorical
	Number of workers	Categorical
	Work process	Categorical
	Location (Province)	Categorical
	Design safety review status	Categorical
Accident information (5)	Accident month	Categorical
	Accident day of the week	Categorical
	Accident time zone	Categorical
	Accident object	Categorical
	Accident type	Categorical

Table 2: Dataset description

Model	Predictive variables	Number of data oversampled
Accident object	Temporary facility, tool, machinery, structure, construction member, soil and rock, materials, others	Train: 19,673 Test: 4,919
Accident type	Crushed, tripped, fell, pinched, struck by an object, hit, slashed, others	Train: 14,886 Test: 3,722

3.2 Model Optimization

In this study, two separate models were developed to predict dangerous objects and accident types. The optimal algorithm for each model was determined by comparing the performances of the four algorithms. Those algorithms are the most commonly accepted in the latest research on the subject. The four algorithms were tuned using training data sampled via a grid search. Five-fold cross-validation was performed to determine the optimal hyperparameter combination using the grid-search best-param libraries. The hyperparameter combinations for each algorithm are explained in Table 3.

Table 3: Grid search value for algorithms

Algorithm	Grid search value (hyperparameter)
XGBoost	Max_depth: 7-11 (interval: 2), Min_child_weight: 1-7 (interval: 2), colsample_bytree: 0.25-0.75 (interval: 0.25), n_estimators: 10-300 (interval: 10)
CatBoost	Max_depth: 7-11 (interval: 2), learning_rate: 0.01-0.5 (interval: 0.1), n_estimators: 10-300 (interval: 10)
LGBM	Max_depth: 7-11 (interval: 2), Num_leaves: 7-13 (interval: 2), learning_rate: 0.01-0.5 (interval: 0.1), n_estimators: 10-300 (interval: 10)
RF	Min_samples_split: 3-7 (interval: 2), Min_samples_leaf: 3-7 (interval: 2), n_estimators: 10-300 (interval: 10)

After training the model with the optimal hyperparameters, an optimal algorithm was selected by comparing the weighted average F1-score. The F1-score was derived from two metrics: precision and recall. The weighted average F1-score was calculated from the mean of all F1-scores per class, taking into account the support of each class's support. The precision is the ratio of positive predictions that are actually positive, the recall is the ratio of actual positive cases that were correctly predicted to be positive, and the support refers to the number of actual occurrences of a class in the dataset. The weighted average F1-score was calculated using Equation 1.

$$Weighted\ average\ F1\ score = \frac{1}{N}\sum_{i=1}^{N} Support_i \times \frac{2}{(\frac{1}{Precision_i})+(\frac{1}{Recall_i})} \quad (1)$$

Data processing was performed using the Python libraries scikit-learn 0.24.2, Lightgbm 3.3.2, CatBoost 1.1, and XGBoost 1.4.2.

3.3 Model Evaluation

Finally, the performance of the selected algorithm was evaluated. Four metrics (accuracy, precision, recall, and F1-score) commonly used to evaluate the performance of machine learning models were used. Next, the authors analyzed the confusion matrix of each model to discuss the prediction results for each detailed variable. The authors then looked into the recall values of each detailed variable to identify confounding variable combinations. Based on the analysis results, data collection and organization problems and the ways to improve the model were discussed.

4. Experimental Results and Discussion

4.1 Model Optimization Results

Table 3 lists the optimal hyperparameters and weighted average F1-score of the four algorithms for both models. For the accident-object prediction model, the optimal hyperparameter combinations were 9 maximum depth, 3.0 minimum child weight, 0.75 colsample by tree, 200 estimators for XGBoost; 11 maximum depth, 0.2 learning rate, 250 estimators for CatBoost; 9 maximum depth, 13 leaves, 0.4 learning rate, 200 estimators for LGBM; 3 minimum sample split, 3 minimum samples leaf, 300 estimators for RF. The optimal hyperparameter combinations of accident-type prediction model were 11 maximum depth, 1.0 minimum child weight, 0.75 colsample by tree, 300 estimators for XGBoost; 11 maximum depth, 0.3 learning rate, 250 estimators for CatBoost; 11 maximum depth, 13 leaves, 0.5 learning rate, 200 estimators for LGBM; 3 minimum sample split, 3 minimum samples leaf, 300 estimators for RF.

After training both models with the optimal hyperparameters, the weighted average F1-score was obtained, as listed in Table 4 and Figure 2. For the accident-object prediction model, XGBoost yielded the highest weighted average F1-score of 0.874 and was selected as the optimal algorithm. Similarly, for the accident-type prediction model, XGBoost had the highest performance of 0.749 and was determined as the optimal algorithm. In summary, XGBoost outperformed other algorithms for both models.

