Automation in Construction 110 (2020) 102974

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Machine learning predictive model based on national data for fatal accidents of construction workers

Jongko Choi[a], Bonsung Gu[b], Sangyoon Chin[c,*], Jong-Seok Lee[b]

[a] Department of Convergence for Future City, Sungkyunkwan University, Suwon 16419, Republic of Korea
[b] Department of Industrial Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea
[c] School of Civil, Architectural Engineering & Landscape Architecture, Sungkyunkwan University, Suwon 16419, Republic of Korea

ARTICLE INFO

Keywords:
Safety management
Machine learning
Logistic regression model
Decision tree
Random forest
AdaBoost
Fatal accident prediction

ABSTRACT

The purpose of this study is to develop a prediction model that identifies the potential risk of fatality accidents at construction sites using machine learning based on industrial accident data collected by the Ministry of Employment and Labor (MOEL) of the Republic of Korea from 2011 to 2016. The data details 137,323 injuries and 2846 deaths, and includes age, sex, and length of service of each accident victim, as well as the type of construction, employer scale, and date of the accident. Upon describing the distribution of the dataset, machine learning methods, such as logistic regression, decision tree, random forest, and AdaBoost analyses were applied with the derivation of major variables influencing classification in each algorithm. A comparison of the performance of each model showed the area under the receiver operating characteristic (AUROC) curve to be highest for the random forest method, at 0.9198, which translates to a 91.98% successful predictive rate in terms of classifying workers who could face a high fatality risk. The random forest analysis of this study indicates that the month (season) and employment size are the most influential factors, followed by age, weekday, and service length based on mean decrease Gini values to predict the likelihood of a fatality accident. Moreover, this analysis generated ensemble predictions based on all the factors contained in the dataset. Hence, this study demonstrates the feasibility of machine learning in the construction safety management area. The results obtained can contribute to the prevention of accidents by raising awareness of potential safety risks, by quantitatively predicting fatal accidents and incorporating the findings with a manpower control system at a construction site.

1. Introduction

According to the International Labor Organization (ILO)'s report [1], while construction workers in advanced countries have three to four times higher probability to experience a fatal accident than workers in other industries, construction workers in less developed countries face a three to six times greater risk. Employment in the construction industry accounts for about 7% of global employment, and 100,000 workers are killed at construction sites every year, which accounts for around 35% of the world's occupational fatalities [2]. A comparison of the number of fatal accidents in several countries since 2010 gives the annual average number of deaths per country as follows: 535 (30.1%) in Korea from 2011 to 2017, as collected by the Ministry of Employment and Labor (MOEL) [3]; 854 (18.2%) in the United States from 2011 to 2014 collected by Occupational Safety and Health Administration (OSHA) [4]; 315 (32.8%) in Japan from 2015 to 2017 by Japan Industrial Safety and Health Association (JISHA) [5]; and 39 (27.7%) in the United Kingdom between 2013 and 2018 by Health and Safety Executive (HSE) [6]. In the majority of countries around the world, the rate of fatal accidents is the highest in the construction industry [2]. Hence, the need for construction companies to reduce fatalities is a matter of pressing urgency.

Goh and Ubeynarayana [7] pointed out that past accidents can serve as a foundation to prevent future accidents. Traditionally, the causes of construction accidents are classified as "social environment and ancestry," "fault of person," or "unsafe act or condition" [8]. Safety management at construction sites is based on subjective data analysis, expert opinions, regulations, and management perspectives, and in Korea, it is based on the annual (quarterly implementation from 2017) report of the MOEL, the experience of the safety manager, and accident statistics collected from the Korea Workers Compensation and Welfare Service (KCOMWEL). However, in the Republic of Korea, despite creating and applying safety management guidelines, the number of accidents occurring in the construction industry has increased [3]. This

* Corresponding author.
E-mail addresses: jkarch@skku.edu (J. Choi), galosa@skku.edu (B. Gu), schin@skku.edu (S. Chin), jongseok@skku.edu (J.-S. Lee).

https://doi.org/10.1016/j.autcon.2019.102974
Received 1 May 2019; Received in revised form 24 September 2019; Accepted 29 September 2019

is because the construction scale became larger and more complicated, the types of construction diversified, various types of workers were introduced to the site, and the management subject became more widespread.

As of December 2017, the number of licensed safety managers in Korea was 24,382, and the number of construction workers was estimated to be above two million [9,10]. A construction company must, by Korea law, employ personnel who meet the minimum appointment criteria for a safety manager as described in Article 12.1 of the Enforcement Decree of the Occupational Safety and Health Act. However, construction workers outnumber safety managers by approximately 84 to 1, and thus it is necessary to improve the effectiveness of safety management and safety awareness by empowering safety managers to manage tens of or even hundreds of construction workers more effectively, and appropriately inform each employee of the risks they are facing.

Recent developments in forecasting technologies using existing data suggest that accidents do not occur randomly, and more empirical and quantitative research should be conducted to prevent them [11]. In this regard, some studies have applied machine learning to safety management as a means of increasing efficiency throughout the industry. In the field of traffic accidents analysis, applying machine learning to collected data is the most widely used method of predicting potential risks [12–17]. Although the same approach has been applied to the study of accidents in the fields of aviation and construction [2,11,18–23], there is still a paucity of research on how to use and apply machine learning in the field of construction safety.

Therefore, the objective of this study is to develop a prediction model of fatal accidents at construction sites by comparing the use and predictive performance of machine learning approaches to safety management using publicly available national data. This study focuses on machine learning and the subsequent testing of data regarding injuries and deaths, thereby, creating a predictive model that assesses the risk of fatality accidents based on manpower information at a construction site on a daily basis. This study is expected to contribute to construction safety for more proactive management by incorporating the prediction model into a manpower management system at the construction site to alert safety managers of the likelihood of fatal accidents at worker, crew, or contractor levels, as described in the utilization plan.

2. Research scope and method

This study utilized accident data collected by the MOEL through the Information Disclosure Portal (www.open.go.kr). The data is categorized into age, sex, length of service, employer scale, construction type, accident date, and disaster level. This study limited the period under consideration to the six years from 2011 to 2016 due to data availability at the time of the study in 2018. The data corresponding to this period consist of 137,323 recorded injuries and 2846 recorded deaths [3]. To construct a model that predicts the risk not only of fatal, but of all accidents, construction-site data on both accidents and non-accidents is necessary. However, due to the data protection laws in Korea [24], there is no nationally accumulated collection of such all-inclusive construction worker and site data including non-accident instances. Therefore, the scope of this study is limited to develop a model that predicts the likelihood of mortality in the event of an accident.

When applying machine learning methods to our data, the following was considered: 1) various types of methods including both linear and nonlinear models, with utilization of both single and ensemble learning; 2) the method must be either interpretable or capable of finding important features to predict a fatality accident. Logistic regression is an interpretable linear classifier. The decision tree is employed as a nonlinear model, since it is likewise interpretable and can effectively handle categorical features. Random forest is an ensemble learning method of the decision tree, and AdaBoost is a model for logistic regression. Both random forest and AdaBoost can identify important features. Apart from the abovementioned methods, k-nearest neighbors (kNN), support vector machines (SVM), and neural networks (NN) are the most frequently used methods in the machine learning field. However, all of these are known to be inappropriate for the handling of many of categorical features, since the features should be encoded using a one-hot encoding scheme, which results in extremely high dimensionality [25]. In addition, these learning methods cannot be interpreted.

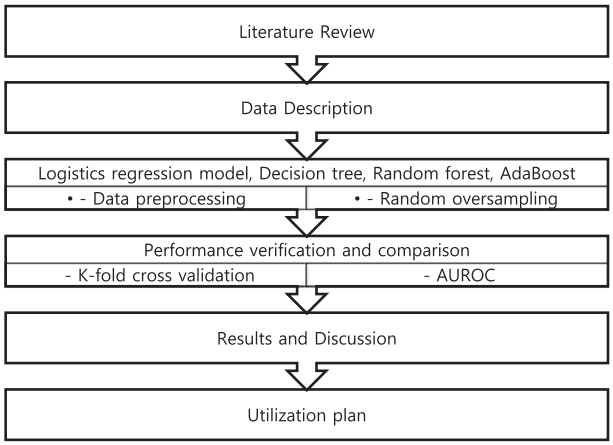

Fig. 1. Research process.

The study was conducted employing the following procedure (Fig. 1): First, prior studies and accident data collected by the MOEL were reviewed. Second, the distribution of data by item within these datasets was explained. Third, data from the dataset was preprocessed, and machine learning and verification were conducted to classify general and fatal accidents through the logistic regression, decision tree, random forest, and AdaBoost models. Fourth, the receiver operating characteristic (ROC) curve and area under receiver operating characteristic (AUROC) were utilized to verify the suitability of the derived models. Fifth, the performances and major items emerging from the predictive models derived by machine learning were compared. Finally, a process is suggested by which the safety manager can classify the risk of serious accidents, dangerous companies, and dangerous work teams, using a predictive model.

3. Literature review

3.1. Machine learning

A study on accident prevention using the national database (DB) was previously attempted in various fields. In the United States, a year's worth of traffic accident data regarding rear vehicular collisions in Florida were analyzed using multiple logistic regression of accident inducers and victims' data. The accident inducers considered were the number of lanes, highway division (divided or undivided), accident time, road surface conditions, region (urban or rural), posted speed, driver age, substance abuse (alcohol, drugs, or none), residence code, sex (male or female), and type of vehicle [12]. In Japan, data on 2.8 million large-scale accidents that had occurred over twelve years were analyzed using the logistic regression model. The approach employed in this model has changed, however at that time, the process involved using 50% of the data to develop the model and the remaining 50% to validate it [17]. In Canada, vehicular crash data gathered over a period of six years from 31 routes across the state of Ontario were compared with the analysis results produced by three logistic regression models (the sequential binary logit model, ordered logit model, and multinomial logit model) and three methods of counting the crash data (passenger-based, vehicle-based, and crash-based). Yannis et al. [13]

analyzed 1300 major accident samples in seven European countries, and applied the logistic regression model by item and step, observing how the coefficients changed with the addition of independent variables. In Turkey, logistic regression and discriminative analysis of 2552 traffic accidents determined damages from injury and non-injury in the Eskisehir province [15]. In Cyprus, accident data were applied to Bayesian networks, and the Bayesian network and traffic situation estimation of the transportation system over time were integrated with the dynamic traffic assignment (DTA) simulation to derive the risk index [14].

