Optimized Big Data Analytics for Health and Safety Hazards Prediction in Power Infrastructure Operations

Abstract

Forecasting imminent accidents in power infrastructure projects require a robust and accurate prediction model to trigger a proactive strategy for risk mitigation. Unfortunately, getting ready-made machine learning algorithms to eliminate redundant features optimally is challenging, especially if the parameters of these algorithms are not tuned. In this study, a particle swarm optimization is proposed both for feature selection and parameters tuning of the gradient boosting machine technique on 1,349,239 data points of an incident dataset. The predictive ability of the proposed method compared to conventional tree-based methods revealed near-perfect predictions of the proposed model on test data (classification accuracy - 0.878 and coefficient of determination - 0.93*) for the two outcome variables ACCIDENT and INJURYFREQ.* The high predictive power obtained reveals that injuries do not occur in a chaotic fashion, but that underlying patterns and trends exist that can be uncovered and captured via machine learning when applied to sufficiently large datasets. Also, key relationships identified will assist safety managers to understand possible risk combinations that cause accidents; helping to trigger proactive risk mitigation plans.

Keywords: Big Data analytics, Particle swarm optimization, Power infrastructure, Safety management.

1. Introduction

Construction is a high-risk industry, very intricate, and a complicated environment (Le, Lee, and Park 2014). Thus, promoting safety at construction sites is crucial for this industry, especially the Power Transmission and Distribution (PT&D) or Power Infrastructure domain. Workers involved in constructing PT&D lines and related infrastructure are at high risk of occupational hazards (e.g., severe burns, musculoskeletal disorders, and deaths) (Albert and Hallowell 2013). Also, an occurrence of a fatal injury leading to the demise of an employee during project execution, for instance, may adversely affect profit because of delays, rescheduling of new resources, and increased severity cost. Beyond time and cost, there is also a reputation problem that could have an impact on winning new bids in the future. Investments in power infrastructure will continue to grow with increased electricity demand since most developed countries are making efforts to cut their carbon emissions. This demand will significantly impact investments in the national electricity transmission and distribution systems. Power infrastructure and contracting companies need to consider strategies to reduce the frequency and severity of health and safety risks (injuries and equipment damage) and associated monetary and non-monetary costs.

Industry statistics provide useful information to help in accident prevention (Fung, Lo, and Tung 2012). The more and accurate the data is, the better for a machine learning (ML) technique for detailed exploration and understanding. Current applications are benefitting from recent ML techniques such as Big Data analytics due to its ability to process massive data to reveal insights that can improve operational performance. For instance, Big Data analytics was used for flight delay forecasting using the massive data collected from 112 airports around the world, with flights operating for a major airline in Hong Kong(Chung, Ma, and Chan 2017). Also, Big Data analytics was used to accurately monitor workers' behavior (Guo et al. 2016), minimize construction wastes (Bilal et al. 2016), and identify optimal weather indices for agricultural food-related weather risk management (Biffis and Chavez 2017).

Parameters attributable to construction site accidents, such as location, poor individual work practices, workplace layout, work pressure, amongst others, can be employed to formulate a strategy to predict injuries. Accurate injury prediction will ensure that occupational risks are well controlled and managed (Silva and Jacinto 2012). Incident datasets, however, are unreliable, unstructured, incomplete, and imbalanced to develop an efficient prediction model. Also, it is often challenging to get ready-made ML algorithms to eliminate redundant features in high dimension data optimally. Many of the conventional ML algorithms have challenges addressing a large amount of irrelevant or redundant attributes.

Based on the preceding, the main goal of this study is to present a robust and efficient technique for finding complex patterns, establishing the statistical cohesion of patterns, and reducing the number of unrelated attributes in datasets for optimal future decision-making. To achieve a reliable prediction model for the prediction of occupational incident outcomes, we employ a particle swarm optimization (PSO) technique for parameter optimization and feature selection in the gradient boosting machine (GBM). PSO has been successfully employed in several areas, especially for feature selection (Unler and Murat 2010; Xue, Zhang, and Browne 2014). Though GBM is efficient and popular, it suffers from long training times when tuning its hyper-parameters. We also identify relevant features attributable to incident outcomes using chi-square statistics and perform an explanation for injuries occurrence on the proposed GBM-PSO models using probability distribution plots and decision trees approaches. In achieving the set objectives, we developed GBM-PSO models to model the relationship of independent and dependent variables of the incident dataset obtained from a UK leading power infrastructure provider. The expected outcomes used in this study are ACCIDENT, a yes or no value indicating the likelihood of an accident, and INJURYFREQ, the number of personnel with hand-related injuries. Appropriate performance metrics are used to benchmark the prediction ability of GBM-PSO with other tree-based ML models, namely, decision trees, random forest (RF), and GBM.

We selected these tree-based methods because of their versatility and ease of use. They are easy to interpret relative to other "black box" techniques, such as neural networks. Furthermore, they are less complicated when compared with other tree-based methods such as Bayesian Additive Regression Trees (BART) that are computationally intensive. Predicting the posterior probabilities of BART via Markov Chain Monte Carlo is time-consuming and requires complex computations. Besides, RF, GBM, and decision trees are accurate approaches (Cheng et al. 2012; Tixier et al. 2016; Goha et al. 2018), allowing interpretation of features importance used for predictions. In the case of construction safety performance, this is a critical property that enables decision-makers to understand and identify trends that impact injuries and deaths. Also, they have been successfully applied in construction safety (Cheng et al. 2012; Patri and Patnaik 2015; Tixier et al. 2016) and a variety of fields including medicine (Oztekin, Kong, and Delen 2011), energy and buildings (Tsanas and Xifara. 2012), and agriculture (Brillante et al. 2015). The results from this study will provide information on probable causes of accidents and offer proactive safety precautions to mitigate those risks.

The rest of the paper is structured as follows: a review of the literature is carried out in Section 2, and the methodology employed is discussed in Section 3. In Section 4, we discuss the design of models and performance evaluation. Discussion on results and implications of the study are made in section 5. Concluding remarks are presented in Section 6.

2. Literature review

In reviewing and assessing literature, prediction analytics methods and accident causes are examined.

2.1. Big Data analytics

Big Data are large and complex datasets that cannot be manipulated using traditional processing techniques. They are platforms where recording, measuring, and capturing of data occurs (Lee 2018). Six defining attributes of Big Data are volume, variety, velocity, veracity, variability and complexity, and value (Gandomi and Haider 2015). Volume represents the magnitude of data usually measured in units such as terabytes and petabytes. Attribute 'variety' is the structural heterogeneity in a dataset, while velocity is the rate at which data are generated. Veracity defines the unreliability traits in data sources, while variability is the variation exhibited in data flow rates. The attribute 'value' is the insight from analysis to aid decision making.

Some of these attributes are evident in a typical large construction incident dataset that is heterogeneous and dynamic (Fenrick et al. 2012). Aside from volume and veracity, value is another key attribute used in this study. Value gives a measure of information extracted from datasets for optimal control decisions to mitigate risks. The Big Data analytics inspect, clean, transform, and model the Big Data to discover useful information to support decision-making (Bilal et al. 2016). It is also a suite of techniques and processes that allow businesses to process, organize, visualize, and analyze data to produce insights for data-driven operational planning, decision-making, and execution (Lee 2018).

Big Data analytics is intellectually rich and borrows from related fields such as statistics, data mining, business analytics, and knowledge discovery from data (KDD). Its forms are descriptive, predictive, prescriptive. A variety of software packages such as R language, MATLAB, Hyperion, and Tableau can be used for the various analytical forms. However, the R language is used in this study for predictive analytics.

2.2. Common attributes for modelling occupational accident

The three biggest safety hazards on construction sites are widely acknowledged to be excavations, working at height, and movement of vehicles and plant machinery (Hinze and Teizer 2011). Variables attributable to construction health and safety risk are enormous. These include environmental conditions, poor work practices, ignorance, work pressure, and time constraint (Törner and Pousette 2009). Others are the working surface condition, human error, harsh temperature, equipment failure action, materials handling equipment, employment contract, experience, and animal or insect attack. The task (operation) to be performed, sex, employee age, day, time amongst others have also been used in estimating the distribution of work accident risk (Bailey, Cordeiro, and Lourenço 2007; Cheng et al. 2012). Tasks in a power infrastructure project may include wiring, excavating, stringing cables, blasting, cutting, pulling, erecting structures, lifting, loading/offloading, and jointing. Table 1 depicts the summary of previous research employing these attributes in occupational safety modeling.

Table 1: Common attributes for safety modelling

Attributes	**Reference**
Project types	(Cheng, Lin, and Leu 2010; Sanchez et al. 2015)
Project complexity	(Törner and Pousette 2009)
Location	(Huang and Hinze 2003; Cheng, Lin, and Leu 2010; Soltanzadeh et al. 2016)
Client	(Liu and Tsai 2012)
Equipment type /and state	(Liu and Tsai 2012; Soltanzadeh et al. 2016)
Employee age	(Bailey, Cordeiro, and Lourenço 2007; Paul and Maiti 2007; Silva and Jacinto 2012)
Employee experience	(Paul and Maiti 2007; Törner and Pousette 2009; Cheng, Lin, and Leu 2010)
Employment contract	(Sánchez et al. 2011; Cheng et al. 2012; Sanchez et al. 2015)
Month	(Liao and Perng 2008; Pinto 2014)
Time of the day	(Huang and Hinze 2003; Sanchez et al. 2015; Soltanzadeh et al. 2016)
Day of the week	(Silva and Jacinto 2012; Sanchez et al. 2015; Tsoukalas and Fragiadakis 2016)
Contract Status	(Sanchez et al. 2015)
Tasks	(Grassi et al. 2009; Cheng, Lin, and Leu 2010; Silva and Jacinto 2012)
Working surface layout condition	(Sánchez et al. 2011; Sanchez et al. 2015; Soltanzadeh et al. 2016)

2.3. Analytics techniques for health and safety risk modelling

Several studies in the literature (Sánchez et al. 2011; Liu and Tsai 2012; Rubio-romero, Rubio, and Carrillo-castrillo 2013; Pinto 2014; Yorio, Willmer, and Haight 2014; Sanchez et al. 2015) have discussed the use of either statistical analysis or machine learning (ML) techniques for modelling occupational accidents in construction projects. For example, statistical analysis proponents have applied a bivariate approach (Paul and Maiti 2007) and Poisson models (Yorio, Willmer, and Haight 2014) for modeling workplace safety. However, due to the huge amount of data, ML techniques supersede traditional statistical counterparts in prediction problems, and in addition to their remarkable results, they have been used in various fields such as engineering, medical science, finance (Witten et al. 2013).

