buildings

MDPI

Article

Determining Critical Cause Combination of Fatality Accidents on Construction Sites with Machine Learning Techniques

Qing Shuang * and Zerong Zhang

Department of Construction Management, School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China

* Correspondence: qings@bjtu.edu.cn

Abstract: The construction industry is fraught with danger. The investigation of the causes of occupational accidents receives considerable attention. The purpose of this research is to determine the hierarchical relationship and critical combination of the fatal causes of accidents on construction sites. The framework for fatal cause attribute was established. Machine learning technologies were developed to predict the different types of accidents. Using feature importance, the hierarchical relationship of fatal causes was extracted. An iterative analysis algorithm was created to quantify the cause combinations. The F1 prediction score was 92.93%. The results revealed that combinations existed in fatal causes analysis, even if they were hierarchical. Furthermore, this study made recommendations for improving safety management and preventing occupational accidents. The findings of this study guide construction participants in providing early warning signs of fatal and unsafe factors, ultimately assisting in the prevention of fatalities.

Keywords: fatal cause combination; fatality accidents; construction industry; machine learning; predictive modeling

check for updates

Citation: Shuang, Q.; Zhang, Z. Determining Critical Cause Combination of Fatality Accidents on Construction Sites with Machine Learning Techniques. *Buildings* **2023**, *13*, 345. https://doi.org/10.3390/buildings13020345

Academic Editor: Agnieszka Leśniak

Received: 30 December 2022
Revised: 17 January 2023
Accepted: 23 January 2023
Published: 26 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Despite the introduction of numerous safety preventive measures in recent decades, occupational safety in the construction industry still requires improvement and progress [1]. The construction industry has the most hazardous working conditions [2–4]. As a result, its risk level is regarded as the highest in many countries [5]. Construction employment accounts for approximately 7% of the global workforce, while near 35% of the world's workers die in fatal accidents each year, resulting in around 100,000 workers killed on construction sites [6]. The occupational fatality rate in the construction industry was about three times the national average for major industries in the United States [7]. The South Korean Ministry of Employment and Labor reported 458 fatal incidents and 24,617 injuries in the construction industry in 2020, accounting for 51.9% and 26.6% of all such accidents, respectively [8]. Construction sites accounted for one-third of all workplace fatalities in the United Kingdom, meaning the fatality rate was four times higher than the overall average [9]. For 9 years, the total number of construction accidents ranked first among industrial, mining, and commercial accidents in China [10]. According to an International Labor Organization report, occupational accidents cost the economy about 4% of its annual gross domestic product (GDP), or USD 1.25 trillion [11]. These accidents not only cause serious safety and health issues, but they also result in massive financial losses [12–14]. As a result, analyzing the causes of construction accidents is critical to improving safety performance.

The high accident rate in the construction industry can be attributed to the industry's complexity. The construction process consists of numerous activities that involve various stakeholders. Project activities rely heavily on human labor. Workers' physical and mental conditions are prone to occupational accidents. Furthermore, the types of construction become more diverse as the scale of construction grows larger and more complex. The

required types and the number of workers grow in tandem, as do widespread management subjects. As a result, the risk associated with construction projects rises [4,15,16]. This is why construction site operations are marked as hazardous, dangerous, complex, uncertain, and labor-intensive [17,18]. Developing advanced approaches is required to comprehend the interdependence and combination of fatality causes or safety factors due to their inherent complexity and abstraction [19,20].

Determining the correct causes of a construction accident is a difficult task. A large amount of information is collected on a regular basis in the form of data pieces [21]. The accident summary report has become one of the most important information sources because it contains a description of the accident process and an analysis of the causes. Learning from accidents is thought to be an effective method of preventing future injuries [22,23]. Important information must be extracted from accident reports in order to improve safety measures and safety management.

In response to the need to predict and analyze the factors underlying occupational fatalities through accident reports, machine learning (ML) technologies have become a popular choice for researchers. ML technologies outperform traditional statistics in prediction, feature selection, and data generation. More importantly, they are capable of interpreting data [24,25]. ML technologies form an important branch of safety research because of their ability to obtain valuable information from large amounts of data that are difficult to understand [26,27]. However, in addition to accident prediction, the underlying safety knowledge that contributes to accident causes remains to be investigated.

An effective ML classifier may not fulfill the decision-making purpose in safety knowledge extraction because it cannot provide sufficient explanations about the interrelationships of accident causation. Furthermore, traditional techniques concentrate on analyzing the weakest link in the event chain [28]. As a result, there is a scarcity of research on the interdependence and combination of various fatal accident causes. It means that some underlying cause combinations are still being ignored or undiscovered. In this situation, the chances of developing effective control measures to actively prevent a recurrence of the fatality event are slim.

The primary goal of this study was to use ML technologies to investigate the hierarchy and combination of fatal accident causes in the construction industry. The associated objectives are to (1) categorize construction fatality types, based on fatal attributes, (2) identify the hierarchical relationships among fatal causes, and (3) determine and quantify the critical cause combinations associated with each fatality accident type. This study is organized as follows. It begins by presenting the research context and then reviews the relevant literature. Following that, the fatal cause attribute framework is established. The model flowcharts are detailed, and the results are presented alongside each research procedure. The findings are discussed, followed by limitations and recommendations for further research.

2. Literature Review

ML technologies are a powerful tool for analyzing massive amounts of multi-attribute data in a timely, accurate, and interpretable manner [29]. Knowledge discovery can be performed by developing effective predictive or descriptive models to explain the original datasets and further generalize new knowledge [30,31]. Many studies on ML technologies in the construction industry have been created since the first discussion of neural networks (NNs) in construction engineering and management, in 1991 [32].

2.1. ML Technology for Construction Safety Management

Internationally, there is a greater emphasis on safety management in the construction industry. Accidents on construction sites not only cause significant financial loss, but they also result in serious injuries or death. Identifying the causes and factors is critical for construction safety, particularly for project managers who want to develop effective safety measures. To determine the causes, traditional research typically employed qualitative methods, such as expert experiences. When the number of causes is large, analyzing critical

causes and their interdependence becomes a time-consuming, laborious, and resource-intensive challenging process [17,29]. ML technologies can automatically process datasets and extract knowledge with the help of domain experts' experience, serving as an effective supplement to qualitative methods.

Tixier et al. [17] applied natural language processing (NLP) to extract attributes from textual construction injury reports. The rank probability skill score was used to assess the predictive skill of the random forest (RF) and stochastic gradient tree boosting models in classifying safety outcomes, such as injury, energy, and injured body parts. Zhang et al. [33] introduced text mining and NLP to process construction accident reports. The ensemble model performed best (68%), in terms of average weighted F1 score. Kim and Chi [34] proposed a construction accident management system that analyzed knowledge automatically based on the user's expectations. In terms of accuracy, the rule-based and conditional random field models matched 93.75% and 84.13%, respectively. Kang and Ryu [35] established the RF model to predict the types of occupational accidents. The F1 score was 71.3%. Sarkar et al. [25] used occupational accident data to create support vector machine (SVM) and artificial neural network (ANN) models to predict injury, near misses, and property damage. The prediction rate reached 89%, in terms of accuracy. Sarkar et al. [24] developed a tree-based classifier using C5.0, classification and regression tree (CART), and RF to predict slip-and-trip accidents. The prediction rate was 88%, in terms of accuracy. Baker et al. [36] used NLP to extract the basic attributes from construction accident reports. The F1 score was adopted to predict safety outcomes with XGBoost and linear SVM. Koc et al. [5] presented four tree-based ML models, including RF, XGBoost, AdaBoost, and an extra tree to predict the disability status of construction workers following accidents in Turkey. The prediction rate was 82.92%, in terms of accuracy and 81.20%, with respect to AUROC.

ML technologies and data mining methods can be adopted for exploring the relationship between construction accidents and their causes. Liao and Perng [37] applied association rule mining to identify occupational injury characteristics in the construction industry. Cheng et al. [38] used the association rule to discover the potential causality of construction accidents. The CART analysis was built on the Taiwan construction industry database to show the cause-and-effect relationships. Tixier et al. [39] developed a framework using NLP, graph mining, and hierarchical clustering on principal component methods to explore factors related to construction injuries. Assaad and El-Adaway [20] employed the spectral clustering algorithm to group the fatal causes of construction accidents. The frequent pattern data mining and the Apriori algorithm were used to further determine the cause combinations and associations.

2.2. Construction Fatality Research in China

The dataset analyzed in this study of fatal construction accidents was collected in China. Several publications on a similar topic, with a focus on construction fatalities in China, are available. They are classified into the following categories, based on the type of research:

1. Statistical analysis and case study: Meng et al. [40] analyzed fatal accidents in China, from 2004 to 2016, to determine the impact of climate factors, period distribution, and provincial distribution. Choi et al. [41] used a statistical analysis to compare fatal occupational injuries in the United States, South Korea, and China. Shao et al. [42] explored fatal accident patterns in China using a frequency analysis, correlation coefficient analysis, and variance analysis. Xu and Xu [43] analyzed the key features of fatal accidents in China using a statistical analysis. The results showed the most likely day, month, province, and accident type. Qi et al. [44] developed the data envelopment analysis method to evaluate the construction safety performance in the regions of Jiangsu, Zhejiang, and Shanghai in China;
2. Interview and survey: Goh and Sa'adon [45] explored workers' unsafe behaviors using surveys collected in Bangladesh, India, and China. Multiple stepwise linear regression,

ANN, and decision tree (DT) techniques were applied to evaluate the survey data. Man et al. [46] conducted a questionnaire with 536 Hong Kong construction workers, to study risk-taking behaviors that lead to fatal accidents. Yu et al. [47] examined the effects of safety behaviors and physiologically perceived control in a field survey of 385 construction workers in China's Yangtze region;

3. Modeling: Zhou and Irizarry [48] integrated the accident energy release model and network theory to identify 11 sub-accidents in the Hangzhou subway construction collapse accident. Jia et al. [49] developed ML technologies to assess the key factors influencing earthquake fatalities in China. Luo et al. [50] addressed a vision-based warning system for detecting worker and excavator statuses in the hazardous areas of a mega-project, the Wuhan rail transit system in China. Zhang et al. [51] introduced an order relationship analysis, decision-making trial, and evaluation laboratory methods to identify the critical causes of tower-crane accidents in China.