In the field of aviation, Yeoum & Lee [20] developed an accident-prediction model using an artificial neural network (ANN) and logistic regression analysis, drawing data from the Korean Air Force's accident records for the past 30 years. They compared and analyzed 24 items, selected 13 items as variables, and subsequently performed logistic regression analysis with 9 of the 13 selected items that were considered most influential in the accidents. Li et al. [18] reported that less experienced pilots tended to have more accidents. However, Bazargan & Guzhva [19] analyzed the effects of sex, age, experience, and major accidents on general flight accidents, and concluded that men with more flight experience and men aged 60 or older tend to be correlated with fatal accidents. These findings likely explain the propensity of older and more experienced male pilots to fly in more dangerous environments and perform more challenging maneuvers.

The machine learning methods have been utilized for hydrological studies and water resource management [26–29]. The adaptive network-based fuzzy inference system (ANFIS) was utilized to predict nanofluid relative viscosity [30], and the extreme learning machine model was adopted to predict river flow [31].

In the construction industry, studies have been performed to make various predictions using machine learning data accumulated in national or corporate DBs. Tixier et al. [11] performed random forest analysis and stochastic gradient tree boosting with 4400 datasets validated in previous studies, and Gerassis et al. [32] analyzed six years of accident data. Employing known or predicted causal factors, a Bayesian network was established to predict the probability of an accident type risk scenario for bank-related accidents. Amiri et al. [22] analyzed five years of data in Iran using multiple-correspondence analysis, decision tree analysis, ensembles of decision tree, and the association rules method. Alizadeh et al. [21] calculated the conditional probability of severe and fatal injury between the parameters of age, marital status, career, accident experience, and accident severity employing Bayesian theory. Furthermore, they attempted to apply the results to worker training to reduce accidents by ranking risks in terms of parameter totals, and improving the perception of vulnerabilities and problems in safety systems on site. Lastly, Chiang et al. [2] conducted a cluster analysis of fatal accidents in Hong Kong.

3.2. Accidents and correlation factors

Personal characteristics have been identified as factors affecting accidents [33], and in the construction sector, the correlation between demographics, field characteristics, and accidents has been analyzed [2,21,32,34,35]. Demographic characteristics are best understood quantitatively; whereas, personal characteristics are best understood qualitatively. Although both affect the occurrence of accidents, in the construction industry it is difficult to perform quantitative and qualitative evaluations of safety training at the construction site, as workers are frequently moved between sites and new workers arrive often. Thus, although qualitative factors are certainly important, the complexities of the construction industry require focusing on the quantitative demographic characteristics and environmental factors, to assess workers and quickly classify accident risk groups in the field.

The relationship between demographic characteristics and the propensity for accidents has been studied by a number of researchers. Some studies have indicated that older people are more likely to have accidents [22,35,36], because physical and mental ability decreases with increasing age along with the capacity to adapt to job requirements [37,38]. In contrast, a study showed that although fatal accidents occur more frequently as people age, younger people tend to have more accidents in general [39].

Considering the role of sex, some studies indicate that men are more likely to be involved in an accident [35,36,39]. The observed sex-difference in accident rates is associated with risk exposure, and also with the mechanism of accident occurrence [40]. However, Villanueva and Garcia [36] criticized sex-dependent risk studies as being of limited value, because of the very small proportion of women present in the sample groups of these fatal accident studies.

Regarding career experience, temporary or atypical workers may have higher accident rates due to inexperience and poor safety behavior [36]. Alizadeh et al. [21] presented a group with high accident rates for their age, marital status, career, and accident experience. Here, the highest accident rate occurred in workers 50 years of age or older, married, with 1 to 5 years of career experience, and no previous accident experience. In addition, another study analyzed the relationship between career, company size, and accident experience as a factor in construction site accidents, and found that unskilled workers in small-to-medium sized enterprises (SMEs) were at the greatest risk of a falling accident [35]. On the contrary, Zhang et al. [41] insists that individual differences among workers had not been statistically significant in the safety behavior of workers, whereas workplace conditions had a negative effect.

To effectively apply these factors to safety education, it is necessary to consider the combination of accident factors to prevent actual accidents, rather than identifying the risk groups through the correlation of individual factors. For this purpose, Chiang et al. [2] collected data from the WiseNews website in Hong Kong. They statistically analyzed the data for 256 accidents according to the following factors: age, sex, time, day, month, ordering organization, and work type. These factors were then classified into four clusters, and further classified into groups with a high probability of accident occurrence through combination. In the present study, clusters were formed by excluding sex, work type, and accident type, because of the observation by Chiang et al. [2] that the sex, work type, and accident type factors dominated cluster formation, obscuring the effects of other factors. Consequently, only the month, age, time, and day of the week represented items in the cluster, and a model was presented to prevent accidents occurring with employees who experience seasonal risk factors based on the derived clusters [2], while excluding important factors, such as sex, work type, and accident type, that reflect the characteristics of the construction industry.

The current study deviates from previous work in a number of important ways. The first is that the researchers did not intentionally collect data; instead, they used existing historical national datasets collected by public agencies. The significantly larger size of the dataset used - 140,169 datum – enables the development of a more reliable prediction model. Second, the model is field-customized rather than managing a rank or cluster by presentation of a statistical table. Third, the research models result in prediction models that are not coordinated by researchers and predictability, by allowing the machine to randomly classify learning and verification data without affecting the data categories. This facilitates better assessment of the model's performance. Finally, the study presents a method to utilize the developed model by employing the existing equipment environment on construction sites.

4. Development of prediction model for fatal accidents

After data preprocessing, machine learning was performed using logistic regression, decision tree, random forest, and AdaBoost analyses. The process involved randomly selecting 80% of the data from the accident dataset to learn and using the remaining 20% of the dataset for

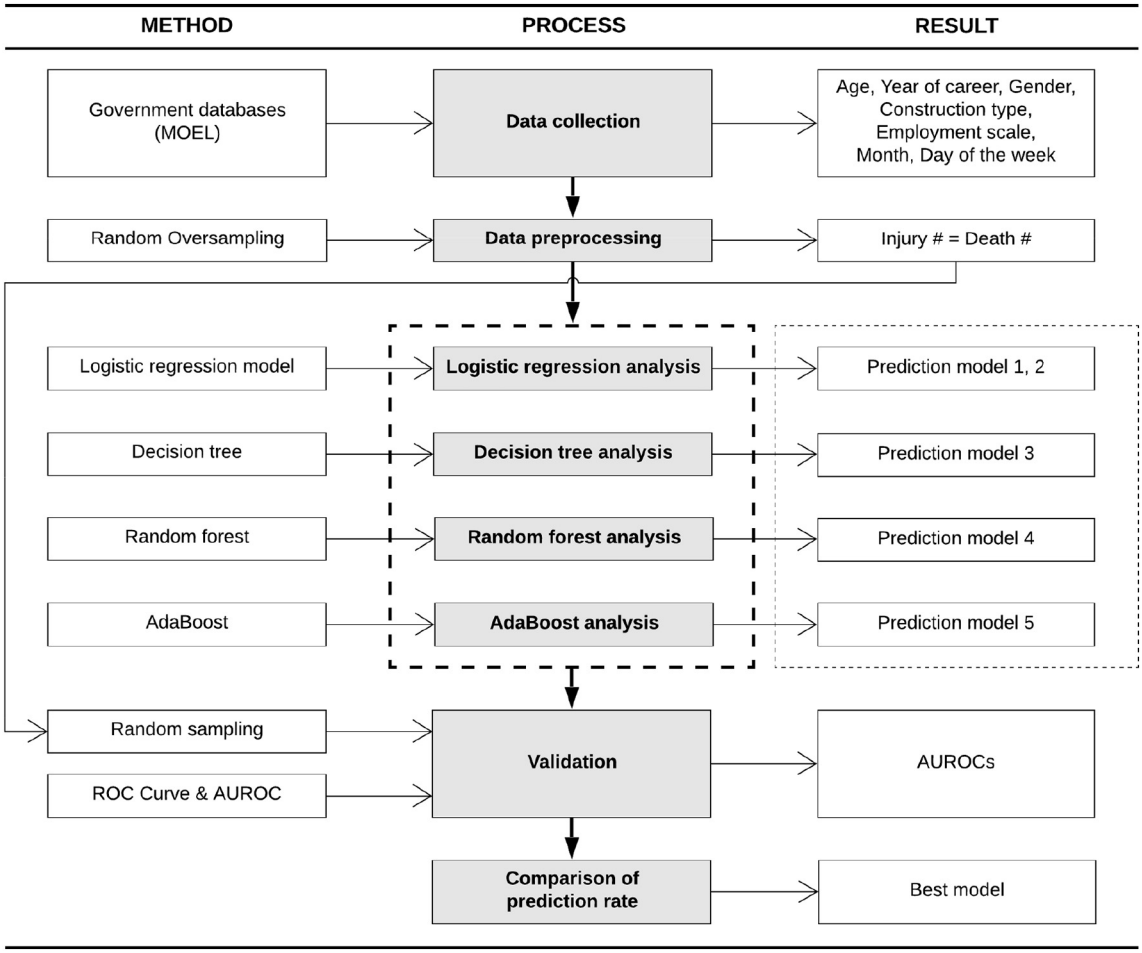

Fig. 2. Machine learning process.

verification. This was repeated 30 times, and the average AUROC was derived as the performance evaluation value, since it is more appropriate for evaluating the model performance from an imbalanced dataset [42]. In contrast, in the logistic regression model, all data are learned to obtain a single regression equation, and the average AUROC is presented as the performance after 30 verifications with 20% randomly selected data. The performance of the algorithms was compared, and the best performance model was determined (Fig. 2).

4.1. Input data description for machine learning analysis

4.1.1. Number of accidents and deaths by year

The number of disaster victims in Korea has increased annually from 21,610 in 2011 to 25,114 in 2016 (Table 1). Although the number of deaths is neither steadily increasing nor steadily decreasing, there are 400–500 fatalities each year, and the number of accidents, including fatal accidents, has increased over the six-year period under consideration [3]. In the light of annual accident statistics (Table 1), the number of accidents has increased regardless of the annual contract amount [43].

4.1.2. Relationship between age and accident propensity

The number of injuries and deaths by age is normally distributed, as shown in Table 2. The total number of accidents, injuries, and deaths were highest in the 55–59 age range, indicating that over 66% of accidents occurred in the age of 50 or older people. As 68% of fatality accidents happened in the age of 50 or older people, the mortality rate in the event of an accident (MA) generally increased in the below 80 age group, showing the highest MA in the age group of 75–79.

4.1.3. Relationship between length of service and accident propensity

Considering the workers' lengths of service, 99,059 workers, who experienced injury or death, had less than one month of experience, accounting for 70.67% of the total, as shown in Table 3. It is notable that the percentage of workers with a career shorter than a year is above 95% in accidents, and above 90% in the MA. The total number of accidents, injuries, and deaths were the highest for people with less than one month of service; however, the highest mortality rate, at 14.29%, was experienced by workers with more than 20 years of service experience. This means that the accident rate is the lowest for those with more than 20 years of career, whereas the rate of death due to

Table 1
Number of injuries and deaths by year - data from [3,43].