Table 4: Optimum hyperparameter and weighted average F1-score of each algorithm

Model	Algorithm	Optimum hyperparameter	Weighted average F1- score
Accident object	XGBoost	Max_depth: 9, Min_child_weight: 3, colsample_bytree: 0.75, n_estimators: 200	0.874
	CatBoost	Max_depth: 11, learning_rate: 0.2, n_estimators: 250	0.864
	LGBM	Max_depth: 9, Num_leaves: 13, learning_rate: 0.4, n_estimators: 200	0.862
	RF	Min_samples_split: 3, Min_samples_leaf: 3, n_estimators: 300	0.859
Accident type	XGBoost	Max_depth: 11, Min_child_weight: 1, n_estimators: 300, colsample_bytree: 0.75	0.749
	CatBoost	Max_depth: 11, learning_rate: 0.3, n_estimators: 250	0.730
	LGBM	Max_depth: 11, Num_leaves: 13, learning_rate: 0.5, n_estimators: 200	0.730
	RF	Min_samples_split: 3, Min_samples_leaf: 3, n_estimators: 300	0.738

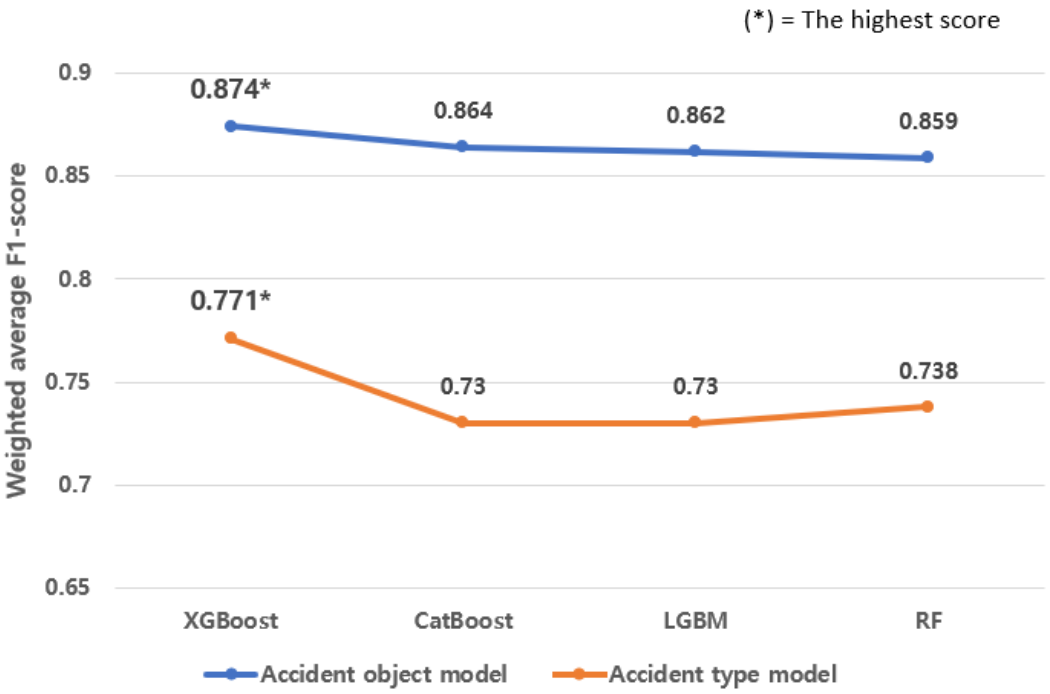

Figure 2: Results of the model performance comparison

4.2 Model Evaluation Results and Discussion

Table 5 summarizes the accuracy, precision, recall, and F1-score of each model. Both models performed acceptable, with the accident-object prediction model performing better with an accuracy of 0.879, precision of 0.873, recall of 0.878, and F1-score of 0.874. The accident-type prediction model achieved an accuracy of 0.776, precision of 0.768, recall of 0.776, and F1 score of 0.771. Both models performed equally well on all four metrics.

