Automation in Construction 69 (2016) 102–114

Contents lists available at ScienceDirect

Automation in Construction

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

Application of machine learning to construction injury prediction

Antoine J.-P. Tixier [a,*], Matthew R. Hallowell [a], Balaji Rajagopalan [b], Dean Bowman [c]

[a] Department of Civil, Environmental, and Architectural Engineering, University of Colorado at Boulder, Boulder, CO 80309, United States
[b] Department of Civil, Environmental, and Architectural Engineering, Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado at Boulder, Boulder, CO 80309, United States
[c] Bentley Systems, United States

ARTICLE INFO

Article history:
Received 10 October 2015
Received in revised form 24 March 2016
Accepted 22 May 2016
Available online 15 June 2016

Keywords:
Machine learning
Construction safety
Predictive modeling
Injury prevention
Random Forest
Boosting
Attribute

ABSTRACT

The needs to ground construction safety-related decisions under uncertainty on knowledge extracted from objective, empirical data are pressing. Although construction research has considered machine learning (ML) for more than two decades, it had yet to be applied to safety concerns. We applied two state-of-the-art ML models, Random Forest (RF) and Stochastic Gradient Tree Boosting (SGTB), to a data set of carefully featured attributes and categorical safety outcomes, extracted from a large pool of textual construction injury reports via a highly accurate Natural Language Processing (NLP) tool developed by past research. The models can predict *injury type*, *energy type*, and *body part* with high skill ($0.236 < RPSS < 0.436$), outperforming the parametric models found in the literature. The high predictive skill reached suggests that injuries do not occur at random, and that therefore construction safety should be studied empirically and quantitatively rather than strictly being approached through the analysis of subjective data, expert opinion, and with a regulatory and managerial perspective. This opens the gate to a new research field, where construction safety is considered an empirically grounded quantitative science. Finally, the absence of predictive skill for the output variable *injury severity* suggests that unlike other safety outcomes, *injury severity* is mainly random, or that extra layers of predictive information should be used in making predictions, like the energy level in the environment. In the context of construction safety analysis, this study makes important strides in that the results provide reliable probabilistic forecasts of likely outcomes should an accident occur, and show great potential for integration with building information modeling and work packaging due to the binary and physical nature of the input variables. Such data-driven predictions had been absent from the field since its inception.

1. Introduction and motivation

Construction is one of the largest industries in the United States, but is also one of the deadliest [12]. Between 1992 and 2010, an average of 730 lives have been claimed each year [20]. Despite the numerous efforts that have been motivated by this alarmingly poor performance, injury statistics have not significantly improved in the past decade [12]. This might be explained by the fact that the construction industry has reached saturation with respect to traditional approaches to safety and that innovations are needed [27]. Risk analysis has emerged as a promising alternative to managerial and regulation-based approaches. However, construction safety risk analyses are currently limited because existing techniques overlook the complex and dynamic nature of construction sites and are not based on empirical data.

To jointly address these limitations, Esmaeili and Hallowell [26,28] laid the groundwork of a new conceptual framework, offering a systematic and comprehensive way to extract safety critical structured information from unstructured injury reports. Unlike traditional safety risk analysis techniques, this attribute-based approach renders construction injuries as the resulting outcome of the joint presence of a worker and the interplay among a finite set of universal descriptors of the work environment that are observable before an injury occurs. These binary attributes, also called injury precursors, make physical sense and are related to construction means and methods, human behavior, and environmental conditions. For instance, in the following excerpt of an injury report: "employee was welding and grinding inside tank and experienced discomfort to left eye", four fundamental attributes can be identified: (1) *welding*, (2) *grinding*, (3) *tank*, and (4) *confined workspace*.

The attribute-based framework derives its strength from the ability to capture and encode the information of every possible construction situation in a finite, standardized format, regardless of trade, project type, or industry sector. Therefore, as illustrated in Fig. 1, extracting attributes and various safety outcomes from injury reports (i.e., objective empirical data) enables the constitution of a structured, consistent

* Corresponding author.
E-mail addresses: antoine.tixier-1@colorado.edu (A.J.-P. Tixier), matthew.hallowell@colorado.edu (M.R. Hallowell), rajagopalan.balaji@colorado.edu (B. Rajagopalan), dean.bowman@bentley.com (D. Bowman).