Safety is a major concern in the construction industry [29]. According to the most recent data from China's Ministry of Housing and Urban-Rural Development (MHURD), there were 689 accidents and 794 deaths in 2020, with an average of 1.8 accidents and 2.1 fatalities per day [52]. However, existing studies have mostly focused on the statistical analysis of fatal accidents, studying time and regional distribution, occurrence trends, and related factors. By reviewing the research on construction fatalities in China, to the best of our knowledge, there is no ML technology for predicting the fatal accident types and further research on exploring the fatal cause combinations.

2.3. Knowledge Gaps and Research Needs

According to the extensive review of the existing literature, many previous research efforts have provided important knowledge about construction accidents with various methods. Nonetheless, some limitations and gaps in the literature remain. Accident prediction and the corresponding causes identification are open to improvement.

First, research on fatal accidents in the construction industry is limited because causality studies typically include both injuries and fatalities [12,16,35,53], or only focus on injuries [5,17,36,54,55]. Choi et al. [16] suggested separating the fatal and nonfatal data because the values on the corresponding features differed. Assaad and El-Adaway [20] argued that it was necessary to conduct research specifically on fatal accidents. The statistics published by The Center for Construction Research and Training [56] also reported that construction fatalities increased in the previous years while injuries decreased. There is an urgent need to focus on fatalities in order to improve safety performance, particularly in the construction industry, which has a high fatality rate. This study collects accident reports involving at least one death to investigate the fatal causes.

Second, a fatal causes analysis, based on ML prediction models to discover the relationship between cause associations and prediction accuracy, is lacking. This relationship can also be used to identify control and management points to reduce the number of accidents. As a result, it is necessary to determine the critical causes among the described attributes that influence prediction performance. This study combines the ML predictive modeling process and the permutation importance method to extract the hierarchical relationship between each fatality accident and its corresponding causes. The relative importance of fatal causes is determined.

Third, the interdependence and combination of fatal causes for various types of fatal accidents remain to be studied. Previous research on construction safety management has focused on identifying the weakest link or providing risk factor rankings. However, there has been little research on the association of fatality causes analyzed from the specific accident type. The combinations of causation and their impact on prediction accuracy remain to be explored. In this regard, this study develops an iterative analysis algorithm, based on the hierarchical relationship to discover fatal cause combinations.

The primary goal of this study is to develop ML models for predicting fatal accidents, based on the identification of interdependence between key fatal causes. It takes into

account the interdependence and combination of various causes of accidents with fatal outcomes. The designed ML tool is capable of providing timely and reliable information to project managers in order to prevent accidents, providing a solid foundation for accident prevention and safety management.

3. Methodology

This study consists of three major steps: (1) the establishment of a fatal cause attribute framework; (2) the predictive modeling process (ML technologies); and (3) the combination determination process (iterative analysis algorithm). The research flowchart is shown in Figure 1.

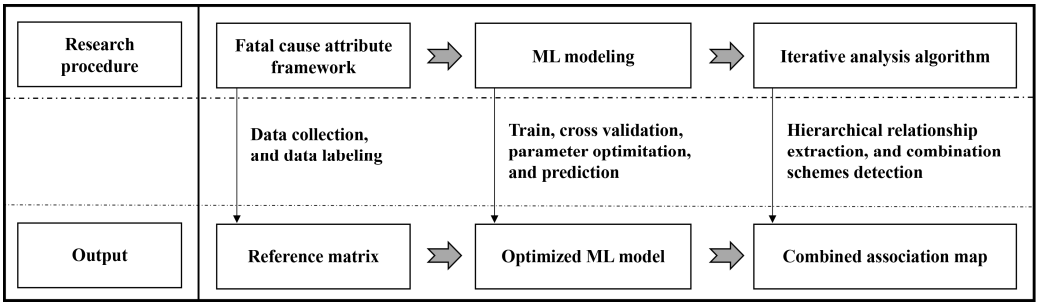


Figure 1. Research flowchart.

3.1. Fatal Cause Attribute Framework

3.1.1. Framework Establishment

The fatal cause attribute framework was established. The safety attribute list developed by Tixier et al. [17,54] and Baker et al. [36] was adopted. Desvignes [57] and Villanova [58] conducted the original attribute list.

Since this study focused research on occupational fatalities in China, there may be differences between the safety attribute list and the accident conditions in China. A word frequency analysis was used to recognize the high-frequency terms in accident reports, which calculates the number of times that a single term appears in a document. Each report's descriptions of accident causes were collected and then entered into the word frequency statistics package for analysis. The obtained word list contained several irrelevant words, such as employee types, organizational structure, and time descriptions. For example, words, such as "construction worker", "manager", "investigation team", "project team", "morning", and "suddenly" had less relation to the fatal causes. Thus, terms in the word frequency list was refined further. Eighteen fatal causes were obtained after eliminating the irrelevant ones, as shown in Table 1.

Table 1. Fatal causes and the corresponding frequency, derived from accident reports in China.

Fatal Causes	Frequency	Fatal Causes	Frequency
Working at height	189	Slope	76
Scaffold	180	Pin roll	74
Tower crane	179	Groove	72
Concrete	145	Lifting	68
Foundation pit	131	Air vent	52
Steel cable	129	Dark	44
Steel pipe	123	Construction hole	34
Suspension coop	88	Elevator shaft	25
Crane	79	Geology	23

There were two main differences between the safety attribute list and word frequency statistical results. First, some of the safety attributes on the list did not appear in the accident reports. Sixteen of the 80 attributes were found in word frequency results. This is

mainly due to the fact that the safety attribute list included accident causes other than the construction industry, such as industrial, energy, infrastructure, and mining. Redundant attributes would be created if all 80 attributes were extracted. Second, some new attributes emerged that were not included in the safety attribute list. The fatal cause attribute framework was established in this study by deleting redundant attributes and adding newly found ones. Finally, 34 cause attributes were obtained. The two major causes of accidents in the construction industry are unsafe construction conditions and unsafe behaviors [25]. Taking into account the impact of weather, this study classified the fatal causes into three categories: unsafe construction condition, unsafe behavior, and weather. Table 2 summarizes the definition of each fatal cause.

3.1.2. Data Labeling

The cause reference matrix was extracted using the human annotation method. Construction site descriptions typically include a variety of synonyms and expressions [59]. This results in a higher correlation of extracted terms, which may leads to a lower prediction accuracy. Baker et al. [36] combined NLP with independent human annotations. The safety outcomes provided by safety professionals effectively improve the predictive accuracy, eliminating the potential source of correlation between the predictors and predictands. Many previous studies performed similar manual transformations [17,20,60,61].

The manual data labeling process was described as follows. The fatality accident reports were collected and downloaded. Each report was thoroughly reviewed for completeness, including the descriptions of project stakeholders, accident process, casualties, and direct economic loss, direct and indirect reasons, accident type, and accident liability analysis.

The data labeling process had two aspects. The different types of fatal accidents were first labeled. The accident types were extracted directly from the accident reports and labeled numerically. For example, C_1 was set to "falls from heights", C_2 to "collapse", C_3 to "lifting injuries", and C_4 to "being struck by objects". Then, the fatal causes were manually entered. Each textual sample collected was converted into a binary vector, based on the cause attributes presented in Table 2 and the associated descriptions in the accident report. The binary format was sufficient for analysis since the same fatal causes could not be repeated multiple times in the same sample. A value of 1 was manually entered in the corresponding cell if one of the fatal cause attributes was mentioned in the accident sample; otherwise, a value of 0 was entered. These binary vectors were then combined to form a binary reference matrix. There were no subjective factors involved in the manual labeling process because each accident report recorded the direct and indirect reasons of the fatal accident. The data labeling process is shown in Figure 2.

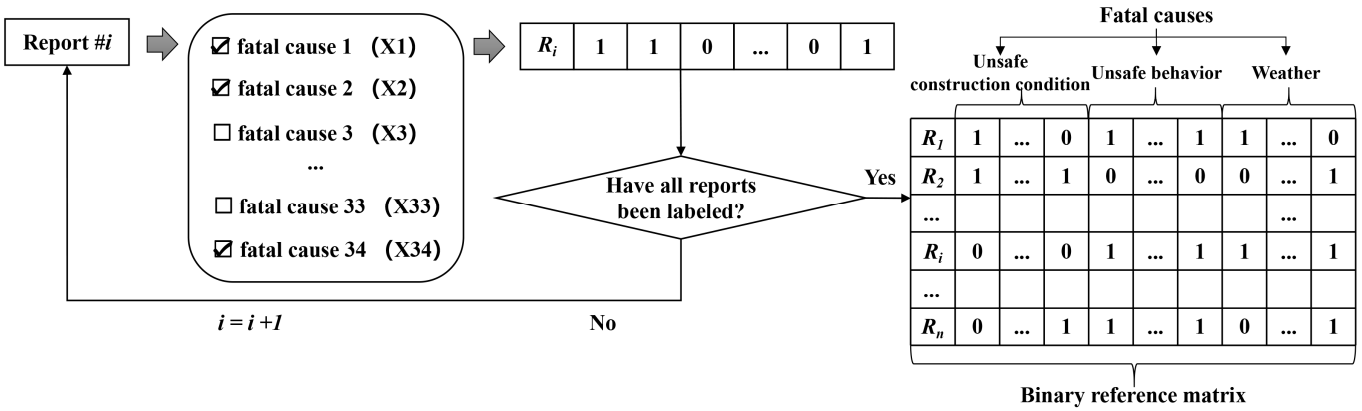


Figure 2. Fatal cause labeling process.

Table 2. Causes of fatal accidents.