Examples of machine learning techniques commonly used for modeling occupational injuries are linear regression, support vector machines, decision trees, RF, and artificial neural networks. A logistic regression model was used to predict roof fall injuries (Soltanzadeh et al. 2016), but the model cannot appropriately capture nonlinear relationships among variables (Tixier et al. 2016). The fuzzy logic was used to model safety risk assessment (Pinto 2014). However, fuzzy systems are incapable of generalizing without alterations to the rule base. Due to its ability to learn from data, artificial neural networks (ANN) have been employed for work-related injury risk analysis (Zurada 2012; Rubio-romero, Rubio, and Carrillo-castrillo 2013; Goha et al. 2018). However, ANN suffers from interpretability functionality and the difficulty in determining the number of layers and neurons. The adaptive neuro-fuzzy inference system has also been employed for work-related risk analysis (Ciarapica and Giacchetta 2009). Support vector machines (SVM) due to their low computational costs and a unique optima solution have been used to classify workers suffering work-related injuries (Sánchez et al. 2011; Zurada 2012). However, its computational complexity grows exponentially with the size of training samples. Bayesian networks are desirable for making inferences in cases where the input data is incomplete. They have been used to study the influence of working conditions on occupational accidents (García-Herrero et al. 2012). A fundamental difficulty in applying Bayesian networks is the computational complexity of evaluating these networks. The K-nearest neighbor (kNN) method, due to its simplicity has been used to classify workers according to their risk of suffering musculoskeletal disorders (Zurada 2012; Sanchez et al. 2015), and to evaluate the relative importance of different cognitive factors in influencing safety behavior (Goha et al. 2018). However, kNN has difficulties in classifying close objects originating from different classes correctly. Other ML techniques such as decision tree (Zurada 2012; Goha et al. 2018), random forest (Zurada 2012; Tixier et al. 2016; Goha et al. 2018), and gradient boosting machine (Tixier et al. 2016) have also been applied to model work-place injuries due to their high accuracies. However, many conventional ML algorithms suffer from over-fitting and have challenges in addressing the massive amount of irrelevant or redundant attributes for Big Data analytics.

In this era of Big Data analytics with numerous data types and advanced information technologies, new challenges are emerging regarding the computing requirements and strategies for data processing and analysis. The advent of Big Data calls for innovative methods for precise estimation of the safety effects of risk factors, and hotspots identification with higher resolution. Besides, in Big Data analytics, it is difficult to get ready-made ML algorithms to eliminate redundant features and achieve a decrease in the signal to noise ratio. Eliminating unrelated attributes will reduce ML algorithms running times and produce a more efficient classifier. Conventional ML techniques do not produce good results if their parameters are not tuned.

Optimization techniques such as particle swarm optimization (PSO), ant colony optimization, and genetic algorithms are often used for tuning ML techniques' parameters. PSO, which has been employed in several areas, especially, for feature selection (Unler and Murat 2010; Xue, Zhang, and Browne 2014), is an evolutionary computing technique depending on swarm intelligence. It has better performance when compared with the genetic algorithm (Chakraborty 2008). Based on the reviewed literature, and to the best of authors' knowledge, optimization techniques are a novel approach for tuning the parameters of the GBM technique to enhance its prediction accuracy. Also, there are limited studies on accident modeling and prevention in the power infrastructure domain. The available studies focused more on construction, mining, and shipbuilding industries. Therefore, in this paper, PSO is selected both for features selection and optimization of GBM parameters. The optimized Gradient Boosting Machine -Particle Swarm Optimization (GBM-PSO) model's prediction ability is benchmarked with the decision trees, random forest, and GBM techniques. We chose the tree-based techniques to benchmark the proposed model because they are highly accurate (Goha et al. 2018). They also require minimum data preprocessing and are capable of fitting highly nonlinear data (Hastie, Tibshirani, and Friedman 2009).

3. Methodology

We discuss in this section, methodology, data, exploratory analysis, and the overview of analytics methods employed. The goal of an exploratory data analysis is to confirm the justification of the proposed model. Initial exploratory data analysis results support findings in the literature.

3.1. Dataset and analysis

The authors obtained a privately maintained health and safety dataset containing 1,607,010 data points. This dataset represents incident cases that occurred over the past seventeen years from a leading UK utility infrastructure company. The dataset has various features about utility infrastructure projects (i.e., overhead lines, underground cabling, and onshore/offshore substations). The performance of classifiers depends on the quality of the data used (Tixier et al. 2016). Thus, we employed string processing techniques to retrieve useful underlying concepts contained in few text-free columns to provide valuable additional information to impact the classifier's performance. String processing techniques manipulate raw texts and convert them to tokens. Tokens are used to build document-term matrix (DTM). The retrieved information is then used to complete columns with missing or null entries (e.g., project type, employee experience, and task). We anonymized data to protect the privacy of the subjects. We follow the recommendation by Sarkar et al. [36] and convert the categorical data to numeric since numerical attributes hold more information than categorical attributes. For columns with missing entries that cannot be completed with the text processing approach, we use the k-nearest neighbor (kNN). In *k*NN, *K* nearest neighbors are selected from the complete cases, so that they minimize a similarity measure. If we assume a data set *D*, defined in Equ. (1), is composed of *N* labeled incomplete patterns or cases,

$$D = \{X, T, M\} = \{x_j, t_j, m_j\}_{j=1}^{M} \qquad (1)$$

where $x_j = [x_{1j}, x_{2j}, \ldots., x_{dj}]^T$ is the *j*th input vector composed of *d* features; labeled as $t_j \in [C_1, C_2, \ldots . C_c]$; C$_i$ represents classes, and $m_j = [m_{1j}, m_{2j}, \ldots., m_{dj}]^T$ indicates which input features are unknown in x$_j$. Then X is an *d*×*N* matrix representing the input data set, T is a row vector (1×*N*) representing the target set, and M is a binary d x N matrix. X can be divided into two parts based on M as $X = \{X_o, X_m\}$, where X$_o$ and X$_m$ represent the complete and incomplete cases. Given an incomplete pattern x, $U = \{u_j\}_{j=1}^{K}$ represents the set of its *K* nearest neighbors (according to a distance metric, computed as $d(x_p, x_q) = \sqrt{\sum_{i=1}^{n}(x_{ip} - x_{ip})^2}$, x_p and x_q are input vectors of an ith feature) arranged in increasing order of their distance. Once the nearest neighbors are found, a replacement value to substitute the missing attribute value is determined using the mean value of their nearest neighbors. In this study, we substituted the missing entries using the means of their k-nearest neighbors determined using the Euclidean distance. We used the kNN technique because of its simplicity and relatively high accuracy (Eskelson et al. 2009). We noticed a low imbalanced data problem as there were fewer accident risks compared to no accident risks in the dataset for the classification problem. We used the SMOTE algorithm (Chawla et al. 2002) to balance the dataset. Outliers are also eliminated using a Box plot (BP) statistical method. BP can graphically convey the level and spread of a distribution of data values at a glance. It also provides information on data's symmetry and skewness and displays outliers, unlike other data display methods. BP presents five-number summary: the minimum, lower quartile (ϑ_1), median (ϑ_2), upper quartile (ϑ_3), and maximum. The range of the middle two quartiles is called the inter-quartile range ($IQR = \vartheta_3 - \vartheta_1$). In detecting outliers, we employ a common rule: outliers are data points higher than $\vartheta_3 + 1.5 * IQR$ or lower than $\vartheta_1 - 1.5 * IQR$. In this study, we identified a few values that were included in this range and eliminated them appropriately. The final dataset after data cleansing has 1,349,239 data points, and a summary of variables in the dataset is given in Table 2.

3.2. Relationship between exploratory variables

The priority of occupational safety and health regulatory bodies (e.g., OSHA) is primarily to reduce injuries. The overall proportion of injuries to body parts due to various injury sources, or the proportion of injury types incurred from varied utility infrastructure projects worth further exploration. To allow exploration and understanding of complicated flow scenarios within a system interactively, Sankey diagrams (Sankey 1896), are often used based on their simplicity and popularity in industrial environments (Pépin et al. 2017). Sources of injuries within a construction project, for instance, can be visualized and interactively explored to help managers better understand trends and injury causes. Fig. 1 shows a simplified form of a distribution of body parts injuries with a proportion of injury represented as weights of flows between sources (i.e., Manual handling, Plant Equipment, Walking) and destinations (i.e., Finger/Hand, Knee, Arm/Elbow). For instance, walking on a muddy or wet/icy surface can result in slip/trip/fall events, which can ultimately result in an injured ankle, or knee, or fingers. Similarly, manual handling tasks (including loading and lifting) by linemen can trap their fingers ("caught in" event) or injure fingers using hand tools.