Year	Total	2011	2012	2013	2014	2015	2016
1. Accidents	140,169	22,109	22,583	22,801	22,851	24,212	25,613
1-1. Injuries	137,323	21,610	22,122	22,285	22,417	23,775	25,114
1-2. Deaths (%)	2846 (2.03)	499 (2.26)	461 (2.04)	516 (2.26)	434 (1.90)	437 (1.80)	499 (1.95)
2. Contract (US$, billion)	667.14	100.64	92.28	83.01	97.70	143.62	149.89

Table 2
Number of accidents and deaths by age - data from [3].

Age	Total	Up to 17	18–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80+	None
1. Accidents	140,169	3	709	1721	3715	6986	13,308	20,832	28,222	29,870	20,392	10,328	3406	617	51	9
Percentage (%)	100	0.00	0.51	1.23	2.65	4.98	9.49	14.86	20.13	21.31	14.55	7.37	2.43	0.44	0.04	0.01
1-1. Injuries	137,323	3	692	1688	3634	6856	13,059	20,432	27,692	29,286	19,952	10,054	3317	598	51	9
Percentage (%)	97.97	0.00	0.49	1.20	2.59	4.89	9.32	14.58	19.76	20.89	14.23	7.17	2.37	0.43	0.04	0.01
1-2. Deaths	2846	–	17	33	81	130	249	400	530	584	440	274	89	19	–	–
Percentage (%)	2.03	–	0.01	0.02	0.06	0.09	0.18	0.29	0.38	0.42	0.31	0.20	0.06	0.01	–	–
2. MA[a] (%)	2.03	–	2.40	1.92	2.18	1.86	1.87	1.92	1.88	1.96	2.16	2.65	2.61	3.08	–	–

[a] MA (Mortality in the event of an accident) = Deaths/Accidents * 100.

accident is the highest.

4.1.4. Relationship between type of construction and accident propensity

Considering accidents and deaths by industry for the six-year study period, there were 90,723 victims in the building construction industry, accounting for 64.72% of all accidents in the overall construction industry, as shown in Table 4. This was followed by other construction sectors (31.93%), such as construction machinery management (CMM) (1.52%), and machinery (1.52%). The MA was 12.99% for the railway and track sector, 4.43% for the machinery sector, 3.83% for the road construction sector, 3.24% for the CMM sector, 2.17% for other construction sectors, and 1.86% for the building construction sector. Although the building construction sector accounts for most of the industrial accidents in the data, the MA is highest in the construction of civil infrastructure such as roads, railways, and tracks.

4.1.5. Relationship between employer size and accident propensity

As shown in Table 5, of the total accidents, 37.99% occurred in projects with less than five people, and 0.15% were in projects with more than 2000 people, presenting that above 92% of the accidents happened in projects with under 100 people. The number of accidents decreased as the employment scale of the project increased. However, the MA was 1.77% at sites with less than five people and 8.10% in projects with more than 2000 people. The MA rose as the employment size at the site increased, which implies that large scale construction projects have less accidents than small or medium scaled projects, however these accidents are more likely to involve fatality.

4.1.6. Relationship between month, day of the week, and accident propensity

The data were classified by month on the basis of the date on which each accident occurred, as shown in Table 6. Most construction accidents occurred in October, and the least in February. Grouping the accidents data by season, accidents occurred more frequently in the summer and fall, while the least number of accidents occurred during winter. However, the MA is highest during winter, while other seasons have almost no differences in MA.

The data were also classified by day of the week, as shown in Table 7. The accident rate was highest on Tuesdays and lowest on Sundays, whereas the MA was the highest on Sundays and lowest on Mondays.

4.1.7. Relationship between sex and accident propensity

Considering victims by sex for the six-year study period, men suffered 134,020 injuries and 2805 deaths, whereas women suffered 3303 injuries and 41 deaths, as shown in Table 8. The data indicated that the mortality rate was approximately 41 times higher for men than for women, herein 2.05% for men and 1.23% for women, and 1.67 times higher for men in the event of an accident.

4.2. Data preprocessing

Data preprocessing was required for the machine method to gain an understanding before learning the data. First, unclassifiable data were deleted. Second, because no linear correlation exists between age, length of service, scale of employer and month, the items were recognized as factors rather than a number. The data in this study were severely imbalanced, with injury data being 48 times more prevalent than fatality data. In this state, when the prediction model classifies all the data as a non-serious accident, the prediction rate emerges erroneously as 98%. Since most machine learning methods were developed under the assumption of balanced class, a class imbalance dataset should be preprocessed to produce the class distribution balance. The most frequently used preprocessing methods are 1) random over-sampling (ROS), 2) random under-sampling (RUS), and 3) synthetic minority over-sampling technique (SMOTE). ROS duplicates positive objects, which are randomly chosen, to increase the number positive objects, whereas RUS arbitrarily deletes negative objects to decrease the number of negative objects. There is no difference in performance for the classification between these sampling methods if the degree of class imbalance is moderate [44,45]. However, if the class distribution is highly skewed, then ROS is preferred. RUS may deteriorate the negative class distribution by deleting negative objects in such a case, whereas the ROS never faces such risk. The SMOTE generates artificial positive objects by interpolating existing positive objects [46]. Since the interpolation can be done on numerical data, this method is not a good choice for our data including many categorical features. Therefore, ROS was performed to address the class imbalance problem. The death objects were extracted from the dataset and duplicated to obtain the same number of objects for the injury and the death. This method offered the advantage of eliminating the possibility of the researcher manipulating the data.

4.3. AUROC

In binary classification, the predictive performance of a classifier is generally measured using four statistics: true positives (TP), the number of correctly predicted positive objects; false negatives (FN), the number of incorrectly predicted positive objects; false positives (FP), the number of incorrectly predicted negative objects; and true negatives (TN), the number of correctly predicted negative objects. In the case of class balance, the accuracy defined by

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

is used to quantify the classification performance of a classifier [47]. However, this metric becomes invalid if the class distribution is skewed. Assuming that a dataset includes 90 negative objects and 10 positive objects, and that we learned a classifier that predicts only a negative class. Then, the accuracy of this classifier would be 90%, although the classifier is not useful at all. In the case of class imbalance, we should therefore simultaneously consider two metrics, namely the true positive rate (TPR) and false positive rate (FPR). They are defined in Eqs. (1) and (2).

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

Table 3
Number of injuries and deaths with respect to length of service - data from [3].

Career	Total	Up to 1 month	1–2 months	2–3 months	3–4 months	4–5 months	5–6 months	6 months to 1 year	1–2 years	2–3 years	3–4 years	4–5 years	5–10 years	10–20 years	20+ years	None
1. Accidents	140,169	99,059	15,411	6950	3750	2284	1583	4344	2743	1098	659	469	1111	447	56	205
Percentage (%)	100	70.67	10.99	4.96	2.68	1.63	1.13	3.10	1.96	0.78	0.47	0.33	0.79	0.32	0.04	0.15
Cumulative %	100	70.67	81.66	86.62	89.30	90.93	92.06	95.16	97.12	97.90	98.37	98.70	99.49	99.81	99.85	100
1-1. Injuries	137,323	97,326	15,133	6791	3644	2224	1521	4178	2621	1054	641	454	1060	423	48	205
Percentage (%)	97.97	69.43	10.80	4.84	2.60	1.59	1.09	2.98	1.87	0.75	0.46	0.32	0.76	0.30	0.03	0.15
1-2. Deaths	2846	1733	278	159	106	60	62	166	122	44	18	15	51	24	8	–
Percentage (%)	2.03	1.24	0.20	0.11	0.08	0.0%	0.04	0.12	0.09	0.03	0.01	0.01	0.04	0.02	0.01	–
Cumulative %	100	60.89	70.66	76.25	79.97	82.08	84.26	90.09	94.38	95.92	96.56	97.08	98.88	99.72	100	–
2. MA[a] (%)	2.03	1.75	1.80	2.29	2.83	2.63	3.92	3.82	4.45	4.01	2.73	3.20	4.59	5.37	14.29	–

[a] MA (Mortality in the event of an Accident) = Deaths/Accidents * 100.

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

TPR is the ratio of correctly classified positive objects to true positive objects, whereas FPR is the ratio of incorrectly classified negative objects to true negative objects. A receiver operating characteristic (ROC) graph is a useful tool for visualizing and comparing classifier performance based on TPR and FPR measures [48]. The ROC curve is drawn on the two dimensional space, where TPR is on the vertical axis, and FPR is on the horizontal axis. The closer the curve is to the upper left corner, the better is the prediction of a classifier. This also means that the area under the ROC (AUROC) becomes large and approaches unity, which indicates perfect classification. Although there are some alternatives to the AUROC, such as the F1 measure and G-mean, recent studies generally use the AUROC, because it can be visualized and is therefore easy to interpret [47]. In general, an AUROC value of 0.9–1.0 is excellent, 0.8–0.9 is good, 0.7–0.8 is fair, 0.6–0.7 is poor, and 0.5–0.6 indicates failure.

4.4. Logistic regression model

The logistic regression model, proposed by Cox [49] and applied in a diverse variety of fields, is a statistical technique used to predict the likelihood of the occurrence of specific events using a linear combination of independent variables as a probability model. The logistic regression and the general regression analysis are the same in that a dependent variable is explained by a linear combination of independent variables, and both are used for future prediction. However, unlike linear regression analysis, logistic regression is a classification technique, because the dependent variable includes categorical values, and the result becomes one of the categories when a new input is provided.

The data used in this study are categorized based on the logistic regression model, as they contain categorical data. This model is useful, because it demonstrates the correlation between items, although it may not be as good as other analytical methods. Logistic regression analysis is easy and enables clear interpretation of the coefficients. Whereas neural networks are almost impossible to explain and interpret, in that they are often referred to as black boxes, logistic regression analysis offers the advantage that users can interpret and utilize the resulting formula expressed in a simple linear form. The basic approach to logistic regression is to use linear regression. The linear prediction function $f(i)$ can be expressed as Eq. (3) for a specific data term. The logistic regression equation ensures that the dependent variable or result is between the probabilities [0,1], regardless of any number between independent variables $[-\infty, \infty]$, and classifies the result as 0 or 1 by the threshold value of 0.5.

$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \tag{3}$$

As a part of this study, the preliminary study using the logistic regression model has been published by Choi et al. [50]. The logistic regression model was tested with 80% of the data, as well as with 20% of the data over 30 repetitions using a 5-fold cross-validation method [51]. AUROC was measured by verifying the optimal model derived from this process with the remaining 20% of the data. However, because this method randomly extracts training, test, and validation data, it poses the problem that the results model and performance are changed every time a new model is created. To solve this problem, we developed a single model by learning and testing the entire dataset only in logistic regression analysis. To find the objective performance of the model, 20% of the random data was extracted and verified 30 times (minimum value), and the average was presented as representative of objective performance.