http://dx.doi.org/10.1016/j.autcon.2016.05.016
0926-5805/

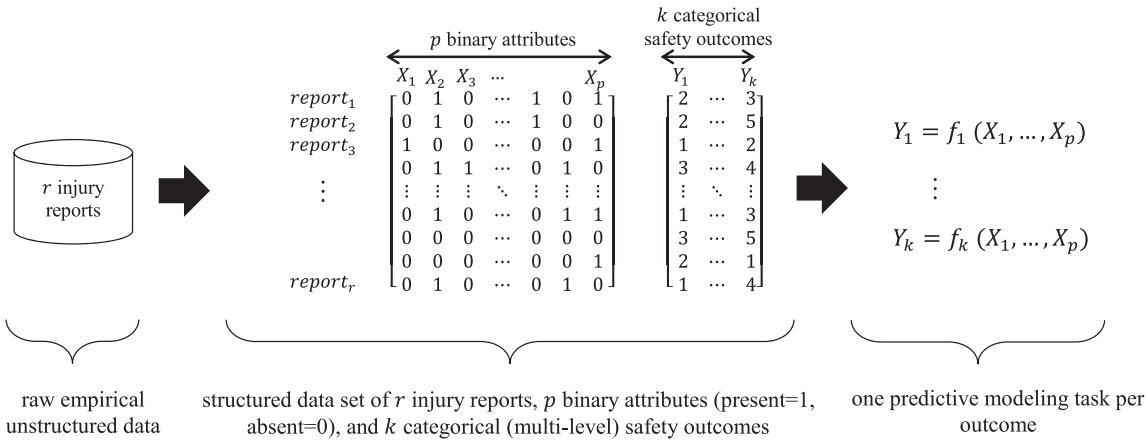

Fig. 1. The derivation of predictive models from injury reports is enabled by the attribute-based framework.

multivariate data set ideally suited for data mining, predictive modeling, and, thus, for knowledge discovery. Such new knowledge can enhance understanding of the underlying mechanisms that shape construction safety risk and create injuries. More precisely, *this study seeks to demonstrate that the workflow illustrated in Fig. 1 is viable and can be used to produce empirically-driven models with high predictive skill.* A fundamental postulate made here is that construction safety is not a strictly managerial outcome, but rather features a non-random component that can be studied by means of observation, like any other natural phenomenon. If this assumption holds, adopting the attribute-based framework would succeed in transforming construction safety research from opinion-based and qualitative to objective, empirically grounded quantitative science.

The effectiveness of the attribute-based framework depends on a number of methodological parameters including: (1) the way attributes are created and defined, (2) the quality and quantity of the injury reports available, (3) the technique with which attributes are extracted from the reports, and (4) the methods used for data mining and predictive modeling. As will be discussed in the background section, all previous work in this emerging research area (e.g., [22,26,28,29,30,70]) is subject to limitations with respect to one or more of the aforementioned parameters.

Building on three recent studies [22,66,70] that respectively addressed the limitations pertaining to the first three of the aforementioned criteria, here we tackle the limitations related to the fourth: predictive modeling. More specifically, two state-of-the-art machine learning (ML) algorithms, Random Forest (RF) and Stochastic Gradient Tree Boosting (SGTB), were used to predict safety outcomes from fundamental construction attributes. As will be shown, the models built outperform that of past research, in terms of predictive skill, variety of outcomes predicted, and actionable feedback that can be used to direct efforts towards targeted preventive actions and corrective measures.

2. Background and point of departure

This section provides the inspiration for our work, a brief description of past research in the domain of attribute-based safety analysis and in the application of ML in the construction industry, and the expected contributions.

2.1. Why does prediction of safety outcome matter?

Many industries, including construction, struggle with decision-making under uncertainty. Making the wrong decisions can have dramatic consequences, especially when lives are at stake. In healthcare, for example, Seera and Lim [58] observed that lack of experience, information overload, and unawareness of the most recent advancements in medical research were the leading causes of misdiagnosis by physicians. In the exact same way, even an experienced construction worker or safety manager has limited personal history with accidents. They may have witnessed, in their entire professional life, hundreds of near misses and first aid injuries, dozens of medical cases and lost work time injuries, and, perhaps, a few permanent disablement injuries and fatalities. Because of this limited experience with incidents, they may misdiagnose the risk of a given construction situation. It is actually well known that poor hazard recognition skill is a proximal cause of risk misperception and injury in construction [2,13]. People working upstream of the construction phase, like designers, face an even greater risk of failing to recognize hazards and misestimating risk[2,8].

Furthermore, without even considering the limited experience problem, human judgment and intuition will always be subject to important biases and fallacies (e.g., [69]). Also, humans have very limited capability of inducing knowledge from large numbers of observations [59]. This is due to the fact that human short-term memory is only capable of handling at most seven items evaluated for seven attributes at the same time [50].