Table 2: Data description (Variables)

Predictors	**Explanation**	**Min**	**Mean**	**Max**
Project type (PT)	The project type (e.g., overhead line, cabling, and substation).	1	2.34	3
Project complexity (PC)	Determines whether the project is a new build, maintenance, or refurbishment.	1	1.81	3
Region (REG)	Regions where projects are constructed.	1	3.57	5
Location (LOC)	The project site or work premises.	1	1527.5	3390
Client (CLT)	Energy company, communications, digital, power supplier contractors.	1	14.52	22
Equipment type (EQP_T)	Machinery or tool for a task (plant equipment, hand tool, fleet/vehicle, etc.).	1	34.75	65
Equipment state (EQP_S)	Equipment state (Good, moderately ok, not ok).	1	2.02	3
Employee age (EMP_A)	Age expressed in a predefined range (16-25,26-44, 45-60).	1	1.15	3
Employee experience (EMP_E)	The length of time on the job (<1 year, 1-3 years, > 3 years).	1	1.85	3
Employment contract (EMP_C)	Employment defined as either temporary or permanent.	1	1.35	3
Month (MONTH)	Also include external factor such as the weather (i.e., winter, spring, autumn and summer).	1	6.19	12
Time (TIME)	Time of incident (6am-12pm early in the day) or (12pm-19pm later in the day).	1	1.49	2
Day of the week (DAY)	Day name (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday).	1	3.89	7
Contract Status (CS)	Employing company as main contractor, or subcontracted, or third-party company.	1	2.43	4
Task (TASK)	The task or operation (lifting, cutting, loading, pushing).	1	4.35	10
Working surface layout condition (WSLC)	Good condition, moderately in good condition, not in good condition.	1	1.72	3
Health Safety Executive control measures or Health and safety risk management (HSRM)	Represents any of these- adequate supervision/control, appropriate risk management policy but no supervision/control, no risk management policy/supervision or control.	1	2.28	3
Outcome variables				
ACCIDENT	Forecasts the occurrence of accidents (Yes or No).	0	-	1
INJURYFREQ	Predicts the number of personnel with hand related injuries	1	2.955	5

Also, a lineman working outdoor in a dusty environment can irritate his or her eyes. This study considers the frequency of hand-related injuries as one of the dependent variables because of their dominance as shown in Fig 1. Hands and fingers are essential for manual and equipment related operations such as lifting, loading, pushing, and pulling. The high proportion of hand-related injuries, as revealed in this study agrees with Albert and Hallowell (2013). There are more occurrences of low severity injuries (first aid and medical cases) when compared to high severity accidents (lost-time injuries and fatalities).

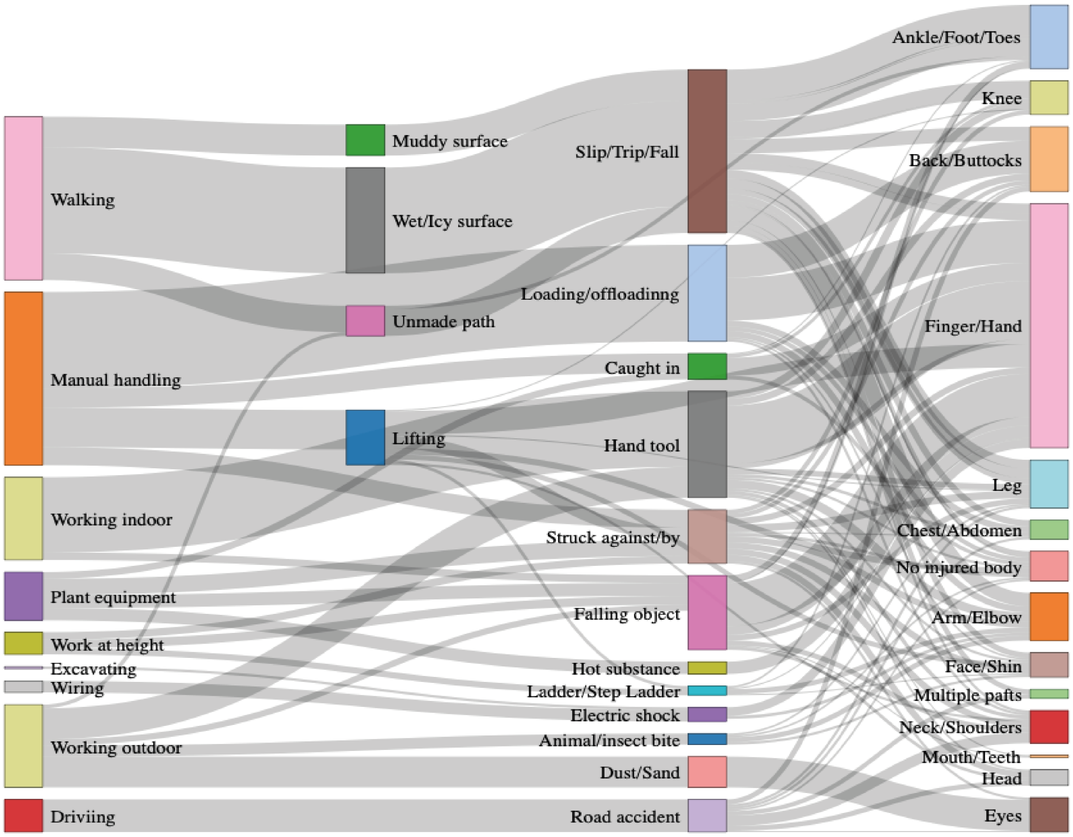


Figure 1. Causes of injury to body parts and distribution

3.3. Variables selection and correlation

The validity of a predictive model is highly dependent on its goodness of fit. Consequently, in predicting accidents in a high-dimensional power infrastructure incident database, it is useful to determine how each predictor will contribute to an optimal and reliable predictive model. Potential benefits of attributes selection include aiding data visualization and understanding, reducing storage requirements, training, and utilization times, and improving prediction performance. For the two dependent attributes (ACCIDENT, INJURYFREQ), relevant techniques are deployed to identify key attributes associated with each safety outcome. For instance, in Fig. 2, we depict a correlation between different attributes. Fig. 2a represents a typical power infrastructure project risk for a probable injury and Fig. 2b represents the number of personnel suffering from an injury. The correlation is calculated using the Chi-square statistic. The residual resulting from the computation can be visualized to comprehend the nature of correlation amongst attributes and help reduce redundancy in datasets. The sign of the standardized residuals can also help in interpreting the association between rows and columns. For instance, positive residuals (in blue) in cells in Fig. 2 specify an attraction between the corresponding row and column attributes. Negative residuals (in red) imply a negative association between the corresponding row and column variables. In Fig 2a, for instance, there are associations (both positive and negative) between the column *Emp_E* and the rows No (injury not likely) and Yes (probable injury). This association can mean that experienced personnel may not likely be involved in accidents while there is a chance that an inexperienced one might get injured. Also shown in Fig. 2b (row "2"), column "EQP_T" is strongly associated with two linemen sustaining hand-related injuries, and frequently associated with a lineman injuring a finger or hand (row "1"). Columns PROJ_C, TASK, WSLC, and MONTH are slightly associated with five linemen sustaining finger/hand injuries (row 5). Also, Columns PROJ_T and EMP_E are slightly associated with a lineman injuring a finger or hand (row 1). We can also compute the contribution (in %) of a given cell to the total Chi-square score as follow:

$$contribution = \frac{r^2}{\aleph^2}, \qquad (2)$$

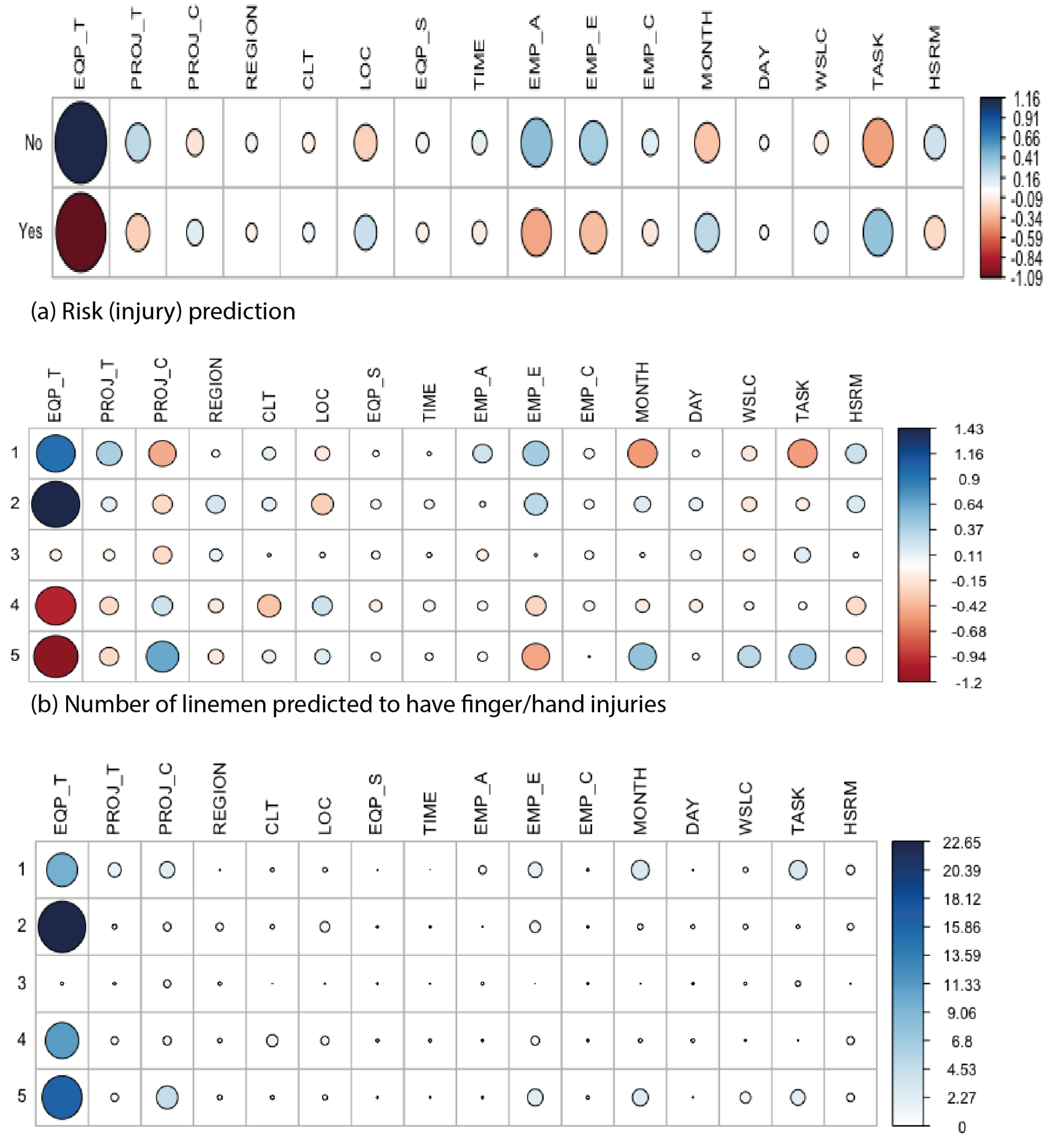
where $r = \frac{o-e}{\sqrt{e}}$, $\chi^2 = \frac{\sum(a-e)^2}{e}$, o=observed value, and e = expected value. Fig. 2c depicts the contribution of each cell for the Pearson residuals given in Fig 2b. The highly correlated attributes are often dropped from the subset feature set, but their correlation values with each other must be analyzed using appropriate algorithms before making such decisions.

