Journal of Construction Engineering, Management & Innovation
2023 6(1):01-15
DOI 10.31462/jcemi.2023.01001015

RESEARCH ARTICLE

Determining the Short-term Susceptibility of Construction Workers to Occupational Accidents Using Stochastic Gradient Boosting

Kerim Koc

Yildiz Technical University, Faculty of Civil Engineering, Civil Engineering Department, İstanbul, Türkiye

Article History

Received 16 December 2022
Accepted 20 February 2023

Keywords

Construction safety management
Explainable artificial intelligence
Explainable machine learning
Stochastic gradient boosting
Worker vulnerability

Abstract

Employees working on construction projects have a higher risk of experiencing occupational accidents compared to workers in other sectors. Newly employed workers might face an occupational accident shortly after they start working due to not being able to notice risky environmental conditions. Despite existing studies in construction safety literature focusing on several output variables such as fatality, accident type or accident severity predictions, no studies have examined the short-term susceptibility of construction workers to occupational injuries. This study aims to develop a model to predict construction workers' susceptibility to short-term occupational accidents using interpretable machine learning (ML) methods. Hence, the primary research objective is to identify construction workers who have high probability of experiencing an occupational accident shortly after their employment. In this respect, a national dataset of occupational accidents encountered in the construction industry in Turkey was collected and subjected to various pre-processing elements (data cleaning, data scaling, and data resampling) to prepare the data for prediction. At the processing step, Stochastic Gradient Boosting (SGB) algorithm was applied for the classification purpose. In the next step, Shapley Additive Explanations (SHAP) was used as an interpretable artificial intelligence algorithm to explain how, to what extent, and in which direction the input variables affect the prediction scheme, which is another distinguishing feature of the present study compared to past studies in the subject matter. Results show that the proposed SGB model is a powerful detector for the classification problem and salary of workers, past accident in the company, and number of workers in the company were the most influencing factors. Overall, this study contributes to practice by improving the safety conditions of the newly employed workers as well as minimizing their accident probability through intensified safety training. Given that contemporary safety management applications demand a new set of data-driven inputs, proposed model is expected to help industry professionals and safety managers apply more robust safety risk mitigation and/or prevention measures.

1. Introduction

Construction projects involve highly complex, dynamic, and intertwined business practices, and are carried out under severe environmental conditions [1, 2]. This makes the construction projects extremely dangerous by their nature [3]. As

Correspondence Kerim Koc kerimkoc@yildiz.edu.tr

eISSN 2630-5771 © 2023 Authors. Publishing services by Golden Light Publishing®.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

a result, a significant number of occupational accidents occur at construction sites [4–6]. Due to challenging environmental and dynamic conditions on construction sites and pressures on construction workers to achieve time, cost and quality objectives, occupational accidents are inevitable in the industry [7]. Although the number of fatality and injury cases encountered in the sector have been reduced by the improvements in the occupational health and safety (OHS) protocols over the years [8, 9], construction industry is still the most dangerous sector in many countries [10]. This indicates that the ameliorations are not fully sufficient to prevent occupational accidents, and the conditions that influence the occurrence of occupational accidents should be addressed diligently by prediction models in advance [11].

Past studies highlight that newly employed workers have a high risk of having an occupational accident [10] or severe accident [12] in the first days of work compared to experienced workers in that company. At this point, it is of principal importance to identify which employees have an occupational accident risk in the short-term period after starting the work in advance. This will enable practitioners to both improve the necessary and relevant OHS training and control the conditions/environment within the framework of OHS protocols. In this context, construction safety management literature draws attention to artificial intelligence (AI), or more specifically machine learning (ML) applications to predict accident outcomes and correspondingly apply necessary prevention or mitigation measures [10, 12–15]. By using data-driven ML methods, various output variables such as whether an occupational accident will occur [16], whether it will result in fatality or injury [17], whether unsafe behaviour will be observed [18], or in which time period it will occur more frequently, can be predicted.

The main objective of the present research is to develop a prediction model to detect workers who have high probability of experiencing an occupational accident on construction sites shortly after their employment. In this respect, the primary focus of investigation is to determine which construction workers will experience an occupational accident in a short-term period after they are employed and identify the factors that affect this classification model. Although various output variables such as fatal/non-fatal [10, 19], severity of accidents [9, 12], occurrence of accidents [16], type of accidents [14, 20], and body part injured [14] have been investigated with AI-oriented prediction models in the construction safety management literature, no previous study has been found on predicting whether workers will face an occupational accident in the short term or not, which constitute the conceptual research gap of this study. Past research also highlights the necessity of examining the most dangerous time periods that construction workers might face an occupational accident [17]. In addition, although the importance of input variables has been determined by using classical feature importance algorithms such as permutation importance [20] or Gini importance [21], the recently developed SHAP (Shapley Additive Explanations) method, which offers a visual output of both the importance weight of features and the direction of their effects [22], has not received significant attention from research society. However, a recent study highlights the necessity of explainable prediction models in the construction safety management field to provide confidence to the potential users and help safety managers anatomize the precursors of accidents [19]. The model developed with such an approach is essential as it not only predicts which employees will have a risk of experiencing occupational accident in the short run, but also shows why they will have such an accident risk. Given that this study proposes a data-driven and interpretable ML model that SHAP method is an auxiliary of it, the model can be easily used by industry professionals and integrated with other data-driven applications or technologies.

2. Literature review

In the current study, a ML-based prediction model was proposed to detect whether construction workers will face an occupational accident in a short-time period. In general, ML phenomenon can

be defined as a prediction tool that implements various algorithms and statistical models, which learns from historical data to predict new observations based on patterns and inference [23]. It involves two main processes: defining a set of features and training a model to estimate potential outcomes from the predefined features [24]. It is based on learning the data insights of the input variables to provide a basis for smarter decisions [23]. ML-based prediction studies include two major output types, regression and classification. In the present research, a binary classification problem was addressed.