Therefore, the complete preprocessed dataset was modeled based on the logistic regression model. As a result, sex, employer scale, length of service, month, and day of the week were determined to be significant factors in fatal accidents, whereas age and construction type were

Table 4
Number of injuries and deaths by type of construction - data from [3].

Type	Total	CMM	Building	Overpass/U.G.R.	Dam	Machinery	Road	Hydroelectric	Railway and track	Other
1. Accidents	140,169	2128	90,723	4	6	2236	235	7	77	44,753
Percentage (%)	100	1.52	64.72	0.00	0.00	1.6	0.17	0.00	0.05	31.93
1-1. Injuries	137,323	2059	89,036	4	6	2137	226	7	67	43,781
Percentage (%)	97.97	1.47	63.52	0.00	0.00	1.52	0.16	0.00	0.05	31.23
1-2. Deaths	2846	69	1687	–	–	99	9	–	10	972
Percentage (%)	2.03	0.05	1.20	–	–	0.07	0.01	–	0.01	0.69
2. MA[a] (%)	2.03	3.24	1.86	–	–	4.43	3.83	–	12.99	2.17

[a] MA (Mortality in the event of an Accident) = Deaths/Accidents * 100.

Table 5
Number of injuries and deaths by employer scale - data from [3].

Employment scale	Total	up to 4	5–9	10–15	16–29	30–49	50–99	100–199	200–299	300–499	500–999	1000–1999	2000+
1. Accidents	140,169	53,256	26,352	14,839	16,328	10,575	8526	4488	1950	1769	1407	469	210
Percentage (%)	100	37.99	18.80	10.59	11.65	7.54	6.08	3.20	1.39	1.26	1.01	0.33	0.15
Cumulative %	100	37.99	56.79	67.38	79.03	86.57	92.66	95.86	97.25	98.51	99.52	99.85	100.00
1-1. Injuries	137,323	52,314	26,017	14,608	16,040	10,350	8247	4265	1847	1674	1327	441	193
Percentage (%)	97.97	37.32	18.56	10.42	11.44	7.38	5.88	3.04	1.32	1.19	0.95	0.31	0.14
1-2. Deaths	2846	942	335	231	288	225	279	223	103	95	80	28	17
Percentage (%)	2.03	0.67	0.24	0.16	0.21	0.16	0.20	0.16	0.07	0.07	0.06	0.02	0.01
2. MA[a] (%)	2.03	1.77	1.27	1.56	1.76	2.13	3.27	4.97	5.28	5.37	5.69	5.97	8.10
Cumulative %		1.77	1.60	1.60	1.62	1.67	1.77	1.88	1.93	1.97	2.01	2.02	2.03

[a] MA (Mortality in the event of an Accident) = Deaths/Accidents * 100.

Table 6
Number of monthly injuries and deaths - data from [3].

Month	Total	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.
		Spring (Mar–May)			Summer (Jun–Aug)			Fall (Sep–Nov)			Winter (Dec–Feb)		
1. Accidents	140,169	11,073	11,594	13,192	13,315	12,768	13,418	11,818	14,075	12,774	11,076	8008	7058
Percentage (%)	100	7.90	8.27	9.41	9.50	9.11	9.57	8.43	10.04	9.11	7.90	5.71	5.04
Seasonal %	100	25.6			28.2			27.6			18.7		
1-1. Injuries	137,323	10,827	11,372	12,953	13,067	12,508	13,158	11,569	13,802	12,531	10,825	7820	6891
Percentage (%)	97.97	7.72	8.11	9.24	9.32	8.92	9.39	8.25	9.85	8.94	7.72	5.58	4.92
Seasonal %	100	25.6			28.2			27.6			18.6		
1-2. Deaths	2846	246	222	239	248	260	260	249	273	243	251	188	167
Percentage (%)	2.03	0.18	0.16	0.17	0.18	0.19	0.19	0.18	0.19	0.17	0.18	0.13	0.12
Seasonal %	100	24.8			27.0			26.9			21.3		
2. MA[a] (%)	2.03	2.22	1.91	1.81	1.86	2.04	1.94	2.11	1.94	1.90	2.27	2.35	2.37
Seasonal %	100	1.97			1.94			1.98			2.32		

[a] MA (Mortality in the event of an Accident) = Deaths/Accidents * 100.

Table 7
Number of injuries and deaths by day of the week - data from [3].

Day of the week	Total	Sun.	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.
1. Accidents	140,169	11,323	21,474	22,011	21,507	22,149	21,744	19,961
Percentage	100%	8.08%	15.32%	15.70%	15.34%	15.80%	15.51%	14.24%
1-1. Injuries	137,323	11,059	21,075	21,562	21,086	21,675	21,309	19,557
Percentage	97.97%	7.89%	15.04%	15.38%	15.04%	15.46%	15.20%	13.95%
1-2. Deaths	2846	264	399	449	421	474	435	404
Percentage	2.03%	0.19%	0.28%	0.32%	0.30%	0.34%	0.31%	0.29%
2. MA[a]	2.03%	2.33%	1.86%	2.04%	1.96%	2.14%	2.00%	2.02%

[a] MA (Mortality in the event of an Accident) = Deaths/Accidents * 100.

determined to be insignificant factors (Tables 9 and 10). The reason why the age (Table 9) and the construction type (Table 10) are inconsequential is because the variables were not completely significant from the beginning of the learning process of the optimum solution. Consequently, the Y intercept was not significant, and the cutoff was not calculated, however the average AUROC was derived as 0.6648. To derive the linear prediction function, the logistic regression model was derived again as shown in Eq. (4), where p is the probability that an accident results in death, excluding the age and construction type, which are not significant variables. Resultantly, Eq. (4) with a cutoff at −0.0710 was derived, and the average AUROC of this model was measured as 0.6326 (Fig. 3).

Table 8
Number of injuries and deaths by sex - data from [3].

Sex	Total	Man	Woman
1. Accidents	140,169	136,825	3344
Percentage	100%	97.61%	2.39%
1-1. Injuries	137,323	134,020	3303
Percentage	97.97%	95.61%	2.36%
1-2. Deaths	2846	2805	41
Percentage	2.03%	2.00%	0.03%
2. Mortality in the event of an accident[a]	2.03%	2.05%	1.23%

[a] Mortality in the event of an Accident = Deaths/Accidents * 100.

$$\log\left(\frac{p}{1-p}\right) = 5.273\times10^{-1}x_1 - 1.549x_2 - 1.847x_3 - 1.648x_4 - 1.530x_5 - 1.358x_6 - 9.185\times10^{-1}x_7 - 5.177\times10^{-1}x_8 - 5.061\times10^{-1}x_9 - 4.655\times10^{-1}x_{10} - 3.255\times10^{-1}x_{11} - 4.034\times10^{-1}x_{12} - 2.168x_{13} - 2.267x_{14} - 2.055x_{15} - 1.837x_{16} - 2.038x_{17} - 1.598x_{18} - 1.604x_{19} - 1.355x_{20} - 1.340x_{21} - 1.675x_{22} - 1.565x_{23} - 1.175x_{24} - 1.021x_{25} - 2.357\times10^{-1}x_{26} - 1.246\times10^{-1}x_{27} - 1.459\times10^{-1}x_{28} - 7.351\times10^{-2}x_{29} - 1.435\times10^{-1}x_{30} - 1.376\times10^{-1}x_{31} - 4.567\times10^{-2}x_{32} - 2.538\times10^{-2}x_{33} - 1.503\times10^{-1}x_{34} - 2.072\times10^{-1}x_{35} - 1.549\times10^{-1}x_{36} - 6.883\times10^{-2}x_{37} - 9.982\times10^{-2}x_{38} - 8.724\times10^{-2}x_{39} - 1.204\times10^{-1}x_{40} - 1.382\times10^{-1}x_{41} - 7.643\times10^{-2}x_{42} + 3.186 \quad (4)$$

X1 = sex (man = 1, woman = 0), X2 = scale of employer (4 or fewer = 1, others = 0), X3 = scale of employment (5–9 persons = 1, others = 0), X4 = scale of employment (10–15 persons = 1, others = 0), X5 = scale of employment (16–29 persons = 1, others = 0), X6 = scale of employment (30–49 persons = 1, others = 0), X7 = scale of employment (50–99 persons = 1, other = 0), X8 = scale of employment (100–199 persons = 1, other = 0), X9 = scale of employment (200–299 persons = 1,

Table 9
Logistic regression model (1/2).

Factor		Code	Estimate	Std. error	z value	Pr(> |z|)	Signif.
(Intercept)			−4.998.E+10	4.169.E+10	−1.199.E+00	2.306.E−01	
Sex	Man	G1	6.990.E−01	3.042.E−02	2.298.E+01	< 2E−−16	***
	Woman	G2	NA	NA	NA	NA	
Employer scale	−4	E1	−1.776.E+00	8.201.E−02	−2.166.E+01	< 2E−16	***
	5–9	E2	−1.994.E+00	8.229.E−02	−2.424.E+01	< 2E−16	***
	10–15	E3	−1.825.E+00	8.266.E−02	−2.208.E+01	< 2E−16	***
	16–29	E4	−1.685.E+00	8.247.E−02	−2.043.E+01	< 2E−16	***
	30–49	E5	−1.483.E+00	8.283.E−02	−1.790.E+01	< 2E−16	***
	50–99	E6	−1.013.E+00	8.283.E−02	−1.223.E+01	< 2E−16	***
	100–199	E7	−5.096.E−01	8.361.E−02	−6.095.E+00	1.090.E−09	***
	200–299	E8	−4.541.E−01	8.617.E−02	−5.269.E+00	1.370.E−07	***
	300–499	E9	−3.790.E−01	8.653.E−02	−4.380.E+00	1.190.E−05	***
	500–999	E10	−3.210.E−01	8.770.E−02	−3.660.E+00	2.520.E−04	***
	1000–1999	E11	−2.725.E−01	9.870.E−02	−2.760.E+00	5.773.E−03	**
	2000–	E12	NA	NA	NA	NA	
Age	−17	A1	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	18–24	A2	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	25–29	A3	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	30–34	A4	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	35–39	A5	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	40–44	A6	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	45–49	A7	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	50–54	A8	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	55–59	A9	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	60–64	A10	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	65–69	A11	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	70–74	A12	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	75–79	A13	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
	80–	A14	2.333.E+10	5.144.E+10	4.530.E−01	6.502.E−01	
Length of service	−1 m	L1	−1.728.E+00	1.557.E−01	−1.110.E+01	< 2E−16	***
	1–2 m	L2	−1.861.E+00	1.561.E−01	−1.192.E+01	< 2E−16	***
	2–3 m	L3	−1.660.E+00	1.566.E−01	−1.060.E+01	< 2E−16	***
	3–4 m	L4	−1.423.E+00	1.572.E−01	−9.049.E+00	< 2E−16	***
	4–5 m	L5	−1.599.E+00	1.583.E−01	−1.010.E+01	< 2E−16	***
	5–6 m	L6	−1.235.E+00	1.591.E−01	−7.761.E+00	8.430.E−15	***
	6 m–1 y	L7	−1.210.E+00	1.569.E−01	−7.714.E+00	1.220.E−14	***
	1–2 y	L8	−9.560.E−01	1.575.E−01	−6.070.E+00	1.280.E−09	***
	2–3 y	L9	−9.496.E−01	1.604.E−01	−5.920.E+00	3.220.E−09	***
	3–4 y	L10	−1.299.E+00	1.646.E−01	−7.894.E+00	2.940.E−15	***
	4–5 y	L11	−1.254.E+00	1.677.E−01	−7.478.E+00	7.530.E−14	***
	5–10 y	L12	−8.212.E−01	1.600.E−01	−5.131.E+00	2.880.E−07	***
	10–20 y	L13	−7.136.E−01	1.663.E−01	−4.292.E+00	1.770.E−05	***
	20 y–	L14	NA	NA	NA	NA	

Table 10
Logistic regression model (2/2).

