buildings

MDPI

Article

Nature of Occupational Incidents among Roofing Contractors: A Data Mining Approach

Ikechukwu Sylvester Onuchukwu [1], Pouya Gholizadeh [2], Gentian Liko [2] and Behzad Esmaeili [3,*]

[1] Lyles School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA; ionuchuk@purdue.edu
[2] Sid and Reva Dewberry Department of Civil, Environmental and Infrastructure Engineering, George Mason University, Fairfax, VA 22030, USA; pgholiz@gmu.edu (P.G.); gliko@gmu.edu (G.L.)
[3] School of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA
* Correspondence: besmaei@purdue.edu

Abstract: Given that roofing contractors in the construction industry have the highest fatality rate among specialty contractors, understanding the root cause of incidents among roofers is critical for improving safety outcomes. This study applied frequency analysis and decision tree data-mining techniques to analyze roofers' fatal and non-fatal accident reports. The frequency analysis yielded insights into the leading cause of accidents, with fall to a lower level (83%) being the highest, followed by incidence sources relating to structures and surfaces (56%). The most common injuries experienced by roofing contractors were fractures (49%) and concussions (15%), especially for events occurring in residential buildings, maintenance and repair works, small projects (i.e., $50,000 or less), and on Mondays. According to the decision tree analysis, the most important factor for determining the nature of the injury is the nonfragile injured body part, followed by injury caused by coating works. The decision tree also produced decision rules that provide an easy interpretation of the underlying association between the factors leading to incidents. The decision tree models developed in this study can be used to predict the nature of potential injuries for strategically selecting the most effective injury-prevention strategies.

Keywords: construction incidents; roofing contractors; fall incidents; decision tree; occupational injuries

Citation: Onuchukwu, I.S.; Gholizadeh, P.; Liko, G.; Esmaeili, B. Nature of Occupational Incidents among Roofing Contractors: A Data Mining Approach. *Buildings* **2024**, *14*, 595. https://doi.org/10.3390/buildings14030595

Academic Editors: Jaemin Jeong, Jaewook Jeong and Jaehyun Lee

Received: 26 November 2023
Revised: 17 January 2024
Accepted: 19 February 2024
Published: 23 February 2024

1. Introduction

The dynamic nature and environmental exposure of operations make the construction industry one of the most dangerous industries [1–3]. Attributes ranging from the variable composition of work crews to unobserved safety regulations and the type of work performed all combine to make preventing occupational accidents in construction a real challenge [4]. While all construction trades face such concerns, roofing is considered one of the most hazardous occupations [5,6]: Roofing workers face a fatal accident rate over three times higher than the construction industry's overall fatal incident average [7]. The risk of injury for roofing workers particularly escalates when they must regularly perform their tasks at heights, an unavoidable factor in this industry. Thus, considering the frequency and severity of incidents among roofers, it is important to examine the causes of occupational accidents in this specialty trade in the construction industry.

One effective technique for discovering the underlying causes of occupational accidents is investigating accident reports [8,9] since analyzing accident reports helps stakeholders better understand the nature and pattern of accidents as well as how to prevent and mitigate their risks [10]. While previous studies have employed different statistical techniques to analyze accidents involving roofers [5,11–13], such studies fall short of fully revealing non-linear patterns and the chain of events that may lead to an incident. To overcome this shortcoming, researchers have suggested using data-mining techniques [14–16], which presents an excellent opportunity for unpacking the various attributes—and their relationships—underlying incidents.

To fill this knowledge gap, this study uses the methodologies of data mining to determine which contributing factors affect occupational accidents among roofers and to determine these factors' relative importance when predicting the nature of occupational injuries (both fatal and non-fatal). To accomplish these objectives, the research team conducted several tasks. First, accident reports related to roofing contractors were collected from the Occupational Safety and Health Administration's (OSHA's) Integrated Management Information System (IMIS) database. Second, the authors conducted a content analysis on the collected data to extract variables that characterize accidents (e.g., *nature of injury*, *event type*, *injured body part*). Third, the results were analyzed using a decision tree—a supervised data-mining technique that is renowned for its ability to reveal hidden trends in categorical and multivariate data [15–19]; to avoid overfitting the decision tree, the research team utilized pruning and extreme gradient boosting techniques. Finally, to test the robustness of the model, a ten-fold cross-validation technique was implemented.

The results of this study advance understanding of the scenarios and chain of events that lead to incidents among roofers. Practitioners can use such knowledge to assess the risk of injuries more accurately (e.g., Esmaeili [10]), implement safety interventions to break the chain of events that lead to an incident, allocate safety resources more strategically, and design personalized safety training programs to increase awareness among construction workers (e.g., Li et al. [20]) and subsequently prevent incidents.

The outline of the study is organized as follows. The paper first presents the background of the occupational incidents among roofing contractors organized into three subsections: survey-based studies, analyzing roofing accident reports using statistical techniques, and analyzing roofing accident reports using machine learning techniques. Then, the research methods and proposed machine learning model are described, followed by the findings. Finally, the results are discussed, limitations are acknowledged, and conclusions are summarized for the readers.

2. Background

The research team conducted a literature review of the current body of knowledge related to occupational accidents among roofers in the construction industry and classified past studies based on their data collection analysis approach: (1) survey-based studies, (2) analyzing accident reports using statistical techniques, and (3) analyzing accident reports using machine learning techniques. Here, we present the salient results.

2.1. Survey-Based Studies

Some studies examining roofing incidents relied on reviews from surveys. Among these studies, Fredericks et al. [5] created a survey based on the Bureau of Labor Statistics (BLS) safety and health statistics database to pinpoint tasks linked to injury, illness, and fatality patterns in roofing contracting between 1999 and 2000. Their study discovered significant contributing factors relating to (mostly non-fatal) injuries—including the *type of injuries*, *event type*, and *causes of injuries*. While informative, the study confronted several limitations: (1) the findings were based on the views of contractors who participated in the survey and not their actual incident history; (2) the geographical allocation of the survey participants was restricted to Michigan; and (3) the study only presented descriptive data without employing any inferential statistics.

In another study, accidents due to falls from height were compared among commercial roofers and residential roofers [13]. The authors recruited 252 roofers in the Midwest (Illinois, Iowa, Michigan, Indiana, and Wisconsin) to participate in their survey. The analysis used various factors—such as the presence of fall-protection programs, enforcement versus actual use of fall-protection devices, company size, work type, and race/ethnicity—to determine whether residential roofers were at a higher risk of falling than commercial roofers. Their finding is consistent with other studies: Fall incidents occur 33% more in residential projects than commercial projects and 67% more in residential projects than other building projects [21].

While these studies all present interesting findings, they each face a major limitation inherent in survey data collection: The subjective judgment of respondents. Survey respondents may suffer numerous cognitive biases when filling out surveys, including how recent events occurred, the availability of information about past events, the proximity or location of events, etc. To address such subjectivity in analyses, studies need to collect and analyze objective data regarding the conditions that led to occupational incidents among roofers—data obtainable via accident reports.

2.2. Analyzing Roofing Accident Reports Using Statistical Techniques

The effects of fall accidents during a roofing project have been investigated through various studies (e.g., Bobick [22]; Sa et al. [13]; Cheng et al. [23]; Dong et al. [6]; Moore and Wanger [7]; Mistikoglu et al. [14]; Kang et al. [3]; Halabi et al. [24]); however, a very few have examined the nature of incidents associated with roofing contractors in general (e.g., Kines [11]; Fredericks et al. [5]). In one of the early studies, Suruda et al. [25] examined the data of 288 death certificates from the National Institute for Occupational Safety and Health National Traumatic Occupational Fatalities database and found that roofing contractors (SIC 1761 [26]) experienced the highest number of fatalities from falls to lower level (average annual rate of 18 fatalities per 100,000 workers). In another study, Stern et al. [27] reviewed the fatality causes among 11,144 roofers and waterproofers (members of the United Union of Roofers, Waterproofers, and Allied Workers) and discovered falls to be the most significant event resulting in fatalities.

Dong et al. [6] investigated CFOI data presenting 20,498 deaths in the construction industry from 1992 to 2009 (in-depth data analyses limited to the years 2003 to 2009) to analyze trends and patterns of fatal falls from roofs in construction. The results indicated that roof fatalities constituted one-third of fatal falls in the construction industry in 1992–2009, 67% of which happened in small construction firms (1–10 employees). Dong et al. [6] confirmed that residential roofers face higher risks than commercial roofers because (1) residential roofing employers were less likely to provide or enforce the use of fall-protection devices and ensure adequately safe work environments, and (2) residential roofers were more likely to be employed by small employers (with fewer than 10 employees) than commercial roofers. Notably, this study focused on only fatal incidents from roofs and did not account for non-fatal scenarios, which are also common among roofing workers.

To better understand the factors involved in residential roofing fatalities—and thereby understand broader factors in fatal events throughout the industry—Moore and Wanger [7] examined 112 fatality reports filed by Occupational Safety and Health investigators for the years 2005–2010. The study highlighted the notion that high risks in the residential roofing industry stem from common factors or characteristics within the industry, including the structure of the industry, the environment of jobs, and the workforce. The study especially identified residential roofing as a highly hazardous occupation that is over nine times riskier than the average occupation and over three times riskier (with respect to fatalities) than the average construction trade. The results of Moore and Wanger's study revealed that the risk and fatality rate in residential roofing could be impacted by the fact that most of the cases examined did not comply with the existing safety standards and involved minimal-to-no adequate use of fall-protection practices or systems, as well as a lack of planning and training among the workforces.

The objective analysis in these studies offset the critique of survey-based studies, but these studies still faced three main limitations. First, the majority of past studies were conditional on a fall event and did not cover other types of accidents (e.g., struck-by, electrocution, caught-in/between) in which roofers may also be involved. Second, the data for these investigations were usually gathered from all trades in the construction industry (e.g., including large residential project construction and big general contractors), and therefore, the findings may not be relevant to small specialty trade contractors such as roofers; this constraint becomes an issue since the outcomes from these past studies may not sufficiently consider the nature of certain tasks and processes nor the more specific and

effective safety programs designed and implemented for specific specialty trades. Third, past studies mainly employed descriptive statistics in their study without supplementing such analyses with inferential statistics; since the relationships between accident-causation factors may not be linear and include higher-order interactions, descriptive statistical methods would fail to uncover hidden trends in the data. Therefore, more advanced inferential statistical methods are required to handle the substantial amounts of high-dimensional data involved in these accidents to uncover hidden trends that may inform safety practitioners' decisions for mitigating the risk of fatalities and injuries.