Cause	Definition	Number	Category
Air vent	A mall hole in the walls or roofs of some houses for ventilation.	X1	I
Bolt	A bolt that helps hold the pieces together.	X2	I
Concrete	A general term for engineering composite material that uses cementitious materials to cement aggregates into a whole.	X3	I
Construction hole	A hole is reserved on the wall to facilitate the transportation of materials and personnel for convenience.	X4	I
Crane	A multi-action hoisting machine used for vertical lifting and horizontal transport of heavy objects within a certain range.	X5	I
Dark	Dim work environment.	X6	I
Electricity	Injuries due to electrical shocks in general, whether they are from an equipment dysfunction or lightning. It can also apply to tasks involving an electrical panel.	X7	I
Elevator shaft	Shaft for installing an elevator.	X8	I
Formwork	Concrete formwork that is constructed or demolished in a construction project.	X9	I
Foundation pit	Pit excavated at the design position of the foundation, according to the foundation elevation and plane size.	X10	I
Geology	Geological conditions, such as rocks, stratigraphic structures, minerals, groundwater, and landforms in a certain area.	X11	I
Groove	The bottom width is less than 7 m, and the bottom length is more than three times the bottom width.	X12	I
Grout	Liquid and dry grout are used by the worker in stirring, applying, or removing.	X13	I
Guardrail/ handrail	Barriers used to prevent workers or equipment from entering a specific area or preventing falls.	X14	I
Heavy material/tool	Any material of substantial weight (>40 lbs). Does not include timber, pipe, steel beam, or concrete beam.	X15	I
Heavy vehicle	A large vehicle other than machinery and light vehicles.	X16	I
Pin roll	Standardized fasteners, mainly used for the hinged connection of two parts to form a hinged connection.	X17	I
Piping	Any type of piping.	X18	I
Scaffold	A work platform built to ensure the smooth progress of each construction process.	X19	I
Slag/Spark	Small steel particles produced by grinding or welding operations and thus may be incandescent or not. The heat source is included in slag/spark in case of spark-related burns.	X20	I
Slope	To ensure the foundation's stability, a slope with a certain degree of slope is created on both sides of the foundation.	X21	I
Steel cable	The multi-layer steel wires are firstly twisted into strands, and then a certain number of strands are twisted into a spiral rope with the core of the rope as the center.	X22	I
Steel pipe	Steel with a hollow section whose length is much larger than the diameter or circumference.	X23	I
Suspension coop	A device for transporting people up and down.	X24	I
Tower crane	Used for the vertical and horizontal transportation of materials and installation of building components in building construction.	X25	I
Unstable support/surface	Any unstable surfaces, usually a temporary support or a loose plank to access a specific workspace.	X26	I
Working at height	Work performed at a height of more than 2 m.	X27	I
Improper body position	When a worker uses improper body position (somehow limited by the poor position of the environment but not by choice).	X28	II
Improper procedure	Any time a worker uses improper procedures.	X29	II
Inattention	A worker cannot concentrate at any time.	X30	II
Lifting	Refers to the behavior of moving heavy objects vertically or horizontally.	X31	II
No/Improper personal protective equipment (PPE)	Absence or incorrect PPE. No/improper PPE should be explicitly mentioned in the accident description, rather than a consequence of explaining how less serious the injury may be, compared to the actual one. Exceptions to this rule are eye injuries and concrete burns, which always call for no/improper PPE.	X32	II
Rain	Natural precipitation phenomenon.	X33	III
Wind	Natural winds, gust of wind, or explosion blasts.	X34	III

Note: I = unsafe construction condition; II = unsafe behavior; and III = weather.

3.2. ML Predictive Modeling

Figure 3 describes the ML modeling processes of DT, SVM [62], k-nearest neighbor (KNN) [63], RF [64], AdaBoost [65], and gradient boosting decision tree (GBDT) [66]. DT, SVM, and KNN are examples of basic and classical ML models, while RF, AdaBoost, and GBDT are examples of ensemble learning methods. These ML technologies were modeled and compared to develop the best classifier for predicting the type of fatality accident. The training and testing datasets were randomly divided. Cross-validation (CV) and parameter optimization were performed on the training set. The optimized models were then run on the testing set. The F1 score was used to evaluate the classification performance.

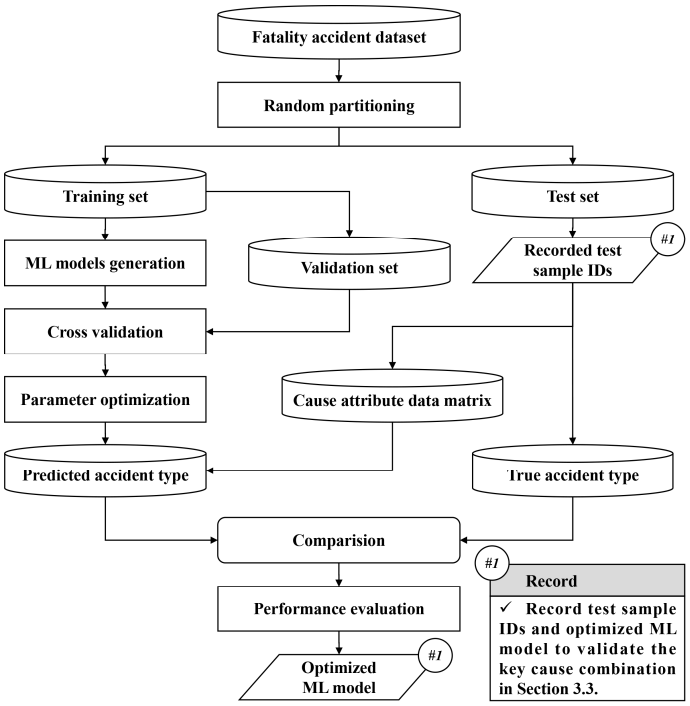


Figure 3. Flowchart of ML modeling.

3.2.1. Class Imbalance and Train/Test Splits

The imbalance problem in ML means that the samples in one class are far fewer than those in others, causing the learning model to treat them as noise. ML technologies perform best when sample sizes are balanced. To address the class-imbalance issue, this study implemented the random undersampling method. It removes multiple classes at random to balance the class distribution. The fatal accident types were removed when samples were less than 1.5% of the total [35,67].

The fatality accident dataset was divided into training and testing sets, with 80% of the samples chosen for training and the remaining 20% prepared for testing. Stratified random sampling was used to extract samples from each class. The test sample IDs were recorded to test further in Section 3.3.3.

3.2.2. Model Validation

The CV technique was introduced to estimate accuracy by dividing the training dataset into k-folds [68,69]. Within each fold, the training dataset was subdivided into a sub-training set and a matched validation set. The split ratio was equal to the training/testing set ratio [70,71]. The model was trained using the sub-training set, and its performance was validated using the validation set. The CV was performed k times, and the average performance of the k times validation was applied to the estimate the classifier's performance. A 5-fold CV was adopted.

GridSearch was then used to optimize the parameters. GridSearch is a Python parameter searching package provided by ScikitLearn. It generates all parameter combinations systematically and determines the best-performing parameters. GridSearch also adopted the 5-fold CV.

3.2.3. Performance and Evaluation Metrics

The performance metrics selected were precision, recall, and F1 score. Precision reflects how many of the positive outcomes were predicted as true. Recall, also recognized as sensitivity, is defined as the proportion of true samples to the total correctly predicted outcomes. The F1 score is a better metric of accuracy because it connects the precision and recall indices. They are expressed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

where TP (true positive) denotes the number of true samples that are also predicted to be positive. FP (false positive) represents the number of false samples that are predicted to be positive. TP + FP was the total positive predicted samples. The values of TN (true negative) and FN (false negative) represent the number of true and false samples that are predicted to be negative, respectively. The negative predicted samples were the sum of TN and FN.

3.3. Iterative Analysis Algorithm

The combination association map was determined by the iterative analysis algorithm. There were two stages. The first stage was concerned with identifying the hierarchical relationship of fatal causes. The preferred ML technology was used to iteratively build classifiers for each fatality accident type to extract cause rankings, based on feature importance. The second stage displayed the combination connections. Schemes were developed, based on the first stage results and re-tested with the optimized ML model to quantify and select the best-performing combination. The flowchart is shown in Figure 4.

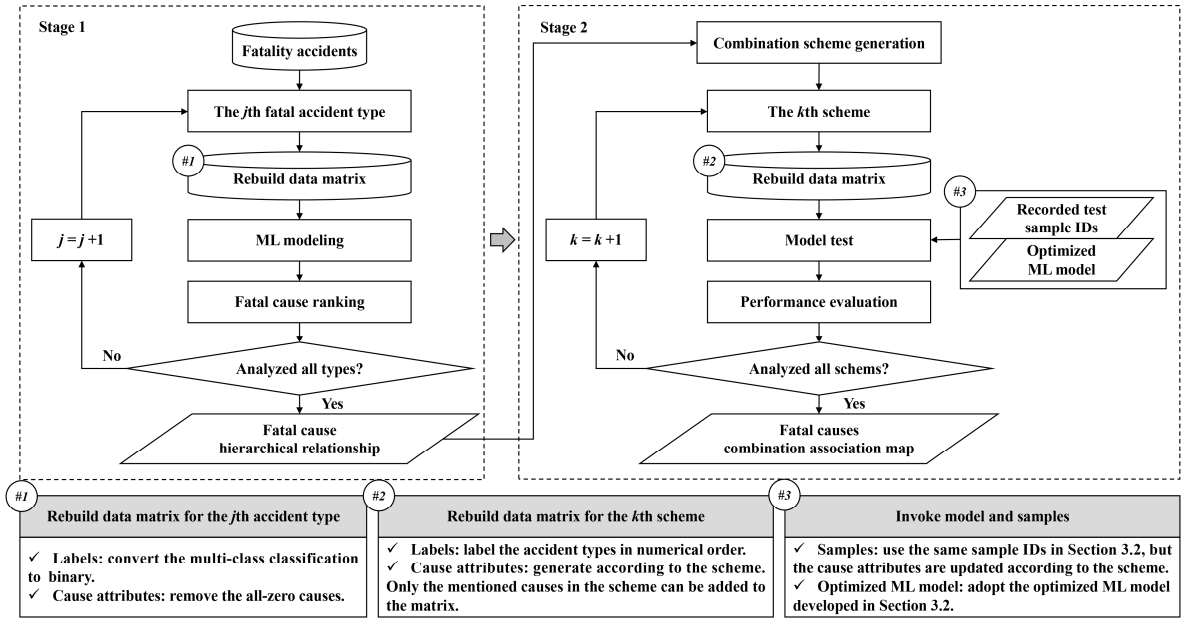


Figure 4. Flowchart of the iterative analysis algorithm.