Factor		Code	Estimate	Std. error	z value	Pr(> |z|)	Signif.
Construction type	Building construction	C1	2.666.E+10	5.115.E+10	5.210.E−01	6.023.E−01	
	Construction machinery management	C2	2.666.E+10	5.115.E+10	5.210.E−01	6.023.E−01	
	Hydroelectric power plant	C3	2.666.E+10	5.115.E+10	5.210.E−01	6.023.E−01	
	Railway and track	C4	2.666.E+10	5.115.E+10	5.210.E−01	6.023.E−01	
	Embankment	C5	2.666.E+10	5.115.E+10	5.210.E−01	6.023.E−01	
	Machinery	C6	2.666.E+10	5.115.E+10	5.210.E−01	6.023.E−01	
	Road	C7	2.666.E+10	5.115.E+10	5.210.E−01	6.023.E−01	
	Overpass and underground railway	C8	2.666.E+10	5.115.E+10	5.210.E−01	6.023.E−01	
	Other construction	C9	2.666.E+10	5.115.E+10	5.210.E−01	6.023.E−01	
Day of the week	Sunday	D1	NA	NA	NA	NA	
	Monday	D2	−2.459.E−01	1.711.E−02	−1.437.E+01	< 2E−16	***
	Tuesday	D3	−1.572.E−01	1.690.E−02	−9.297.E+00	< 2E−16	***
	Wednesday	D4	−1.517.E−01	1.701.E−02	−8.918.E+00	< 2E−16	***
	Thursday	D5	−7.911.E−02	1.679.E−02	−4.711.E+00	2.460.E−06	***
	Friday	D6	−1.615.E−01	1.698.E−02	−9.512.E+00	< 2E−16	***
	Saturday	D7	−1.386.E−01	1.721.E−02	−8.053.E+00	8.080.E−16	***
Month	1	M1	−5.153.E−02	2.362.E−02	−2.182.E+00	2.913.E−02	*
	2	M2	NA	NA	NA	NA	
	3	M3	−4.316.E−02	2.207.E−02	−1.955.E+00	5.055.E−02	.
	4	M4	−1.449.E−01	2.219.E−02	−6.530.E+00	6.600.E−11	***
	5	M5	−1.821.E−01	2.174.E−02	−8.376.E+00	< 2E−16	***
	6	M6	−1.404.E−01	2.164.E−02	−6.486.E+00	8.820.E−11	***
	7	M7	−7.213.E−02	2.164.E−02	−3.333.E+00	8.580.E−04	***
	8	M8	−1.063.E−01	2.153.E−02	−4.935.E+00	8.020.E−07	***
	9	M9	−5.135.E−02	2.187.E−02	−2.347.E+00	1.891.E−02	*
	10	M10	−1.442.E−01	2.140.E−02	−6.738.E+00	1.610.E−11	***
	11	M11	−1.557.E−01	2.182.E−02	−7.136.E+00	9.580.E−13	***
	12	M12	−5.850.E−02	2.214.E−02	−2.642.E+00	8.238.E−03	**

Signif.: '.' Pr(Z > |z|) < 0.1, '*' Pr(Z > |z|) < 0.05, '**' Pr(Z > |z|) < 0.01, '***' Pr(Z > |z|) < 0.001
AUROC = 0.6648.

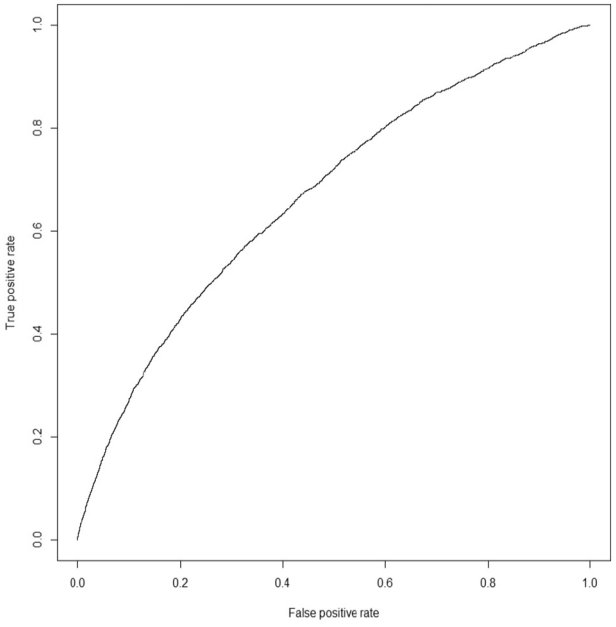

Fig. 3. AUROC of logistic regression model.

other = 0), X10 = scale of employment (300–499 persons = 1, other = 0), X11 = scale of employment (500–999 persons = 1, others = 0), X12 = scale of employment (1000–1999 persons = 1, others = 0), X13 = length of service (less than 1 month = 1, others = 0), X14 = length of service (1 month to 2 months = 1, others = 0), X15 = length of service (2 months to 3 months = 1, others = 0), X16 = length of service (3 months to 4 months = 1, others = 0), X17 = length of service (4 months to 5 months = 1, others = 0), X18 = length of service (5 months to 6 months = 1, others = 0), X19 = length of service (6 months to 1 year = 1, others = 0), X20 = length of service (1 year to 2 years = 1, others = 0), X21 = length of service (2 years to 3 years = 1, others = 0), X22 = length of service (3 years to 4 years = 1, others = 0), X23 = length of service (4 years to 5 years = 1, others = 0), X24 = length of service (5 years to 10 years = 1, others = 0), X25 = length of service (10 years to 20 years = 1, others = 0), X26 = day of the week (Monday = 1, others = 0), X27 = day of the week (Tuesday = 1, others = 0), X28 = day of the week (Wednesday = 1, others = 0), X29 = day of the week (Thursday = 1, others = 0), X30 = day of the week (Friday = 1, others = 0), X31 = day of the week (Saturday = 1, others = 0), X32 = month (January = 1, other = 0), X33 = month (March = 1, other = 0), X34 = month (April = 1, other = 0), X35 = month (May = 1, other = 0), X36 = month (June = 1, other = 0), X37 = month (July = 1, other = 0), X38 = month (August = 1, other = 0), X39 = month (September = 1, other = 0), X40 = month (October = 1, other = 0), X41 = month (November = 1, other = 0), X42 = month (December = 1, others = 0).

In the model, the coefficient varies widely depending on the employer scale and length of service, indicating that these factors are significant determinants of fatal accidents. Additionally, the greater the number of regular workers and the longer the years of service, the higher the risk of serious disasters. The larger the number of regular workers, the higher the mortality rate in the event of an accident. Hence, the absolute number of accidents in a large-scale employment site is low, however the death rate is nevertheless high. As such, legally arranging safety and health managers is important. Under the Korea Occupational Safety and Health Act [52], if a construction site has more than 300 construction workers per day on average, the construction company must hire a safety manager, and once the number of construction workers per day rises to 600 or more, the company must hire a health manager as well.

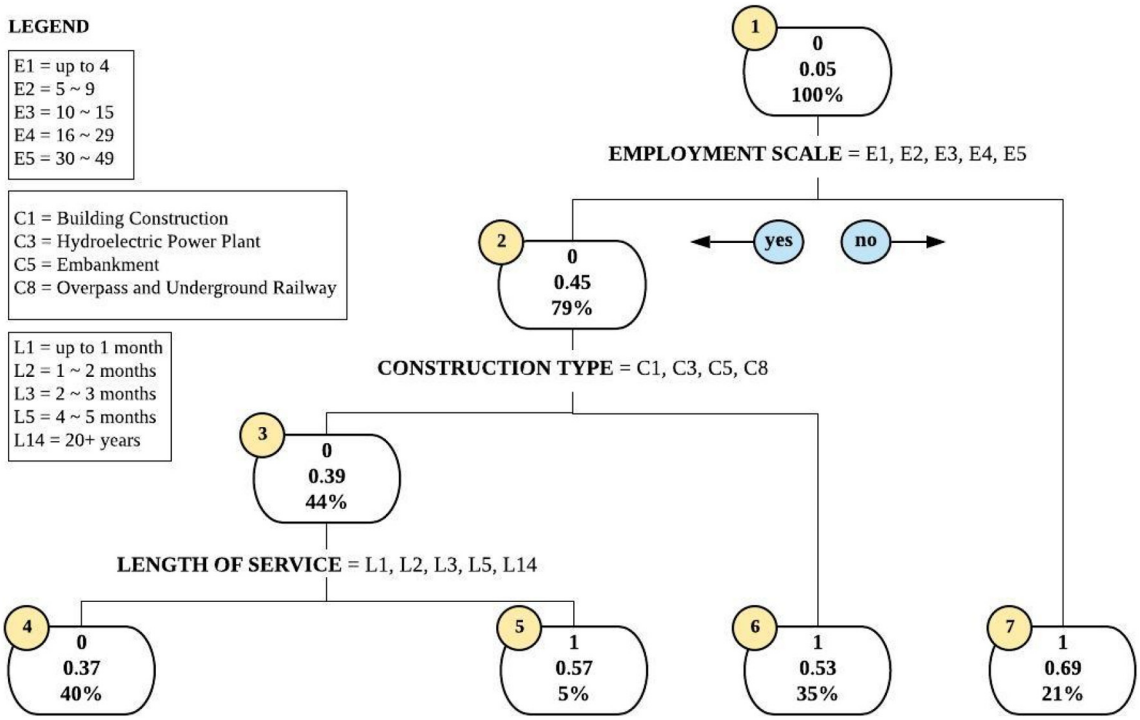

Fig. 4. Decision tree.