2.3. Analyzing Roofing Accident Reports Using Machine Learning Techniques

To address these limitations, researchers are employing more sophisticated data analytics, such as machine learning, to search for cause-and-effect relationships in highly complex data [28]. While several such studies showed the benefits of machine learning techniques in investigating accident reports in the construction industry (e.g., Rivas et al. [29]; Cheng et al. [23]; Nenonen [30]; Mistikoglu et al. [14]; Amiri et al. [31]; Gholizadeh and Esmaeili [15]; Gholizadeh et al. [16]; Kang and Ryu [32]; Sarkar and Maiti [33]; Li et al. [34]; Li et al. [28]), few have applied these techniques to analyze roofing accidents. In one of the only studies in this field, Mistikoglu et al. [14] used the decision tree data-mining technique to analyze 1413 of OSHA's fall accident reports (594 fatalities and 819 non-fatal injuries) for roofing contractors' injuries occurring during the period between 1994 and 2008. They used the C5.0 and CHAD algorithms to investigate the impact of a multitude of input variables on the degree of injury (fatality versus non-fatal). The techniques enabled drawing decision trees and developing decision rules to explain the relationship between the output and input features of the roofers' fall accident reports, and the findings of the research study indicated the possibility of fatality increased with greater fall distance and decreased with provisions for safety training. Additionally, using the proposed models, the study concluded that the most important features for predicting whether a roofing contractor's accident would involve a fatality or nonfatality were the fall distance, fatality/injury cause, safety training, and the construction operation during which the fall originated.

While the research study conducted by Mistikoglu et al. [14] demonstrated the significant potential of decision trees in identifying hidden patterns in accident reports, the study still faced some limitations that need to be addressed. First, similar to some of the previous studies, Mistikoglu et al. [14] only focused on fall hazards and did not include other types of accidents roofers may face like struck-by, electrocution, and caught-in/between; such a focus limits the predictive capacity since un-/under-trained workers may, e.g., face electrocution or struck-by fatalities. Second, the prediction accuracy of models developed in Mistikoglu et al.'s [14] study was relatively low (67–68%), and the authors did not use any bias-reduction techniques. Finally, the authors themselves suggested the use of CART to determine the importance of other variables in predicting the degree of injury. The study reported in this manuscript aims to address these limitations.

3. Points of Departure

Regarding the literature review, three main limitations manifest among previous research studies analyzing roofing accident reports. First, the majority of past studies only focused on fatal injuries (e.g., Suruda et al. [25]; Stern et al. [27]; Dong et al. [6]; Moore and Wanger [7]) and therefore reveal few insights regarding the variables influencing fatal versus non-fatal injuries. Considering that non-fatal injuries are very common and, in the past, have impacted the direct and indirect costs of roofing contractors [4], investigating these injuries and how to mitigate their risks is important since preventing non-fatal and fatal injuries will enable roofing contractors, especially smaller ones, to achieve improved job site safety while increasing profits. Second, while many of the past studies concentrated on fall accidents, they mainly ignored other types of accidents among roofers, including struck-by, electrocution, and caught-in/between; problematically, these events are still major risks, and for studies seeking to prevent injuries among roofers via training, analyzing the im-

pact of these alternative events is relevant. Third, the use of machine learning techniques (e.g., decision trees) in analyzing accident reports among roofing contractors is limited and presents significant room for improvement regarding accuracy and bias reduction.

To address these limitations and contribute to the body of knowledge supporting improved safety outcomes for roofers, this study includes both fatal and non-fatal accidents, analyzes all types of roofing accidents, and applies data-mining methods to the OSHA IMIS accident report dataset to predict the nature of occupational injuries among roofers and identify the most important factors for predicting the nature of occupational injuries among roofers. Our results reveal the important contributing factors affecting roofer safety outcomes, which support future efforts in isolating, mitigating, and/or training-to-offset these factors to prevent fatal and non-fatal injuries within this high-risk industry.

4. Research Methodology

To investigate the various factors leading to accidents among roofers, this study uses a supervised data-mining technique known as classification and regression trees (CART). Additionally, known as decision trees, CART is renowned for its ease of use and interpretability [35], particularly with the development of decision rules [29]. This method provides a distinct and precise indication of the association between variables [14] and can be used to reveal hidden trends in continuous (regression) and categorical (classification) variables, a consideration that is especially valuable in our case as most of the variables presented here are categorical. CART has been successfully used by other researchers in similar studies [14,23,29–31] to analyze construction accident reports. In this study, the research team formed the classification using a decision tree technique and the accident reports collected from OSHA. Thereafter, we employed cross-validation techniques to avoid overfitting, and we applied a reserved testing dataset to validate the results. The details of these steps appear in the following sections.

4.1. Accident Database

To investigate the various characteristics contributing to incidents among roofers, this research study collected accident reports from the OSHA IMIS online database and used the Occupational Injury and Illness Classification Manual (OIICM), developed by the U.S. Department of Labor Bureau of Labor Statistics [36]. The research team filtered the OSHA IMIS online database of accident investigation reports (fatal and non-fatal injuries) to isolate accidents among SIC 1761 "Roofing Contractors" between 1 January 2007 and 31 December 2013; the filtered dataset yielded 679 accident reports. While most of these incidents only involved one worker, some instances included multiple workers' injuries; thus, in total, 734 roofing workers' occupational injuries were recorded in the database during the seven-year period examined in this research study.

Each accident report includes case summaries, company information, project context (e.g., *project cost*, *type*, *end-use*), a limited number of variables used to describe an accident (e.g., *event type*, *source*, and *cause of injury*), and its consequences (e.g., *nature* and *degree of injuries*, *injured part of body*), so following protocols established by Neuendorf [37] and Krippendorff [38], the study team carried out a thorough content analysis on the collected data to extract variables for the analysis. Similar studies have applied content-analysis methods to identify construction accidents' attributes, outcomes, and energy sources from textual accident reports (e.g., Esmaeili [10], Villanova [39], and Desvignes [40]). Through an iterative process, the authors identified nine attributes (variables) in the content analysis. These variables were divided into two main categories: Independent variables—variables that could be identified before the incident (i.e., pre-incident characteristics such as *project end-use*, *cost*, *type*, *cause of injury*, and *day of incident*)—and dependent variables—variables that can be identified only after the incident (e.g., *degree of injury*, *nature of injury*, *injured body part*, and *source of injury*).

To ensure the consistency of variables across the data and to minimize the ambiguity of the reported values, the research team adopted the Occupational Injury and Illness

Classification System (OIICS) for "use in coding the case characteristics of injuries, illnesses, and fatalities in the Survey of Occupational Injuries and Illnesses (SOII) and the Census of Fatal Occupational Injuries (CFOI) programs" [36] (p. 4). In total, 32 accidents were removed from the original data due to incomplete data or non-occupational causes of injury (e.g., heart attack). OIICM describes a hierarchical classification system and provides a four-level ontology to classify accident attributes, which helped authors combine small categories to better represent the data.

4.2. Decision Tree Model

We implemented the classification and regression trees (CART) algorithm with the recursive partitioning and regression trees (RPART) [41] to (1) predict the nature of occupational injuries due to an accident during a roofing project and (2) identify which factors are most important in predicting the nature of occupational injuries. To build the decision trees, the research team used the classification and regression training (CARET) packages [42] in R [43].

The decision tree model was developed with the *nature of the injury* as the response variable. The *nature of the injury* is split into two categories: non-fracture and fracture. Each of these categories represents about 50% of the data. On the other hand, the *injured body part* was a predictor (explanatory) and was categorized into *fragile* (*head/neck* and *body system*) and *nonfragile* (*lower/upper extremities*, *trunk*, and *multiple body parts*) to avoid data imbalance. The other predictor (explanatory) variables included *source of injury*, *cause of injury*, *project details* (*cost*, *end use*, and *type*), and *day of injury*.

4.2.1. Data Partitioning

The process of subjecting the raw data to CART involved splitting the dataset into training and testing datasets in ratios such as 70:30, 75:25, or 80:20, respectively. In this experiment, a random selection of 560 (i.e., 80%) of the 699 collected observations was trained with the use of the CART algorithm in R [43]. The remaining 139 (i.e., 20%) were used for testing. The training and testing datasets (in the ratio 80:20) were independent of each other. The training dataset was further divided into ten folds for cross-validation. In ten-fold cross-validation, as commonly seen in other studies [29], the training dataset is split into ten parts ("folds"), whereby exactly one fold is iteratively used for testing, and the remaining nine are used for training. The details of the ten-fold cross-validation analysis appear below.

4.2.2. Confusion Matrix

The performance assessment of a classifier is typically and initially presented in the form of a confusion matrix. In its most basic form, the confusion matrix is a report that provides a 2 × 2 array of the performance evaluation results of a proposed two-class (such as a positive and a negative class) classification model. Figure 1 depicts a typical example of a two-class 2 × 2 confusion matrix. Furthermore, the results presented in a confusion matrix can be used to compute several performance measures of a proposed classifier. Table 1 presents some performance measures commonly computed in the literature and their associated formulas.

Predicted values	True values	
	Positive (P)	Negative (N)
Positive	True Positive (TP)	False Positive (FP) *
Negative	False negative (FN) **	True Negative (TN)

Figure 1. Confusion matrix (* Type I error; ** Type II error).

Table 1. Measures of performance *.

Measure	Formula
Accuracy	$(TP + TN)/(P + N)$
True Positive Rate (TPR) **	TP/P
False Positive Rate (FPR) ***	FP/N
Precision	$TP/(TP + FP)$
F1 Score	$2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$
Kappa coefficient ****	$(p_0 - p_e)/(1 - p_e)$

* Except for the FPR, a value close to 1 indicates a good classification performance. ** Additionally, known as sensitivity, recall, and/or hit rate. *** Equals to 1− specificity, where specificity is (TN/N). **** p_0 is identical to accuracy and p_e is the expected agreement by chance.

4.2.3. Cross-Validation Analysis

The robustness of the model developed by CART in this research was tested using the ten-fold cross-validation technique. This method checks for overfitting, selection bias, and how well the model generalizes to new and independent data, as seen in the real world. This technique, sometimes referred to as "out-of-sample testing," is a resampling method and was applied in this analysis using ten different portions or resamples of the training dataset to train and test the model in ten distinct iterations.

The ten fittings/resamples of the model in the ten-fold cross-validation yielded ten Kappa values and ten prediction sub-testing accuracies, as listed in the results section. The expected or average accuracy computed from the ten resamples was used to develop the final decision tree model [16]. The selection of the ten folds in the ten-fold cross-validation was randomized and iterated 50 times using R's random generator set.seed function. The cross-validation prediction accuracies were averaged to get the mean accuracy of the training dataset. The summary of the results appears in the results section.

4.2.4. The Receiver Operating Characteristic (ROC) Curve

The receiver operating characteristics (ROC) curve offers a visual aid for comparing a classifier's performance across thresholds and for illustrating the trade-offs between Type I and Type II errors [44]. This enables the selection of an appropriate error combination for the given situation. They are commonly used to interpret and improve the results of machine learning and data mining techniques [45] more accurately. They can be drawn using a two-dimensional graph with the sensitivities shown on the Y axis and the specificity values represented on the X axis. The area under the ROC curve (AUC) gives an indication of the performance of the proposed classification model [45]. At random performance, the curve crosses the straight diagonal and reaches an AUC of 0.5. An AUC value of 1 denotes flawless performance. The AUC is close to 1 if the ROC plot passes close to the upper left corner [45,46]. The details of the ROC curve and AUC obtained in this study are presented in Section 5.