3.3.1. ML Modeling for the Specific Fatality Type

This study ranked the importance of the causes within each type of fatal accident. If the total fatal causes attributes were analyzed together, it would be difficult to distinguish the differences in fatal causes among various accident types. Therefore, the multi-class problem was converted into several binary-class problems, each focusing on a different type of fatal accident. To generate the specific ML model, the optimized ML technology recognized in Section 3.2 was used. The label of the target accident type was reset to one, while the other types were set to zero, converting the multi-class classification into binary.

The cause attribute dataset was updated as well. When focusing on a specific type of accident, all-zero causes existed, indicating that these causes did not occur. In fact, not all of the cause attributes may be present in a specific type of fatality accident. In "falls from heights", for example, the fatal causes of piping, electricity, heavy vehicle, concrete, groove, slope, geology, and pin roll were not presented. The disparity in fatal causes is mainly because of the differences in the accident process, construction conditions, and worker behavior. These all-zero causes would have a significant impact on the feature selection results, resulting in a confusing evaluation. The ML model may identify the non-occurred all-zero cause as the most critical classification feature because it has a clear discrimination effect. As a result, the all-zero cause was designed to eliminate.

The optimized ML technology was introduced to generate the model, which used 80% of samples for training and 20% for testing. This process was repeated until models for all types of fatalities were developed.

3.3.2. Hierarchical Relationship Extraction

This step's purpose was to establish the hierarchical relationship of fatal causes for each accident type. The permutation importance was introduced to evaluate the fatal cause importance. In ML technologies, the permutation importance method is part of the feature selection. There are two feature importance measures in ML technologies. Gini importance is one, and permutation importance is another [72]. The Gini importance, also known as the "mean decrease impurity", is quick to compute but does not always yield an accurate significant value [73]. The permutation importance, also known as the "mean decrease accuracy", necessitates more calculation time but can provide reliable feature importance by measuring how much the score drops when the feature is unavailable [35].

Following the establishment of each specific model, the hierarchical relationship between each fatality accident and its corresponding causes was obtained using permutation importance. The causes were arranged in descending order in line with their importance. The first was the most critical fatal cause.

3.3.3. Combination Identification

The fatal cause combinations were quantified in three looped steps.

First, the combination schemes were designed. The causes of each accident type were combined with other types in the order of the first to the first kth. A list of possible cause combinations was generated. The test cause set for each accident type covered the causes from the most important to the kth. It was a combination of several causes rather than a single cause. When only one cause was selected, the chosen cause was the one with the highest importance value; when two causes were selected, the chosen causes were the first and second cause with the highest priority.

The fatal cause attribute dataset was then updated in accordance with the scheme. The top k causes for each fatality accident type were extracted, and the data matrix was generated sequentially.

Thirdly, the effectiveness of each scheme was determined. The test sample IDs recorded in Section 3.2.1 were introduced. The same test samples were extracted along with these IDs, but their cause attributes were updated according to each scheme. To test the data, the optimized ML model obtained in Section 3.2.3 was invoked. The performance was calculated using the F1 score.

The critical cause combination was finally identified. The F1 scores of the total combination schemes were generated after the above three looped steps were completed. The schemes with the highest F1 score were selected because they could express the prediction effect with more compact cause attributes. Similarly, among these highest-scoring schemes, the one with the fewest fatal causes was considered the most critical.

Python, an interpreted, high-level, general-purpose programming language, was used to implement the ML technologies. This study also made use of Python packages, such as Pandas, Numpy, Jieba, and ScikitLearn.

4. Results and Analysis

4.1. Data Preprocessing

This study focused on fatality accidents in China. The fatality accident reports, which implied at least one death, were obtained from MHURD [74], which was responsible for reporting occupational injuries in the construction industry. A total of 304 fatality reports were collected from 2017 to 2021, since regular accident investigations began in 2016 [75]. The fatal cause attribute framework provided 34 fatal causes across three categories: unsafe construction condition, unsafe behavior, and weather. A binary reference matrix with 304 samples and 34 causes was developed through manual labeling. The accident reports are reliable and valid because MHURD has strict investigation and monitoring procedures, and these reports are always cited by other national authorities, such as the National Bureau of Statistics and the National Development and Reform Commission.

Fatal accident types were predicted in this study. The random undersampling method was used to eliminate accident types with samples under 1.5% of the total, to avoid deviation [35]. Four accident types, namely "falls from heights", "collapse", "lifting injuries", and "being struck by objects", remained as the classification predictive problem. They were labeled numbers 1, 2, 3, and 4, respectively. The total sample count was reduced from 304 to 289. Table 3 shows the types of fatal accidents and their proportions.

Table 3. Types of fatal accidents and their proportions.

Accident Type	Prior to Being Processed		Once Processed	
	Number of Samples	Frequency (%)	Number of Samples	Frequency (%)
Falls from heights	104	34.21%	104	35.99%
Collapse	78	25.66%	78	26.99%
Lifting injuries	63	20.72%	63	21.80%
Being struck by objects	44	14.47%	44	15.22%
Vehicle injuries	4	1.32%		
Mechanical injuries	4	1.32%		
Electric injuries	2	0.66%		
Drowning	2	0.66%		
Poisoning and asphyxiation	2	0.66%		
Fire and explosion	1	0.33%		
Total	304	100%	289	100%

The correlation analysis was performed to identify the best subset of fatal causes and to remove the impact of multicollinearity, which may lead to incorrect conclusions about the relationship between the independent and dependent variables. A threshold of 0.6 is recommended for detecting and eliminating multicollinearity [76,77].

Fatal cause attributes X3 (concrete) and X13 (grout) were highly correlated with a value of 0.81. Furthermore, with a value of 0.65, fatal cause attributes X5 (crane) and X31 (lifting) were relatively highly correlated. Either of these two correlated attributes could be removed to avoid multicollinearity [77]. Accordingly, X13 (grout) and X31 (lifting) were removed. Therefore, the best subset relevant to the fatal accident classification predictive model was 32 fatal causes.

4.2. Results of the Classification Prediction

Six ML technologies were adopted in this study. Their training and CV results are shown in Figure 5. The training and CV accuracy are represented by the blue and yellow bars, respectively. The mean and standard deviation of the CV scores are displayed by the black circles and lines. All of the training scores were above 75%. How to select the appropriate model depends on the CV score. The AdaBoost model had the lowest average CV score. The RF and GBDT models had the highest CV scores, slightly over 80% with little difference between them. Variance of the RF model was a little higher than that of GBDT. Hence, these two ML models were advanced to parameter optimization.

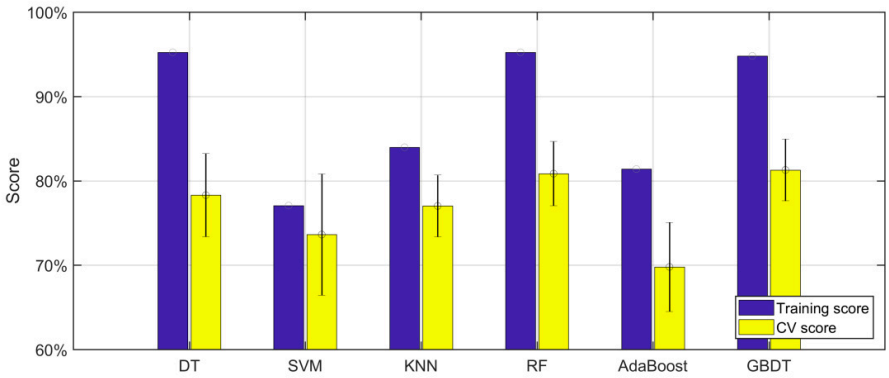


Figure 5. Training and CV scores of the ML models.

The parameter optimization process was divided into two steps since the RF and GBDT technologies all belong to the ensemble learning methods. The first step was to optimize the parameters of the ensemble learning framework, and the second was to focus on the base DT learners.

Bagging and boosting are two frameworks for ensemble learning. RF adheres to the bagging framework. Hence the number of trees was chosen for parameter optimization. This parameter limits the number of DTs to be created. GBDT follows the boosting framework, the number of trees and learning rate were chosen. Learning rate is a weight reduction coefficient associated with each base learner. A lower coefficient means that more base learners are involved to iterate. The number of trees ranged from 10 to 500 in five increments. The learning rate was varied from 0.05 to 0.3 with a 0.05 step.

Four parameters were considered for DT parameter optimization: maximum depth, minimum splitting samples, minimum number of leaf nodes, and maximum features. The maximum depth limits the depth of each DT to avoid over-fitting. The minimum split samples represent the minimum number of training datasets required to be split in an internal node. The minimum sample leaves limit the minimum number of samples on leaf nodes. If there are fewer leaf nodes than samples, they will be pruned together with sibling nodes. Finally, the maximum features limit the number of features that can be used to generate the best split. The maximum depth, the minimum sample leaves, and the maximum features all ranged from 1 to f_n (f_n represents the number of input variables) with a step of 1. The minimum splitting samples ranged from 2 to 150 with a step of 1. The evaluation was performed by Python package ScikitLearn GridSearch with 5-fold CV to determine the optimal parameters. Table 4 shows the optimized parameters.

Table 4. The optimized parameters for the RF and GBDT models.

Parameter	RF	GBDT	Parameter	RF	GBDT
Number of trees	80	85	Minimum sample leafs	1	5
Maximum depth	20	2	Maximum features	4	32
Minimum splitting samples	4	120	Learning rate		0.2

The optimized RF and GBDT models were tested on the testing set to verify the model performance. C_1 (falls from heights), C_2 (collapse), C_3 (lifting injuries), and C_4 (being struck by objects) were the target values. The classification reports for the RF and GBDT models are summarized in Table 5. Equations (1)–(3) calculated precision, recall, and F1 score metrics. The RF model achieved an average precision of 77.58%, recall of 77.28%, and F1 score of 80.72%. The GBDT model had an average precision of 81.80%, recall of 81.77%, and F1 score of 84.04%. The GBDT performed better than the RF. Hence, the GBDT model was selected as the preferred model for predicting fatality accident types.

Table 5. Classification reports for the RF and GBDT models.