There is a growing body of literature to predict different output variables in the construction safety research field, such as accident severity, accident type, body part affected, and injury type. Many researchers developed a binary classification model to predict whether an accident will result in fatality or injury. In this vein, Choi et al. [10] compared the performance of decision tree (DT), logistic regression (LR), Adaboost and random forest (RF) methods based on national dataset of Korea. The researchers found that the RF model achieved the highest prediction accuracy and month was the most influential attribute compared to employment size, age, day, service length of the worker, construction type, and gender. Similarly, Koc et al. [17] compared the performance of a distance-based, probability-based, tree-based, and neural network-based ML methods based on national dataset of Turkey and found that the RF method outperformed its counterparts. Mistikoglu et al. [21] suggested a DT classifier to predict the same output label (i.e., fatal/non-fatal) by using the national dataset of United States. As a different binary classification scheme, Ayhan and Tokdemir [9] classified accidents as low severe and high severe and developed an artificial neural network (ANN) model to predict the severity of occupational accidents.

Type of accident was another commonly investigated output domain in the construction safety management literature. For instance, Zhang et al. [25] adopted five ML methods such as support vector machine (SVM), LR, k-nearest neighbour (KNN), DT, and Naïve Bayes (NB) after applying text mining and natural language processing (NLP) subjected to the dataset obtained from the US to classify accident types (e.g., struck by falling object, collapse of object, exposure to extreme temperature, falls, etc.). Likewise, Goh and Ubeynarayana [11] also used narratives to perform NLP analysis and compared the performance of several ML algorithms (i.e., SVM, LR, RF, KNN, DT, NB) and found that the classifications were more powerful for the accident type of electrocution. Similarly, Kang and Ryu [26] captured the national data of Korea and performed a multi-label classification using RF algorithm for the type of accident and found that assailing materials was the most influential attribute in the prediction.

Some researchers considered multi outputs related to occupational accident outcomes. For instance, Baker et al. [24] compared the performances of DT, RF, SVM and XGBoost for classifying accident severity, accident type, injury type, and affected body part. Ajayi et al. [16] adopted deep neural network approach for the prediction of the occurrence of accidents, injured body part, lost time, and damage due to accidents. Lee et al. [27] considered seven output variables such as year of the accident, type of work, assailing materials, severity of accident, body part affected, accident cause, and accident type, and each was predicted by taking other six as input variables. Comprehensive literature review shows that there is a variety of research in the subject matter that adopted ML methods. Still, no previous study has leveraged a ML model to predict the short-term susceptibility of construction workers, which is the main motivation to perform the present research.

3. Research methodology

This study followed nine consecutive steps to predict short-term susceptibility of construction workers to occupational accidents (Fig. 1). In the first step, the required data comprising national accident records between 2012 and 2020 was collected from Social Security Institution.