On the other hand, ML can induce general rules from very large amounts of cases belonging to highly dimensional spaces, and is therefore a way to ground safety-related decisions under uncertainty on empirical knowledge. This could lead to improved decision-making and save lives. Indeed, other industries have begun to realize great benefits by transitioning from subjective to objective decision-making thanks to statistical learning. For instance, Seera ad Lim [58] trained ML models on large numbers of health records to automatically diagnose new patients, providing physicians with an opportunity to reconsider initial decisions and improve diagnosis accuracy.

2.2. Limitations of previous work on attribute-based construction safety

Although Esmaeili and Hallowell [26,28] made important strides by introducing and using the attribute-based framework for the first time, some serious limitations remained. In particular, some of the attributes identified via manual content analysis were not in full accordance with the framework as they were outcomes (e.g., *structure collapse*, *falling from roof*). By nature, an injury precursor should be observable *before* an injury occurs. Some other attributes were overlapping (e.g., *working underground*, *working in a confined space*), or loosely defined (e.g., *not considering safety during site layout*). Finally, the content analysis had rather low consistency (76% of inter-coder agreement), and only 300 reports all related to high severity struck-by injuries were analyzed, so only part of the picture was captured.

Esmaeili et al. [29] took the research a step further by using commercial software to automatically extract attributes from a larger amount of reports (1450). However, the low accuracy of the procedure (21%

disagreement between manual and automated coding on average) was a significant limitation, as it compromised the reliability of the data set obtained. In addition, the usefulness of the models built was restricted by the fact that only high severity struck-by injuries were taken into account. It should also be noted that only 22 attributes were considered.

Finally, Esmaeili et al. [30] used the data set obtained by Esmaeili et al. [29] to predict a binary severity outcome (fatality/no fatality) via a logistic regression model taking principal component scores as input variables. On the full training data set, the best model obtained a Rank Probability Skill Score (RPSS) of 0.116, which indicates modest skill [37]. In addition, this score was an overly optimistic estimate of the true predictive skill, as the model was tested on the very same observations that were used for training. To ensure unbiased estimation of a model's true ability to extrapolate, testing should always be conducted against unseen observations, using a separate test set when there is enough data, or cross-validation else ([41], pp. 222–223). Another limitation of Esmaeili et al. [30] is the use of logistic regression, a parametric, linear and global model which is by definition unable to capture the nonlinear and local relationships that may exist among predictors and targets [55,67]. Also, because these relationships are unknown, parametric models are not best suited for skillful prediction.

To address the above-mentioned limitations, we first used a broadened and more robust list of 80 attributes engineered and validated by a team of 8 researchers [22,70] and slightly modified by Tixier et al. [66]. This list is provided in Table 2. Second, we used a rather large database of 5298 injury reports that featured all types of injuries and was representative of the true distribution of injury severity. Third, a large and reliable data set of attributes and outcomes was automatically extracted from the database of injury reports by a highly accurate (96% in F1 score) natural language processing (NLP) program developed by Tixier et al. [66], ensuring high data quality. Finally, we used RF and SGTB, two cutting edge statistical learning algorithms, to predict safety outcomes from attributes with high skill. Since RF and SGTB both use decision trees as their base models, these two techniques can capture both nonlinear and linear; local and global relationships between input and output variables.

2.3. Previous use of ML in construction

Construction research has considered ML for more than two decades. Moselhi et al. [51] first discussed the potential applications of neural networks in construction engineering and management and developed a prototype providing optimum markup estimates from attributes describing bid situations, such as the number of competitors or the contractor's estimated cost. Later, Skibniewski et al. [59] applied the AQ15 algorithm on a collection of 31 training examples to automatically learn the mapping between constructability (poor, good, excellent) and 7 predictors, such as the reinforcement ratio of the beam and the number of walls attached to it. Soibelman and Kim [60] applied decision trees and neural networks to a construction management database to identify the causes of delays.

More recently, Lam et al. [45] found that support vector machines could produce accurate forecasts of contractor prequalification using input variables such as financial strength, current workload, quality management, and environment, health and safety considerations. Also, Cheng et al. [17,18] used a support vector machine optimized via a fast messy genetic algorithm to estimate building cost and loss risk from ten input variables, such as change orders, number of rainy days, number of floors, season, and geological conditions. Finally, Yang et al. [76] developed an algorithm to automatically track workers in digital videos; Tsanas and Xifara [68] used RF to predict heating and cooling loads of residential buildings from wall area, glazing area, overall height, and other input variables; and Son et al. [61] used a support vector machine to detect concrete structural components in color images from actual construction sites.