RF	Precision	Recall	F1 Score	GBDT	Precision	Recall	F1 Score
C_1	90.91%	95.24%	80.72%	C_1	91.30%	100.00%	84.04%
C_2	80.00%	75.00%		C_2	84.62%	68.75%	
C_3	76.92%	83.33%		C_3	84.62%	91.67%	
C_4	62.50%	55.56%		C_4	66.67%	66.67%	

4.3. Results of the Hierarchical Relationship Extraction

The preferred GBDT model was further adopted to develop a specific predictive model of fatal accident types. The reference data matrix was rebuilt, with updated class labels and the removal of all-zero causes. The F1 scores were all above 85%, that is, "falls from heights" (98.28%), "collapses" (87.51%), "lifting injuries" (91.25%), and "being struck by objects" (91.17%) (as shown in Figure 6).

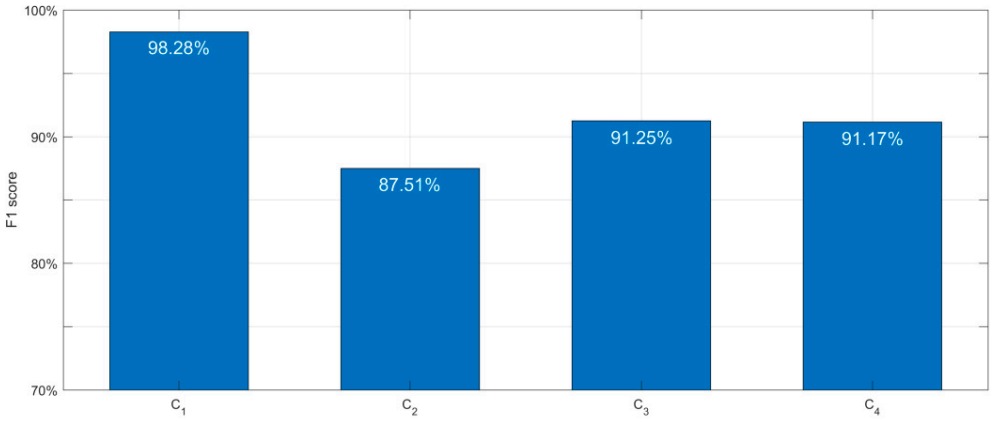


Figure 6. F1 scores for each fatal accident type.

The hierarchical relationship between each fatality accident and its corresponding causes was obtained by adopting the permutation importance, as shown in Figure 7. Working at height, slope, crane, and steel cable were the most critical fatal cause for C_1 (falls from heights), C_2 (collapse), C_3 (lifting injuries), and C_4 (being struck by objects), respectively. It was also noticed that the cumulative importance of the top five fatal causes accounted for 96.39% (C_1), 77.78% (C_2), 89.01% (C_3), and 56.67% (C_4) of all causes. This meant that the top five causes were responsible for more than half of all causes. Hence, k was selected as five to extract the critical fatal causes. That is, X27, X25, X32, X14, and X10 for C_1 (falls from heights); X21, X29, X26, X23, and X3 for C_2 (collapse); X5, X27, X24, X23, and X25 for C_3 (lifting injuries); and X22, X29, X15, X24, and X5 for C_4 (being struck by objects). These top k causes were prepared for the development of the combination scheme.

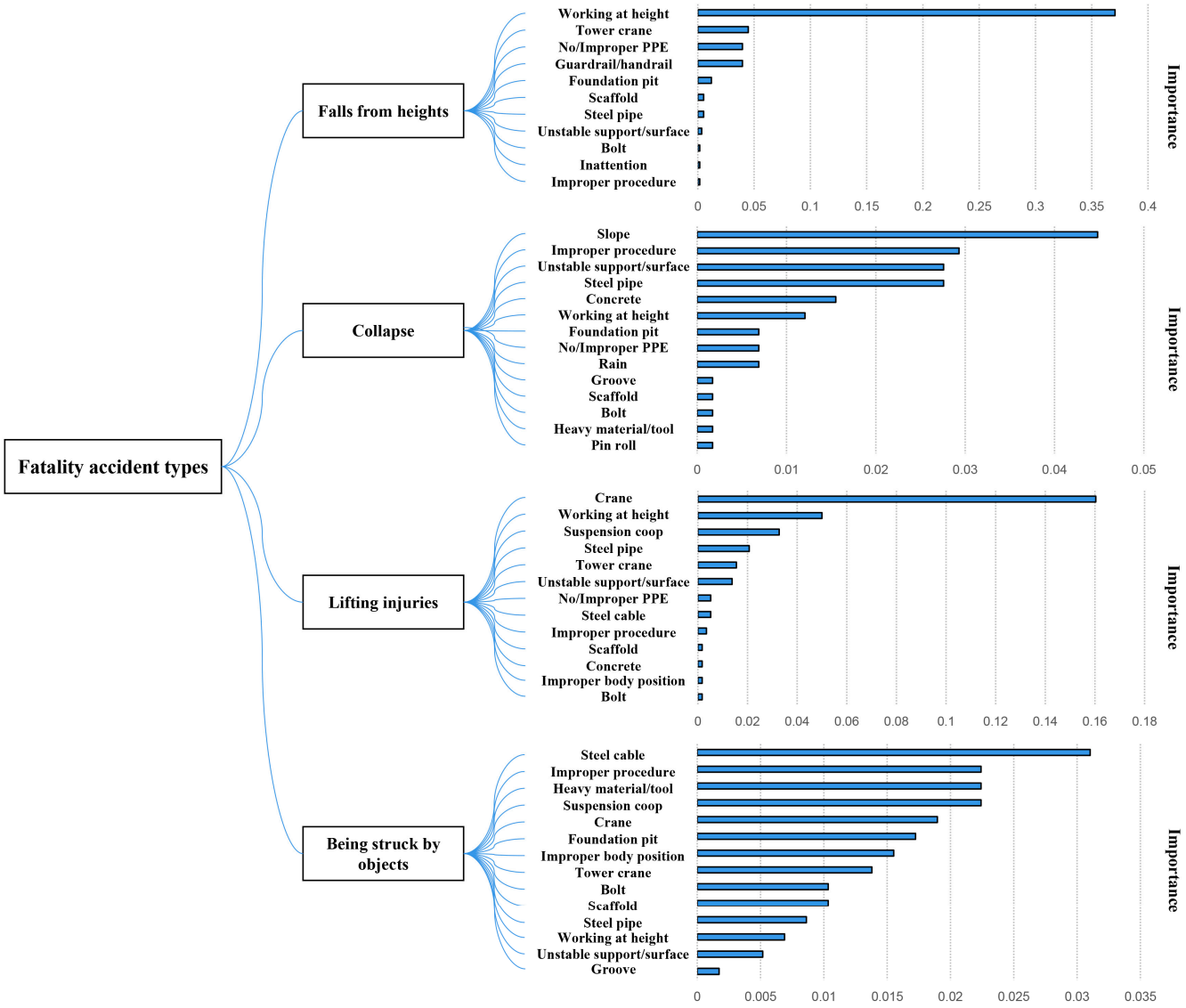


Figure 7. The hierarchical relationship between each fatality accident and related causes.

4.4. Results of the Critical Combinations

There were a total of 5^4 = 625 schemes when all of the top five fatal causes and four fatality accident types were considered. The fatal cause attribute set was rebuilt according to the scheme. The preferred GBDT model, as well as the recorded test sample IDs, were used to re-test the attribute data.

Figure 8 shows the F1 scores of the various combination schemes. The combined schemes had the highest F1 of 92.93% and the lowest of 78.32%. There were ten schemes with the highest F1 score, No. 351–355 and No. 476–480. The combination patterns of these ten schemes are shown in Table 6. It was noticed that the fatal cause schemes for C_2 and C_3 were consistently the first and the top five causes, respectively, which remained constant. On the contrary, there were differences in schemes for C_1 and C_4. The model accuracy for C_1 remained unchanged as the fatal cause combinations changed from one to five. In addition, the model accuracy for C_4 stayed the same with the top three and four causes.

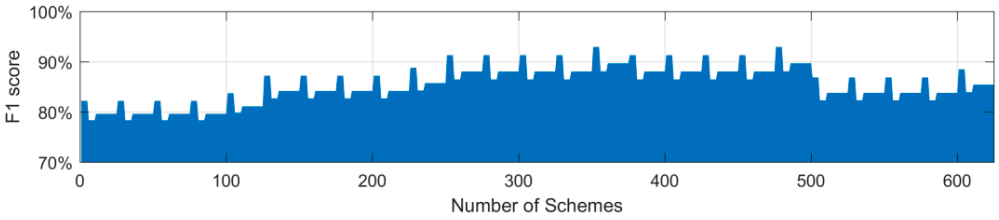


Figure 8. F1 scores of the combination schemes.

Table 6. Combination patterns with the highest F1 score.

No.	C_1	C_2	C_3	C_4	No.	C_1	C_2	C_3	C_4
351	1	1	5	3	476	1	1	5	4
352	2	1	5	3	477	2	1	5	4
353	3	1	5	3	478	3	1	5	4
354	4	1	5	3	479	4	1	5	4
355	5	1	5	3	480	5	1	5	4

Scheme 351 only had nine causes. The ML feature selection principle [78] stated that the scheme with the highest accuracy and the fewest features (i.e., fatal causes) ensured the prediction accuracy and feature compactness. As a result, Scheme 351 was selected as the best scheme to demonstrate the predictive performance. The confusion matrix and the classification report for Scheme 351 were compared to the preferred GBDT model (shown in Table 7). The TP for C_2 (collapsed) increased significantly, from 69% to 100%, indicating that the proposed iterative analysis algorithm improved the probability of correctly judging the types within true samples. The TP for C_4 did not increase, and the TP and FN values were nearly identical to the preferred GBDT model. The precision of C_1, C_3, and C_4 greatly increased to 100%, while the recall of C_2 rose from 68.75% to 100%. The F1 score improved from 84.04% to 92.93%. It revealed that the fatal cause combination schemes and the iterative analysis algorithm efficiently determined the critical cause combination, providing explanations for the interrelationships of the accident causation. It also reduced the issue of multicollinearity between the predictors and predictands. The predictive ability of C_4 needs to be improved further in future research.

Table 7. Comparison of the confusion matrix and classification report.