4.2.5. The No-Information Rate

The no-information rate supports evaluating a model since the no-information rate is an indication of one's best guess, or prediction accuracy when provided with no information other than the number of observations in the classes one wants to predict [47]. A model with a lower prediction accuracy than the no-information rate indicates instances where predictive models would not be helpful since such inaccurate models would not provide any more insight than mere guesswork. Therefore, the research team considered the no-information rate in assessing the utility of our developed models.

In the accident reports analyzed in this study, 356 observations (51%) were in the class labeled *non-fracture*, which is slightly the majority group. The remaining 343 observations (49%) were in the class labeled as *fracture*, which is the slight-minority group. Hence, when no other information is available, our best guess would be to classify a report in the majority group, which should result in a base prediction accuracy of 51%—in other words, if we

were to only predict events based on the majority class, we would expect to be right 51% of the time. This expectation provided an evaluation metric for our assessment.

4.2.6. Pruning

There are typically two phases in the development of a tree: the growing phase and the pruning phase [48]. First, a sufficient-sized tree (called T_{max} in this study), or an overfitted tree, is typically developed in the growing phase. Second, the branches of the grown tree are pruned to a fitted and right-sized model in the pruning phase. Pruning a tree is a systematic way of controlling its size by cutting off subtrees that contribute little or no statistical validity [49–52]. The resulting pruned tree is smaller, and while the smaller tree may increase misclassifications in the training dataset, this smaller tree would likely decrease the misclassifications in the independent testing dataset [49]. Such changes avoid overfitting the decision tree to training data such that the model's performance improves when applied to a new and independent dataset. In this study, the research team pruned the nodes to avoid overfitting the decision tree model [41,53].

4.2.7. Variable Importance

Decision trees are powerful machine learning tools renowned not only for their great predictive abilities but also for their rich variable-importance information [54]. Researchers have especially prioritized the evaluation of relative variable importance in the last two decades [54] and have devised variable importance scores, or rankings, to support such applications as model interpretability and model selection [55]. In such contexts, variables with high-ranking scores may be selected for deeper exploration and for developing a more fitting model.

In this study, the research team used the CART technique to pinpoint the most relevant predictor or explanatory variables for a given problem in the process of growing a tree. In contrast to a linear regression model, CART may not visibly include a variable (node splitter) in generating a tree even if the variable is very important [56] because the variables shown in the structure of the tree are the primary splitters. CART can shift splitting choices to different parts of the tree since, behind the scenes, whenever a variable is missing, surrogate splitters are moving a record down the tree to its appropriate leaf node. Therefore, visual inspection of the tree may be insufficient in ascertaining variable importance rankings because the predictive contributions of a variable may manifest in surrogate splits when growing a tree.

With the use of the summary function in R, the RPART (an R version of CART) package offers an evaluation of variable importance. This score is calculated by tabulating and summing the reduction in the loss function (e.g., mean squared error) attributed to each variable at each split. An overall evaluation of variable importance involves summing the goodness of split measures for each split serving as the primary variable in addition to the goodness (adjusted argument) for all splits in which that split served as a surrogate.

If two variables happen to be essentially duplicates of each other and their surrogates are not added, they would split their importance. This situation would prevent the two variables from showing up as important variables. In other words, surrogates are also included in the calculation of variable importance since even variables that never split a node or do not show up in a tree may be assigned a high importance score, a factor that allows variable importance rankings to reveal variable masking and non-linear correlations among attributes [57]. Hence, CART enabled this study to prioritize variables based on variable importance rankings and provided a better understanding of how importance is attributed to variables based on their predictive contributions to the proposed model.

4.2.8. Extreme Gradient Boosting Machine

The extreme gradient boosting (XGBoost) algorithm [58] in R [43] is helpful in controlling overfitting and produces a prediction model in the form of a group of prediction models, typically decision trees. This package, referred to as the advanced application of

gradient boosting machine by Tang et al. [59], reduces the error rate of a current model by building on the model (i.e., developing and adding other supporting models). The overall predicted result of the boosted model is the summation of the various scores of all the individual models. Therefore, according to Tang et al. [59], given an F space of boosted trees, XGBoost prediction is the sum of the scores of K boosted classification trees, as defined by:

$$\hat{y}_i = \varnothing(x_i) = \sum_{k=1}^{k} f_k(x_i),\ f_k\ \epsilon\ F \quad (1)$$

where x_i represents the ith sample, and the score of x_i at the ith boosted tree is denoted by $f_k(x_i)$. Hence, in an attempt to push the accuracy of the proposed decision tree model to the limit and to check for overfitting, this study applied the extreme gradient boosting machine to improve the accuracy of the proposed model. However, the boosting machine results obtained did not provide improvement to the proposed model, indicating its effectiveness.

5. Results

The findings of this study are organized into two main sections: (1) frequency analysis and (2) decision tree analysis. The first section explains the descriptive statistics of occupational incidents among roofing contractors while concentrating on the *degree of injury* as the most apparent consequence of these accidents. In the second section, the results of the decision tree are presented to provide insights into the association between the target variable (*nature of injury)* and the predictor variables—such as the *injured body part*, *source of injury*, *cause of injury*, *project details* (*cost*, *end use*, *type*), and *day of injury.*

5.1. Frequency Analysis

The frequency analysis indicated that 276 (39%) of occupational accidents among roofing contractors resulted in a fatality. The rest (non-fatal) of the injuries were filed into two categories: 391 (56%) *hospitalized injuries* and 32 (5%) *non-hospitalized injuries*. Table 2 displays the frequency results and the ratios of *degree of injury* for each of the nine identified categories identified from the content analysis. One thing to bear in mind is that the fatality ratio in this study accounts for the proportion of fatal incidents from the total number of catastrophic injuries in the current database (i.e., 699 reports) and should not be confused by estimated rates that are computed by applying full-time equivalent workers. Otherwise stated, a fatality ratio of 39% implies that 39% of all injuries in the record resulted in a fatality, not that 39% of roofers would die on the job site.

Table 2. Accident characteristics among roofing contractors.

	Variable	Frequency (% [1])	Degree of Injury (%)		
			Fatality [1]	Hospitalized	Non-Hospitalized
End-use	Non-residential building	326 (47)	141 (43)	174 (53)	11 (3)
	Residential building	367 (53)	132 (36)	214 (58)	21 (6)
	Utility and heavy and civil engineering	6 (1)	3 (50)	3 (50)	0 (0)
Project type	Alteration or rehabilitation	181 (26)	72 (40)	103 (57)	6 (3)
	Demolition	5 (1)	1 (20)	4 (80)	0 (0)
	Maintenance or repair	314 (45)	137 (44)	167 (53)	10 (3)
	New project or new addition	176 (25)	59 (34)	103 (59)	14 (8)
	Other	23 (3)	7 (30)	14 (61)	2 (9)

Table 2. *Cont.*

	Variable	Frequency (% [1])	Degree of Injury (%)		
			Fatality [1]	Hospitalized	Non-Hospitalized
Project cost	$50,000 and less	401 (57)	157 (39)	227 (57)	17 (4)
	$50,000–$250,000	109 (16)	47 (43)	60 (55)	2 (2)
	$250,000–$500,000	51 (7)	26 (51)	23 (45)	2 (4)
	$500,000–$1,000,000	46 (7)	13 (28)	29 (63)	4 (9)
	$1,000,000–$5,000,000	42 (6)	16 (38)	25 (60)	1 (2)
	$5,000,000–$20,000,000	20 (3)	8 (40)	11 (55)	1 (5)
	$20,000,000 and more	30 (4)	9 (30)	16 (53)	5 (17)
Source of injury	Machinery	24 (3)	10 (42)	12 (50)	2 (8)
	Parts and materials	115 (16)	31 (27)	77 (67)	7 (6)
	Structures and surfaces	391 (56)	177 (45)	201 (51)	13 (3)
	Tools, instruments, and equipment	133 (19)	44 (33)	81 (61)	8 (6)
	Vehicles	16 (2)	8 (50)	7 (44)	1 (6)
	Other sources	20 (3)	6 (30)	13 (65)	1 (5)
Cause of injury	Assembling/Installing roofing elements (insulation, shingles, plywood, etc.)	204 (29)	79 (39)	111 (54)	14 (7)
	Coating Works (asphalt, painting, spraying)	44 (6)	9 (20)	33 (75)	2 (5)
	Dismantling/Demolition	70 (10)	31 (44)	36 (51)	3 (4)
	Metal Works (decking, rainwater system)	52 (7)	21 (40)	29 (56)	2 (4)
	Preparation (inspection, cutting, measurement, marking, cleaning, etc.)	67 (10)	27 (40)	38 (57)	2 (3)
	Repair and renovation	89 (13)	37 (42)	50 (56)	2 (2)
	Transporting/handling materials	86 (12)	33 (38)	49 (57)	4 (5)
	Not reported	87 (12)	39 (45)	45 (52)	3 (3)
Event type	Caught in/between	12 (2)	3 (25)	6 (50)	3 (25)
	Exposure to electricity	33 (5)	25 (76)	7 (21)	1 (3)
	Fall (to lower level)	579 (83)	237 (41)	318 (55)	24 (4)
	Struck-by	39 (6)	5 (13)	32 (82)	2 (5)
	Other	36 (5)	6 (17)	28 (78)	2 (6)
Nature of injury	Bruises, contusions	32 (5)	10 (31)	14 (44)	8 (25)
	Burns	42 (6)	1 (2)	39 (93)	2 (5)
	Concussions	105 (15)	85 (81)	20 (19)	0 (0)
	Cuts, lacerations, puncture	31 (4)	5 (16)	22 (71)	4 (13)
	Electrocutions, electric shocks	34 (5)	26 (76)	7 (21)	1 (3)
	Fractures	343 (49)	75 (22)	255 (74)	13 (4)
	Non-specified injuries and disorders	81 (12)	65 (80)	16 (20)	0 (0)
	Other	22 (3)	8 (36)	12 (55)	2 (9)
Injured part of body	Body system	97 (14)	58 (60)	35 (36)	4 (4)
	Head	239 (34)	165 (69)	69 (29)	5 (2)
	Lower extremities	88 (13)	1 (1)	80 (91)	7 (8)
	Multiple body parts	98 (14)	34 (35)	60 (61)	4 (4)
	Trunk	81 (12)	13 (16)	65 (80)	3 (4)
	Upper extremities	96 (14)	5 (5)	82 (85)	9 (9)

Table 2. *Cont.*

	Variable	Frequency (% [1])	Degree of Injury (%)		
			Fatality [1]	Hospitalized	Non-Hospitalized
Day of incident	Monday	145 (21)	52 (36)	88 (61)	5 (3)
	Tuesday	118 (17)	52 (44)	62 (53)	4 (3)
	Wednesday	135 (19)	47 (35)	80 (59)	8 (6)
	Thursday	118 (17)	49 (42)	60 (51)	9 (8)
	Friday	109 (16)	43 (39)	60 (55)	6 (6)
	Saturday	47 (7)	22 (47)	25 (53)	0 (0)
	Sunday	27 (4)	11 (41)	16 (59)	0 (0)

[1] The percentages may not add up to 100% in certain circumstances because they were rounded to the nearest integer.