Table 5: Model evaluation results

Model	Accuracy	Precision	Recall	F1-score
Accident object	0.879	0.873	0.878	0.874
Accident type	0.776	0.768	0.776	0.771

The confusion matrix for the final model is shown in Figure 3. A confusion matrix analysis of the accident-object prediction model showed that ‘temporary facilities’ and ‘materials’ had low recall values (0.590 and 0.748, respectively). These two variables are often confused with each other, which affects the predictive performance. For example, 86 of the actual accidents caused by ‘temporary facility’ were predicted by ‘materials,’ while 79 of the actual accidents caused by ‘materials’ were predicted by ‘temporary facility.’ Because ‘temporary facility’ is a broad term, 60 of the actual accidents caused by ‘temporary facility’ were incorrectly predicted as ‘others.’ In the

accident-type prediction model, three variables (i.e., tripped, fell, and struck by an object) yielded low recall values (0.415, 0.532, and 0.652, respectively). Unlike the other variables, these three variables were confused with each other, probably because of the similarity of the accident event. For instance, if a person trips over a material on the floor and falls out of the building, this would include two accident types: 'tripped' and 'fell.' Also, falling after being hit by a collapsed structure makes it difficult to specify one accident type. Thus, data entry issues were noted in such cases. In the case of an accident where a worker tripped and injured his ankle, the accident was reported as 'struck by an object,' even though it was 'tripped.' In another accident, where a worker who fell to the ground while installing system scaffolding and was moving without wearing a safety harness, there were two accident types: 'tripped,' 'fell,' but only entered as 'fell.'

In summary, the accident-object and accident-type prediction models performed well, but there were some issues that needed to be resolved. For both models, there were ambiguities in the prediction of some variables. In particular, the accident-type prediction model has data input problems. Thus, the project data entry process for construction projects needs to be improved. In addition, data quality must be enhanced by providing better data entry guidelines to accident reporters. This will allow the models to perform better than current models.

(a)

		Predicted							
		Temporary facility	Tool	Machinery	Materials	Construction member	Structure	Soil & rocks	Others
Actual	Temporary facility	352	32	31	86	25	9	2	60
	Tool	10	548	5	10	0	3	0	6
	Machinery	6	3	614	6	1	1	0	4
	Materials	79	15	15	462	6	6	4	31
	Construction member	0	0	0	15	586	6	0	0
	Structure	0	0	0	6	13	642	0	0
	Soil & rocks	0	0	0	5	0	0	641	10
	Others	47	18	15	47	5	7	1	498

(b)

		Predicted							
		Crushed	Pinched	Tripped	Falling	Struck by object	Hit	Slashed	Others
Actual	Crushed	467	0	0	0	10	5	0	0
	Pinched	1	370	29	21	22	6	4	3
	Tripped	1	37	191	115	58	23	14	21
	Falling	4	26	85	242	67	15	6	10
	Struck by object	1	30	52	57	319	11	10	9
	Hit	0	4	9	3	3	402	0	0
	Slashed	1	2	18	17	7	1	453	1
	Others	0	8	13	5	4	1	0	443

Figure 3: Results of confusion matrix: (a) accident object model; (b) accident type model

5. Conclusion

In this study, a machine learning model is proposed to predict accident-causing objects and accident types based on construction project data. The authors determined the optimal algorithm for each prediction model by comparing the performances of four algorithms. Consequently, XGBoost was found to be the optimal algorithm for both prediction models for accident-causing objects and accident types, showing good performance. This study contributes to the prevention of construction accidents by providing practitioners with information about dangerous conditions and potential accident types. For example, safety managers working in the construction site have the capability to input field information into the model, enabling them to proactively identify and prepare for hazardous objects and potential accident scenarios. In addition, the results of this study can help improve current construction project data collection systems and establish guidelines for accident reporters in data entry. These results also provide further research opportunities, such as the

development of a model to evaluate the safety level of construction sites and the impact analysis of accident-causing factors on site safety.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2003696) and also supported by the National R&D Project for Smart Construction Technology funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport (No. 22SMIP-A158708-03). This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-20241758).

References

Chang, T., Lee, G., & Chi, S. (2023). Development of an Optimized Condition Estimation Model for Bridge Components Using Data-Driven Approaches. *Journal of Performance of Constructed Facilities*, *37*(3). https://doi.org/10.1061/JPCFEV.CFENG-4359

Cheng, C. W., Leu, S. Sen, Cheng, Y. M., Wu, T. C., & Lin, C. C. (2012). Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. *Accident Analysis and Prevention*, *48*, 214–222. https://doi.org/10.1016/j.aap.2011.04.014

Cheng, C. W., Leu, S. Sen, Lin, C. C., & Fan, C. (2010). Characteristic analysis of occupational accidents at small construction enterprises. *Safety Science*, *48*(6), 698–707. https://doi.org/10.1016/j.ssci.2010.02.001