Scheme 351		True Accident Type Condition C_1	C_2	C_3	C_4	TP	FN	Precision	Recall	F1 Score
Predicted	C_1	100.00%	0.00%	0.00%	0.00%	100.00%	0.00%	100.00%	100.00%	92.93%
	C_2	0.00%	100.00%	0.00%	0.00%	100.00%	0.00%	80.00%	100.00%	
	C_3	0.00%	8.33%	91.67%	0.00%	91.67%	8.33%	100.00%	91.67%	
	C_4	0.00%	33.33%	0.00%	66.67%	66.67%	33.33%	100.00%	66.67%	
Preferred GBDT model										
Predicted	C_1	100.00%	0.00%	0.00%	0.00%	100.00%	0.00%	91.30%	100.00%	84.04%
	C_2	13.00%	69.00%	6.00%	13.00%	69.00%	31.00%	84.62%	68.75%	
	C_3	0.00%	0.00%	92.00%	8.00%	92.00%	8.00%	84.62%	91.67%	
	C_4	0.00%	22.00%	11.00%	67.00%	67.00%	33.00%	66.67%	66.67%	

It was worth noting that Scheme 480 achieved the same F1 score as Scheme 351 but had more combination fatal causes. It displayed the safety management control elements on a larger scale, which is important in practice. Figure 9 shows the combined association map of fatal accident types, based on these two schemes.

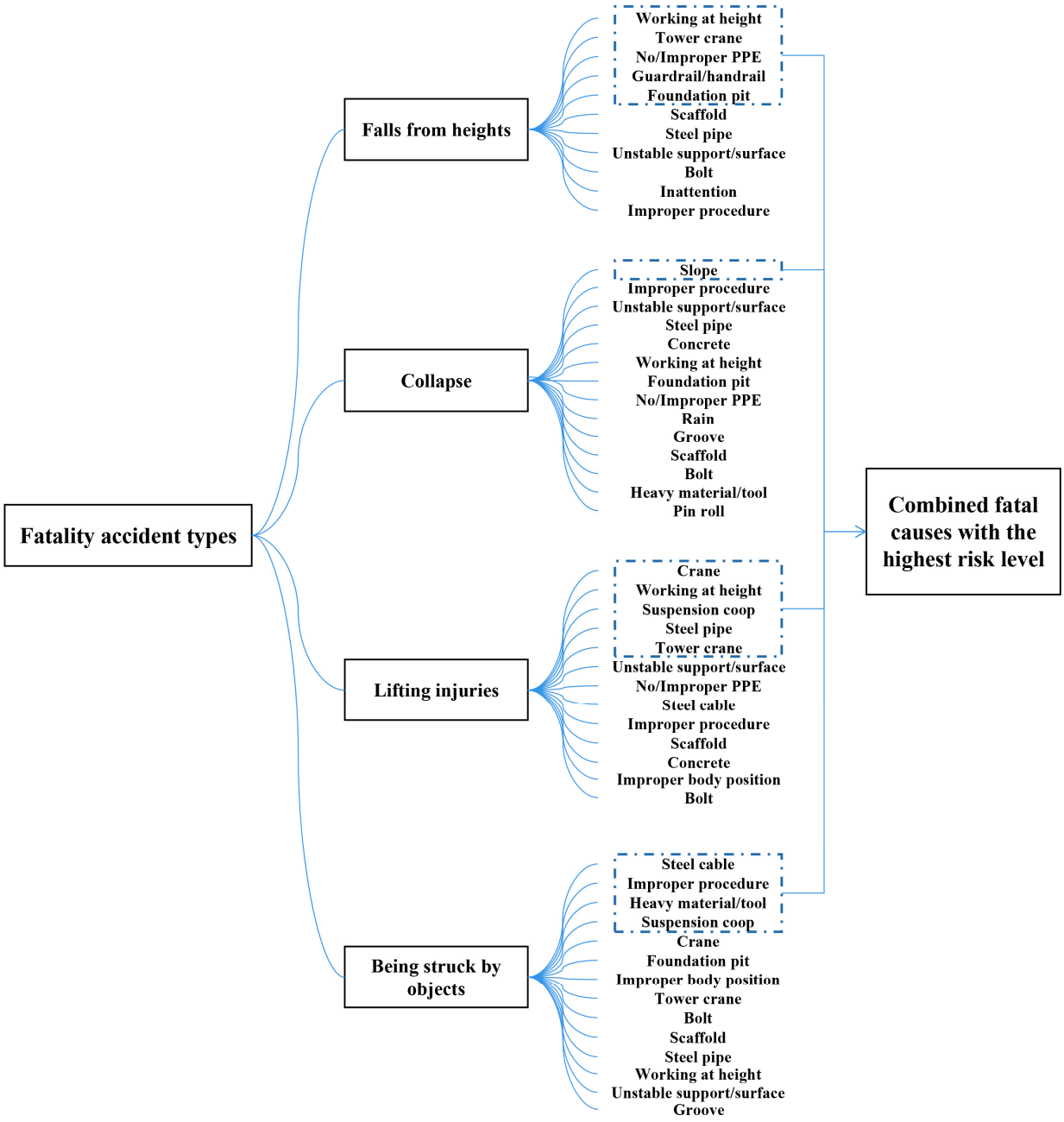


Figure 9. The combined association map of fatal accident types and related key causes.

5. Discussion

5.1. Discussion of the Findings

This study proposed an ML model for predicting accidents with fatal consequences, based on the discovery of interdependence between key fatal causes. It contributes to the risk of casualties and fatalities in the construction industry by taking interdependence and the combination of various causes of accidents with fatal outcomes into account. Although a few researchers have identified cause-and-effect iterations, the combination of cause groups within each specific fatality type has not been sufficiently investigated. This study found critical risk factors and groups in various fatality types, using the designed modeling and analysis procedure. Only focusing on the most critical cause may result in ignoring groups of the causes with the same risk level. It leads to a risky safety management strategy on construction sites. As a result, this research provides a better understanding of fatal cause combinations among risk factors.

We first combined the ML predictive modeling process and the permutation importance method in this article to investigate the fatal cause hierarchical relationship associated with fatality accident types. According to the findings, twenty-one causes have an effect on the predictive accuracy, accounting for about 62% of the original thirty-four attributes. The highest risk factors were identified as working at height, slope, crane, and steel cable. The modeling process provides an efficient method for extracting causes among various attributes and ranking their importance. It is worth noting that the hierarchical relationship was not generated for all fatalities, but rather for each fatality type. As a result, it emphasized the distinction that exists in construction fatality accidents. The derived hierarchical relationship establishes a clear order of cause importance.

Second, we highlight the interdependence and combination characteristics of the identified fatal causes to investigate whether there were risk groups with the same highest risk level or only one cause with the highest risk among different types of fatalities. This emphasizes the significance of incorporating combination detection of fatal causes in safety management. Accordingly, we developed an iterative analysis algorithm to quantify the impact of the prediction effect by selecting each individual cause into the combination groups. The results showed the existence of combination groups in the identified causes. "Falls from heights", "lifting injuries", and "being struck by objects" each had a causative combination of five, five, and four key causes. Only "collapses" had a single key factor. There were a total of 12 critical causes with the same highest risk level among these three cause combinations and one critical cause. In addition to the hierarchical structure obtained above, eight more causes with high risk levels were investigated to be noticed. Previous studies did not give much thought to the combination of cause interaction. For example, [5,35,36] provided the feature importance of accidents without taking the combination groups into account. Moreover, Lee et al. [20] emphasized the presence of combinations in fatal accidents. However, rather than specific fatality accident types, their combinations were derived from clusters calculated by the spectral clustering computational algorithm. The critical fatal cause combinations in this study were determined, based on the hierarchical structure of each fatality type and their corresponding cause importance. According to the results, if attention was only paid to the first-ranked fatal factor, other critical factors with the same risk level were ignored, leading to omissions in construction safety supervision and training. As Goldberg pointed out that "once we do find the weak link, we tend to stop looking for any other sources of the problem" [20,28]. Hence, this study extends the traditional safety analysis by examining the interdependence and combination of the fatal causes.

When compared to the original ML model, the predictive rate with consideration of the combination causes increased by 8.89%, reaching 92.93%, in terms of the F1 score. It demonstrated that the proposed algorithm is capable of revealing the combined properties of fatal causes. The result also showed that the impact of cause combination had significantly changed the identified cause importance ranking, emphasizing the importance of incorporating combination groups into the safety assessment. In comparison to previous studies on prediction in the construction industry, the XGBoost model predicted 82.92% [5], the RF model predicted 71.3%, 78%, and 78.82% [24,35,79], and the SVM model predicted 89.33% [25]. Despite the fact that the compared studies used different datasets and attributes, the prediction model with cause combinations proposed in this study outperformed these classifiers.

We further analyze the causes in the identified combinations and provide corresponding safety practice guidelines. Accidents in the construction sector are typically caused by unsafe construction conditions or unsafe behaviors [25]. Improper procedures and no/improper PPE were regarded as risky behaviors among these major fatal causes, making up about 40% (2/5) of all unsafe behaviors. The Occupational Safety and Health Administration in the US cites lack of fall protection as the most common [80]. Additionally, in 2022, MHURD published rules for identifying hidden dangers in the construction industry [81], suggesting that the risk of serious accidents includes a lack of anti-overturning

devices, a lack of anti-instability measures, and an unreliable connection to a stable structure. The following safety precautions shall be taken in order to reduce construction fatalities: (1) wearing personal protective equipment, such as safety belts, helmets, and anti-skid shoes; (2) equipping guardrails and safety nets around the edge of holes (such as foundations, balconies, and unloading platforms) with solid connections; (3) identifying the potential hazard zones prior to construction using the virtual construction scenes presented by building information modeling; and (4) setting up real-time wearable devices to monitor physical fatigue and locations of the on-site construction employees.

The outcomes demonstrated the value of safety training and education initiatives in lowering fatalities. For construction workers who performed high-risk tasks, the general safety training might not be sufficient. They should have access to specialized, qualified safety education. Additionally, labor-intensive operations and crossover works have also expanded along with a widening range of construction types. Safety education should also be pushed to employers and supervisors to ensure that they can transfer safety knowledge and awareness to construction workers.

Unsafe construction conditions included working at height, tower crane, guardrail/handrail, foundation pit, slope, crane, suspension coop, steel pipe, steel cable, and heavy materials/tools. These 10 factors accounted for about 37% (10/27) of all unsafe construction conditions. This demonstrates the importance of thorough inspection and the installation of intelligent devices on construction sites to decrease the number of fatalities: (1) regularly checking the daily used machinery, equipment, and safety system; (2) introducing intelligent devices to reduce the risk of unsafe conditions (such as the visualization system of tower cranes and the early warning system of lifts); and (3) setting up a remote monitoring center to record daily operational data. Artificially intelligent algorithms can be applied to track unsafe construction conditions to inform workers of the dangers.