Although not exhaustive, this short review of the literature shows that ML has a quite long history of being used in construction research for a variety of applications. However, to the best of our knowledge, this is only the second time that supervised learning algorithms are used to predict construction safety-related outcomes from empirical data (after [30]).

2.4. Goal of this study

The goal of the present research effort is to apply RF and SGTB, two widely used and highly successful ML algorithms, to attribute and outcome data extracted from a large body of injury reports. Note that we could have included other supervised classifiers in our comparison, like Support Vector Machines or Neural Networks, but we were mainly interested in testing whether fundamental construction attributes could make good features at all in predicting safety-related outcomes, not about comparing all major classification algorithms. The predictive models obtained can be used to augment the experience of construction professionals with lessons learned from empirical data representing millions of worker-hours, far exceeding the exposure of even the largest and most experienced group of experts. This extensive amount of empirical knowledge can be used with profit to improve safety management in the design, work packaging, and execution phases of a construction project.

In practice, the models developed assign a probability of occurrence to each level of each safety outcome from a simple description of the work environment in terms of attributes. An example is given in Fig. 2 for the safety outcome *body part injured*. Such probabilistic forecasts provide some insight as to which preventive and/or corrective actions to take, allowing for better-informed, safer proactive decision-making. Providing a risk estimate (green, orange, red) for a given combination of observed attributes such as in Prades Villanova [70] is useful, but predicting the most likely categories of various safety outcomes is a complementary and equally valuable strategy.

2.5. Characteristics of the data set

We had access to a raw database of 5298 injury reports gathered from more than 470 contractors involved in industrial, energy, infrastructure, and mining work throughout the world and representing millions of worker-hours. More details about these data can be found in Prades Villanova [70], Desvignes [22], and Tixier et al. [66]. These reports were automatically scanned for the attributes shown in Table 1 and the safety outcomes listed in Table 2 by Tixier et al.'s [66] NLP system.

As summarized in Table 2, the safety outcomes predicted in this study were the (1) *type of energy* involved in the accident, (2) *injury type*, (3) *body part* affected, and (4) *injury severity*. The outcome *energy type* was taken into account based on the theory that any injury can be associated with the release of some form of energy [31,39]. For *injury type*, *body part*, and *injury severity*, the classification scheme is consistent with that of the Bureau of Labor Statistics [53] and the Occupational Safety and Health Administration (OSHA) [40,53].

It should be noted that Prades Villanova [70] and Desvignes [22] ensured the validity and relevance of the attributes created via content analysis by adhering to a strict coding scheme, implementing an iterative process with team-based calibration meetings, and using peer reviews and random checks by external reviewers with a stringent 95% agreement threshold. Such great care was taken because this procedure, called *feature engineering*, is of paramount importance to ML success [24]. Tixier et al. [66] also tuned their NLP tool by adopting an iterative process involving at each step careful reviews by 7 researchers of 140 randomly selected reports scanned by the system. At each round, lessons learned from examining the errors made by the tool were used to improve skill. A harsh 95% threshold in accuracy was exceeded after 4 iterations (96%). In particular, the NLP system attained precision and

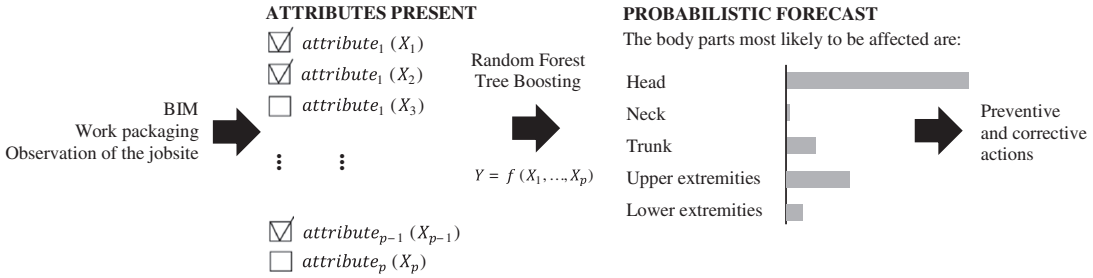

Fig. 2. Practical use of the predictive models built in this study.

recall rates of 95% and 97% for attributes, and error rates of 5.7% for both *energy type* and *injury code*. The NLP tool was designed to return "not detectable" when multiple body parts are detected in a given report, or when the information is missing. However, on the 93.75% of reports it could label, the tool proved 100% accurate [66].