3.4. Predictive analytics models

In this subsection, we present a brief outline of selected predictive analytics techniques namely, the decision trees, random forest, and gradient boosting machine.

3.4.1. Decision tree algorithm

Decision tree, a variant of the classification and regression tree (CART) is a supervised ML algorithm developed by Breiman et al. (Breiman et al. 1984). It is a recursive partitioning technique that builds a tree with created and split nodes using splitting criteria. The best split point is determined before applying split rules, and a standard function for defining the splits is Gini (Breiman et al. 1984). Building a decision tree involves recursively building a tree by splitting nodes where splitting criteria are reached. The building of trees is stopped when the learning dataset is fitted using predictors. Other necessary steps include pruning the trees to produce simpler forms and finally selecting an optimal tree from the generated pruned trees.



(a) Risk (injury) prediction

(b) Number of linemen predicted to have finger/hand injuries

(C) Contribution of a cell to the total Chi-square score for Pearson residuals in Fig 2(b)

Figure 2. Correlation plot from Chi-square statistic.

3.4.2. Random forest algorithm

Random forest (RF) is a popular ML technique that uses numerous independent decision trees created from randomly selected variables. The algorithm used to build independent trees guarantees the distinctness of all trees in the forest. The trees vote after creation to determine the most famous class (Breiman 2001). Randomness in RF is achieved by (1) using different bootstrap sample data to build each tree and randomly choosing a subset of predictors; (2) splitting each node of trees with the best subset. Two reasons for having a bootstrap step are to improve the prediction accuracy and to reduce generalization error (Breiman 2001). The random selection of splitting performs better than bagging in respect to the generalization error (Dietterich 2000). The algorithm is known to achieve better results than support vector machines, neural networks, and discriminant analysis (Liaw and Wiener 2002).

3.4.3. Gradient boosting machine

A gradient boosting machine is a powerful ML approach with impressive success in a range of applications (Hastie, Tibshirani, and Friedman 2009; Patri and Patnaik 2015). GBM is built from a combination of several weakly predictive models that relate predictors to an outcome (Friedman 2002). Primarily, GBM trees are constructed sequentially to reduce errors of the previous trees. This method of constructing trees will make the current model focus on data the previous models failed to predict or capture when fitting a succeeding model. However, residuals are resampled with a fraction used for modeling at each iteration with the learning function regularized through the shrinkage parameter. The shrinkage parameter slows the learning rate by using a fraction of the value for each residual. Parameter tuning of GBM is made using cross-validation or bootstrap techniques to address overfitting and minimize performance on testing data. In mathematical notations, we develop GBM models is as follows: assume a function f(x) is to approximate response y, where x is a vector of predictors. A loss function of the form, $\mathcal{L}(y, f(x)) = (y - f(x))^2$, is usually used to estimate a linear regression function $f(x) = x\psi$, where ψ is a matrix of parameters. For CART models, additive models (Hastie, Tibshirani, and Friedman 2009) can be used to define f(x) as a sum of basis function $b(x, y_m)$, as follows:

$$f(x) = \sum_m f_m(x) = \sum_m \psi_m b(x; y_m) \qquad (3)$$

For GBM, the function $b(x, y_m)$ represents individual trees, with y_m describing the split variables, their values at each node, and the predicted values. The ψ_m values are the weights given to the nodes of each tree in the collection. These weights, ψ_m, are used to determine how predictions from the individual trees are combined. To estimate parameters, the gradient boosting technique is applied (Friedman 2002). The following summarizes the procedure (De'ath 2007):

1) Initialization of $f_0(x) = 0$.
2) For m = 1 to n:
   - i) Residuals calculation, i.e., $r = -(|\delta\mathcal{L}(y, f(x))| / |\delta f(x)|)_{f(x)=f_{m-1}(x)}$.
   - ii) Fitting a least-squares regression tree to r to obtain the estimate of α_m of $\psi b(x; \alpha)$.
   - iii) Estimating ψ_m by minimizing $\mathcal{L}(y, f_{m-1}(x) + \psi_b(x; \alpha_m))$.
   - iv) Updating $f_m(x) = f_{m-1}(x) + \xi\, \psi_m b(x; \alpha_m), where\ 0 < \xi < 1$
3) Calculating $f(x) = \sum_m f_m(x)$

In Step 2(i), the residuals are calculated as the negative of the first derivative of the loss function evaluated for the current value of f(x). Step 2(ii) uses a least-squares regression tree to estimate α_m. Step 2(iii) estimates the values ψ_m assigned to the nodes of the tree to minimize the overall loss. To reduce overfitting, GBM applies a shrinkage strategy, and a learning rate ξ in Step 2(iv) when updating the estimated function $f_m(x)$.

3.4.4 Features optimization with particle swarm optimization

There is a need to reduce the size of the historical accident datasets to improve predictions. The feature selection algorithms are usually applied to select the unique feature set, thereby reducing the dataset and paving the way for convenient, effective, and accurate prediction models. This study uses PSO (Poli, Kennedy, and Blackwell 2007) to optimize the problem of feature subset selection. This algorithm iteratively improves the candidate solution concerning a given measure of quality. PSO uses the population candidate solution to the given problem, called particles. The particles are guided within the search space towards global optimum known positions, while also being pulled towards their best-known local positions. An evaluation function is used to determine the values of both positions and the best positions in each iteration. Mathematically, given that a swarm of particles fly through an N-dimensional search space, where the position of each particle represents a potential solution to an optimization problem. Each particle α in the swarm, $\xi = \{x_1, \ldots x_\alpha, \ldots., x_S\}$, has the following components:

i) $x_{\alpha,j}(t)$: jth dimensional component of the position of particle α, at time t

ii) $v_{\alpha,j}(t)$: jth dimensional component of the velocity of particle α, at time t

iii) $y_{\alpha,j}(t)$: jth dimensional component of the personal best (*pbest*) position of particle α, at time t

iv) $\hat{y}_j(t)$: jth dimensional component of the global best (*gbest*) position of swarm, at time t

Let us also assume Φ denote the fitness function to be optimized with the objective of finding the minimum of Φ in N-dimensional space. Then the *pbest* of particle α, updated in iteration t + 1 is given as,

$$y_{\alpha,j}(t+1) = \left\{\begin{matrix} y_{\alpha,j}(t) & if\ \Phi\big(x_\alpha(t+1)\big) > \Phi\big(y_\alpha(t)\big) \\ x_{\alpha,j}(t+1) & else \end{matrix}\right\} \forall_j \in [1, N] \quad (4)$$

The global best position (*gbest*), is then calculated using $\hat{y}(t) = y_{gbest}(t) = min\big(y_1(t), \ldots, y_S(t)\big)$. The positional updates for each iteration in a PSO process are then performed for each particle, $\alpha \in [1, S]$ and along each dimensional component, $j \in [1, N]$, as follows:

$$v_{\alpha,j}(t+1) = \varpi(t)v_{\alpha,j}(t) + c_1 r_{1,j}(t)\left(y_{\alpha,j}(t) - x_{\alpha,j}(t)\right) + c_2 r_2(t)\left(\hat{y}_j(t) - t_{\alpha,j}(t)\right) \quad (5)$$

$$x_{\alpha,j}(t+1) = x_{\alpha,j}(t) + v_{\alpha,j}(t+1) \quad (6)$$

where ϖ is the inertia weight, c_i are the acceleration constants. $r_{1,j} \sim U(0,1)$ and $r_{2,j} \sim U(0,1)$ are random variables with a uniform distribution. The following is a pseudocode for the PSO feature subset selection.

```
Define the dataset -  D(M,N), D is a dataset with M samples and N attributes
Encode N into numerical particles that are uniformly distributed
Generate initial population from parental chromosomes
Calculate the fitness value for updating particles position
Do while not termination condition
        If termination condition not reached then
                Update a particle's current position if it enhanced its previous best position
                Determine the best particle according to the particle's previous positions
                Update particles' velocities
        Otherwise
                Move particles to their new positions
        End [if]
End [while]
Rank all the selected and fittest attributes obtained
Select the top K optimized attributes from the ranked set
```

4. Modelling and evaluation

We randomly split the dataset into a training set (70%) to train the models and a hold-out test set (30%) for testing the models on new data not previously seen. We construct four predictive analytics models for the outcome variables ACCIDENT and INJURYFREQ. ACCIDENT is a yes or no value, which indicates the likelihood of an accident, while INJURYFREQ represents the number of personnel with hand-related injuries. We consider hand-related injuries because of their dominance in the dataset. Except for the GBM-PSO model that uses PSO for feature selection, other tree-based techniques use their inbuilt feature selection mechanisms. We use the parameter optimization strategy to search through the parameter space for a parameter combination that minimizes the prediction error. A random search technique is most appropriate for searching the parameter space when one is unsure of the initial values of these parameters. The optimal parameter values determined by the random search strategy for RPART, RF, and GBM are shown in Table 3. Similarly, for the GBM-PSO model, suitable values of the PSO parameters (i.e., *pbest* and *gbest*) were used and with 400 iterations, each with 5-fold cross-validation, the optimized GBM (GBM-PSO) produces the best accuracy of 96.23% with the following GBM parameters (see Table 3) n.trees = 600, shrinkage = 0.01, interaction depth = 7, and *n.minobsinnode* = 6. The predictive ability of each predictive model (RPART, RF, GBM, and GBM-PSO) on the test data (30% of the dataset) is measured using the correlation coefficient (R^2) define as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(t_i - y_i)^2}{\sum_{i=1}^{N}(t_i - \bar{y})^2} \tag{7}$$

where t_i denotes target i, y_i denotes prediction i, N is the number of testing observations, and $\bar{y}$ is the mean of predicted values. A high R^2 value indicates more significant similarities between the measured and predicted values. A perfect model has R^2= 1.