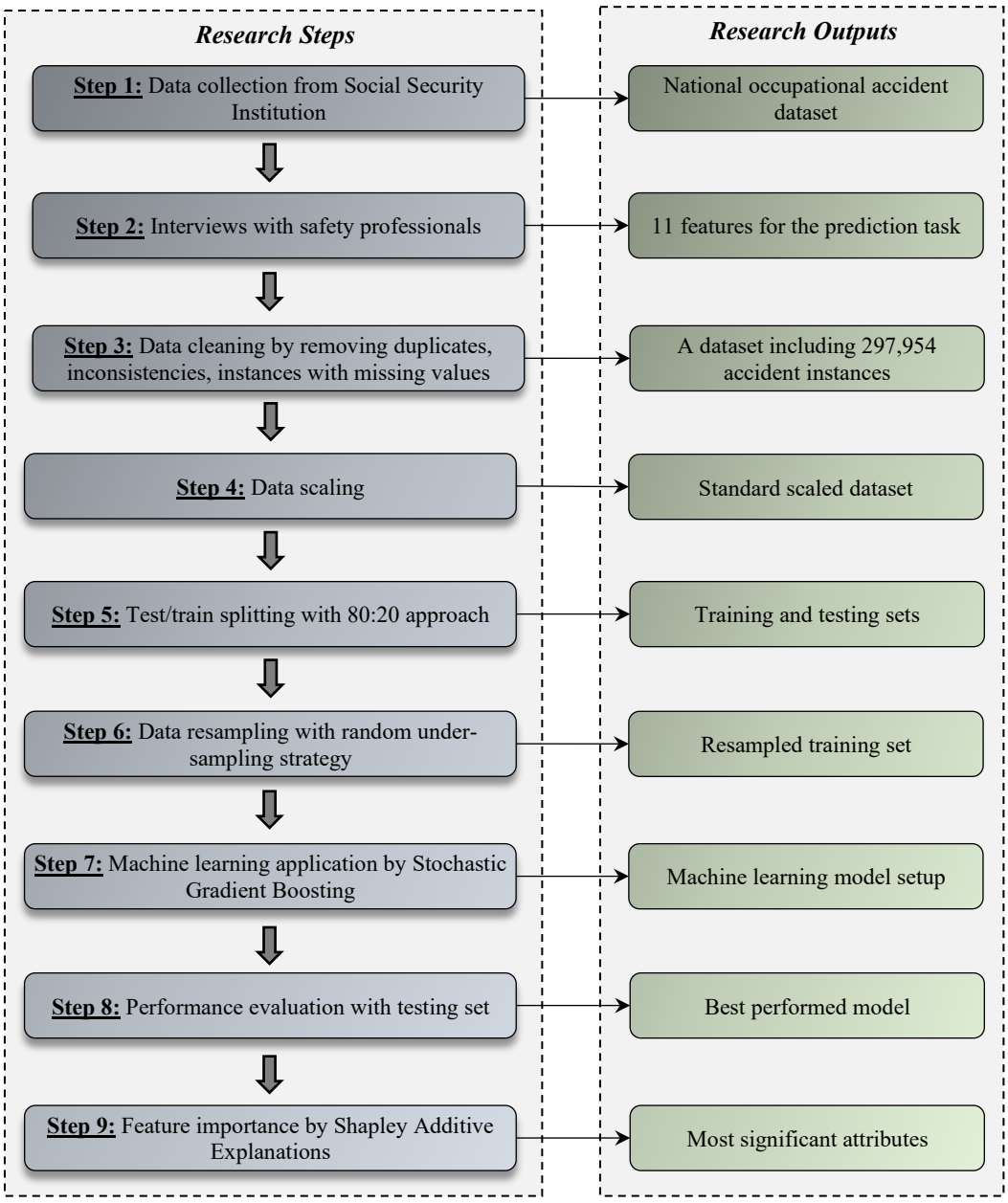


Fig. 1. Research steps

In the next step, three semi-structured interviews were conducted with industry professionals to identify important features In step 3, data cleaning was conducted by means of removing duplicates, inconsistencies, and instances that contain missing data points. Then, data scaling was employed in the 4th step to represent numerical features more explicitly. After dividing the dataset into two major groups as training and testing sets in step 5, imbalanced class distribution was handled through the random under-sampling (RUS) strategy. With the pre-processed dataset, ML implementation was employed through Stochastic Gradient Boosting (SGB) algorithm, and performance of the model was assessed using several performance evaluation indices. At the final step of the present research, Shapley Additive Explanations (SHAP) analysis was performed to explain the contributions of

attributes to the prediction rationale. All these steps were run in Python 3.8 with Spyder code execution environment. The following sub-sections explain the procedures applied in this study in more detail.

3.1. Dataset used in this study

The collected dataset initially contained 395,777 accident records took place in the construction industry in Turkey between 2012 and 2020 (up to October). On the one hand, the time (i.e., duration) elapsed between the starting day of the employee and the day that worker experienced an occupational accident was considered as the output variable of this study. On the other hand, the dataset included many attributes such as date of the accident, age of the workers, type of accident, type of injury, body part exposed to accidents, etc. However, some of these features were not precursors of the output variables while some of them cannot be used as input variables as they are known only after the occurrence of accidents. Therefore, three semi-structured interviews were performed with safety professionals to determine input variables used in the current research. As a result, 11 attributes (8 numerical and 3 binary) were considered suitable to this study by the interviewees. During the interviews, respondents were also asked to determine the time frame for considering accidents as short-term. In this context, 1-day ahead, 7-day ahead, 30-day ahead, and 90-day ahead were the alternatives, and all the respondents stated that considering 7-day ahead can clearly represent the short-term susceptibility of workers to injuries. Therefore, the output domain was transformed into a binary format, which is whether a worker experiences an occupational accident within 7 days of employment or not (short-term accident vs no short-term accident). With reference to 11 attributes and a binary output variable, data cleaning was performed through removing inconsistent data labels, instances that contain missing data points, duplicated instances, etc. As a result, 297,954 accident records (38,021 short-term accidents and 259,933 no short-term accidents) were attained. Statistical features of the numerical attributes are shown in Table 1, and Table 2 illustrates the number of labels in binary attributes (i.e., past accident (worker)), night shift, and gender) by target variable. Figure 2 also presents the number of output labels. Accordingly, nearly 13% of the accidents in the construction industry in Turkey occurred within 7-day period after the employees started working.

"Employed day of the month", "employed day of the week", and "employed month" attributes were captured from the employed date of the victims, which ranges between 1 and 31, 1 (Monday) and 7 (Sunday), and 1 (January) and 12 (December), respectively. "Age", "education level", "salary", and "gender" represents the social and demographic characteristics of the workers. While "past accident (company)" is the number of past accidents in the employed company, "past accident (worker)" indicates whether the worker had an accident experience before the employment or not. Another binary attribute, i.e., "night shift" explains whether the worker will be working at the night shift or not. Past studies that developed ML models based on national datasets to predict different output variables [10, 12] were also checked to ensure the adequacy of the included input variables

Table 1. Statistical features of the adopted numerical attributes

Numerical attributes	Min	Mean	Max	Median	Skewness	Std. Dev.
Employed day of the month	1	14.1327	31	14	0.13624	9.00509
Employed day of the week	1	4.1797	7	4	0.08638	1.67529
Employed month	1	6.3827	12	6	0.02609	3.34036
Age	13	33.3527	79	31	0.54861	11.11374
Education level	0	1.4498	4	1	0.47742	1.06655
Past accident (company)	1	36.1601	1804	3	8.74518	150.35118
Salary (*Turkish Lira*)	9.75	79.3067	5412.37	67.65	7.65043	48.27981
Number of workers	1	261.1080	9980	66	8.24962	845.97101

Table 2. Number of binary feature instances by target variable

Binary features	Binary conditions	No short-term accident (0)	Short term accident (1)	Sum
Past accident (worker)	No (0)	210,226	33,399	243,625
	Yes (1)	49,707	4,622	54,329
Night shift	No (0)	232,361	35,644	268,005
	Yes (1)	27,572	2,377	29,949
Gender	Male (0)	258,295	37,901	296,196
	Female (1)	1,638	120	1758

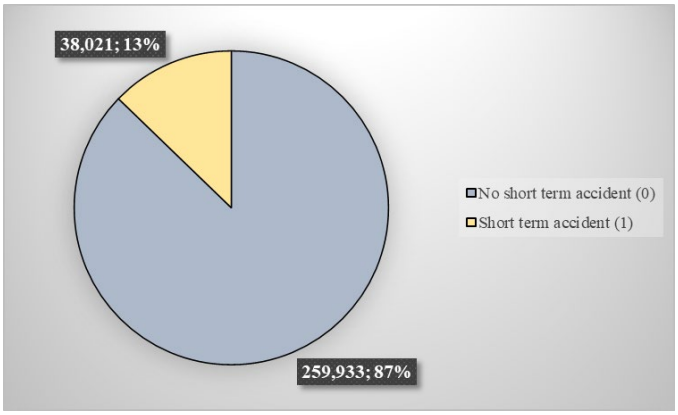


Fig. 2. Distribution of output labels

3.2. Data preprocessing

This study applied several preprocessing steps to prepare the data for the classification task. First, since the attributes used in this study differs from each other in terms of scales, units, and magnitudes, input variables needed to be normalized to improve model performance [28]. This study applied Standard-Scaler, which transforms all features into a distribution with a mean and standard deviation of 1 and 0 [29], as a data normalization technique since it is one of the most frequently applied data scaling methods [28] and was proven to have better performance in detecting minority class instances (short-term risk cases in this study) in the construction safety research field [12].