Apart from the variables mentioned above, OSHA reports also provide information on other variables, including human and environmental factors. Even though a large ratio of these two variables were categorized under the "Other" division (i.e., 21% of human factors and 35% of environmental factors), Table 3 presents some of the most frequent factors for each variable. The outcomes display that while *misjudgment* is by far the most common human factor, problems with *engineering control* and *personal protective equipment or clothing* were more dangerous, resulting in higher fatality ratios. Regarding environmental factors, *work surface, and facility layout conditions* are the contributing factors with the highest frequency among roofing workers and are also the factors causing the highest fatality ratios. We present these factors in Table 3, but they were not included in the following statistical analyses due to a large number of missing values.

Table 3. Frequency and fatality ratio for human and environmental factors.

		Frequency	Fatality Ratio (%)
Human factors	Misjudgment in a hazardous situation	224	44
	Insufficient or lack of personal protective equipment or clothing	71	58
	Inappropriate equipment for operation	37	43
	Malfunction in securing or warning operation	32	28
	Inappropriate position for the task	30	37
	Inappropriate material-handling procedure	17	29
	Insufficient or lack of engineering controls	15	67
Environmental factors	Work surface or facility layout condition	319	45
	Material-handling equipment or method	34	26
	Overhead moving- or falling-object action	24	42
	Flying-object action	15	20
	Temperature tolerance	9	44

5.2. Decision Tree Analysis

5.2.1. Model Interpretation

To predict roofing contractors' *nature of the injury* (conditional on an accident occurring), this study built a decision tree using the 560 training data points and seven project attributes: *source of injury*, *cause of injury*, *project cost*, *project end-use*, *project type*, *day of injury*, and *injured body part*. The target variable, *nature of the injury*, included two categories: *non-fracture* (NF) and *fracture* (F). The tree treated *head/neck injuries* and *injuries within the body system* as *fragile body parts*, whereas the rest of the *injured body parts* were labeled *nonfragile body parts*, as shown in Table 2. To avoid overfitting the model, the decision tree was pruned using tuning parameters, and accuracy helped select the optimal model in terms of the largest values. Figure 2 shows the proposed decision tree model for predicting roofing contractors' *nature of the injury* from workplace accidents.

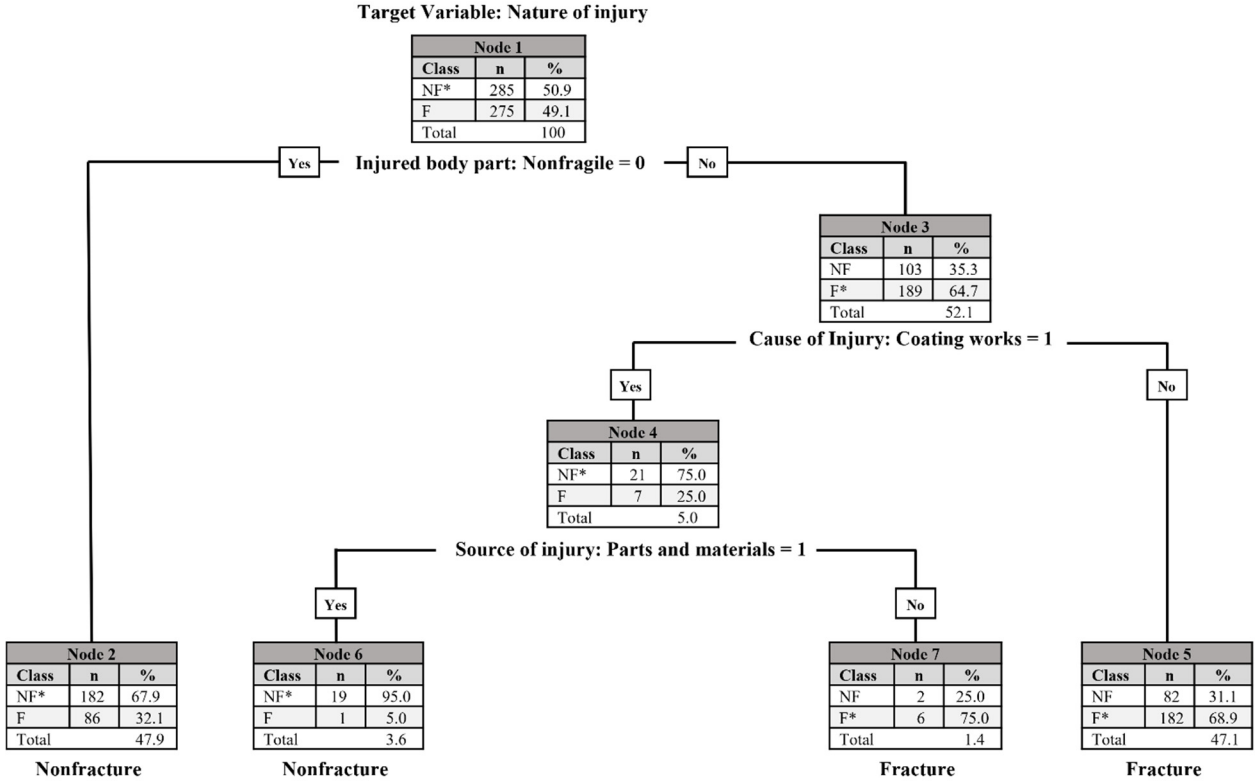

Figure 2. Decision tree for predicting *nature of injury* (* *nature of injury* represented by the node and with the higher proportion of observation). Labels follow the majority: Node 1 is labeled *non-fracture* (NF) since the proportion of NF is greater than that of F in node 1; node 5 is labeled *fracture* (F) since the proportion of F is greater than that of NF present in the leaf node 5.

As shown in Figure 2, the tree has a total of seven nodes, four of which are leaf nodes. The nodes represent split points, where decision criteria were used to classify the accident reports in the nodes into two classes: *fracture* (F) and *non-fracture* (NF). The number of accident reports classified as F and NF in each node is outlined under the number of observations (n) for F and NF, respectively. The sum of the accident reports classified as F and NF gives the total count of observations in that node.

The first-level splitting attribute is the *nonfragile* category for *injured body part*. The second-level factor is *coating works* (which includes asphalt, painting, and spraying) as a *cause of injury*, and the third-level factor is *source of injury*, which is split by *parts and materials* (a category including such building materials as solid elements; containers; electric parts; fasteners, connectors, ropes, and ties; furniture and fixtures; hoisting accessories; structural and nonstructural metal materials; roofing parts; and tars, sealants, caulking, and insulating material). The variable importance section below provides the list of important variables arranged in the order of their relevance; this ranking indicates the extent to which the variables contributed to predicting the *nature of injury* in the development of the decision tree model. Notably, the proposed decision tree in Figure 2 correctly classified 389 accident reports (69.46%) out of the 560 reports in the training dataset.

Generally, the travel starts from the root node and makes a branch decision at every node until the travel terminates at a leaf node or terminal node. For example, in Figure 2, splitting the root node by injured body part: nonfragile = 0 could be interpreted as determining whether "the injured body part is fragile," so if this statement is true (i.e., Yes), one traverses the decision tree by branching to the left, and if this statement is false (i.e., No), one traverses the model by branching to the right; here, a branch to the left is always a Yes-turn while a branch to the right is always a No-turn. Subsequently, decision node 3 splits by cause of injury: coating works = 1, which specifies whether "the cause of injury

is coating works." Hence, if this statement is true (i.e., Yes), one traverses the decision tree by branching to the left, and if it is false (i.e., No), one traverses the model by branching to the right. These terms were used in the following section for the decision rules.

The nodes in Figure 3 represent split points, where decision criteria were used to classify the accident reports in the nodes into two classes: *fracture* (F) and *non-fracture* (NF). The number of accident reports classified as F and NF in each node is outlined under the number of observations (*n*) for F and NF, respectively. The sum of the accident reports classified as F and NF gives the total count of incident reports in that node. The proportion of accident reports classified as F and NF in each node are also listed under the percentage of observations (%) for F and NF, respectively; notably, the *Total* % listed in each node represents the percentage of the 560 observations in the training set that sits within a given node. For example, in the first decision node (root node) in Figure 3, 285 NF and 275 F accident reports travel this node, so the number of observations sum to 560 (i.e., 285 + 275 = 560 accident reports), and the percentage of observations in the training set is 100%. Comparatively, decision node 5 (which is a leaf or terminal node) has 82 NF and 182 F accident reports, which sum to 264 (i.e., 82 + 182 = 264) observations—or about 47.1% of the 560 incident reports in the training dataset.

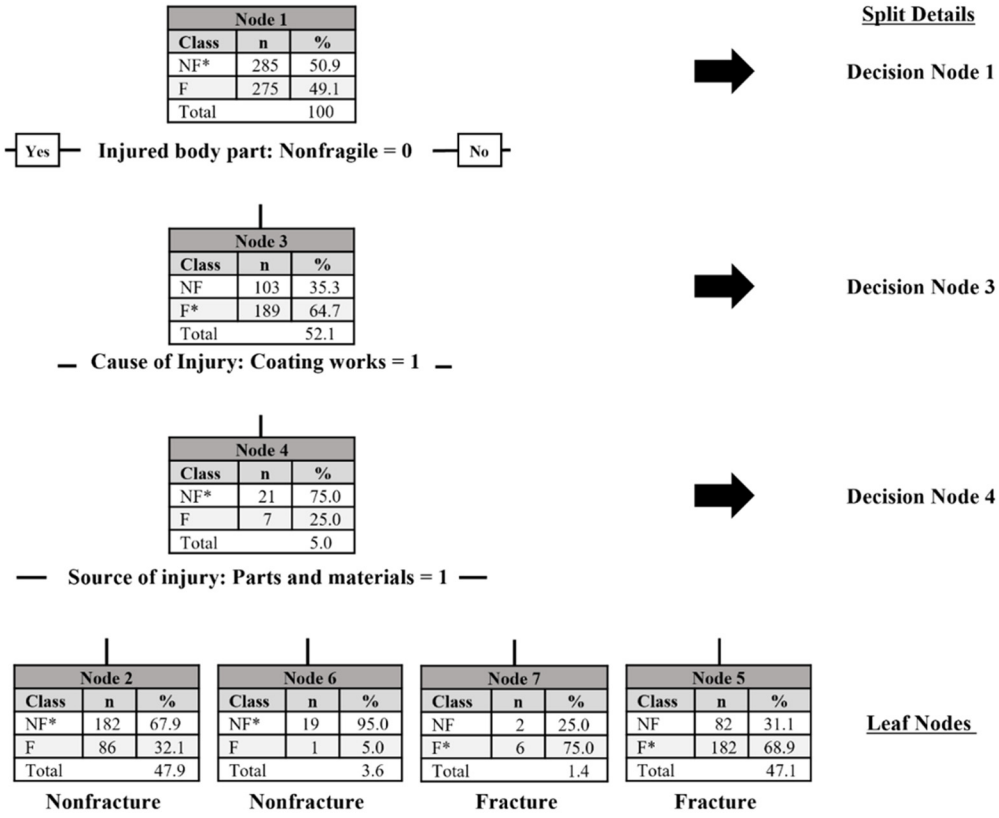

Figure 3. Explanatory notes on decision tree representations. Number (*n*) and percentage (%) of observations classified as *non-fracture* (NF) and *fracture* (F) in each node (* *nature of injury* represented by the node and with the higher proportion of observation). *Total* % signifies the percentage of the 560 training set observations that appear in each leaf node.