Table 11
Comparison of prediction performance and major items by algorithm.

Method/relevant research		AUROC	Factors							Remark
			Month	Employ size	Age	Day	Service length	Const. type	Sex	
Methods	Logistic regression model	0.6326	√	√		√	√		√	Significant factors
	Decision tree	0.6316		1			3	2		Ranking
	Random forest	0.9198	22.1	21.4	18.3	15.7	14.7	6.6	1.2	MDG (%)
	AdaBoost	0.6696	0.66	62.8	3.7	0.29	13.8	17.3	1.39	MDG (%)
Relevant research	Alizadeh et al. [21]				√		√			Significant factors
	Amiri et al. [22]				√					
	Chi et al. [35]			√	√				√	
	Chiang et al. [2]		√		√	√				
	Jiang et al. [39]				√				√	
	Villanueva & Garcia [36]				√		√		√	
	Zhang et al. [41]							√		

4.5. Decision tree

The decision tree is an analytical method that classifies or predicts data. It has the advantage of easily understanding and explaining complex problems in a simple hierarchical form. Decision trees are usually top-down algorithms, where in each step, a variable that divides a given dataset into subsets is selected by a splitting criterion, such as the Gini Index [53]. This method derives a tree with a set of hierarchically structured nodes and edges as its result. The derived nodes consist of properties, purity, and the proportion of instances. The nodes do not refer to correlated items, but to items used for the best prediction performance. Therefore, correlations among variables cannot be confirmed.

In this study, a decision tree analysis was performed on the preprocessed data, and the average AUROC was 0.6316. When the derived decision tree in Fig. 4 was analyzed, an employment scale of 50 or more (No E1, E2, E3, E4, E5) accounted for 21%, which means that 69% of deaths occurred in the event of an accident. Breaking down the data by employer scale, 79% of the total data was from organizations with fewer than 50 employees, in which case the death rate was 45%; 35% of the total data was from organizations with fewer than 50 employees, where the type of construction was not building/embankment, dam/hydroelectric power plant/highway, nor underground roadway construction (C1, C3, C5, C8), in which case the death rate was 53%; 44% of the total data was from organizations with fewer than 50 employees where the type of construction was one of building/embankment, dam/hydroelectric power plant/highway, or underground roadway construction, and here the death rate was 39%. Among the 44% of total data mentioned above, 40% of the accidents occurred in organizations of more than 50 people, with the type of construction being one of building/embankment and dam/hydroelectric power plant, with the employment length falling to the categories of less than 1 months, 1 months to 2 months, 2 months to 3 months, 4 months to 5 months, or more than 20 years (L1, L2, L3, L5, L14). In this case, the death rate was 37%, whereas in the other 5% of data, the accident death rate was 57%.

4.6. Random forest

The random forest is a popular classification technique in the field of machine learning. It is grounded in the research of Amit and Geman [54] and Ho [55], although the current concept of random forest was created by Breiman [56]. Random forest is an ensemble method, in

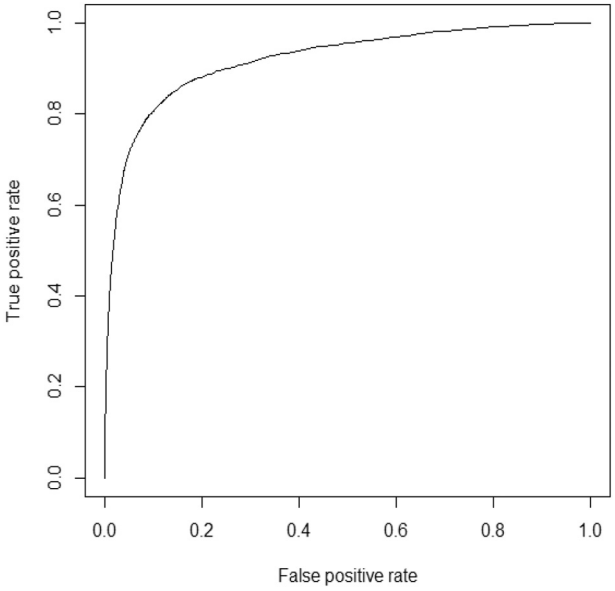

Fig. 5. AUROC of random forest.

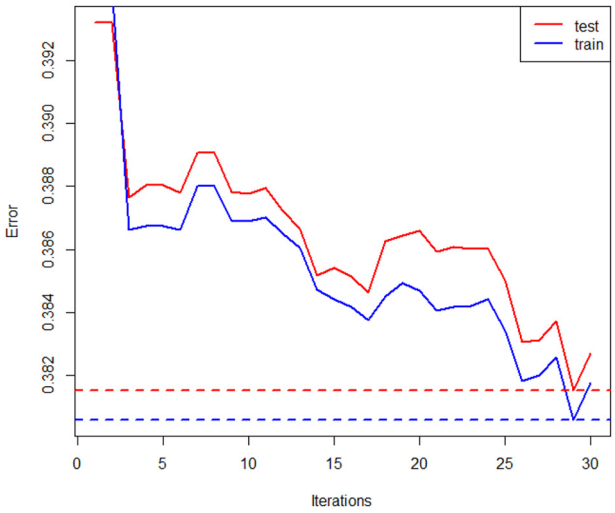

Fig. 6. AdaBoost error with respect to number of trees.

which several decision trees are randomly learned. This method trains the data to construct multiple decision trees (minimum = 300), and because the dataset learned for each decision tree is different, the constructed model and the predicted value are different from each other. When a new input data is given, Y = 0 or Y = 1 are voted by a number of decision trees. Random forest is often used to deal with categorical predictions [57]. In particular, the random forest has been reported to have the best predictive power in the prediction of unequal binomial classifications [58].

In the random forest algorithm, an index called the mean decrease Gini (MDG) index is used to measure the significance of independent variables. This indicator uses the average value from the whole tree to measure the impurity reduction amount of the selected variables as each branch extends out of the tree. Hence, a high MDG value for a specific variable implies that the impurity is low in binding the same categories. Table 11 shows the MDG indices of the model in this study. Although the upper and lower values of this indicator indicate the main items used for the best prediction performance, these do not imply a correlation between the variables and fatal accidents. The random forest analysis performed used the entire set of item variables to classify the data in this study, including the accident month, employer scale, age, day of the week, length of service, construction type, and sex (Table 11). The average AUROC of this model is 0.9198, which means the model can be evaluated as excellent (~0.9–1.0) (Fig. 5).

4.7. Adaptive boosting (AdaBoost)

AdaBoost is a boosting algorithm developed by Schapire et al. [59]. The key idea is to update the instance weights based on the results of previously learned weak classifiers. If the T-th weak classifier, which is the maximum number of learning times set in advance, is learned, then learning is no longer performed. Finally, a strong classifier with a linear combination of weak classifiers has better performance than a single classifier [59]. The weak classifier outputs the class variable when the independent variables X are input, and the strong classifier $H(X)$ that combines the results with the weighted majority voting method returns the final predicted class.

The performance of $H(X)$ is 0.6696 when AdaBoost was applied to the data with over 30 iterations ($T = 30$). Fig. 6 shows the learning and testing performances improving as the iterations progress. AdaBoost can provide the relative importance of independent variables by taking into account the Gini index gain. The mean decrease Gini values for the entire set of independent variables, which are the employer scale, construction type, years of service, age, sex, month, and day of the week were computed from the AdaBoost model used in this study, as shown in Table 11.

4.8. Results and discussion

In this study, the dataset was mechanically learned using four approaches, including logistic regression, decision tree, random forest, and AdaBoost analyses. AUROC and major variables for each algorithm are summarized in Table 11. Out of the four kinds of machine learnings performed, the random forest analysis presents the highest prediction rate applicable to the field, with a performance of 0.9198. This exhibits large differences with other machine learning methods in the AUROC evaluation. The factors selected in each method are different as well. The month, employment size, day, service length, and sex are selected in the logistic regression model, while employment size, construction type, and service length are selected in the decision tree method. Meanwhile, AdaBoost, which presents a relatively higher prediction power than the two methods mentioned previously, and the random forest that shows the most powerful prediction power in this study, include all the factors of the dataset used in this study. However, interestingly, the MDG values of the random forest and AdaBoost methods are different in each factor, which influences to produce the different prediction power between them.

In the random forest analysis, the month factor exhibits the highest MDG. This seems related to the fact that the MA is evidently higher during the winter season, as shown in Table 6. Compared with the Hong Kong area study [2], where most fatality accidents have occurred during the hot and humid summer season, the authors agree that seasonal factors are important factors, however the risk of death accidents varies with the climate, depending on the region or country.

The employment size factor, exhibiting the second highest MDG in the random forest method, was also selected in all the other methods used in this research while being represented as the most impactful factor in AdaBoost. Table 5 indicates that most accidents occur in less than 100 employment size projects. However, with the fatality accident aspect, higher MAs are observed in more than 100 employment size projects, as 8.10% of MA is observed in more than 2000 people-employed sites. This indicates that as the project size grows, efforts are made to reduce accidents through various safety devices and management, however when an accident occurs, it is more likely to be a fatal accident.