Table 3: Optimal control parameters used

Model	Optimal parameter
RPART	cp = 0.01
RF	mtry = 9

GBM	n.trees= 240,
	shrinkage=0.01,
	interaction.depth=5,
	n.minobsinnode=10

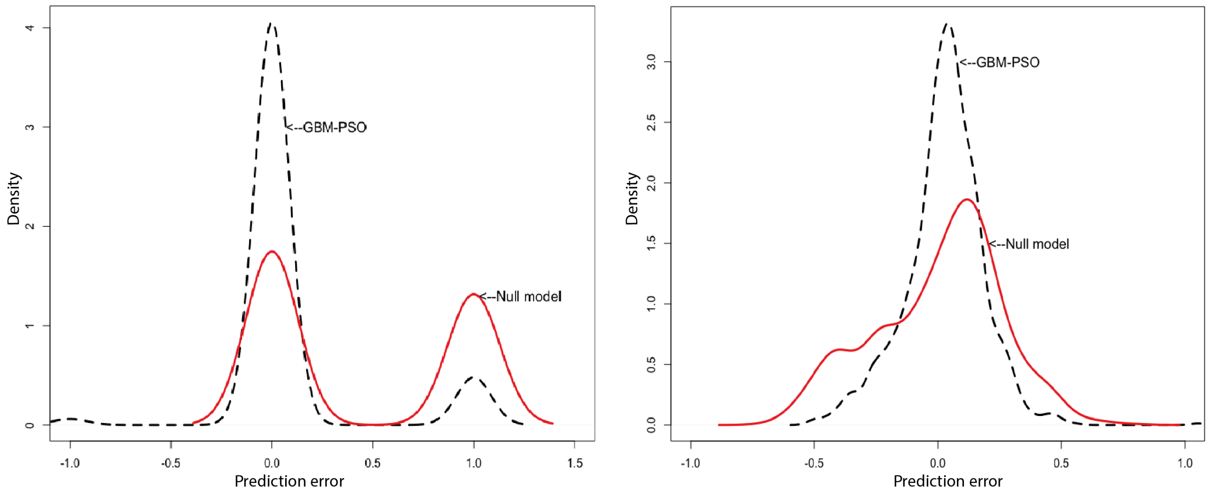
GBM-PSO	n.trees=60,
	shrinkage=0.01,
	interaction.depth=4,
	n.minobsinnode=8

ratio of the total number of correctly classified infrastructure projects with accidents to the total projects in the dataset. We assess the classification accuracy of the fitted GBM-PSO, GBM, RF, and RPART models by comparing their predictions using the predictor values of the test dataset with the actuals. The classification accuracy of models is assessed using the Accuracy, sensitivity-specificity analysis, area under the receiver operating characteristic curve (AUC), Kappa statistic, and learning time cost. Classification accuracy is computed as ${}^{nc}/_{N}$, where nc is the number of correctly classified infrastructure projects with accidents, and N is the total projects in the dataset. Sensitivity is the proportion of power infrastructure projects (cases) correctly classified that result in accidents; Specificity is the proportion of cases correctly classified as no accidents. The AUC measures the discriminatory capacity of classification models, and a model is considered to discriminate better than chance if its AUC value is higher than 0.5. Also, in order to check the robustness of the regression models built from GBM-PSO, GBM, RF, and RPART, a subsample of 300 test cases was randomly selected from the testing set. The average predictions made by GBM-PSO, GBM, RF, and RPART are recorded and compared with the means of the actual linemen with injured hand-related injuries (INJURYFREQ). We repeated this process twenty times, and the average predictions by models and actuals were noted and recorded. The box-plot analysis was performed on the values obtained.

5. Results analysis and discussion

5.1 Prediction ability

Firstly, before evaluating the predictive performance of GBM-PSO with GBM, RF, and RPART, we compared the predictive ability of the GBM-PSO models (classification and regression) with equivalent null models (an untuned GBM model with a deficient number of trees). Fig. 3 shows the residual density for the null and GBM-PSO models (for classification and regression problems). The residuals for GBM-PSO models (Fig 3a and Fig. 3b) are concentrated near zero, while they are relatively dispersed for the null models. Thus, the prediction ability of GBM-PSO is superior and distinct from the null models. The density of GBM-PSO residuals in the two plots is extensively peaked, with more density at the center of the distribution.



(a) Density plot of residuals for the outcome "ACCIDENT"

(b) Density plot of residuals for the outcome "INJURYFREQ"

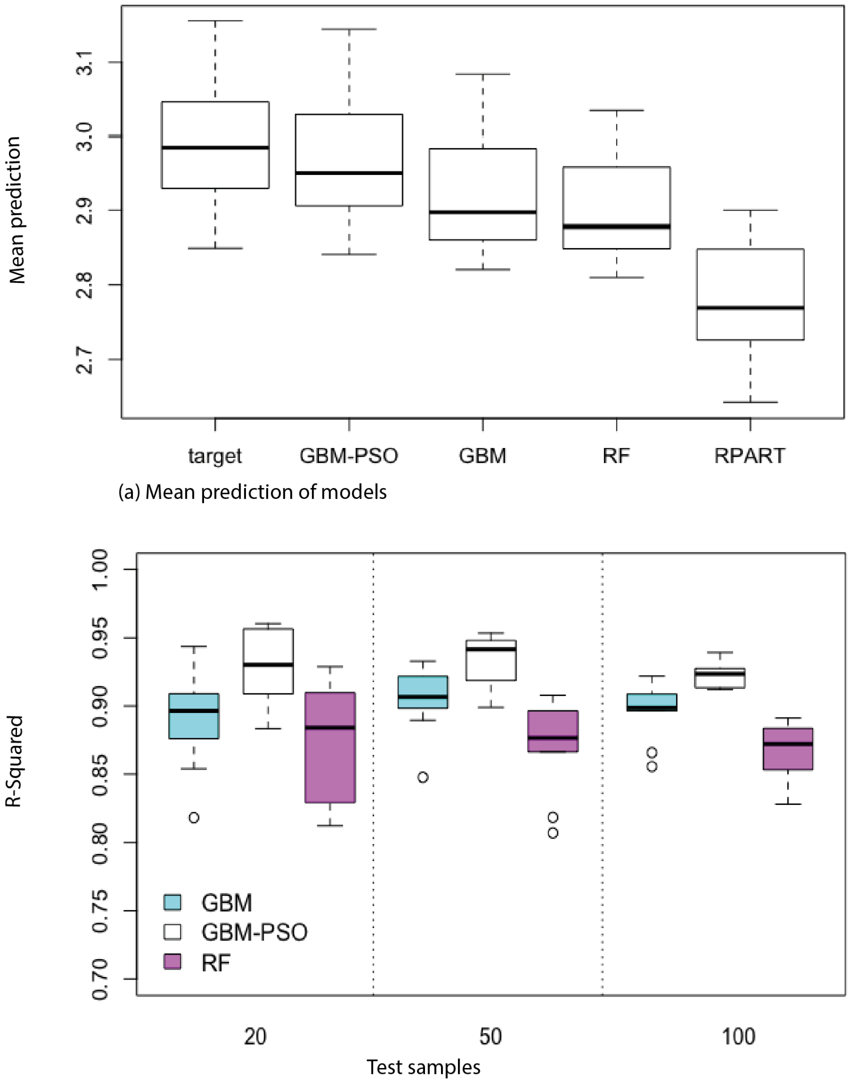
Figure 3. Prediction abilities of GBM-PSO and Null models

Also, the predictive ability of GBM-PSO, GBM, RF, and RPART on the testing set for ACCIDENT is depicted in Table 4 with 95% confidence intervals (CI), comprising of seven columns representing the classification model, accuracy, sensitivity, specificity, AUC, Kappa, and the training time. Each row in the table has the value of the specific metric (i.e., accuracy, Kappa). Standard deviation, and p-value obtained by a statistical test. Though the sensitivity of the GBM model (0.983) in Table 4 is the highest, the GBM-PSO has the highest values of classification accuracy (0.878), specificity (0.831), AUC (0.933), and Kappa statistic (0.745). GBM and RF follow GBM-PSO while RPART has the lowest classification accuracy (0.793), sensitivity (0.936), specificity (0.714), AUC (0.825), and Kappa (0.579) values. However, Kappa values for the four models are comparably higher (≥0.57), indicating an acceptable degree of agreement of their predictions beyond chance.