Once the features were normalized, they were split in two parts, training and testing sets, by taking 80 and 20% of the data randomly. Despite there are other test/train split ratios applied in many disciplines for the prediction purposes such as 60:40 [28], 70:30 [17], and 75:25 [30], this study considered 80:20 approach to train the model in order to consider as many instances as possible. The remaining 20% did not pose any limitation during the performance evaluation step thanks to the size of the dataset (297,954 instances). In the next step, random under-sampling (RUS) was utilized to balance the class distribution of the output variable. Imbalanced class distribution is a critical issue as some output labels are more prevalent than others observed in many studies on construction safety management. The reason to apply RUS algorithm instead of other commonly considered data resampling methods such as random over-sampling (ROS) or synthetic minority over-sampling technique (SMOTE) is its high performance addressed in a comparative study [17]. Besides, artificial cases are not generated in the RUS method unlike ROS or SMOTE algorithms [31]. It is important to state that resampling algorithm was not performed for the testing set to examine the performance of the model, in case of an imbalanced class distribution commonly observed in real-life conditions [32].

3.3. Machine learning application

This study implemented Stochastic Gradient Boosting (SGB) algorithm for the ML application. SGB is developed based on gradient boosting (also referred to as gradient boosting decision tree (GBDT)) method to improve model performance and lower computational load [33]. In the GBDT, additive models are formed through fitting base learner with respect to least squares at every iteration [33]. Despite GBDT has widely been used in many disciplines, addressing the trade-off between prediction power and efficiency becomes a critical issue in this method as it considers each data instance in each feature to calculate the attained information [34]. Besides, overfitting might be a significant burden in the GBDT algorithm, particularly once the robust regularization is not applied during the iteration process [35]. Therefore, this study applied SGB method proposed by Friedman [33]. The method is similar to the gradient boosting, and the main difference was that randomness was incorporated into the procedure. In other words, random sub-sample was taken from the training set in each iteration rather than full samples in order to fit base learners [33]. Therefore, SGB was used in this study. Detailed explanation of the SGB algorithm can be noticed in the study of Friedman [33].

The hyperparameters of SGB algorithm was identified through an iterative trial-error approach. Trial-error parameter tuning strategy was selected as it has commonly been adopted in the construction safety management literature [9, 18, 36–38]. Hence, parameters of SGB, i.e., "number of trees", "minimum sample of split", "learning rate", "maximum depth", "subsample", and "random state" were searched sequentially to identify optimum parameter configuration scenario. At the last step of the ML application, performance of the models was evaluated by using two performance evaluation metrics, area under the receiver operating characteristic (AUROC) curve and confusion matrix. AUROC was considered suitable to this study as it is a reliable instrument in case of an imbalanced class distribution. In other words, if there is imbalanced distribution of the classes, the AUROC is more representative than accuracy [10], which is the most commonly used performance evaluation metric in classification problems. AUROC is calculated by taking the true positive rate (TPR, also known as recall) and false positive rate (TPR) placed at the x and y axes, respectively [39]. In order to distinguish the correct classifications of each output labels, confusion matrix was also used. With confusion matrix, how well each output label is classified correctly can be summarized [40]. It is worth mentioning that performance assessments were conducted with the imbalanced testing set to show the applicability of the proposed model in real-life applications.

3.4. Shapley additive explanations

One of the most critical concerns in ML applications is their black-box nature, hindering the interpretation of the model. This research adopted Shapley Additive Explanations (SHAP) technique to explain the prediction rationale. Despite there are other feature evaluation algorithms such as permutation importance [20] and Gini importance [21], SHAP algorithm is more powerful than these methods. Unlike Gini and permutation importance, SHAP provides which attributes affect the prediction, along with the importance of them [41]. That is to say, the effects of each attribute in each classification outcome is calculated and additively accumulated based on coalition game theory [42]. In this method, attributes having the largest absolute Shapley values are the most significant attributes and vice versa [22].

4. Results and discussion

This study investigates the short-term vulnerability of construction workers to occupational accidents using SGB method. Table 3 shows the parameters of SGB along with corresponding search domain, step size, and optimum parameters. The parameters of SGB were configured through an iterative trial-error approach. Hence, the search space was determined as 50-600, 2-10, 0.01-0.2, 2-10, 0.5-1, and 0-42, for "number of trees", "minimum sample of split", "learning rate", "maximum depth", "subsample", and "random state", respectively.

According to the analysis results, the optimum hyperparameters were identified as 350, 3, 0.1, 5, 0.6, and 42, while searching the parameters with a step size of 50, 1, 0.01, 1, 0.1, and 1, respectively. Figure 3 also illustrates the changes in the AUROC values based on the number of trees, which is one of the most determinant parameters of tree-based ML methods [31]. AUROC curve of the proposed SGB model shows that the model achieved an AUROC value of 0.741 (Fig. 4). As addressed by Choi et al. [10], the threshold value of an acceptable AUROC value is 0.7, indicating that the proposed model is a good predictor of short-term susceptibility. In order to observe the prediction power of the model regarding separated output labels, confusion matrix of the classifications is provided in Figure 5. Findings show that the model detected 69.81% of the short-term accident cases (i.e., 1) correctly, while the remaining 30.19% of the cases were misclassified. Similarly, 66.98% of the non short-term accident cases (i.e., 0) were classified accurately, which left 33.92% of the classifications inaccurate. Results show that the model performance was quite higher for predicting minority class (class 1: 69.81%) compared to majority class (class 2: 66.98%).