5.2.2. Decision Rules

Each leaf node forms a distinct decision rule. Hence, the full set of decision rules is developed with the use of all the leaf nodes in the decision tree. The decision rules formed from the decision tree analysis appear in Table 4. The decision rules are further explained in the Discussion Section.

Table 4. Decision rules are derived from the proposed decision tree model.

Node	Decision Rule
2	If the *injured body part* is *fragile*, then the *nature of the injury* is *non-fracture*
6	If the *injured body part* is *nonfragile*, the *cause of injury* is *coating works*, and the source of injury is *parts and materials*, then the *nature of the injury* is *non-fracture*
7	If the *injured body part* is *nonfragile*, the *cause of injury* is *coating works*, and the *source of injury* is not *parts and materials*, then the *nature of the injury* is *fracture*
5	If the *injured body part* is *nonfragile*, the *cause of injury* is not *coating works*, then the *nature of the injury* is *fracture*

5.2.3. Accuracy of Decision Tree Model

Before using a decision tree model to predict the *nature of injury* for roofing construction workers, it is essential to evaluate the prediction accuracy of the model. Therefore, we further evaluated the decision tree model's accuracy by applying the model to the testing dataset that was independent of the training dataset. The results are summarized in Table 5. The results suggest that the decision tree model in Figure 2 is reliable since it was 77.7% successful (accuracy = 0.777) when predicting the *nature of injury* classifications in the testing dataset. Another evaluation of the accuracy of the proposed decision tree model is the no-information rate. Table 5 shows that the accuracy of the model is higher than the no-information rate 95% of the time; in other words, there is a 95% chance that the true accuracy of the proposed decision tree model is between 69.9% and 84.3%. However, with the use of the proposed model, we obtained a 77.7% accuracy, which is significantly greater and better than the 51.1% accruing to chance. Hence, the proposed decision tree model offers significantly higher performance than the no-information rate, as buttressed by a p-value < 0.001. The kappa statistic is another assessment of the proposed model presented in Table 5. It reflects the extent to which the model prediction matches the actual classifications. The kappa for this decision tree model is 0.554, which is a moderate score [60].

Table 5. Overall accuracy evaluation of decision tree.

Evaluation Statistic	Result
Accuracy	0.777
95% Confidence Interval	(0.699, 0.843)
No Information Rate	0.511
p-Value [Acc > NIR]	<0.001
Kappa	0.554
Precision	0.768
Sensitivity/Recall	0.779
Specificity	0.775

The ROC curve in Figure 4 provides more evaluation of the proposed machine learning model in this study. The AUC is 0.775 (which is close to 1 and a good indication of a reliable model) [46] with balanced sensitivity and specificity values of 0.7794 and 0.7746, respectively. It is based on the results of the ROC plot and other model performance evaluation statistics such as Kappa, precision, and accuracy, and the optimum decision tree model was selected in this study.

5.2.4. Evaluation of Variable Importance

Analyzing the decision tree and variable importance scores reveals that the *injured body part: nonfragile* (root node), is the most important predictor attribute, followed by *cause of injury: coating works* in predicting the *nature of injury*. When traversing from top to bottom, these are the first two variables displayed in the decision tree. Additionally, out of the nine identified most important predictor variables, the two least important are *project end-use: residential* and *project type: new project*, as shown in Table 6. Even though these

two predictor variables did not appear to play a major role in the tree and did not show up in the proposed decision process, they were part of the predictor variables used in forming the tree. Such importance follows the surrogate logic described in the methodology because (for instance) neither the least important variable—*project type: new project*, with a variable importance score of 2%—nor the higher-importance variable *source of injury: structures* (13%) appear as splitters in the decision tree, both had a nonzero importance due to their function as surrogates for other splitting variables. Consequently, even lower-importance variables are worthy of note.

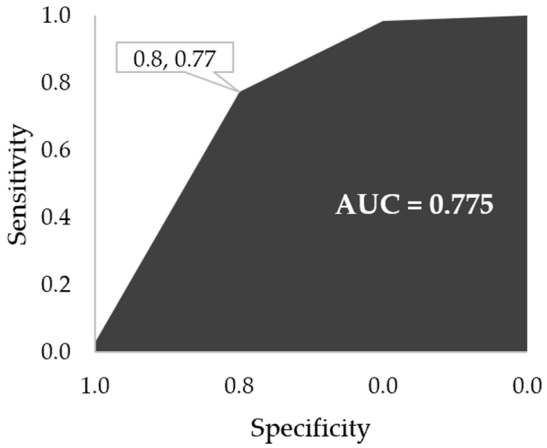

Figure 4. ROC curve and area under the ROC curve (AUC).

Table 6. Variable importance of the proposed decision tree model.

Variable/Attribute	Importance Score
Injured body part: Nonfragile	49
Cause of injury: Coating works	16
Source of injury: Structures	13
Source of injury: Parts and materials	9
Source of injury: Tools, instruments, and equipment	5
Project type: Maintenance or repair	2
Cause of injury: Repair and renovation	2
Project end-use: Residential	2
Project type: New Project	2

The authors wish to underscore that the software R (version 3.6.1) only reports the rounded variable importance scores, which it scales to sum up to 100%, and omits variables whose proportions are below 1%. Consequently, Table 6 summarizes the importance scores of the nine identified primary variables affecting the development of the decision tree model in Figure 2.

5.2.5. Cross-Validation Results

Table 7 gives the summary statistics of the ten-fold cross-validation. The main purpose of this cross-validation is to test the model's prediction or performance ability in practice when applied to new data not used previously in the training process. This process serves to check whether the proposed model, as in this study, generalizes well to new and independent data such as those (accident reports) obtainable in the roofing industry. Additionally, this cross-validation serves to check for overfitting and selection bias [61]. Notably, the prediction accuracy of the overall training set employed in this research is 0.69 (as shown by the leaf nodes in Figure 2) and is approximately equal to the mean prediction accuracy of the ten-fold cross-validation. Since these two accuracies are consistent, this validation indicates that the proposed model has an average fit and, hence, is not overfitted.

Table 7. Summary statistics of the ten-fold cross-validation.

Statistic	Accuracy	Kappa
Minimum	0.632	0.265
Median	0.676	0.348
Mean	0.689	0.378
Standard deviation	0.046	0.091
Maximum	0.768	0.536

We compared the output of the XGBoost package in R with the cross-validation results calculated without the XGBoost. The XGBoost results differed a little from the cross-validation results, as shown in Table 8. In fact, the decision tree model performed slightly better, illustrating the effectiveness of our model.

Table 8. Compared overall accuracy of the extreme gradient boosting and decision tree model.

Evaluation Statistic	Xgboost Result	Decision Tree Model Result
Accuracy	0.712	0.777
95% Confidence Interval	(0.629, 0.786)	(0.699, 0.843)
No Information Rate	0.511	0.511
p-Value [Acc > NIR]	<0.001	<0.001
Kappa	0.423	0.554
Precision	0.733	0.768
Sensitivity/Recall	0.647	0.779
Specificity	0.775	0.775

5.2.6. Testing and Training Dataset Accuracy Summary Statistics

To further confirm the robustness of our model, as previously explained in the methodology, the set.seed function in R randomly generated and shuffled the data based on seeds set consecutively from 1 to 50. Each of these 50 seeds then generated a decision tree, and the summary statistics of the outcomes are given in Table 9. As can be observed, the mean of the 50 prediction accuracies of the ten-fold cross-validation is 0.696 (i.e., the mean training accuracy = 0.696). The tree formed with each of the 50 randomizations was then applied each time to the *testing* dataset, and 50 testing accuracies were obtained and averaged. As shown in Table 9, the mean of the 50 testing/prediction accuracies is 0.704 (i.e., the mean testing accuracy = 0.704). These values are above average, acceptable, and consistent with other studies [14,62].

Notably, the training accuracy obtained from the decision tree model proposed in this investigation is 0.695, and the testing accuracy obtained from the proposed model is 0.777 (as seen in Table 5). Hence, the results of the decision tree model proposed in this study are similar and consistent with (and validated by) the 50 randomization results (as summarized in Table 10), signaling that the accuracy was not obtained by chance.

Table 9. Prediction accuracy summary of 50 decision trees from 50 randomizations of the dataset.

		Training Set		Testing Set	
S/N	Statistic	Cross-Validation Average Accuracy	Cross-Validation Average Kappa	Accuracy	Kappa
1	Mean	0.696	0.391	0.704	0.408
2	Standard Error	0.002	0.004	0.005	0.010
3	Median	0.696	0.392	0.698	0.396
4	Standard Deviation	0.013	0.025	0.035	0.070
5	Sample Variance	0.0002	0.0006	0.0012	0.0048
6	Kurtosis	−0.437	−0.428	0.098	0.122
7	Skewness	−0.242	−0.238	0.206	0.192
8	Range	0.054	0.107	0.158	0.319

Table 10. Comparison between the mean values of prediction accuracy for the 50 decision trees based on randomized data and the proposed decision tree model.

Model	Cross-Validation Average Accuracy (Training Set)	Cross-Validation Average Kappa (Training Set)	Accuracy (Testing Set)	Kappa (Testing Set)
Fifty randomizations/shuffles	0.696	0.391	0.704	0.408
Proposed decision tree model	0.689	0.378	0.777	0.554

6. Discussions

6.1. Frequency Analysis

The statistical analysis and findings of this research study are in line with previous studies [14,23,29–31] and revealed that 39% of occupational accidents among roofing contractors resulted in a fatality, and almost all roofing incidents occurred in building projects (99%)—with *residential buildings* (53%) being the prevalent *end-use*. Kang et al. [3] suggested that part of the higher proportion of fall incidents in residential projects might be due to the higher number of residential projects compared to non-residential projects in the past 10 to 15 years. However, another explanation could be that *residential buildings* are frequently constructed by small contractors who do not usually offer adequate PPE or formal safety training due to budget limitations [63]. Such uncertainty remains a study-worthy open question.