Table 4: Predictive ability of models for the dependent variable ACCIDENT. Boldface means no statistical difference from the best one (p-value ≥ 0.05)

	Classification					Training
Model	Accuracy	Sensitivity	Specificity	AUC	Kappa	Time (s)
RPART	0.793 ±0.014 2.2e-16	0.936 ±0.012 1.6e-13	0.714 ±0.018 2.2e-16	0.825 ±0.011 2.2e-16	0.579 ±0.026 2.2e-16	**0.863** ±0.023 **1.0e00**
RF	0.814 ±0.012 1.2e-15	0.965 ±0.010 2.6e-04	0.730 ±0.017 2.2e-16	0.931 ±0.010 3.3e-14	0.635 ±0.024 7.1e-14	1.594 ±0.009 4.7e-09
GBM	0.841 ±0.012 7.9e07	**0.983** ±0.007 **1.0e00**	0.761 ±0.016 4.9e-12	0.932 ±0.009 6.0e-05	0.664 ±0.024 9.6e-06	7.995 ±0.100 2.2e-16
GBM-PSO	**0.878** ±0.011 **1.0e00**	0.965 ±0.011 3.7e-05	**0.831** ±0.014 **1.0e00**	**0.933** ±0.009 **1.0e00**	**0.745** ±0.023 **1.0e00**	4.514 ±0.040 2.2e-16

RPART has the fastest training time (0.863) but GBM-PSO has the lesser training time (4.514) when compared with GBM (7.995). Though recent developments in parallel computing make it possible to train models based on large datasets, the processing time is still a decisive factor when choosing algorithms. Thus, using the PSO technique for feature selection has appreciably reduced the training times of the conventional GBM technique. GBM-PSO is approximately two times faster than GBM. Similarly, the predictive abilities of regression models on the testing data to forecast the number of linemen with hand-related injuries (INJURYFREQ) are depicted in Fig. 4. It is worth noting from the box-plot (Fig. 4a) that the average prediction obtained by the RPART model (2.780) was significantly less than the other tree-based models, 2.967, 2.922, and 2.902 for GBM-PSO, GBM, and RF, respectively. Also, paired Student's t-tests were used to test the sample mean (actuals) versus mean response. F-tests were also used to test for differences in the variance of predictions between the modeling techniques. In comparing differences in variances of the prediction means, we observe that there is no significant difference in the four sets of data. For instance, comparing prediction means from GBM-PSO and GBM, the p-value of the F-test is p=0.8466, which is greater than the significance level alpha=0.05. Also, the GBM-PSO model is not significantly distinct from the RF model (F=1.3844, df=19, p-value = 0.48451). Similarly, the predictions' means data set produced by RF and RPART models were not significantly distinct (F = 0.84737, df = 19, p-value = 0.7218). For the paired t-test, the GBM-PSO's average prediction is significantly different from RF's and RPART's average predictions with a p-value = 2.26e-09 and p=0.0078, respectively. However, there is no significant difference in the average prediction of GBM-PSO and GBM (p-value=0.0737). In comparing the sample mean with the mean response by GBM-PSO, the p-value = 0.3926, is greater than the significance level alpha =0.05 (Fig. 4a). We, therefore, conclude that the average number of linemen with hand-related injuries predicted by GBM-PSO is not significantly different from the average value of the targets. That is, the predicted INJURYFREQ values are very close to the actual values.



(a) Mean prediction of models

(b) R-Squared values obtained on the test data

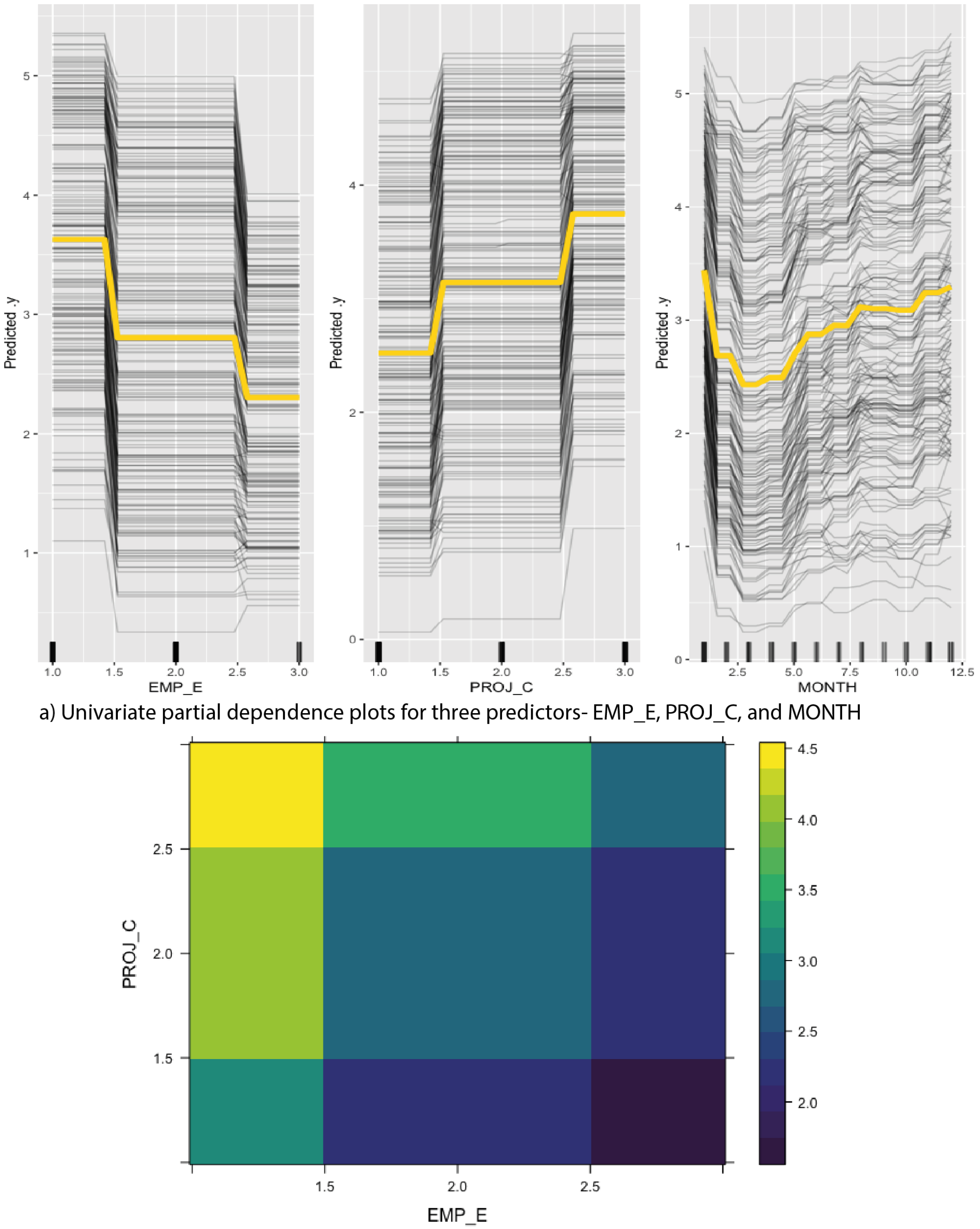
Figure 4. Prediction accuracies of ML algorithms as measured by R-Squared

Similarly, the distribution of the R-Squared value for the best three models is assessed on the sample sizes of 20, 50, and 100is depicted in Fig 4b. For each sample size (20, 50, 100), we randomly select cases from the test data and note R-Squared values for INJURYFREQ predictions made by GBM-PSO, GBM, and RF using the test dataset. We omitted RPART since it has the lowest prediction ability, as revealed in Fig. 4a. Based on the results obtained using the test data, the average R-Squared (Fig. 4b) of GBM-PSO for 20 sample cases (0.93), 50 sample cases (0.93), and 100 sample cases (0.92) are higher than the average R-Squared values of GBM and RF models. Thus, the GBM-PSO's average R-Squared is significantly different from GBM's and RF's average R-Squared with a p-value = 3.51e-06 and p=8.37e-11, respectively. Furthermore, the high R-Squared value shows that GBM-PSO appears to provide an alternative approach for predicting occupational injuries. Consequently, the performance of ensemble models (GBM-PSO, GBM, and RF) in Fig. 4, indicates the flexibility and the advantage of combining multiple base learners for prediction problems. The performance of RPART is inferior to GBM-PSO, GBM, and RF due to its high variance across samples, thus, making prediction unstable for new examples. Similarly, the GBM model consistently outperforms RF due to the way it reduces errors. The error reduction process in RF is through decreasing variance (Hastie, Tibshirani, and Friedman 2009).

In general, RPART, RF, and GBM, with their inbuilt mechanism for feature selection, had modest performance in modeling the dependent outcomes in this study. However, GBM-PSO, evaluation results indicate that GBM-PSO demonstrates comparable or better classification and regression performance than the other tree-based algorithms. Also, it is computationally efficient in times of the training time required when compared with GBM since optimal predictors are used. Training in GBM generally takes longer because the trees are built sequentially. Lastly, the proposed models are easy to implement. Thus, GBM-PSO, due to its predictive ability, will be most useful in situations where feature combinations are more complicated, with counterintuitive outcomes. Overall, GBM-PSO predicts ACCIDENT and INJURYFREQ with a high correlation coefficient and classification accuracy. Thus, it proved to be the most appropriate predictive analytics technique for modeling accident risks in power infrastructure projects.

5.2. GBM-PSO Interpretation

We use partial dependency plots (PDP) and individual conditional expectation (ICE) to investigate and interpret the occurrence rules of outcomes (ACCIDENT and INJURYFREQ) in power infrastructure projects using key features extracted by GBM-PSO models. Fig. 5a depicts the ICE and PDP curves (thick yellow lines) for comparing the marginal impact of three predictors (EMP_E, PROJ_C, and MONTH) on the number of linemen to suffer "finger/hand" injuries. These predictors were selected for illustration based on their relative influence. The univariate PDP for the predictor "EMP_E" shows that inexperienced linemen or workers will likely be involved in hand-related injuries. Also, the univariate PDP for the predictor "PROJ_C" indicates that complicated power infrastructure projects will result in more injuries. The univariate PDP for "MONTH" reveals low injuries for the months (February - May), a bit high for months (November – January). The bivariate partial dependence plot in Fig 5b captures the relationship between EMP_E and PROJ_C on the frequency of personnel with hand-related injuries. The relationship indicates that more injuries occur when new linemen are working on complex projects, whereas low to moderate injuries result when mid-level or experienced linemen are used.



a) Univariate partial dependence plots for three predictors- EMP_E, PROJ_C, and MONTH

b) A bivariate partial dependence plot for predictors "EMP_E" and "Proj_C"

Figure 5. Partial dependency plots for the GBM-PSO model

Also, to interpret and simplify the operations of GBM-PSO, a surrogate model (decision tree) was applied to the data points to generate rules. These rules can then be used to explain predictors causing accidents at sites. These rules are valuable in helping the management carry out preliminary proactive measures to minimize accidents occurrence on sites. Key predictors identified in this study agree with results from previous studies. These predictors are experience (Törner and Pousette 2009; Cheng, Lin, and Leu 2010), task (Grassi et al. 2009; Silva and Jacinto 2012; Sanchez et al. 2015), month (Liao and Perng 2008), project complexity (Törner and Pousette 2009), workplace design and layout condition (Sanchez et al. 2015), and age (Silva and Jacinto 2012). The identified risk relationships summarized in Table 5 will assist safety managers in understanding risk combinations that contribute to accidents or the number of hand-related injuries on sites. Thus, it can help in reducing the higher occurrence of low severity injuries at work sites, as observed in Albert and Hallowell (2013). This information will enable them to balance risk factors and develop a strategic risk mitigation plan to prevent accidents and ensure site safety. For instance, safety managers can reduce hand-related injuries caused by manual handling tasks in cold weather by enforcing warning operations to minimize workers' misjudgment. They can provide facilities for warming up, educate, and train linemen to prevent improper equipment handling or operations. Other strategies may include undertaking construction projects at warmer times, ensuring equipment is safe, and providing quality and appropriate personal protective equipment.