Cheng, M. Y., Kusoemo, D., & Gosno, R. A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in Construction*, *118*. https://doi.org/10.1016/j.autcon.2020.103265

Choi, J., Gu, B., Chin, S., & Lee, J. S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. *Automation in Construction*, *110*. https://doi.org/10.1016/j.autcon.2019.102974

Cho, M., Lee, D., Park, J., & Park, S. (2022). Development of Machine Learning-based Construction Accident Prediction Model Using Structured and Unstructured Data of Construction Sites. *KSCE Journal of Civil and Environmental Engineering Research*, *42*(1), 127–134. https://doi.org/10.12652/Ksce.2022.42.1.0127

Chong, H. Y., & Low, T. S. (2014). Accidents in Malaysian construction industry: Statistical data and court cases. *International Journal of Occupational Safety and Ergonomics*, *20*(3), 503–513. https://doi.org/10.1080/10803548.2014.11077064

Douzas, G., & Bacao, F. (2019). Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. *Information Sciences*, *501*, 118–135. https://doi.org/10.1016/j.ins.2019.06.007

Feng, S., Keung, J., Yu, X., Xiao, Y., & Zhang, M. (2021). Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction. *Information and Software Technology*, *139*. https://doi.org/10.1016/j.infsof.2021.106662

Kang, K., & Ryu, H. (2019). Predicting types of occupational accidents at construction sites in Korea using random forest model. *Safety Science*, *120*, 226–236. https://doi.org/10.1016/j.ssci.2019.06.034

Kim, T., & Chi, S. (2019). Accident Case Retrieval and Analyses: Using Natural Language Processing in the Construction Industry. *Journal of Construction Engineering and Management*, *145*(3). https://doi.org/10.1061/(asce)co.1943-7862.0001625

Koc, K., & Gurgun, A. P. (2022). Scenario-based automated data preprocessing to predict severity of construction accidents. *Automation in Construction*, *140*. https://doi.org/10.1016/j.autcon.2022.104351

KOSIS. (2022). Disaster statistics and analytics by industry. In *Korean Statistical Information Service*.

Lee, J. Y., Yoon, Y. G., Oh, T. K., Park, S., & Ryu, S. il. (2020). A study on data pre-processing and accident prediction modelling for occupational accident analysis in the construction industry. *Applied Sciences (Switzerland)*, *10*(21), 1–23. https://doi.org/10.3390/app10217949

López Arquillos, A., Rubio Romero, J. C., & Gibb, A. (2012). Analysis of construction accidents in Spain, 2003-2008. *Journal of Safety Research*, *43*(5–6), 381–388. https://doi.org/10.1016/j.jsr.2012.07.005

Molenaar, K. R., Park, J.-I., & Washington, S. (2009). Framework for Measuring Corporate Safety Culture and Its Impact on Construction Safety Performance. *Journal of Construction Engineering and Management*, *135*(6). https://doi.org/10.1061/ASCE0733-93642009135:6488

Poh, C. Q. X., Ubeynarayana, C. U., & Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach. *Automation in Construction*, *93*, 375–386. https://doi.org/10.1016/j.autcon.2018.03.022

Sadeghi, H., Mohandes, S. R., Hosseini, M. R., Banihashemi, S., Mahdiyar, A., & Abdullah, A. (2020). Developing an ensemble predictive safety risk assessment model: Case of Malaysian construction projects. *International Journal of Environmental Research and Public Health*, *17*(22), 1–25. https://doi.org/10.3390/ijerph17228395

Shapira, A., Asce, F., & Lyachin, B. (2009). Identification and Analysis of Factors Affecting Safety on Construction Sites with Tower Cranes. *Journal of Construction Engineering and Management*, *135*(1). https://doi.org/10.1061/ASCE0733-93642009135:124

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Application of machine learning to construction injury prediction. *Automation in Construction*, *69*, 102–114. https://doi.org/10.1016/j.autcon.2016.05.016

Wu, X., Liu, Q., Zhang, L., Skibniewski, M. J., & Wang, Y. (2015). Prospective safety performance evaluation on construction sites. *Accident Analysis and Prevention*, *78*, 58–72. https://doi.org/10.1016/j.aap.2015.02.003

Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, *99*, 238–248. https://doi.org/10.1016/j.autcon.2018.12.016

Zhang, Z., Guo, H., Gao, P., Wang, Y., & Fang, Y. (2023). Impact of owners' safety management behavior on construction workers' unsafe behavior. *Safety Science*, *158*. https://doi.org/10.1016/j.ssci.2022.105944