At the final step of the research, SHAP analysis was performed to enhance the explainability of the ML model. Model interpretability is an essential element of prediction studies and highlighted by many scholars [43–45]. SHAP summary plot is illustrated in Figure 6.

Table 3. Parameter ranges and optimum values of SGB application

Parameter	Search domain	Step size	Optimum value
Number of trees	50 – 600	50	350
Minimum sample of split	2 – 10	1	3
Learning rate	0.01 – 0.2	0.01	0.1
Maximum depth	2 – 10	1	5
Subsample	0.5 – 1	0.1	0.6
Random state	0 – 42	1	42

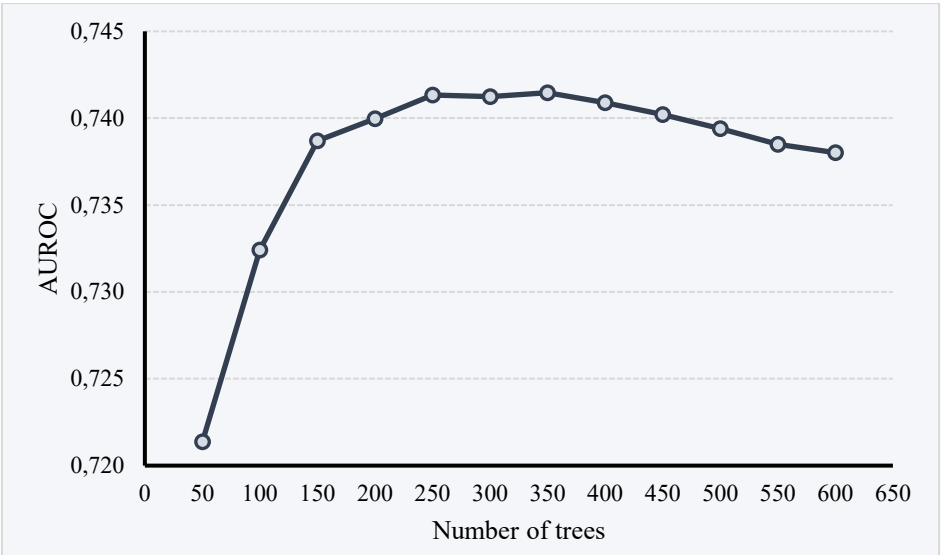


Fig. 3. Changes in AUROC value by number of trees

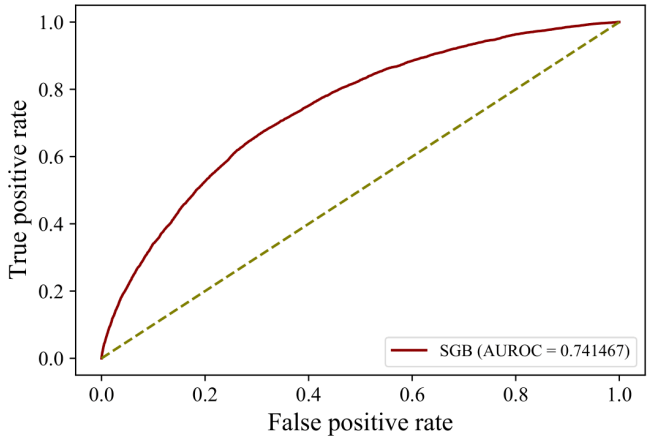


Fig. 4. AUROC curve of the configured model

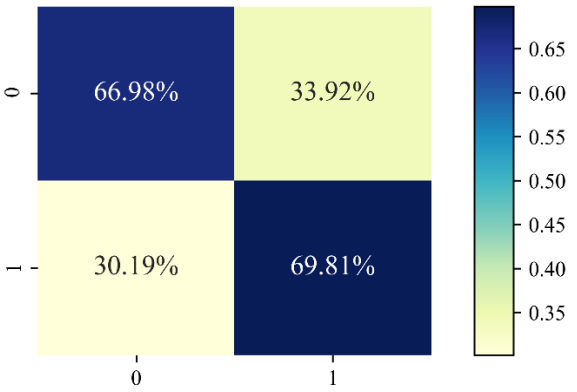


Fig. 5. Recall values of the model

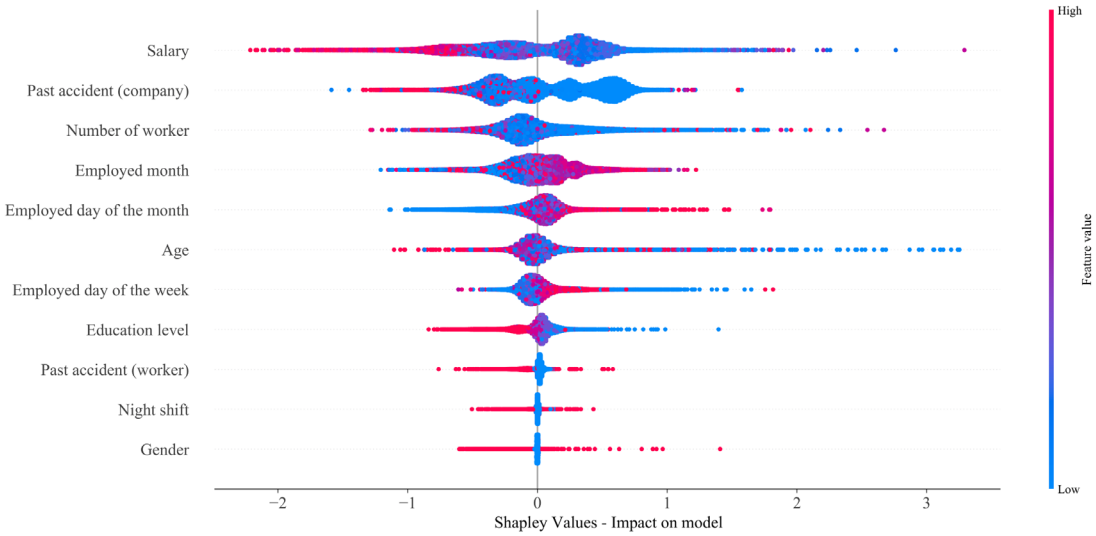


Fig. 6. Feature importance by SHAP analysis

In Figure 6, each point represents a single classification. Red dots indicate higher values of the features, whereas blue dots stand for the lower feature values. On the one hand, if the points are placed at the right side of the figure, then it means that the classifications arein favor of the short-term accident (class value = 1). On the other hand, if the points are located at the left side of the figure, it indicates that the predictions are in favor of non short-term accident (class value = 0). Attributes are provided at the y-axis in descending order regarding their importance weights, which is calculated by the sum of absolute values in each observation (i.e., point).

From the methodological point of view, this study presents several contributions to the literature. First, despite the applicability of many ML methods such as KNN [25], NB [17], RF [10], SVM [11], and ANN [9] has been evaluated in the construction safety management literature, the use of SGB method is examined in this research. Besides, despite some researhers considering during- or post-accident attributes to predict several output variables in the construction safety management research field [27, 31], the current study eliminated this limitation by only considering the pre-accident attributes. This approach is particularly beneficial as proactive safety management framework can only be developed by incorporating pre-accident inputs. Additionally, this study applied several pre-processing methods to improve performance of the model. Particularly handling the class imbalance issue (model training with a balanced dataset and model testing with an imbalanced dataset) is an essential part of classification models as imbalanced class distribution is commonly observed in many real-life applications [17]. Moreover, by using the SHAP method, model explainability was also augmented in this study to improve the applicability and visibility of the proposed model. Hence, the model can be used on construction sites proactively and confidently.

SHAP results indicate that salary, past accident in the company, and number of workers in the company were the most influential attributes in the short-term vulnerability detection. Findings further show that an increase in the salary is associated with a decrease in the shorth-term (first week) accident risk (many of the red dots are at the left side of Figure 6), while construction workers who earn less are more likely to experience an accident within first week after their employment (many of the blue dots are at the right side of Figure 6). Concerning the age of the worker, it was determined that younger employees had a higher risk of having an occupational accident shortly after they start working compared to older employees. These findings can be related to relationships between salary and experience and salary and age, such that more experienced workers might be more familiar with the safety risks in construction sites reducing their short-term accident risk [46]. Similarly, low number of past accidents in the company and number of workers in the