Table 5: Risk rules for power infrastructure projects

GBM-PSO model for outcome ACCIDENT

No	RULE	Predicted probability of accidents
1	EMP_E ≥3 & EMP_A < 2 & TASK ≥5	0.00
2	EMP_E ≥ 2 & EMP_A ≥ 2	0.00
3	EMP_E ≥ 2 & EMP_A < 2 & TASK <5 & MONTH IS 2 TO 11 & PROJ_C <3	0.14
4	EMP_E IS 2 TO 3 & EMP_A < 2 & TASK ≥5	0.76
5	EMP_E ≥ 2 & EMP_A < 2 & TASK <5 & MONTH≥11	0.78
6	EMP_E ≥ 2 & EMP_A < 2 & TASK <5 & MONTH IS 2 TO 11 & PROJ_C≥3	0.89
7	EMP_E<2	0.99
8	EMP_E ≥ 2 & EMP_A < 2 & TASK <5 & MONTH< 2	1

GBM-PSO model for outcome INJURYFREQ

No	RULE	Predicted number of linemen with hand related injury
9	TASK < 5 & MONTH IS 2 TO 7	1
10	TASK ≥ 5 & HSRM ≥ 3	2
11	TASK < 5 & MONTH ≥ 7	2
12	TASK ≥ 5 & HSRM < 3 & WSLC < 2	2
13	TASK ≥ 5 & MONTH IS 2 TO 5 & HSRM < 3 & WSLC ≥ 2	2
14	TASK < 5 & MONTH < 2	2
15	TASK ≥ 5 & MONTH IS 5 TO 8 & HSRM < 3 & WSLC ≥ 2	3
16	TASK ≥ 5 & MONTH < 2 & HSRM < 3 & WSLC ≥ 2	4
17	TASK ≥ 5 & MONTH ≥ 8 & HSRM < 3 & WSLC ≥ 2	4

It is interesting to observe that the reasoning behind some predictions, as shown in Table 5, is clear. For instance, Rule 7 means accidents happen when inexperienced linemen are involved in project operations. However, for other cases, feature combinations are more complicated with counterintuitive outcomes. Thus, predictive models presented in this study, especially GBM-PSO, will be useful for such situations in leveraging empirical data to guide decision-making under uncertainties, and therefore, seem to be effective for power infrastructure construction sites.

In this study, we have adopted an inductive inference approach for our scientific inferencing. Deductive inference, another form of scientific inferencing, formulates a hypothesis for the system behavior a priori and tests the validity of the hypothesis. Consequently, the deductive scientific inference is objective, and it limits our knowledge in the limited scope of the *a priori* hypotheses. Big Data inference generally encompasses broad, diverse, and unique findings that may not be foreseen, and so deductive strategies are too restrictive in this context. Inductive scientific inference, on the other hand, depends on the observed data itself to determine the most plausible conclusion within the problem domain. Inductive inference about unobserved states of a process has a broader scope than the observed data used to derive the conclusions (Dinov et al. 2016). The inductive approach to our Big Data analytics provides versatility in widening inquiries and is suitable for large, complex, and heterogeneous data. Lastly, it provides the means of classifying and predicting outcomes by obtaining efficient data-driven estimates describing the joint variability of the dataset.

5.3. Implication for practice

In this study, we observe the association of work-related accidents to manual handling operations, a period of construction (month), and the employee's experience. Also, principal sources of safety risks revealed in this study are manual handling of equipment, working surface/facility, workplace layout, lifting, cable stringing, wiring, and plant/equipment operations. Misjudgment, inappropriate materials or equipment handling procedures, electrical shock, removal of safety devices, and improper equipment operation are the leading causes of these accidents. The identified vital relationships will assist safety managers in understanding possible risk combinations that can result in accidents. This information will trigger a proactive strategic risk mitigation plan for accident prevention.

The high predictive performance resulting from the proposed model also revealed the existence of underlying patterns and trends in large data sets uncovered using ML to forecast injuries. This finding suggests that construction safety should be studied empirically and scientifically rather than using a qualitative approach (e.g., analysis of subjective, aggregated secondary data; and expert-opinion). The direct benefits to power infrastructure companies are lower costs and increased revenues. Other indirect benefits include reduced absenteeism; improved productivity, reduced replacement and training costs, more motivated linemen, and reduced contract penalties resulting from delays.

6. Conclusion

Health and safety risk in any organization, especially power infrastructure construction, is a thing of concern due to its risky operations, which represent a significant source of injury. Moreover, the health and safety datasets in power infrastructure are huge and synonymous with complex variables interaction, unreliable, data imbalanced, and presence of empty areas (holes). Thus, getting ready-made ML algorithms to optimally eliminate unrelated or redundant features, especially in Big Data analytics, is challenging. Many conventional ML algorithms are unable to address the massive amount of redundant attributes and produce good results if their parameters are not tuned.

Based on the preceding, particle swarm optimization, a technique within the family of evolutionary optimization algorithms is employed both for feature selection and tuning of GBM parameters to analyze explanatory variables within the power infrastructure domain. The optimized particle swarm optimization (GBM-PSO) model's accuracy is then benchmarked with the decision trees, random forest, and GBM techniques. An association between different attributes defining a typical power infrastructure project risks (e.g., accidents and injury frequencies) is also calculated using the Chi-square statistic. The resulting residuals are visualized appropriately to aid understanding of the nature of correlation amongst attributes.

The result of this study has shown that GBM-PSO has a better predictive ability (measured using Correlation coefficient and kappa statistic) than other ML techniques. Similarly, GBM-PSO has the potentials to capture the complex nonlinear relationship in datasets. The results also indicate key predictors, which are in agreement with the literature (Silva and Jacinto 2012; Soltanzadeh et al. 2016) such as the task, employee's age, month, experience, workplace layout, and control measures for modeling occupational risks. The GBM was used to investigate and interpret the occurrence rules of the outcome variables extracted using GBM-PSO models. Consequently, the results obtained have implications for practice in academia and industry. For instance, after a prediction, a safety manager is equipped with relevant information that will trigger a proactive strategic risk mitigation plan for accident prevention. Also, the powerful, robust, and computationally light optimized prediction model is appropriate when computational resources are insufficient. It can also be extended and adapted in related fields.

This study considers data from a single source; in subsequent research, we will attempt to carry out cross-organization studies. Data from multiple organizations and domains will be used to investigate the impact of organizational complexity on society. We considered three Big Data attributes due to time constraints and the unavailability of relevant multimedia data. Work is in progress to include two additional big-data attributes (velocity and variety) for real-time safety outcome prediction. Also, in this study, predictive analytics techniques are used; future research should examine prescriptive analytics. Also, the rules obtained in this study can be deemed as preliminary hypotheses for future studies. Finally, new research areas such as recursive neural networks and convolutional neural networks should be investigated to process multimedia data of safety events for enhanced occupational safety management models.

Acknowledgement

The authors appreciate Innovate UK for the Research funding (File Ref: 102061 Application number: 44746–322224) and Balfour Beatty plc for the support.

References

Albert, Alex, and Matthew R. Hallowell. 2013. “Safety Risk Management for Electrical Transmission and Distribution Line Construction.” *Safety Science* 51 (51): 118–26.

Bailey, Trevor. C., Ricardo Cordeiro, and Roberto.W Lourenço. 2007. “Semiparametric Modeling of the Spatial Distribution of Occupational Accident Risk in the Casual Labor Market, Piracicaba, Southeast Brazil.” *Risk Analysis* 27 (2): 421–31.

Biffis, E, and E Chavez. 2017. “Satellite Data and Machine Learning for Weather Risk Management and Food Security.” *Risk Analysis* 13 (8): 1508–21.

Bilal, Muhammad, Lukumon O. Oyedele, Olugbenga O. Akinade, Saheed O. Ajayi, Hafiz A. Alaka, Hakeem A. Owolabi, Junaid Qadir, Maruf Pasha, and Sururah A. Bello. 2016. “Big Data Architecture for Construction Waste Analytics (CWA): A Conceptual Framework.” *Journal of Building Engineering 6: 144-156.*

Breiman, L. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Breiman, L., J.H. Friedman, R.A. Ohlsen, and C.I. Stone. 1984. “Classification and Regreesion Tree.” In *Wadsworth International Group*, 43–49. Belmont, CA.

Brillante, Luca, Federica Gaiotti, Lorenzo Lovat, Simone Vincenzi, Simone Giacosa, Fabrizio Torchio, Susana Río Segade, Luca Rolle, and Diego Tomasi. 2015. “Investigating the Use of Gradient Boosting Machine, Random Forest and Their Ensemble to Predict Skin Flavonoid Content from Berry Physical-Mechanical Characteristics in Wine Grapes.” *Computers and Electronics in Agriculture* 117: 186–93.

Chakraborty, Basabi. 2008. “Feature Subset Selection by Particle Swarm Optimization with Fuzzy Fitness Function.” *Proceedings of 2008 3rd International Conference on Intelligent System and Knowledge Engineering*, 1038–42.