It is worth noting that weather factor did not exist in the critical fatal causes. This indicated that predicting fatal accidents was not greatly influenced by weather. Weather similarly had little to no impact on the prediction of the type of occupational accident, according to the results of the Korea Occupational Safety and Health Agency inquiry dataset from 2008 to 2014 [35]. However, it does not mean to imply that weather-related elements are not important or cannot cause the fatal accident. On the contrary, one can never be too careful when it comes to weather concerns, especially extreme weather on construction sites. The historical records of extreme weather, such as high temperatures and storms shall be fully investigated. These weather factors may affect construction safety. To avoid potential problems, precautions must be taken. For example, the construction work schedule can be planned with weather and geological survey results in mind. To prevent slips and falls, the foundation pit support system can be strengthened, and work at heights can be stopped before a heavy rain. Given warming and the frequency of extreme weather, there is a need to collect occupational accident reports and weather data on a regular basis to see if there are new challenges and variations between them.

In practice, this study provides a powerful tool for analyzing and determining the combined relationship between the uncertain causes that may result in unexpected fatalities. It provides a data mining-based risk combination detection procedure which is effective. It not only reduces expert workload but also enhances the accuracy of risk combination assessment. To analyze and update the cause combination groups over time, the model can be integrated into the accident statistics system. The cause combinations identified by the ML model can also be transferred to project managers in order to provide them with timely and reliable information to prevent accidents. Because the risk factors are transmitted in the form of a combination pattern, the project manager is prompted to pay attention to these risk factors with the same risk level, reducing the possibility that the project managers may only focus on the most vulnerable link and ignores other equally important risk factors. Combined safety strategies, such as increasing construction workers' safety awareness, providing them with safety protection devices, and implementing intelligent supervision systems, can effectively and efficiently help prevent construction site fatalities.

The combination identification ML model proposed in this study can also be used to analyze interaction issues in construction site near-miss reports to provide combination instances where an accident may occur.

5.2. Limitations and Recommendations

There are some limitations to this study. One limitation is the scarcity of data. This study's report samples are all from China. This is detrimental to the generalizability of the findings. The authors acknowledge that different countries' results may differ depending on the types of datasets available and the data content. An important direction of future research is to compare similarities and differences using an accident data analysis, which will be very useful research in the future. This study focused on four types of fatal accidents. The other six types were not discussed due to the small sample sizes. The dataset used in this study contains a limited number of accident reports. More accident reports and information should be gathered. The analysis of large amounts of data can improve the proposed model's generalizability.

Furthermore, a significant amount of manual labor was required to extract attributes during the data labeling procedure. NLP can be used to extract accident causes and automatically generate structured attribute datasets for future work, in order to better understand unstructured text data. Other methods, such as stacking, optimization techniques, and advanced deep learning approaches (such as bert, convolutional neural network, recurrent neural network, and bi-directional long short-term memory), can be used in future research to improve prediction accuracy.

6. Conclusions

This study investigated the interdependence and combination relationship between fatal causes in the construction industry. A fatality accident prediction model, based on 34 fatal attributes was developed to derive the relationship. The optimized ML model was further developed to determine the hierarchical relationship between each fatality accident and its corresponding causes using the permutation importance method. A total of twenty-one causes with an influence on fatalities were identified. The iterative analysis algorithm quantified the combination relationship further. Three cause combinations and one critical cause were determined as having the highest risk, for a total of 12 causes. The predictive F1 score with the combination groups was 92.93%, which was 8.89% higher than the ML model without combination factors. It proved that the developed model effectively extracted the combination of various accident causes with fatal outcomes. The proposed method assists project managers in identifying critical fatal cause groups with timely and reliable information.

Author Contributions: Conceptualization, Q.S. and Z.Z.; methodology, Q.S.; software, Q.S.; validation, Q.S.; formal analysis, Q.S.; investigation, Z.Z.; resources, Z.Z.; data curation, Z.Z.; writing—original draft preparation, Q.S. and Z.Z.; writing—review and editing, Q.S.; visualization, Q.S.; supervision, Q.S.; project administration, Q.S.; funding acquisition, Q.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities, grant number 2019JBW007; the National Natural Science Foundation of China, grant number 71501008; the Beijing Municipal Social Sciences Foundation, grant number 18GLC070; the Ministry of Education of Humanities and Social Science Foundation, grant number 20YJC630121.

Data Availability Statement: The data used in this study can be downloaded from https://gitee.com/qs_bjtu/Fatal-Accident-Reports.git. Models or code generated or used are available from the corresponding author by request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ayhan, B.U.; Tokdemir, O.B. Accident analysis for construction safety using latent class clustering and artificial neural networks. *J. Constr. Eng. Manag.* **2020**, *146*, 4019114. [CrossRef]
2. Kang, Y.; Siddiqui, S.; Suk, S.J.; Chi, S.; Kim, C. Trends of fall accidents in the U.S. construction industry. *J. Constr. Eng. Manag.* **2017**, *143*, 04017043. [CrossRef]
3. Rubio-Romero, J.C.; Gámez, M.C.R.; Carrillo-Castrillo, J.A. Analysis of the safety conditions of scaffolding on construction sites. *Saf. Sci.* **2013**, *55*, 160–164. [CrossRef]
4. Chong, H.Y.; Low, T.S. Accidents in Malaysian construction industry: Statistical data and court cases. *Int. J. Occup. Saf. Ergon.* **2014**, *20*, 503–513. [CrossRef] [PubMed]
5. Koc, K.; Ekmekcioğlu, Ö.; Gurgun, A.P. Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers. *Autom. Constr.* **2021**, *131*, 103896. [CrossRef]
6. Chiang, Y.-H.; Wong, F.K.-W.; Liang, S. Fatal construction accidents in Hong Kong. *J. Constr. Eng. Manag.* **2018**, *144*, 4017121. [CrossRef]
7. Wiatrowski, W.; Janocha, J. Comparing fatal work injuries in the United States and the European Union. *Mon. Labor Rev.* **2014**, *137*, 1. [CrossRef]
8. Jeong, J.; Jeong, J. Quantitative risk evaluation of fatal incidents in construction based on frequency and probability analysis. *J. Manag. Eng.* **2022**, *38*, 4021089. [CrossRef]
9. Health and Safety Executive Construction Division. Phase 1 Report: Underlying Causes of Construction Fatal Accidents—A Comprehensive Review of Recent Work to Consolidate and Summarise Existing Knowledge. Available online: https://www.hse.gov.uk/construction/resources/phase1.pdf (accessed on 20 September 2022).
10. People's Daily Online. The Ministry of Emergency Management Requires Strict Implementation of the Responsibility to Effectively Curb the Rising Trend of Construction Accidents. Available online: http://politics.people.com.cn/n1/2018/0713/c1001-30144324.html (accessed on 4 July 2022).
11. International Labour Organization. Promoting Safe and Healthy Jobs: The ILO Global Programme on Safety, Health and the Environment (Safework). Available online: http://www.ilo.org/global/publications/world-of-work-magazine/articles/WCMS_099050/lang--en/index.htm (accessed on 16 September 2022).
12. Cheng, M.-Y.; Kusoemo, D.; Gosno, R.A. Text mining-based construction site accident classification using hybrid supervised machine learning. *Autom. Constr.* **2020**, *118*, 103265. [CrossRef]
13. Guo, B.H.; Yiu, T.W.; Gonzalez, V.A. Predicting safety behavior in the construction industry: Development and test of an integrative model. *Saf. Sci.* **2016**, *84*, 1–11. [CrossRef]
14. Xu, N.; Ma, L.; Wang, L.; Deng, Y.; Ni, G. Extracting domain knowledge elements of construction safety management: Rule-based approach using chinese natural language processing. *J. Manag. Eng.* **2021**, *37*, 4021001. [CrossRef]
15. Lee, H.-S.; Kim, H.; Park, M.; Teo, E.A.L.; Lee, K.-P. Construction risk assessment using site influence factors. *J. Comput. Civ. Eng.* **2012**, *26*, 319–330. [CrossRef]
16. Choi, J.; Gu, B.; Chin, S.; Lee, J.-S. Machine learning predictive model based on national data for fatal accidents of construction workers. *Autom. Constr.* **2020**, *110*, 102974. [CrossRef]
17. Tixier, A.J.-P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Application of machine learning to construction injury prediction. *Autom. Constr.* **2016**, *69*, 102–114. [CrossRef]
18. Lee, W.; Lin, K.-Y.; Seto, E.; Migliaccio, G.C. Wearable sensors for monitoring on-duty and off-duty worker physiological status and activities in construction. *Autom. Constr.* **2017**, *83*, 341–353. [CrossRef]
19. Liu, M.; Chong, H.-Y.; Liao, P.-C.; Xu, L. Probabilistic-based cascading failure approach to assessing workplace hazards affecting human error. *J. Manag. Eng.* **2019**, *35*. [CrossRef]
20. Assaad, R.; El-adaway, I.H. Determining critical combinations of safety fatality causes using spectral clustering and computational data mining algorithms. *J. Constr. Eng. Manag.* **2021**, *147*, 4021035.
21. Ubeynarayana, C.U.; Goh, Y.M. An ensemble approach for classification of accident narratives. In *Computing in Civil Engineering 2017*; ASCE: Reston, VA, USA, 2017; pp. 409–416.
22. Lukic, D.; Littlejohn, A.; Margaryan, A. A framework for learning from incidents in the workplace. *Saf. Sci.* **2012**, *50*, 950–957. [CrossRef]
23. Sanne, J.M. Incident reporting or storytelling? Competing schemes in a safety-critical and hazardous work setting. *Saf. Sci.* **2008**, *46*, 1205–1222. [CrossRef]
24. Sarkar, S.; Raj, R.; Vinay, S.; Maiti, J.; Pratihar, D.K. An optimization-based decision tree approach for predicting slip-trip-fall accidents at work. *Saf. Sci.* **2019**, *118*, 57–69. [CrossRef]
25. Sarkar, S.; Vinay, S.; Raj, R.; Maiti, J.; Mitra, P. Application of optimized machine learning techniques for prediction of occupational accidents. *Comput. Oper. Res.* **2019**, *106*, 210–224.
26. Liao, S.-H.; Chu, P.-H.; Hsiao, P.-Y. Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Syst. Appl.* **2012**, *39*, 11303–11311. [CrossRef]
27. Xu, Z.; Saleh, J.H. Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. *Reliab. Eng. Syst. Saf.* **2021**, *211*. [CrossRef]