In addition, age, day, and service length exhibited remarkable MDGs in the random forest analysis. Age and service length are based

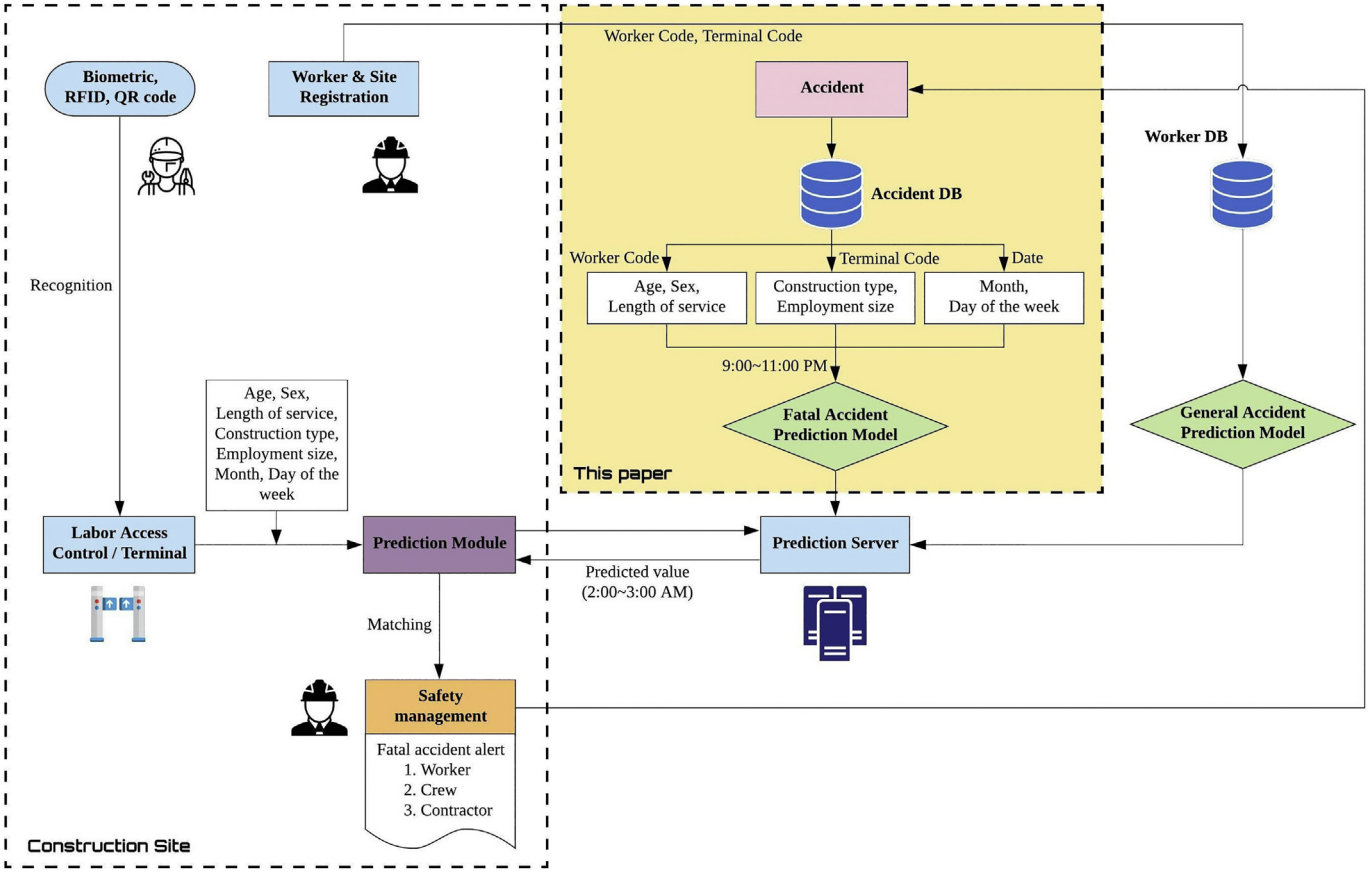

Fig. 7. Utilization of research results.

on individual characteristics. As shown in Table 2, most accidents including fatality are observed in the age group of 50 or older people, as derived from the existing studies [22,35,36].

High MA is also observed in the age of 24 or younger people, however it is believed that this could be the combined results due to their short service lengths, as observed in Jiang et al. [39] and Villanueva & Garcia [36]. As shown in Table 3, over 97% of accidents and over 95% of fatality accidents happened to those whose service lengths were less than 3 years. However, it is also notable that the highest MA (14.29%) is observed in those whose service length is more than 20 years as well.

The weekday factor related to the fatality accidents was mentioned by Chiang et al. [2]. In this study, the accidents rates are similar among most weekdays except Saturday and Sunday. In Korea, a lot of construction sites have executed construction on Saturday biweekly basis, while there have been few work days on Sundays, which tend to reduce works on Saturday according to the fifty-two working hours per week law in Korea. The MA of Monday, the starting day of a week, is relatively low at 1.86%, however the MA of Sunday is the highest at 2.33%. Based on this, the authors believe that the absence of managers or a loose mental state on holidays likely leads to fatality accidents.

Based on the results of this study, the random forest method with ensemble prediction based on various factors showed the most powerful predictive power. This indicates that there are limitations in predicting the risk of fatality accidents by only a few factors and that various factors must be applied in a highly complex way to increase prediction power. Comparing with the existing studies depicted in Table 11, this study shows academic value and uniqueness by developing the machine learning based prediction model, based on a 140,169 accidents and 2846 fatality accidents dataset.

4.9. Utilization plan

The use of machine learning models enables the identification of safety managers, workers, contractors, and work teams who are on a daily basis likely at high risk of serious accidents in the field. The use of the developed model will, therefore, contribute to lowering the number of annual accidents and deaths in the construction industry. The results of this study can be applied to an access control system or a manpower management system at a construction site, as shown in Fig. 7.

Moreover, access-control systems using electronic cards or biometrics are increasingly being adopted worldwide, not only for access control to a site, but also to secure social benefits for construction workers. The models presented in this study could, therefore, contribute more proactively to the prevention of accidents through continuous data acquisition and learning by integrating with such manpower access-control systems.

Initially, data can be collected from government agencies' databases to create a fatal accident prediction model for a construction site. When a construction worker checks into a construction site through an access control system, the prediction model can be utilized to provide the safety manager with information on any critical or potentially risky group at the individual, work team, and company levels. When the system is in operation, accident data from all regions are collected daily at the government level. The data are then included in a dataset that is used to create a new model at the end of each day. Thereafter, the process of classifying the information is registered as a new model, and transmitting the information to the field terminal, it can be repeated. Through this process, construction workers' and accident data can be gathered every day into a large dataset that can provide a basis for the development of a future disaster prediction model.

5. Conclusions

This study analyzed national data of about 140,169 industrial accident victims in the construction industry from 2011 to 2016 (comprising data on 137,323 injuries and 2846 deaths), to create a predictive model to classify serious accident risk groups for the purpose of increasing the efficiency of on-site safety management. The study is considered significant, as the data were collected over a period of 6 years from official national databases, ensuring consistent data acquisition over a number of variables, and high reliability. Moreover, the study applies the power of machine learning techniques to the construction field and suggests a method to apply the developed system in practice on construction sites.

The use of historical data imposes the limitation that the available item variables can only be drawn from what has been recorded. The introduction of new variables, such as textual description of the accident situation and involved personal information, was not available in this study because of the privacy information law in Korea. This information could be quite valuable as a future study if the relevant data becomes available in the future.

The methodology employed in this study involved preliminary exploratory data analysis prior to the development of the fatal accident prediction model. Subsequently, logistic regression analysis, decision tree analysis, random forest analysis, and AdaBoost analysis were performed to identify workers with a high likelihood of mortality by learning injury-and-death data associated with previous fatal accidents, in order to predict further fatal accidents. Finally, the performances of each model were compared, and a utilization method was suggested. The performance of the random forest analysis, reported to have the best predictive power in predicting unbalanced binomial classifications, was 91.98%, showing that random forest was indeed the best performing model, indicating that month, employment size, age, day, and service length are important factors to predict the likelihood of a fatality accident. This study showed the creativity and uniqueness by demonstrating that the machine learning based prediction model is feasible in construction safety management, and that the model comprehensively incorporates various factors that were separately identified in the existing studies mentioned above.

Despite allowing the machine to randomly classify learning and verification data without affecting the data category, the authors recognize that different countries may have different results depending on the category of datasets available, as well as their data contents. It is expected that the analysis of similarities and differences through the analysis and comparison of accident data in each country will be a very meaningful study in the future.

This study showed that that the difference between injury and death can be featured by the independent variables in our data with a predictive power of 92%. In the space of the same variables, accident and non-accident data must be more separable, because the value difference on these features must be bigger. This would arise with a better predictive power as well. From this point of view, this study acknowledges that there is a limit to the developed model to predict fatality accidents based on accident-based data. However, this study is expected to aid the government to recognize the necessity of safety relevant information at a wider level, where a prediction model could provide the likelihood of safety accidents based on accident and non-accident data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was financially supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2018R1A2B6003564). A part of this work is also supported by a grant (19AUDP B127891 03) from the Architecture & Urban Development Research Program funded by the Ministry of Land, Infrastructure and Transport of the Korean government.