As far as project types are concerned, *maintenance or repair* projects contribute to the majority of roofing accidents (45%), followed by *alteration or rehabilitation* (26%), and *new projects or new addition* (25%). Comparing these ratios about roofing contractors to those about all contractors—as reported by Huang and Hinze [63]—indicates that roofers are involved in maintenance and repair work at significantly higher rates and involved in *new projects* at significantly lower rates. We also found that 92% of the maintenance projects in the data had budgets less than $250,000, whereas this budget occurred for *new projects* only 31% of the time. In other words, maintenance projects appear to be much smaller than *new projects* and, therefore, may have a smaller safety budget and/or be supervised by individual clients who may not have the capacity to monitor safety practices. Whatever the fundamental cause, this finding signifies the importance of considering the project context when analyzing construction accidents.

The authors also found that "misjudging the hazardous situation" occurred at a higher level (37%) in *maintenance/repair projects* than in *new projects* (25%), an outcome that may be due to the temporary nature of maintenance/repair projects since workers may not have enough time to explore the construction site to understand potential unsafe conditions. A similar finding in a study by Mistikoglu et al. [14] revealed that project types involving alteration, rehabilitation, maintenance, and demolition resulted in more fatalities among roofing contractors than new projects or new additions. Perhaps for such reasons, an investigation by Kim [64] recommended that during the design stage of new projects, the construction design and management team should draw up a health and safety plan and hand it over to the project owner to prepare for such future work such as maintenance, repair, alteration, and rehabilitation. Such a recommendation is interesting since maintenance or repair projects are strongly related to design. Some of these design plans could include the design and construction of a safe passage to the roof to enable safe access to the roof during the execution of future works. Another design regulation could include placing the electrical control boxes at a reachable/reduced level for easy access, which would reduce the need to work on ladders during future work.

This study's results show that projects with lower costs (i.e., less than $50,000) tend to have a higher frequency of occupational accidents (57%) than projects with higher costs (e.g., 7% for $250,000 to $500,000). This result is almost identical to the proportion of fall accidents in low-cost projects reported by Huang and Hinze [63] and Kang et al. [3] and underscores the importance of appropriate safety training, culture, supervision, and

equipment in preventing construction accidents, especially among smaller companies. To address this issue, Siddiqui [65] recommended stricter enforcement of regulations and standards to control the non-compliance problem among low-budget residential projects to improve the overall safety among small specialty contractors.

With respect to the *source of injury*, a large portion of accidents have occurred in *structures and surfaces* (56%), which may be unsurprising as this category involves all roofs and roof skylights/openings—primary locations for fall accidents, as mentioned by Kang et al. [3]. Previous studies have examined methods that can reduce the risk of fall accidents near skylights and roof openings. Bobick et al. [66] have introduced an injury reduction matrix to establish organizational duties and work requirements during three construction phases (i.e., pre-construction, construction, and post-construction) to ensure proper management of fall-protection programs near skylights and roof openings; the authors recommended that the building owner "must implement reporting requirements during the construction phase" (p. 36) and "an employee of the owner or a consultant should monitor effectiveness of the safety and health program during construction" (p. 36). Implementing this small requirement could have a considerable effect, particularly in small maintenance projects. Similarly, Prevention through Design (Ptd) protocols, such as those conducted by Ho et al. [67] for solar panel installers, may help small roofing contractors understand the hazardous situations near roof openings and introduce some easy-to-implement solutions to reduce the risks of working near these hazards. Furthermore, new technologies such as building information modeling (BIM) can be adapted to better design and implement fall-protection systems on roofs, scaffolds, and near skylights [68]. For instance, Zhang et al. [69] have implemented a rule-based engine on a commercial BIM program to automatically check OSHA's rules and regulations on a building's 3D model and inform construction and safety managers about safety measures that are required to avoid fall-related accidents. Such efforts may prove particularly helpful in new projects where developing 3D models during the design phase is a prevalent practice.

The event type *fall (to a lower level)* represents the highest portion (83%) of accidents in our data, and *exposure to electricity* accidents generates the highest fatality ratio (76%) among the event types. A high number of falls among roofers is anticipated, as their tasks mainly involve working on elevated structures. In a report on workers' health, the National Institute for Occupational Safety and Health (NIOSH) disclosed that for roofers, the fatal fall rate was seven times the average rate for all construction [70], which is mainly due to the higher elevations at which roofers work as compared to other trades. However, of the 24 fatal *exposures to electricity*, 20 of them were caused by contacting a power line, and in 10 out of these 20 cases, an aluminum ladder contacted the line. Such data suggest using fiberglass ladders instead of aluminum ones and keeping a safe distance from powerlines could have easily saved the lives of these workers and may represent a best practice for similar situations. There were also instances of touching energized metal parts on transformers and roof parapets, which could have been avoided by adding a best practice to make sure the building's electrical systems are turned off before starting a job. These opportunities raise examples of why studying accident reports for all types of event types may lead to better safety practices across industries.

While *fractures* occurred nearly half the time (49%), in terms of the nature of the injuries, *concussions* contributed the most to the fatal injuries (81%) of the cases considered in this study. These two natures of injury, which are also the most common types in Huang and Hinze [63] study, are highly correlated with falls: 88 of 90 *concussions* (98%), and 323 of 335 *fractures* (96%) in this study were caused by a fall accident. This result is in accordance with Betsis et al.'s [71] correlation analysis, which found fall accidents have the highest correlation with *fractures*, *dislocations*, and *concussions*. Designing more effective PPEs and including further protections in the form of personal fall-arrest systems should help reduce the frequency of fracture injuries. Additionally, thorough fall prevention training is highly recommended [63] because providing workers with effective training may significantly reduce at-risk behaviors. Traditional safety training that is limited to verbal

and written explanations of OSHA requirements may not be enough to teach workers how to identify and eliminate all fall hazards [63]. Innovative training methods must be considered and carefully examined in order to help reduce the frequency of concussions and fracture injuries.

Injured body parts involving the *head* represented the largest number of injuries (34%) and the highest fatality ratio (69%), indicating how fall from height can critically result in fatality. This further shows that, in addition to safety helmets, workers need to be protected by preventive fall-protection systems and receive adequate training on their use in order to prevent serious injuries among roofers as they usually work in higher elevations. This study also discovered that accidents happened mostly on *Mondays* (21%), with a lower frequency of accidents happening during the rest of the week and weekend. One needs to interpret this result with caution as the higher ratios of accidents among roofing contractors on Mondays may be due to the "Monday Effect" [72,73]: some injuries during the weekend may not be reported until Monday. Having toolbox talks, particularly at the start of the week or after any other closings due to holidays or inclement weather, may help raise workers' awareness of safety risks on a job site, particularly among small contractors with limited resources and less-experienced workers [74]. In particular, a toolbox safety meeting that has been tailored to the specific conditions of working at elevation and considers the distinct hazards of a building may improve the safety performance of roofing contractors. For instance, Olson et al. [75] conducted three case studies to evaluate different construction toolbox talks on fatalities. The results showed that brief, scripted toolbox talks could help supervisors share historical incidents and explain suggested prevention methods. Adding action items, discussion prompts, and line drawings also could improve workers' understanding of the material.

6.2. Decision Tree Analysis

The research illustrated the possibility of analyzing the nature of occupational injury of roofing contractors with the use of the data-mining method referred to as decision trees. The algorithm successfully classified the data into groups of the target attribute (*nature of injury*) and took note of both the associations between project information (predictor attributes) and the level of importance assigned to these attributes.

This study presents decision rules to reveal some associations that exist between the attributes of roofers' occupational injuries. Four decision rules that correspond with the four root nodes of the proposed decision tree model were formed and appear in Table 4. For example, Rule 1 (that corresponds with leaf node 2 in Figure 2) states that if the *injured body part* is *fragile (e.g., head/neck and body system),* then the *nature of the injury* is *non-fracture*. This rule suggests that injuries involving the *fragile parts of the body*, like the *head/neck* and *body system*, will most likely (about 67.91% of the time) not lead to a *fracture*. As illustrated by the decision rules presented in this study, most of the injuries involving the *nonfragile body parts* (e.g., *lower/upper extremities*, *trunk*, and *multiple body parts*) lead to a *fracture*. The variable importance ranking in Table 6 also identifies *nonfragile injured body parts* as the most important attribute in predicting the nature of roofers' injury on a job site. It is, therefore, essential and wise to emphasize and recommend personal protective equipment for roofing contractors [76,77], particularly those that are worn around the nonfragile body parts, as a strict safety measure during roof construction. It is also necessary to ensure that workers are adequately trained on the correct use of personal protective equipment on the job so that they can carry out their work safely. Furthermore, innovative training methods should be considered and carefully examined [20,78,79].

Leaf node 5 (which corresponds with decision rule 4) explains that if the injured body part is nonfragile and the cause of injury is not coating works, then the injury is likely to involve a fracture (with an accuracy of 68.94%). This rule is because the roofers' accident reports analyzed in this study recorded more fractures in the nonfragile body parts for all work other than coating works. Additionally, 47.1% of the 560 records in the training dataset fall under this category (and were grouped into this class by our model). In other

words, according to the proposed decision tree model provided in this research, 47.1% of the accident reports in the training dataset led to a fracture that involved the nonfragile body parts and was not from activities involving coating works. Furthermore, the second most important attribute listed in Table 6 is coating works as a cause of injury. This intersection indicates that even though safety measures (especially towards the nonfragile body parts) are important in all works carried out by roofers, deficiencies in safety procedures are more pronounced and detrimental in such roofing activities as assembling/installing roofing elements, dismantling/demolition, preparation, metal works, repair and renovation, and transporting/handling materials than in coating works.

Leaf node 6 (which corresponds with decision rule 2) indicates that if the *injured body part* is *nonfragile,* the *cause of injury* is *coating works,* and the *source of injury* is *parts and materials,* then it may not lead to an injury involving a *fracture* (with a 95% accuracy). 3.6% of the 560 records in the training dataset fall under this category. Decision rule 2 corresponds with decision rule 4 above, except that here, the injury involves *parts and materials* as the source of injury. However, *parts and materials* are also some of the most important predictor attributes in the variable importance ranking listed in Table 6. This observation highlights the need to encourage roofers to pay close attention to their environment when working on a job site and when moving parts and materials. It is also important to depower electric parts and equipment and adhere to lockout/tagout procedures when they are idle or under repair [80]. Additionally, accidents can also be prevented on roofing job sites with the employment of a competent person who can assist in identifying and eliminating potential surrounding hazardous situations, components, parts, or materials [81].