company implies that workers just started working in these companies have higher probability to face occupational accidents in the short-term. The findings can be explained by the fact that safety training procedures are more clear and applied more frequently in relativly larger projects where a higher number of workers work. Besides, if there is a low number of workers, it means that newly employed workers might be requested to perform several different tasks without proper training, which may increase their vulnerability to occupational accidents [47]. Interestingly, it was found that workers who were employed towards the end of the month were more likely to face a short-term accident. This finding is rational as workers are provided with safety trainings at the beginning of months in some companies, which makes workers employed lately more vulnerable due to not having adequate safety training. Similar to the age of the worker, education level was also found to have negative relationship with short-term accident risk, such that workers with low education level have high short-term accident risk. Despite Koc and Gurgun [12] addressed that education level is not a significant attribute in the severity sevel of accidents, SHAP summary plot shows substantial distinction between workers with less education level and high

education level (blue and red points located at the right and lest sides of the figure distinctively). Given that occupational accidents and conditions may change considerably based on occupational groups [48], safety conciousness of workers with higher educational level might be higher and as a result they may avoid risky behaviors at the first days of their employment.

Despite significant attention from research society, data-driven prediction models are hardly adopted by construction companies due to several limitations [19]. First of all, many construction firms, particularly small and medium-sized companies, do not have adequate human and/or technical resources to apply a robust prediction model. Additionally, there are only a few number of accidents in many construction sites, making it hard to train the proposed model with limited number of accidents. These facts further signify the importance and necessity of this study, because the proposed model does not require additional data collected by companies. All the required data is collected by the Social Security Institution mandatorily. Therefore, if there is a nationally-led tool, such as a web-based tool, implementing the proposed model, then any company can directly use the proposed model by introducing the data of newly employed workers, and correspondingly can detect whether these workers will face an occupational accident in the short-term period or not. In other words, although many companies may not be willing to develop these models, they can easily use these models if they are available and provided by official institutions. Accordingly, when a new dataset (or even a single data representing the features of a worker) is introduced to the model, then the proposed model can achieve an acceptable prediction accuracy (i.e., 0.741 of AUROC) as performance evaluation in the present study was conducted by using the blinded and imbalanced testing set. In this way, practitioners and safety managers can focus more on the workers who have high risk of experiencing a short-term accident in terms of safety training and monitoring. This approach will not only increase the effectiveness of safety management practices, but also enhance the productivity of construction projects by reducing short-term accident risks of workers. Because short-term accidents occur just before the workers become familiar with the environmental conditions of the construction sites and are expected to increase their productivity with the augmented know-how. Hence, the proposed model is further projected to increase the productivity of construction projects and effectiveness of safety management applications.