References

[1] International Labour Organization, Safety and health in the construction sector – overcoming the challenges, https://www.ilo.org/empent/Eventsandmeetings/WCMS_310993/lang–en/index.htm (Sep. 21, 2018).
[2] Y.-H. Chiang, F.K.-W. Wong, S. Liang, Fatal construction accidents in Hong Kong, J. Constr. Eng. Manag. 144 (3) (2017) 04017121, , https://doi.org/10.1061/(ASCE)CO.1943-7862.0001433.
[3] Ministry of Employment and Labor, Analysis of industrial accidents in 2011–2017, https://www.kosha.or.kr/english/index.do (Sep. 21, 2018).
[4] Occupational Safety and Health Administration, Commonly used statistics, https://www.osha.gov/oshstats/commonstats.html (June 5, 2017).
[5] Japan Industrial Safety and Health Association, OSH statistics in Japan, industrial accident in 2017, 〈https://www.jisha.or.jp/english/statistics/ (Sep. 21, 2018).
[6] Health and Safety Executive, Workplace fatal injuries in Great Britain 2018, https://www.hse.gov.uk/statistics/pdf/fatalinjuries.pdf (Sep. 21, 2018).
[7] Y.M. Goh, C.U. Ubeynarayana, Construction accident narrative classification: an evaluation of text mining techniques, Accid. Anal. Prev. 108 (2017) 122–130, https://doi.org/10.1016/j.aap.2017.08.026.
[8] H.W. Heinrich, D.C. Petersen, N.R. Roos, S. Hazlett, Industrial Accident Prevention: A Safety Management Approach, fifth ed., McGraw-Hill, New York, 1980 (ISBN-10:0070280618).
[9] Korea Construction Engineers Association, Construction engineer statistics 2017 - statistic status by qualification of technical grade, https://homenet.kocea.or.kr.
[10] Statistics Korea, Monthly report on economic activities 2017, https://kostat.go.kr (Sep. 21, 2018).
[11] A.J.P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Application of machine learning to construction injury prediction, Autom. Constr. 69 (2016) 102–114, https://doi.org/10.1016/j.autcon.2016.05.016.
[12] X. Yan, E. Radwan, M. Abdel-Aty, Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model, Accid. Anal. Prev. 37 (6) (2005) 983–995, https://doi.org/10.1016/j.aap.2005.05.001.
[13] G. Yannis, E. Papadimitriou, E. Dupont, H. Martensen, Estimation of fatality and injury risk by means of in-depth fatal accident investigation data, Traffic Injury Prevention 11 (5) (2010) 492–502, https://doi.org/10.1080/15389588.2010.492536.
[14] A. Gregoriades, K.C. Mouskos, Black spots identification through a Bayesian networks quantification of accident risk index, Transportation Research Part C: Emerging Technologies 28 (2013) 28–43, https://doi.org/10.1016/j.trc.2012.12.008.
[15] M. Karacasu, B. Ergul, A.A. Yavuz, Estimating the causes of traffic accidents using logistic regression and discriminant analysis, Int. J. Inj. Control Saf. Promot. 21 (4) (2014) 305–313, https://doi.org/10.1080/17457300.2013.815632.
[16] T. Usman, L. Fu, L.F. Miranda-Moreno, Injury severity analysis: comparison of multilevel logistic regression models and effects of collision data aggregation, Journal of Modern Transportation 24 (1) (2016) 73–87, https://doi.org/10.1007/s40534-016-0096-4.
[17] T. Nishimoto, K. Mukaigawa, S. Tominaga, N. Lubbe, T. Kiuchi, T. Motomura, et al., Serious injury prediction algorithm based on large-scale data and under-triage control, Accid. Anal. Prev. 98 (2017) 266–276, https://doi.org/10.1016/j.aap.2016.09.028.
[18] G. Li, S.P. Baker, J.G. Grabowski, Y. Qiang, M.L. McCarthy, G.W. Rebok, Age, flight experience, and risk of crash involvement in a cohort of professional pilots, Am. J. Epidemiol. 157 (10) (2003) 874–880, https://doi.org/10.1093/aje/kwg071.
[19] M. Bazargan, V.S. Guzhva, Impact of gender, age and experience of pilots on general aviation accidents, Accid. Anal. Prev. 43 (3) (2011) 962–970, https://doi.org/10.1016/j.aap.2010.11.023.
[20] S.J. Yeoum, Y.H. Lee, A study on prediction modeling of Korea millitary aircraft accident occurrence, International Journal of Industrial Engineering: Theory, Applications, and Practice 20 (9–10) (2013) 562–573 https://journals.sfu.ca/ijietap/index.php/ijie/article/view/1138 (Aug. 31, 2018).
[21] S.S. Alizadeh, S.B. Mortazavi, M.M. Sepehri, Assessment of accident severity in the construction industry using the Bayesian theorem, Int. J. Occup. Saf. Ergon. 21 (4) (2015) 551–557, https://doi.org/10.1080/10803548.2015.1095546.
[22] M. Amiri, A. Ardeshir, M.H.F. Zarandi, E. Soltanaghaei, Pattern extraction for high-risk accidents in the construction industry: a data-mining approach, Int. J. Inj. Control Saf. Promot. 23 (3) (2016) 264–276, https://doi.org/10.1080/17457300.2015.1032979.
[23] C. Cho, K. Kim, J.W. Park, Y.K. Cho, Data-driven monitoring system for preventing the collapse of scaffolding structures, J. Constr. Eng. Manag. 144 (8) (2018) 04018077, , https://doi.org/10.1061/(asce)co.1943-7862.0001535.
[24] G. Greenleaf, Global data privacy laws: 89 countries, and accelerating, Privacy Laws & Business International Report, Queen Mary School of Law Legal Studies Research Paper No. 98/2012, Issue 115, Special Supplement, February 2012 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2000034 (June 2, 2019).
[25] Y. Bengio, S. Bengio, Modeling high-dimensional discrete data with multi-layer neural networks, Proceedings of Advances in Neural Information Processing

Systems, 2000, pp. 400–406 (ISBN:0262194503;978-026219450-1).
[26] C. Chuntian, K.W. Chau, Three-person multi-objective conflict decision in reservoir flood control, Eur. J. Oper. Res. 142 (3) (2002) 625–631, https://doi.org/10.1016/S0377-2217(01)00319-8.
[27] C.L. Wu, K.W. Chau, Rainfall-runoff modeling using artificial neural network coupled with singular spectrum analysis, J. Hydrol. 399 (3–4) (2011) 394–409, https://doi.org/10.1016/j.jhydrol.2011.01.017.
[28] R. Moazenzadeh, B. Mohammadi, S. Shamshirband, K.W. Chau, Coupling a firefly algorithm with support vector regression to predict evaporation in northern Iran, Engineering Applications of Computational Fluid Mechanics 12 (1) (2018) 584–597, https://doi.org/10.1080/19942060.2018.1482476.
[29] S. Samadianfard, A. Majnooni-Heris, S.N. Qasem, O. Kisi, S. Shamshirband, K.W. Chau, Daily global solar radiation modeling using data driven techniques and empirical equations in a semi-arid climate, Engineering Applications of Computational Fluid Mechanics 13 (1) (2019) 142–157, https://doi.org/10.1080/19942060.2018.1560364.
[30] A. Baghban, A. Jalali, M. Shafiee, M.H. Ahmadi, K.W. Chau, Developing an ANFIS based swarm concept model for estimating relative viscosity of nanofluids, Engineering Applications of Computational Fluid Mechanics 13 (1) (2019) 26–39, https://doi.org/10.1080/19942060.2018.1542345.
[31] Z.M. Yaseen, S.O. Sulaiman, R.C. Deo, K.W. Chau, An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction, J. Hydrol. 569 (2019) 387–408, https://doi.org/10.1016/j.jhydrol.2018.11.069.
[32] S. Gerassis, J.E. Martín, J.T. García, A. Saavedra, Bayesian decision tool for the analysis of occupational accidents in the construction of embankments, J. Constr. Eng. Manag. 143 (2) (2017) 04016093, , https://doi.org/10.1061/(ASCE)CO.1943-7862.0001225.
[33] J. Neeleman, J. Ormel, R.V. Bijl, The distribution of psychiatric and somatic ill health: associations with personality and socioeconomic status, Psychosom. Med. 63 (2) (2001) 239–247 (Electronic).
[34] C.-W. Cheng, S.-S. Leu, C.-C. Lin, C. Fan, Characteristic analysis of occupational accidents at small construction enterprises, Saf. Sci. 48 (6) (2010) 698–707, https://doi.org/10.1016/j.ssci.2010.02.001.
[35] C.F. Chi, T.C. Chang, H.I. Ting, Accident patterns and prevention measures for fatal occupational falls in the construction industry, Appl. Ergon. 36 (4) (2005) 391–400, https://doi.org/10.1016/j.apergo.2004.09.011.
[36] V. Villanueva, A.M. Garcia, Individual and occupational factors related to fatal occupational injuries: a case-control study, Accid. Anal. Prev., 43(1), 123–127. doi:https://doi.org/10.1016/j.aap.2010.08.001.
[37] A. Garg, Ergonomics and the older worker: an overview, Exp. Aging Res. 17 (3) (1991) 143–155, https://doi.org/10.1080/03610739108253894.
[38] D.R. Davies, G. Matthews, C.S.K. Wong, Ageing and work, Int. Rev. Ind. Organ. Psychol. 6 (1991) 149–211 https://psycnet.apa.org/record/1991-98004-005 (Aug. 03, 2019).
[39] G. Jiang, B.C.K. Choi, D. Wang, H. Zhang, W. Zheng, T. Wu, et al., Leading causes of death from injury and poisoning by age, sex and urban/rural areas in Tianjin, China 1999–2006, Injury 42 (5) (2011) 501–506, https://doi.org/10.1016/j.injury.2009.10.050.
[40] D.P. Loomis, D.B. Richardson, S.H. Wolf, C.W. Runyan, J.D. Butts, Fatal occupational injuries in a southern state, Am. J. Epidemiol. 145 (12) (1997) 1089–1099, https://doi.org/10.1093/oxfordjournals.aje.a009071.
[41] L. Zhang, Q. Liu, X. Wu, M.J. Skibniewski, Perceiving interactions on construction safety behaviors: workers' perspective, J. Manag. Eng. 32 (5) (2016) 04016012, , https://doi.org/10.1061/(ASCE)ME.1943-5479.0000454.
[42] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recogn. 30 (7) (1997) 1145–1159, https://doi.org/10.1016/S0031-3203(96)00142-2.
[43] Construction Association of Korea, Annual domestic construction orders, http://www.cak.or.kr/stat/statisticsMain.do?menuId=7, (2018) (Sep. 21, 2018).
[44] G.E. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD Explorations Newsletter 6 (1) (2004) 20–29, https://doi.org/10.1145/1007730.1007735.
[45] N.V. Chawla, N. Japkowicz, A. Kotcz, Editorial: special issue on learning from imbalanced data sets, ACM SIGKDD Explorations Newsletter 6 (1) (2004) 1–6, https://doi.org/10.1145/1007730.1007733.
[46] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, https://doi.org/10.1613/jair.953.
[47] Q. Gu, L. Zhu, Z. Cai, Evaluation measures of the classification performance of imbalanced data sets, International Symposium on Intelligence Computation and Applications, ISICA 2009, Communications in Computer and Information Science 51 (2009) 461–471, https://doi.org/10.1007/978-3-642-04962-0_53.
[48] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874, https://doi.org/10.1016/j.patrec.2005.10.010.
[49] D.R. Cox, The regression analysis of binary sequences, J. R. Stat. Soc. Ser. B Methodol. 20 (2) (1958) 215–232, https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.
[50] J. Choi, B. Gu, S. Chin, Development of the prediction model of workers with fatal accident at construction site using machine learning, Proceeding of the Creative Construction Conference, 2018, pp. 870–875, , https://doi.org/10.3311/CCC2018-113.
[51] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, International Joint Conference on Artificial Intelligence, 14(2) 1995, pp. 1137–1145 https://www.researchgate.net/publication/2352264 (Sep. 21, 2018).
[52] Ministry of Government Legislation, Occupational Safety and Health Act, Act No. 3532, https://www.moleg.go.kr/english/korLawEng?pstSeq=57986 (Sep. 30, 2018).
[53] L. Rokach, O. Maimon, Top-down induction of decision trees classifiers-a survey, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 35 (4) (2005) 476–487, https://doi.org/10.1109/TSMCC.2004.843247.
[54] Y. Amit, D. Geman, Shape quantization and recognition with randomized trees, Neural Comput. 9 (7) (1997) 1545–1588, https://doi.org/10.1162/neco.1997.9.7.1545.
[55] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844, https://doi.org/10.1109/34.709601.
[56] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[57] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (3) (2002) 18–22 https://www.researchgate.net/profile/Andy_Liaw/publication/228451484_Classification_and_Regression_by_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression-by-RandomForest.pdf (Sep 28, 2018).
[58] I. Brown, C. Mues, An experimental comparison of classification algorithms for imbalanced credit scoring data sets, Expert Syst. Appl. 39 (3) (2012) 3446–3453, https://doi.org/10.1016/j.eswa.2011.09.033.
[59] R.E. Schapire, Y. Freund, P. Bartlett, W.S. Lee, Boosting the margin: a new explanation for the effectiveness of voting methods, Ann. Stat. 26 (5) (1998) 1651–1686 https://projecteuclid.org/euclid.aos/1024691352.