7. Limitations

There are some limitations in this research that need to be mentioned. First, unfortunately, the OSHA IMIS database used in this study is biased toward more severe—although non-fatal—accidents as minor injuries are not usually reported in most public databases. Future studies should include minor incidents or even near-misses to provide a more comprehensive understanding of safety incidents among roofing contractors. Second, the authors have classified the degree of injuries into three groups (fatality, hospitalized, and non-hospitalized); future studies can incorporate the cost of injuries or the number of days away from work for non-fatal cases to better quantify the severity of accidents. Third, the sample size of the database used in this research can be increased in future studies. Additionally, future extensions of this research may include the use of other data-mining techniques (e.g., support vector machine, random forest, regression analysis) to discover hidden patterns. Furthermore, future studies could explore whether the results of the study will vary extensively from region to region based on practices and level of training. Moreover, future studies could examine other factors that could be potential confounding variables—such as safety budgets, experience and education levels of the workers, protection measures, and environmental and human factors—to explore more trends visible in incident outcomes. Unfortunately, these factors were unavailable in OSHA's database and could not be investigated in this research. Such continuing studies would provide insight to help practitioners better understand the contributing factors leading to roofers' occupational injuries to reduce the frequency, severity, and risk of accidents.

8. Conclusions

Roofing contractors encounter a variety of dangerous circumstances in their line of work, such as working from heights, which results in a high rate of serious injuries and fatalities. There has been an increase in workplace deaths among roofing contractors in recent years [77]. An initial step in reducing the risks of construction incidents in this trade is to find statistically significant relationships between these hazardous situations and a few well-defined contributory factors. Despite its significance, little is known about the factors that contribute to the incidence of occupational accidents among roofing contractors. To address this gap in knowledge, the main objective of the current study is to determine

which contributing factors affect occupational accidents among roofers and to determine these factors' relative importance when predicting the nature of occupational injuries (both fatal and non-fatal). To achieve this objective, this study analyzed roofing accident reports from OSHA's IMIS accidents database using both descriptive statistics and data-mining techniques.

It was revealed that *fall to the lower level* (83%) was the major cause of both fatal and non-fatal accidents in roofing projects. Additionally, the majority of these accidents occurred in *residential buildings*, *maintenance and repair works*, *small projects* (i.e., $50,000 or less), and on *Mondays*. Furthermore, the most frequent injuries recorded by roofers were *fractures* (49%) and *concussions* (15%), while the most injured body part was the *head* (34%). Moreover, the three main sources of injury were *structures and surfaces* (56%), *tools, instruments, and equipment* (19%), and *parts and materials* (16%). Hence, the overall fatality ratios and costs associated with incidents can be reduced by enforcing the appropriate use of PPE, especially the effective implementation of fall protection systems among small residential roofing construction firms.

We also proposed a decision tree model for predicting the nature of occupational injuries that could result from an incident during a roofing project and examined the underlying association that exists between attributes. This work analyzed whether the *nature of injury* (fracture or non-fracture) among roofing contractors could be predicted for roofing accidents using limited project information (such as the *source of injury*, *cause of injury*, *project cost*, *project end-use*, *project type*, *day of injury*, and *injured body part*). The results of this analysis successfully predict the nature of occupational injuries due to a roofing accident and highlight which factors are the most important in determining the *nature of injury*. Such a model will help safety managers better understand the underlying associations between a limited number of project attributes and determine how to mitigate such associations to alleviate the severity of the injuries among roofers.

The research findings identified *nonfragile injured body parts* as the most important factor for predicting the nature of roofers' injuries. This outcome is followed by *coating works* as a *cause of injury*, which is the second most important predictor variable. The least important factor in the presented variable importance score table is the *new project type*. Four decision rules that explain the associations between attributes/variables are also given in this analysis, most significantly that most of the injuries involving *nonfragile body parts* (e.g., *lower/upper extremities*, *trunk*, and *multiple body parts*) resulted in a *fracture*. This association highlights the need to strictly enforce the use of PPEs, particularly those that are worn around the nonfragile body parts, and to train roofing construction workers on the correct use of PPEs. Since roofers typically work from heights, it is also essential for safety managers to incorporate PPEs with preventive fall protection systems to mitigate the risk of occupational injuries on roofers' job sites. Furthermore, workers' deaths and injuries can also be prevented by implementing other methods, such as fabricating roofing components on the ground before lifting them up for installation on the roof to reduce the risk of exposure to working at height.

Finally, it is notable that since the accident databases used in this study are from the United States of America, the readers should be cautious about generalizing findings to geographical locations that have limited similarity to the construction industry in the United States. Nonetheless, the proposed analysis approach can be used to analyze similar accident report databases in the construction industry and beyond.

Author Contributions: All authors contributed to the idea and concept of this study. Conceptualization, I.S.O. and B.E.; methodology, I.S.O.; software, I.S.O.; validation, I.S.O. and B.E.; formal analysis, I.S.O.; investigation, I.S.O., P.G., G.L. and B.E.; resources, B.E.; data curation, I.S.O. and G.L.; writing—original draft preparation, I.S.O. and B.E.; writing—review and editing, I.S.O. and B.E.; visualization, I.S.O. and B.E.; supervision, B.E.; project administration, B.E.; funding acquisition, B.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation (NSF) through the Decision, Risk, and Management Sciences (DRMS) program, grant number 2049842. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Publication of this article was funded in part by Purdue University Libraries Open Access Publishing Fund.

Data Availability Statement: All data, models, or code generated or used during the study are available from the corresponding author by request.

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

References

1. Webster, T. Workplace Falls. In *Compensation and Working Conditions*; Bureau of Labor Statistics: Washington, DC, USA, 2000; pp. 28–38.
2. Aroke, O.M.; Onuchukwu, I.S.; Esmaeili, B.; Flintsch, A.M. Countermeasures to Reduce Truck-Mounted Attenuator (TMA) Crashes: A State-of-the-Art Review. *Future Transp.* **2022**, *2*, 425–452. [CrossRef]
3. 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]
4. Choi, S.D.; Griinke, D.; Lederer, M. Fall protection equipment effects on productivity and safety in residential roofing construction. *J. Constr. Res.* **2006**, *7*, 149–157. [CrossRef]
5. Fredericks, T.K.; Abudayyeh, O.; Choi, S.D.; Wiersma, M.; Charles, M. Occupational Injuries and Fatalities in the Roofing Contracting Industry. *J. Constr. Eng. Manag.* **2005**, *131*, 1233–1240. [CrossRef]
6. Dong, X.S.; Choi, S.D.; Borchardt, J.G.; Wang, X.; Largay, J.A. Fatal falls from roofs among U.S. construction workers. *J. Saf. Res.* **2013**, *44*, 17–24. [CrossRef]
7. Moore, J.R.; Wagner, J.P. Fatal events in residential roofing. *Saf. Sci.* **2014**, *70*, 262–269. [CrossRef]
8. Bellamy, L.J.; Ale, B.J.M.; Geyer, T.A.W.; Goossens, L.H.J.; Hale, A.R.; Oh, J.; Mud, M.; Bloemhof, A.; Papazoglou, I.A.; Whiston, J.Y. Storybuilder—A tool for the analysis of accident reports. *Reliab. Eng. Syst. Saf.* **2007**, *92*, 735–744. [CrossRef]
9. Ale, B.J.M.; Bellamy, L.J.; Baksteen, H.; Damen, M.; Goossens, L.H.J.; Hale, A.R.; Mud, M.; Oh, J.; Papazoglou, I.A.; Whiston, J.Y. Accidents in the construction industry in the Netherlands: An analysis of accident reports using Storybuilder. *Reliab. Eng. Syst. Saf.* **2008**, *93*, 1523–1533. [CrossRef]
10. Esmaeili, B. Identifying and Quantifying Construction Safety Risks at the Attribute Level. Ph.D. Dissertation, University of Colorado at Boulder, Boulder, CO, USA, 2012.
11. Kines, P. Construction workers' falls through roofs: Fatal versus serious injuries. *J. Saf. Res.* **2002**, *33*, 195–208. [CrossRef]
12. Baradan, S.; Usmen, M.A. Comparative Injury and Fatality Risk Analysis of Building Trades. *J. Constr. Eng. Manag.* **2006**, *132*, 533–539. [CrossRef]
13. Sa, J.; Seo, D.C.; Choi, S.D. Comparison of risk factors for falls from height between commercial and residential roofers. *J. Saf. Res.* **2009**, *40*, 1–6. [CrossRef] [PubMed]
14. Mistikoglu, G.; Gerek, I.H.; Erdis, E.; Mumtaz Usmen, P.E.; Cakan, H.; Kazan, E.E. Decision tree analysis of construction fall accidents involving roofers. *Expert. Syst. Appl.* **2015**, *42*, 2256–2263. [CrossRef]
15. Gholizadeh, P.; Esmaeili, B. Applying Classification Trees to Analyze Electrical Contractors' Accidents. In Proceedings of the Construction Re-Search Congress, San Juan, Puerto Rico, 31 May–2 June 2016.
16. Gholizadeh, P.; Onuchukwu, I.S.; Esmaeili, B. Trends in Catastrophic Occupational Incidents among Electrical Contractors, 2007–2013. *Int. J. Environ. Res. Public. Health* **2021**, *18*, 5126. [CrossRef] [PubMed]
17. Chou Philip, A. Optimal partitioning for classification and regression trees. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 340–354. [CrossRef]
18. Chen, Y.L.; Hsu, C.L.; Chou, S.C. Constructing a multi-valued and multi-labeled decision tree. *Expert Syst. Appl.* **2003**, *25*, 199–209. [CrossRef]
19. Piramuthu, S. Input data for decision trees. *Expert. Syst. Appl.* **2008**, *34*, 1220–1226. [CrossRef]
20. Li, W.; Huang, H.; Solomon, T.; Esmaeili, B.; Yu, L.F. Synthesizing Personalized Construction Safety Training Scenarios for VR Training. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 1993–2002. [CrossRef] [PubMed]
21. Hatipkarasulu, Y. Project level analysis of special trade contractor fatalities using accident investigation reports. *J. Saf. Res.* **2010**, *41*, 451–457. [CrossRef]
22. Bobick, T.G. Falls through Roof and Floor Openings and Surfaces, Including Skylights: 1992–2000. *J. Constr. Eng. Manag.* **2004**, *130*, 895–907. [CrossRef]
23. Cheng, C.; Leu, S.; Cheng, Y.; Wu, T.C.; Lin, C.C. *Applying Data Mining Techniques to Explore Factors Contributing to Occupational Injuries in Taiwan's Construction Industry*; Elsevier: Amsterdam, The Netherlands, 2012.
24. Halabi, Y.; Xu, H.; Long, D.; Chen, Y.; Yu, Z.; Alhaek, F.; Alhaddad, W. Causal factors and risk assessment of fall accidents in the U.S. construction industry: A comprehensive data analysis (2000–2020). *Saf. Sci.* **2022**, *146*, 105537. [CrossRef]
25. Suruda, A.; Fosbroke, D.; Braddee, R. Fatal work-related falls from roofs. *J. Saf. Res.* **1995**, *26*, 1–8. [CrossRef]