5. Conclusions

Occupational accidents are a critical issue on construction sites and many accidents take place at the very initial days of the employment of workers, incurring a significant number of injuries. Hence, predictive models developed by machine learning algorithms can be useful tools in detecting workers who have high accident risk in a short run. Therefore, this study aims at developing a prediction model to detect construction workers' susceptibility to short-term accident, as an initial attempt in the literature. That is to say, although there have been many studies predicting different output variables (such as fatality, severity or accident type) in the construction safety management research domain, no single study has examined the short-term susceptibility of construction workers to occupational accidents. This projection scheme is particularly important as it not only contributes to more effective safety training by focusing on the most vulnerable workers but also helps practitioners enhance productivity in construction projects through minimizing the occurrence of accidents shortly after the employment of workers. To achieve this goal, the dataset provided by the Social Security Institution was first manually pre-processed by eliminating the cases with missing values and inconsistencies, scaling the dataset with standard scaler, and resampling the dataset with random under sampling. Gradient stochastic boosting algorithm was employed for the machine learning application and performance evaluation was conducted using AUROC and confusion matrix. As another discriminating feature of the present research,

SHAP analysis was employed to investigate the impacts of the included variables on the classification scheme.

Analysis results show that workers' susceptibility to experience an occupational accident in a short-term period can be detected through a data-driven prediction model fitted based on the national dataset. Hence, the intensity of safety trainings can be increased if newly employed workers are predicted to face an occupational accident in a short-term period. SHAP results also highlight that salary of workers, past accidents and number of workers in the company are the most contributor factors, such that an increase in these features are associated with a decrease in the short-term susceptibility of construction workers and vice versa. These can be further supported by the captured dataset as more short-term accidents occurred when the wage of the workers, past accidents in the company, and number of employed workers in the company were lower. These findings are critical as they point to a focus on low-wage employees, and companies with low number of workers and past accidents. Therefore, construction companies and safety management departments of these companies can benefit from the findings by focusing on the most risky employee and company profiles determined by the SHAP analysis results. Social Security Institution can also use the results of the present study to identify the most vulnerable worker/company characteristics to short-term accidents. In this respect, the relevant institutions can develop a tool running the proposed model nationally to help companies use the collected historical dataset of occupational accidents.

This study poses some limitations that need to be addressed. First, some of the important precursors of occupational accidents were not incorporated in the analysis as the utilized dataset includes only the features captured by Social Security Institution. Addition of features can therefore be considered to improve the performance of the model. Second, the present research applied only the SGB method without comparing its performance with other machine learning algorithms. Since the SGB algorithm (a tree-based machine learning method) is determined as a powerful algorithm for the classification task, the comparison of other tree-based methods (such as random forest, XGBoost, AdaBoost and extra trees) might be a potential research direction to improve the accuracy of the prediction. Moreover, this study considered the short-term susceptibility of workers as a binary classification problem. Increasing the output label to predict the time horizon that each worker has higher probability to experience an occupational accident can also contribute a lot to the body of knowledge. Overall, proposed data-driven model is expected to enhance the awareness in safety management, which could affect the perception of workers, capability of safety professionals, mechanisms to cope with hazardous events, and corresponding mitigation strategies to reduce the number of occupational accidents.

Declaration

Funding

This research received no external funding.

Author Contributions

K. Koc: Conceptualization, Methodology, Software, Resources, Data Curation, Writing-Original draft, Writing- Review & Editing.

Acknowledgments

The author would like to thank the Social Security Institution for their support and providing the dataset.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Ethics Committee Permission

Not applicable.

Conflict of Interest

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] Laitinen H, Päivärinta K (2010) A new-generation safety contest in the construction industry - A long-term evaluation of a real-life intervention. Safety Science 48(5):680–686.

[2] Koc K, Gurgun AP (2021) Stakeholder-Associated Life Cycle Risks in Construction Supply Chain. Journal of Management in Engineering 37(1):04020107.

[3] Tözer KD, Gürcanlı GE, Zalihe Yarkıner (2022) Analysis of workday losses due to falls from scaffoldings in the construction industry. Journal of Construction Engineering, Management & Innovation 5(1):15–27.

[4] Nnaji C, Gambatese J, Karakhan A, Osei-Kyei R (2020) Development and Application of Safety Technology Adoption Decision-Making Tool. Journal of Construction Engineering and Management 146(4):04020028.

[5] Yılmaz F (2021) Analysis of the interaction between safety measures and workers' safety awareness from the construction workers' perspective. Eng Constr Archit Manag. doi: 10.1108/ECAM-07-2021-0564.

[6] Mostofi F, Toğan V, Ayözen YE, Tokdemir OB (2022) Construction Safety Risk Model with Construction Accident Network: A Graph Convolutional Network Approach. Sustainability 14(23):15906.

[7] Koc K, Ekmekcioğlu Ö, Gurgun AP (2022) Accident prediction in construction using hybrid wavelet-machine learning. Automation in Construction 133(February 2021):103987.

[8] Toğan V, Mostofi F, Ayözen Y, Tokdemir OB (2022) Customized AutoML: An Automated Machine Learning System for Predicting Severity of Construction Accidents. Buildings 12(11):1933.

[9] Ayhan BU, Tokdemir OB (2020) Accident Analysis for Construction Safety Using Latent Class Clustering and Artificial Neural Networks. Journal of Construction Engineering and Management 146(3):04019114.

[10] Choi J, Gu B, Chin S, Lee JS (2020) Machine learning predictive model based on national data for fatal accidents of construction workers. Automation in Construction 110(September 2019):102974.

[11] Goh YM, Ubeynarayana CU (2017) Construction accident narrative classification: An evaluation of text mining techniques. Accident Analysis and Prevention 108(May):122–130.

[12] Koc K, Gurgun AP (2022) Scenario-based automated data preprocessing to predict severity of construction accidents. Automation in Construction 140(May):104351.

[13] Ayhan BU, Tokdemir OB (2019) Predicting the outcome of construction incidents. Safety Science 113(November 2018):91–104.

[14] Baker H, Hallowell MR, Tixier AJ (2020) AI-based prediction of independent construction safety outcomes from universal attributes. Automation in Construction 118(June):103146.