28. Goldberg, A. Rethinking the Chain of Events Analogy for Incidents. In Proceedings of the Proceeding of ASSE Professional Development Conference and Exposition, Des Plaines, IL, USA, 22 June 2003.
29. Yan, H.; Yang, N.; Peng, Y.; Ren, Y. Data mining in the construction industry: Present status, opportunities, and future trends. *Autom. Constr.* **2020**, *119*, 103331. [CrossRef]
30. Hui, S.C.; Jha, G. Data mining for customer service support. *Inf. Manag.* **2000**, *38*, 1–13. [CrossRef]
31. Chen, F.; Deng, P.; Wan, J.; Zhang, D.; Vasilakos, A.V.; Rong, X. Data mining for the internet of things: Literature review and challenges. *Int. J. Distrib. Sens. Netw.* **2015**, *11*, 431047. [CrossRef]
32. Moselhi, O.; Hegazy, T.; Fazio, P. Neural networks as tools in construction. *J. Constr. Eng. Manag.* **1991**, *117*, 606–625. [CrossRef]
33. Zhang, F.; Fleyeh, H.; Wang, X.; Lu, M. Construction site accident analysis using text mining and natural language processing techniques. *Autom. Constr.* **2019**, *99*, 238–248. [CrossRef]
34. Kim, T.; Chi, S. Accident case retrieval and analyses: Using natural language processing in the construction industry. *J. Constr. Eng. Manag.* **2019**, *145*.
35. Kang, K.; Ryu, H. Predicting types of occupational accidents at construction sites in Korea using random forest model. *Saf. Sci.* **2019**, *120*, 226–236. [CrossRef]
36. Baker, H.; Hallowell, M.R.; Tixier, A.J.-P. AI-based prediction of independent construction safety outcomes from universal attributes. *Autom. Constr.* **2020**, *118*, 103146. [CrossRef]
37. Liao, C.-W.; Perng, Y.-H. Data mining for occupational injuries in the Taiwan construction industry. *Saf. Sci.* **2008**, *46*, 1091–1102. [CrossRef]
38. Cheng, C.-W.; Lin, C.-C.; Leu, S.-S. Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry. *Saf. Sci.* **2010**, *48*, 436–444. [CrossRef]
39. Tixier, A.J.-P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Construction Safety Clash Detection: Identifying Safety Incompatibilities among Fundamental Attributes using Data Mining. *Autom. Constr.* **2017**, *74*, 39–54. [CrossRef]
40. Meng, W.-L.; Shen, S.; Zhou, A. Investigation on fatal accidents in Chinese construction industry between 2004 and 2016. *Nat. Hazards* **2018**, *94*, 655–670. [CrossRef]
41. Choi, S.D.; Guo, L.; Kim, J.; Xiong, S. Comparison of fatal occupational injuries in construction industry in the United States, South Korea, and China. *Int. J. Ind. Ergon.* **2019**, *71*, 64–74. [CrossRef]
42. Shao, B.; Hu, Z.; Liu, Q.; Chen, S.; He, W. Fatal accident patterns of building construction activities in China. *Saf. Sci.* **2019**, *111*, 253–263. [CrossRef]
43. Xu, Q.; Xu, K. Analysis of the characteristics of fatal accidents in the construction industry in China based on statistical data. *Int. J. Environ. Res. Public Health* **2021**, *18*, 2162. [CrossRef] [PubMed]
44. Qi, H.; Zhou, Z.; Li, N.; Zhang, C. Construction safety performance evaluation based on data envelopment analysis (DEA) from a hybrid perspective of cross-sectional and longitudinal. *Saf. Sci.* **2022**, *146*, 105532. [CrossRef]
45. Goh, Y.M.; Binte Sa'adon, N.F. Cognitive factors influencing safety behavior at height: A multimethod exploratory study. *J. Constr. Eng. Manag.* **2015**, *141*, 4015003. [CrossRef]
46. Man, S.S.; Chan, A.H.S.; Alabdulkarim, S.; Zhang, T. The effect of personal and organizational factors on the risk-taking behavior of Hong Kong construction workers. *Saf. Sci.* **2021**, *136*, 105155. [CrossRef]
47. Yu, X.; Mehmood, K.; Paulsen, N.; Ma, Z.; Kwan, H.K. Why safety knowledge cannot be transferred directly to expected safety outcomes in construction workers: The moderating effect of physiological perceived control and mediating effect of safety behavior. *J. Constr. Eng. Manag.* **2021**, *147*, 4020152. [CrossRef]
48. Zhou, Z.; Irizarry, J. Integrated framework of modified accident energy release model and network theory to explore the full complexity of the hangzhou subway construction collapse. *J. Manag. Eng.* **2016**, *32*. [CrossRef]
49. Jia, H.; Lin, J.; Liu, J. An Earthquake fatalities assessment method based on feature importance with deep learning and random forest models. *Sustainability* **2019**, *11*, 2727. [CrossRef]
50. Luo, H.; Liu, J.; Fang, W.; Love, P.E.; Yu, Q.; Lu, Z. Real-time smart video surveillance to manage safety: A case study of a transport mega-project. *Adv. Eng. Informatics* **2020**, *45*, 101100. [CrossRef]
51. Zhang, X.; Zhang, W.; Jiang, L.; Zhao, T. Identification of critical causes of tower-crane accidents through system thinking and case analysis. *J. Constr. Eng. Manag.* **2020**, *146*, 4020071. [CrossRef]
52. MHURD. Circular on the Production Safety Accidents of Housing and Municipal Engineering in 2020. Available online: https://www.mohurd.gov.cn/gongkai/fdzdgknr/zfhcxjsbwj/202210/20221026_768565.html (accessed on 18 January 2023).
53. Zhong, B.; Pan, X.; Love, P.E.; Ding, L.; Fang, W. Deep learning and network analysis: Classifying and visualizing accident narratives in construction. *Autom. Constr.* **2020**, *113*, 103089. [CrossRef]
54. Tixier, A.J.-P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Autom. Constr.* **2016**, *62*, 45–56. [CrossRef]
55. Baker, H.; Hallowell, M.R.; Tixier, A.J.-P. Automatically learning construction injury precursors from text. *Autom. Constr.* **2020**, *118*, 103145. [CrossRef]
56. Center for Construction Research and Training (CPWR). The Construction Chart Book—The U.S. Construction Industry and Its Workers (Sixth Edition). Available online: https://www.cpwr.com/research/data-center/the-construction-chart-book/ (accessed on 17 September 2022).

57. Desvignes, M. Requisite Empirical Risk Data for Integration of Safety with Advanced Technologies and Intelligent Systems. Master's Thesis, University of Colorado at Boulder, Boulder, CO, USA, 2014.
58. Villanova, M. Attribute-Based Risk Model for Assessing Risk to Industrial Construction Tasks. Ph.D. Thesis, University of Colo-rado at Boulder, Boulder, CO, USA, 2014.
59. Zou, Y.; Kiviniemi, A.; Jones, S.W. Retrieving similar cases for construction project risk management using Natural Language Pro-cessing techniques. *Automat. Constr.* **2017**, *80*, 66–76. [CrossRef]
60. Assaad, R.; El-Adaway, I.H. Enhancing the knowledge of construction business failure: A social network analysis approach. *J. Constr. Eng. Manag.* **2020**, *146*, 4020052. [CrossRef]
61. Nabi, M.A.; El-Adaway, I.H. Modular construction: Determining decision-making factors and future research needs. *J. Manag. Eng.* **2020**, *36*, 4020085. [CrossRef]
62. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
63. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
64. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
65. Freund, Y.; Schapire, R. Experiments with a New Boosting Algorithm. In Proceedings of the ICML, New York, NY, USA, 24 June 1996.
66. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
67. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [CrossRef]
68. Vanwinckelen, G.; Blockeel, H. On Estimating Model Accuracy with Repeated Cross-Validation. In Proceedings of the 21st Belgian-Dutch Conference on Machine Learning, Ghent, Belgium, 24–15 May 2012.
69. Ayhan, M.; Dikmen, I.; Birgonul, M.T. Predicting the occurrence of construction disputes using machine learning techniques. *J. Constr. Eng. Manag.* **2021**, *147*, 4021022. [CrossRef]
70. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of the In-ternational Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995.
71. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
72. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef] [PubMed]
73. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef] [PubMed]
74. General Office of the Ministry of Housing and Urban-Rural Development Notice on Production Safety Accidents in Housing and Municipal Construction in 2019. Available online: https://www.mohurd.gov.cn/ (accessed on 24 February 2022).
75. MHURD. Disclosure Work Report in 2016. Available online: https://www.mohurd.gov.cn/gongkai/gknb/201704/20170418_231537.html (accessed on 16 September 2022).
76. Subbarayudu, M. An effective approach to resolve multicollinearity in agriculture data. *Int. J. Res. Electron. Comput. Eng.* **2013**, *1*, 27–30.
77. Ahmed, M.O.; Khalef, R.; Ali, G.G.; El-Adaway, I.H. Evaluating deterioration of tunnels using computational machine learning algorithms. *J. Constr. Eng. Manag.* **2021**, *147*, 4021125. [CrossRef]
78. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
79. Poh, C.Q.; Ubeynarayana, C.U.; Goh, Y.M. Safety leading indicators for construction sites: A machine learning approach. *Autom. Constr.* **2018**, *93*, 375–386. [CrossRef]
80. Occupational Safety and Health Administration. Top 10 Most Frequently Cited Standards for Fiscal Year 2021. Available online: https://www.osha.gov/top10citedstandards (accessed on 16 September 2022).
81. MHURD. Notice of the MHURD on Printing the "Standards for Determining Hidden Hazards of Major Accidents in Housing and Municipal Engineering Production Safety (2022 Edition)". Available online: http://www.gov.cn/zhengce/zhengceku/2022-04/26/content_5687357.htm (accessed on 28 June 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.