26. Standard Industrial Classification [SIC CODE]. SIC Code 1761 Roofing, Siding, and Sheet Metal Work 2019. Available online: https://www.naics.com/sic-industry-description/?code=1761 (accessed on 19 May 2023).
27. Stern, F.; Ruder, A.; Chen, G. Proportionate mortality among unionized roofers and waterproofers. *Am. J. Ind. Med.* **2000**, *37*, 478–492. [CrossRef]
28. Li, H.; Ren, X.; Yang, Z. Data-driven Bayesian network for risk analysis of global maritime accidents. *Reliab. Eng. Syst. Saf.* **2023**, *230*, 108938. [CrossRef]
29. Rivas, T.; Paz, M.; Martín, J.E.; Matías, J.M.; García, J.F.; Taboada, J. Explaining and predicting workplace accidents using data-mining techniques. *Reliab. Eng. Syst. Saf.* **2011**, *96*, 739–747. [CrossRef]
30. Nenonen, N. Analysing factors related to slipping, stumbling, and falling accidents at work: Application of data mining methods to Finnish occupational accidents and diseases statistics database. *Appl. Ergon.* **2013**, *44*, 215–224. [CrossRef]
31. Amiri, M.; Ardeshir, A.; Hossein, M.; Zarandi, F.; Soltanaghaei, E. Pattern extraction for high-risk accidents in the construction industry: A data-mining approach. *Int. J. Inj. Control. Saf. Promot.* **2016**, *23*, 264–276. [CrossRef]
32. 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]
33. Sarkar, S.; Maiti, J. Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. *Saf. Sci.* **2020**, *131*, 104900. [CrossRef]
34. Li, X.; Zhu, R.; Ye, H.; Jiang, C.; Benslimane, A. MetaInjury: Meta-learning framework for reusing the risk knowledge of different construction accidents. *Saf. Sci.* **2021**, *140*, 105315. [CrossRef]
35. Kassambara, A. Machine Learning Essentials: Practical Guide in R. 2013. Available online: https://www.amazon.com/Machine-Learning-Essentials-Practical-Guide/dp/1986406857 (accessed on 18 May 2023).
36. Bureau of Labor Statistics (BLS). *Occupational Injury and Illness Classification Manual (OIICM)*; BLS: Washington, DC, USA, 2012.
37. Neuendorf, K.A. *The Content Analysis Guidebook*; SAGE: New York, NY, USA, 2022.
38. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*; SAGE: New York, NY, USA, 2004.
39. Villanova, M.P. Attribute-Based Risk Model for Assessing Risk to Industrial Construction Tasks. Ph.D. Dissertation, University of Colorado at Boulder, Boulder, CO, USA, 2014.
40. Desvignes, M. Requisite Empirical Risk Data for Integration of Safety with Advanced Technologies and Intelligent Systems. 2014. Available online: https://www.semanticscholar.org/paper/Requisite-empirical-risk-data-for-integration-of-Desvignes/f4432ee2481b5f2ccfb529d99d9e790b0a713ddd (accessed on 18 May 2023).
41. Therneau, T.; Atkinson, B.; rpart: Recursive Partitioning and Regression Trees. R Package Version 4.1-15. 2019. Available online: https://cran.r-project.org/web/packages/rpart/rpart.pdf (accessed on 19 May 2023).
42. Kuhn, M.; caret: Classification and Regression Training. R Package Version 6.0-85n. 2020. Available online: https://cran.r-project.org/web/packages/caret/caret.pdf (accessed on 18 May 2023).
43. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
44. Gholizadeh, P.; Esmaeili, B.; Memarian, B. Evaluating the performance of machine learning algorithms on construction accidents: An application of ROC curves. In *Construction Research Congress*; American Society of Civil Engineers: New Orleans, LA, USA, 2018; pp. 8–18.
45. Zhang, X.; Akber, M.Z.; Zheng, W. Predicting the slump of industrially produced concrete using machine learning: A multiclass classification approach. *J. Build. Eng.* **2022**, *58*, 104997. [CrossRef]
46. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef] [PubMed]
47. Lipinski, J. Tutorial: How to Assess Model Accuracy. *HR Anal.* **2017**, *101*, 5–7.
48. Luo, L.; Zhang, X.; Peng, H.; Lv, W.; Zhang, Y. A new pruning method for decision tree based on structural risk of leaf node. *Neural Comput. Appl.* **2013**, *22* (Suppl. S1), 17–26. [CrossRef]
49. Mingers, J. An Empirical Comparison of Pruning Methods for Decision Tree Induction. *Mach. Learn.* **1989**, *4*, 227–243. [CrossRef]
50. Quinlan, J.R. Simplifying decision trees. *Int. J. Man-Mach. Stud.* **1987**, *27*, 221–234. [CrossRef]
51. Wang, X.; Chen, B.; Qian, G.; Ye, F. On the optimization of fuzzy decision trees. *Fuzzy Sets Sys.* **2000**, *112*, 117–125. [CrossRef]
52. Fürnkranz, J. Pruning Algorithms for Rule Learning. *Mach. Learn.* **1997**, *27*, 139–172. [CrossRef]
53. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. Classification and Regression Trees. *Biometrics* **1984**, *40*, 874.
54. Loh, W.-Y.; Zhou, P. Variable importance scores. *J. Data Sci.* **2021**, *19*, 569–592. [CrossRef]
55. Kazemitabar, S.J.; Amini, A.A.; Bloniarz, A.; Talwalkar, A. Variable Importance using Decision Trees. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
56. Steinberg, D. What Is the Variable Importance Measure? Salford Systems—Dan Steinberg's Blog. 2018. Available online: https://www.salford-systems.com/blog/dan-steinberg/what-is-the-variable-importance-measure (accessed on 18 May 2023).
57. Steinberg, D. CART: Classification and Regression Trees. In *The Top Ten Algorithms in Data Mining*; Chapman and Hall/CRC Press, Taylor & Francis Group: Boca Raton, FL, USA, 2009; pp. 193–216.
58. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H. *Xgboost: Extreme Gradient Boosting*; R Package Version 1.0.0.1; The Comprehensive R Archive Network: Vienna, Austria, 2020.
59. Tang, J.; Zheng, L.; Han, C.; Liu, F.; Cai, J. Traffic Incident Clearance Time Prediction and Influencing Factor Analysis Using Extreme Gradient Boosting Model. *J. Adv. Transp.* **2020**, *2020*, 6401082. [CrossRef]

60. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Medica* **2012**, *22*, 276–282. [CrossRef]
61. Cawley, G. On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
62. Lakshmi, T.; Martin, A.; Begum, R.M. An analysis on performance of decision tree algorithms using student's qualitative data. *Int. J. Mod. Educ. Comput. Sci.* **2013**, *5*, 18. [CrossRef]
63. Huang, X.; Hinze, J. Analysis of Construction Worker Fall Accidents. *J. Constr. Eng. Manag.* **2003**, *129*, 262–271. [CrossRef]
64. Kim, K. *Statistical Approach to Design for Fall Prevention in Construction*; Wayne State University: Detroit, MI, USA, 2015.
65. Siddiqui, S.U.S. Construction Worker Fall Accidents: Their Causes and Influential Factors. Master's Thesis, FIU Electron, Miami, FL, USA, 2014; p. 27. Available online: https://digitalcommons.fiu.edu/etd/1157 (accessed on 19 May 2023).
66. Bobick, T.G.; Stanevich, R.L.; Pizatella, T.J.; Keane, P.R.; Smith, D.L. Preventing falls through skylights and roof openings. *Prof. Saf.* **1994**, *39*, 33.
67. Ho, C.; Lee, H.W.; Gambatese, J.A. Application of Prevention through Design (PtD) to improve the safety of solar installations on small buildings. *Saf. Sci.* **2020**, *125*, 104633. [CrossRef]
68. McGraw-Hill Construction. *Safety Management in the Construction Industry: Identifying Risks and Reducing Accidents to Improve Site Productivity and Project ROI*; McGraw-Hill Construction: Bedford, MA, USA, 2013.
69. Zhang, S.; Teizer, J.; Lee, J.K.; Eastman, C.M.; Venugopal, M. Building Information Modeling (BIM) and Safety: Automatic Safety Checking of Construction Models and Schedules. *Autom. Const.* **2013**, *29*, 183–195. [CrossRef]
70. NIOSH (National Institute for Occupational Safety and Health). Worker Health Chartbook. *Anal. Bioanal. Chem.* **2004**, *379*, 368–374.
71. Betsis, S.; Kalogirou, M.; Aretoulis, G.; Pertzinidou, M. Work Accidents Correlation Analysis for Construction Projects in Northern Greece 2003–2007: A Retrospective Study. *Safety* **2019**, *5*, 33. [CrossRef]
72. Card, D.; McCall, B.P. Is workers' compensation covering uninsured medical costs? Evidence from the "monday effect". *Ind. Labor. Relations. Rev.* **1996**, *49*, 690–706.
73. Campolieti, M.; Hyatt, D.E. Further Evidence on the "Monday Effect" in Workers' Compensation. 2006. Available online: https://www.jstor.org/stable/25067541 (accessed on 18 May 2023).
74. Eggerth, D.E.; Keller, B.M.; Cunningham, T.R.; Flynn, M.A. Evaluation of toolbox safety training in construction: The impact of narratives. *Am. J. Ind. Med.* **2018**, *61*, 997–1004. [CrossRef]
75. Olson, R.; Varga, A.; Cannon, A.; Jones, J.; Gilbert-Jones, I.; Zoller, E. Toolbox talks to prevent construction fatalities: Empirical development and evaluation. *Saf. Sci.* **2016**, *86*, 122–131. [CrossRef]
76. Olbina, S.; Hinze, J.; Rube, M. Safety in Roofing: Practices of Contractors That Employ Hispanic Workers. *Prof. Saf.* **2011**, *56*, 44–52.
77. Onuchukwu, I.S.; Esmaeili, B. Random Forest Analysis of Occupational Accident Reports among Roofing Contractors. In *Construction Research Congress 2022: Health and Safety, Workforce, and Education—Selected Papers from Construction Research Congress*; American Society of Civil Engineers: Arlington, VA, USA, 2022; pp. 561–570. Available online: https://ascelibrary.org/doi/10.1061/9780784483985.057 (accessed on 26 May 2023).
78. Li, W.; Esmaeili, B.; Yu, L.F. Simulating Wind Tower Construction Process for Virtual Construction Safety Training and Active Learning. In Proceedings of the 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Christchurch, New Zealand, 12–16 March 2022; pp. 369–372.
79. Lee, K.; Hasanzadeh, S.; Esmaeili, B. Assessing hazard anticipation in dynamic construction environments using multimodal 360-degree panorama videos. *J. Manag. Eng.* **2022**, *38*, 04022039. [CrossRef]
80. Occupational Safety and Health Administration. *Control of Hazardous Energy: Lockout/Tagout (OSHA 3120 (Revised))*; Occupational Safety and Health Administration: Washington, DC, USA, 2002.
81. Reese, C.D.; Eidson, J.V. *Handbook of OSHA Construction Safety and Health*; CRC Press: Boca Raton, FL, USA, 2006. Available online: https://www.taylorfrancis.com/books/9781420006230 (accessed on 19 May 2023).