[15] Gondia A, Ezzeldin M, El-Dakhakhni W (2022) Machine Learning–Based Decision Support Framework for Construction Injury Severity Prediction and Risk Mitigation. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 8(3):1–17.

[16] Ajayi A, Oyedele L, Owolabi H, Akinade O, Bilal M, Davila Delgado JM, Akanbi L (2020) Deep Learning Models for Health and Safety Risk Prediction in Power Infrastructure Projects. Risk Analysis 40(10):2019–2039.

[17] Koc K, Ekmekcioğlu Ö, Gurgun AP (2022) Prediction of construction accident outcomes based on an imbalanced dataset through integrated resampling techniques and machine learning methods. Eng Constr Archit Manag. doi: 10.1108/ECAM-04-2022-0305.

[18] Goh YM, Ubeynarayana CU, Wong KLX, Guo BHW (2018) Factors influencing unsafe behaviors: A supervised learning approach. Accident Analysis and Prevention 118(June):77–85.

[19] Koc K, Ekmekcioğlu Ö, Gurgun AP (2023) Developing a national data-driven construction safety management framework with interpretable fatal accident prediction. Journal of Construction Engineering and Management 149(4):04.

[20] Kang K, Ryu H (2019) Predicting types of occupational accidents at construction sites in Korea using random forest model. Safety Science 120:226–236.

[21] Mistikoglu G, Gerek IH, Erdis E, Mumtaz Usmen PE, Cakan H, Kazan EE (2015) Decision tree analysis of construction fall accidents involving

roofers. Expert Systems with Applications 42(4):2256–2263.

[22] Chakraborty D, Elhegazy H, Elzarka H, Gutierrez L (2020) A novel construction cost prediction model using hybrid natural and light gradient boosting. Advanced Engineering Informatics 46(April):101201.

[23] Pan Y, Zhang L (2021) Roles of artificial intelligence in construction engineering and management: A critical review and future trends. Automation in Construction 122(August 2020):103517.

[24] Baker H, Hallowell MR, Tixier AJ-P (2020) AI-based prediction of independent construction safety outcomes from universal attributes. Automation in Construction 118:103146.

[25] 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(December 2018):238–248.

[26] Kang K, Ryu H (2019) Predicting types of occupational accidents at construction sites in Korea using random forest model. Safety Science 120:226–236.

[27] Lee JY, Yoon YG, Oh TK, Park S, Ryu SI (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.

[28] Ekmekcioğlu Ö, Koc K (2022) Explainable step-wise binary classification for the susceptibility assessment of geo-hydrological hazards. Catena 216(February):106379.

[29] Luo M, Xie J, Yan Y, Ke Z, Yu P, Wang Z, Zhang J (2020) Comparing machine learning algorithms in predicting thermal sensation using ASHRAE Comfort Database II. Energy and Buildings 210:109776.

[30] Rahmati O, Pourghasemi HR (2017) Identification of Critical Flood Prone Areas in Data-Scarce and Ungauged Regions: A Comparison of Three Data Mining Models. Water Resources Management 31(5):1473–1487.

[31] Koc K, Ekmekcioğlu Ö, Gurgun AP (2021) Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers. Automation in Construction 131:103896.

[32] Ekmekcioğlu Ö, Koc K, Özger M, Işık Z (2022) Exploring the additional value of class imbalance distributions on interpretable flash flood susceptibility prediction in the Black Warrior River basin, Alabama, United States. Journal of Hydrology 610(March):127877.

[33] Friedman JH (2002) Stochastic gradient boosting. Computational Statistics & Data Analysis 38(4):367–378.

[34] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 2017-Decem(Nips):3147–3155.

[35] Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review 54(3):1937–1967.

[36] Ayhan BU, Tokdemir OB (2019) Safety assessment in megaprojects using artificial intelligence. Safety Science 118(March):273–287.

[37] Ayhan BU, Tokdemir OB (2019) Predicting the outcome of construction incidents. Safety Science 113(November 2018):91–104.

[38] Poh CQX, Ubeynarayana CU, Goh YM (2018) Safety leading indicators for construction sites: A machine learning approach. Automation in Construction 93(October 2017):375–386.

[39] Canbek G, Taskaya Temizel T, Sagiroglu S (2021) BenchMetrics: a systematic benchmarking method for binary classification performance metrics. Neural Computing and Applications 33(21):14623–14650.

[40] Zhu R, Hu X, Hou J, Li X (2021) Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Safety and Environmental Protection 145:293–302.

[41] Fryer D, Strumke I, Nguyen H (2021) Shapley Values for Feature Selection: The Good, the Bad, and the Axioms. IEEE Access 9:144352–144360.

[42] Naser MZ (2021) An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating causality, forced goodness, and the false perception of inference. Automation in Construction 129(July):103821.

[43] Wang R, Kim JH, Li MH (2021) Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. Science of the Total Environment 761:144057.

[44] Wang R, Asghari V, Cheung CM, Hsu SC, Lee CJ (2022) Assessing effects of economic factors on construction cost estimation using deep neural networks. Automation in Construction 134(December 2021):104080.

[45] Dikshit A, Pradhan B (2021) Interpretable and explainable AI (XAI) model for spatial drought prediction. Science of the Total Environment 801:149797.

[46] Alizadeh SS, Mortazavi SB, Sepehri MM (2015) Assessment of accident severity in the construction industry using the Bayesian theorem. International Journal of Occupational Safety and Ergonomics 21(4):551–557.

[47] Hidaka A, Miyake A (2015) Selection of Suitable Safety Activities and Experience of Safety Condition Establishment. Open Journal of Safety Science and Technology 05(01):1–9.

[48] Ustabaş İ, Sarı H, Fandaklı S (2021) The effect of occupational choice on job satisfaction and occupational accidents in the construction sector: A case study on Rize and Çorum in Turkey. Journal of Construction Engineering, Management & Innovation 4(4):198–209.