Chawla, N.V., K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. 2002. “Smote: Synthetic Minority over-Sampling Technique.” *Journal of Artificial Intelligent Research* 16: 321–357.

Cheng, CW, SS Leu, YM Cheng, TC Wu, and CC Lin. 2012. “Applying Data Mining Techniques to Explore Factors Contributing to Occupational Injuries in Taiwan’s Construction Industry.” *Accident Analysis and PreventionPrevention* 48: 214–22.

Cheng, CW, CC Lin, and SS Leu. 2010. “Use of Association Rules to Explore Cause-Effect Relationships in Occupational Accidents in the Taiwan Construction Industry.” *Safety Science* 48 (4): 436–44.

Chung, CH, HL Ma, and HK Chan. 2017. “Cascading Delay Risk of Airline Workforce Deployments with Crew Pairing and Schedule Optimization.” *Risk Analysis* 37 (8): 1443–58.

Ciarapica, F E, and G Giacchetta. 2009. “Classification and Prediction of Occupational Injury Risk Using Soft Computing Techniques : An Italian Study.” *Safety Science* 47 (1): 36–49.

De’ath, Glenn. 2007. “Boosted Regression Trees for Ecological Modeling and Prediction.” *Ecology* 88 (1): 243–51.

Dietterich, T.G. 2000. “An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization.” *Machine Learning* 40 (2): 139–57.

Dinov, Ivo D., Ben Heavner, Ming Tang, Gustavo Glusman, Kyle Chard, Mike Darcy, Ravi Madduri, et al. 2016. “Predictive Big Data Analytics: A Study of Parkinson’s Disease Using Large, Complex, Heterogeneous, Incongruent, Multi-Source and Incomplete Observations.” *PLoS ONE* 11 (8): 1–28.

Eskelson, B N. I., H.M. Temesgen, V. Lemay, T M. Barrett, Crookston N.L, and A. T. Hudak. 2009. “The Roles of Nearest Neighbor Methods in Imputing Missing Data in Forest Inventory and Monitoring Databases.” *Scandinavian Journal of Forest Research* 24: 235–46.

Fenrick, Steve A., Lullit Getachew, L. Fenrick, and S. Getachew. 2012. “Cost and Reliability Comparisons of Underground and Overhead Power Lines.” *Utilities Policy* 20 (1): 31–37.

Friedman, J. 2002. “Stochastic Gradient Boosting.” *Computational Statistics & Data Analysis* 38 (4): 367–378.

Fung, Ivan W H, Tommy Y Lo, and Karen C F Tung. 2012. “Towards a Better Reliability of Risk Assessment: Development of a Qualitative and Quantitative Risk Evaluation Model (Q 2 REM) for Different Trades of Construction Works in Hong Kong.” *Accident Analysis and Prevention* 48: 167–84.

Gandomi, M., and A Haider. 2015. “Beyond the Hype: Big Data Concepts, Methods, and Analytics.” *International Journal of Information Management* 35 (2): 137–44.

García-Herrero, Susana, M. A. Mariscal, Javier García-Rodríguez, and Dale O. Ritzel. 2012. "Working Conditions, Psychological/Physical Symptoms and Occupational Accidents Bayesian Network Models." *Safety Science* 50 (9): 1760–74.

Goha, Y.M., C.U. Ubeynarayanaa, K.L Wong, and B.H. Guo. 2018. "Factors Influencing Unsafe Behaviors: A Supervised Learning Approach." *Accident Analysis and Prevention* 118: 77–88.

Grassi, Andrea, Rita Gamberini, Cristina Mora, and Bianca Rimini. 2009. "A Fuzzy Multi-Attribute Model for Risk Evaluation in Workplaces." *Safety Science* 47 (5): 707–16.

Guo, S.Y., L.Y. Ding, H.B. Luo, and X.Y. Jiang. 2016. "A Big-Data-Based Platform of Workers' Behavior: Observations from the Field." *Accident Analysis and Prevention* 93: 299–309.

Hastie, T., R. Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd editio. New York: Springer-Verlag.

Hinze, Jimmie W., and Jochen Teizer. 2011. "Visibility-Related Fatalities Related to Construction Equipment." *Safety Science* 49 (5): 709–18.

Huang, X., and J. Hinze. 2003. "Analysis of Construction Worker Fall Accidents." *Journal of Construction Engineering and Management* 129: 262–71.

Le, Quang Tuan, Do Yeop Lee, and Chan Sik Park. 2014. "A Social Network System for Sharing Construction Safety and Health Knowledge." *Automation in Construction* 46: 30–37.

Lee, Hau L. 2018. "Big Data and the Innovation Cycle." *Production and Operations Management* 0 (0): 1–5. https://doi.org/10.1111/poms.12845.

Liao, CW., and YH. Perng. 2008. "Data Mining for Occupational Injuries in the Taiwan Construction Industry." *Safety Science* 46 (7): 1091–1102.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by Random Forest." *R News* 2 (3): 18–22.

Liu, H, and Y Tsai. 2012. "A Fuzzy Risk Assessment Approach for Occupational Hazards in the Construction Industry." *Safety Science* 50 (4): 1067–78.

Oztekin, A., Z. J. Kong, and D Delen. 2011. "Development of a Structural Equation Modeling-Based Decision Tree Methodology for the Analysis of Lung Transplantations." *Decision Support Systems* 51 (1): 155–166.

Patri, Ashutosh, and Yugesh Patnaik. 2015. "Random Forest and Stochastic Gradient Tree Boosting Based Approach for the Prediction of Airfoil Self-Noise." *Procedia Computer Science* 46: 109–21.

Paul, P.S., and J. Maiti. 2007. "The Role of Behavioral Factors on Safety Management in Underground Mines." *Safety Science* 45 (4): 449–71.

Pépin, Lambert, Pascale Kuntz, Julien Blanchard, Fabrice Guillet, and Philippe Suignard. 2017. "Visual Analytics for Exploring Topic Long-Term Evolution and Detecting Weak Signals in Company Targeted Tweets." *Computers & Industrial Engineering* 112 (October): 450–58.

Pinto, Abel. 2014. "QRAM a Qualitative Occupational Safety Risk Assessment Model for the Construction Industry That Incorporate Uncertainties by the Use of Fuzzy Sets." *Safety Science* 63: 57–76.

Poli, R., J. Kennedy, and T. Blackwell. 2007. "Particle Swarm Optimization: An Overview." *Swarm Intelligence* 1: 33–57.

Rubio-romero, Juan Carlos, M Carmen Rubio, and Jesús Antonio Carrillo-castrillo. 2013. "Analysis of the Safety Conditions of Scaffolding on Construction Sites." *Safety Science* 55: 160–64.

Sánchez, A. S., P.R Fernández, F. S Lasheras, F.J. de Cos Juez, and P.J.G Nieto. 2011. "Prediction of Work-Related Accidents According to Working Conditions Using Support Vector Machines." *Applied Mathematics and Computation* 218 (7): 3539–52.

Sanchez, AS., FJ Iglesias-Rodriguez, RP. Fernandez, FJ. de Cos Juez, PR Fernandez, FJ. de Cos Juez, F J Iglesias-rodríguez, P Riesgo Fern, F J De Cos Juez, and A Su. 2015. "Applying the K-Nearest Neighbor Technique to the Classification of Workers According to Their Risk of Suffering Musculoskeletal Disorders." *International Journal of Industrial Ergonomics* 52: 92–99.

Sankey, H. 1896. "The Thermal Efficiency of Steam-Engines." In *Minutes of the Proceedings of the Institution of Civil Engineers*, edited by Thomas Telford, 182–212.

Silva, Joaquim F., and Celeste Jacinto. 2012. "Finding Occupational Accident Patterns in the Extractive Industry Using a Systematic Data Mining Approach." *Reliability Engineering & System Safety* 108: 108–22.

Soltanzadeh, Ahmad, Iraj Mohammadfam, Abbas Moghimbeigi, and Mahdi Akbarzadeh. 2016. "Analysis of Occupational Accidents Induced Human Injuries: A Case Study in Construction Industries and Sites." *Journal of Civil Engineering and Construction Technology* 7 (1): 1–7.

Tixier, A.J.P., M.R. Hallowell, B. Rajagopalan, and D. Bowman. 2016. "Application of Machine Learning to Construction

Injury Prediction." *Automation in Construction* 69: 102–14.

Törner, Marianne, and Anders Pousette. 2009. "Safety in Construction - a Comprehensive Description of the Characteristics of High Safety Standards in Construction Work, from the Combined Perspective of Supervisors and Experienced Workers." *Journal of Safety Research* 40 (6): 399–409.

Tsanas, Athanasios, and Angeliki Xifara. 2012. "Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools." *Energy and Buildings* 49: 560–567.

Tsoukalas, V D, and N G Fragiadakis. 2016. "Prediction of Occupational Risk in the Shipbuilding Industry Using Multivariable Linear Regression and Genetic Algorithm Analysis." *Safety Science* 83: 12–22.

Unler, Alper, and Alper Murat. 2010. "A Discrete Particle Swarm Optimization Method for Feature Selection in Binary Classification Problems." *European Journal of Operational Research* 206 (3): 528–39.

Witten, I.H., E. Frank, M.A. Hall, and C.J. Pal. 2013. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Xue, Bing, Mengjie Zhang, and Will N. Browne. 2014. "Particle Swarm Optimisation for Feature Selection in Classification: Novel Initialisation and Updating Mechanisms." *Applied Soft Computing Journal* 18: 261–76.

Yorio, Patrick. L, Dana. R Willmer, and Joel. M Haight. 2014. "Interpreting MSHA Citations through the Lens of Occupational Health and Safety Management Systems: Investigating Their Impact on Mine Injuries and Illnesses 2003-2010." *Risk Analysis* 34 (8): 1538–53.

Zurada, Jozef. 2012. "Classifying the Risk of Work Related Low Back Disorders Due to Manual Material Handling Tasks." *Expert Systems With Applications* 39 (12): 11125–34.