900 reports out of the 5298 available were not associated with any attribute, and were therefore removed. An inspection of these reports showed that they were very short and did not contain any attribute-related information. The attributes *poor housekeeping* and *electricity* were discarded due to their absolute rarity (2 and 3 observations only), as well as the energy type *electricity* (3), and the injury types *transportation accident* (4) and *fall to lower level* (18). This made for a final data set of $r = 4398$ observations, $p = 78$ attributes, and $k = 4$ safety outcomes (using the notation from Fig. 1). The attribute counts for this data set are shown in Table 2. The safety outcome *body part affected* could not be inferred for 831 reports, so for this particular target, only 3556 observations were available for training. Also, at the time of writing, Tixier et al.'s [66] NLP tool could not extract the safety outcome *injury severity*, so for this prediction task, the 1829 reports manually analyzed by Prades Villanova [70] and Desvignes [22] had to be used. Finally, the levels *permanent disablement* and *fatality* were removed (respectively one and no observation), and *pain* (159 observations) was combined with *first aid* (1362) since the difference between these two severity levels appeared to be very tenuous. The counts of each category of the safety outcomes in the final data sets are presented in Table 3.

As one can see from Table 3, four multi-class prediction tasks were to be tackled in this study (i.e., there were four categorical safety outcomes to predict). Using the notation from Fig. 1, the four output variables were $Y_1 =$ *energy source* (7 levels), $Y_2 =$ *injury type* (5 levels), $Y_3 =$ *body part* (5 levels), and $Y_4 =$ *injury severity* (3 levels). For each safety outcome (i.e., each Y_k), the goal was to determine the best f_k such that $Y_k = f_k(X_1,\ldots,X_p)$, where $(X_1,\ldots,X_p)$ are the fundamental construction attributes presented in Table 2. The methods used and procedure followed to accomplish these tasks are presented next.

3. Application of ML

We used the $r = 4398$ by $p = 78$ structured data set of attributes and outcomes shown in Fig. 1 ($p = 78$ since *poor housekeeping* and *electricity* were removed as previously explained). The features, or input variables, were the fundamental construction attributes $(X_1,\ldots,X_p)$ listed in Table 2, such as *welding*, *uneven surface*, or *adverse low temperatures*, and the targets, or output variables, were the four categorical safety outcomes $(Y_1,\ldots,Y_4)$, listed in Table 3: *energy type*, *injury type*, *body part*, and *injury severity*. Each injury report, also referred to as an observation or training example in what follows, associated a specific combination of attributes to a specific combination of safety outcomes. Based on such training data, ML algorithms could infer rules mapping combinations of attributes to levels of safety outcomes, and use these rules later on to predict the most likely outcomes for brand new observations.

ML was preferred over parametric modeling because the latter is not optimal when little knowledge is available about the phenomenon studied. Indeed, parametric modeling imposes a model a priori to the data, either arbitrarily or based on some knowledge about the underlying process. Therefore, if the model selected is a poor representation of the phenomenon studied in the first place, it may be nothing more than "the right answer to the wrong question" [9]. On the other hand, ML algorithms do not assume that the data have been generated by any parametric model prescribed by the user. Rather, the assumption is that independent and dependent variables are related in a totally complex and unknown manner. Both linear and nonlinear relationships can be captured, as well as complex high-order interactions among variables, without imposing any formal model and its inherent suite of limitations.

More specifically, we used RF and SGTB as our ML algorithms. These two techniques were used as they currently stand among the most popular and successful supervised ML methods available. The rationale for using two different algorithms stemmed from (1) the exploratory nature of this research, (2) the absence of general rule saying that SGTB is always better than RF and vice versa (performance really depends on the data and on the problem at hand), and (3) the interest in comparing predictive skill. As already remarked, we could have included other ML models in our comparison like Support Vector Machines or Neural Networks, but we were mainly interested in testing the extent to which fundamental construction attributes would carry predictive skill, independently of the classification algorithm. After briefly introducing RF and SGTB, we present and justify the methodological choices made to address class imbalance and parameter optimization, and discuss the application of the procedures in practice.

3.1. RF

The RF algorithm Breiman [10] grows many decision trees built via CART [11] and aggregates their output (majority vote here, in the case of classification). Using binary splits, decision trees recursively partition the predictor space by identifying the regions that have the most homogeneous responses to predictors [25]. Then, a constant is locally fit to each final region (or leaf): for a categorical outcome variable, it is the most probable category. As opposed to *global* models such as logistic regression, where the same equation holds over the entire data space, trees are *local* models, enabling them to adapt to and truly represent the multiple domain-specific facets of the relationships between input and output variables. RF inherits many of the advantages of trees, such as the ability to capture complex nonlinear high-order interactions among predictors, to handle highly dimensional data sets with large numbers of observations, and the robustness to outliers and to the inclusion of irrelevant predictors [63,65]. Furthermore, by growing each tree on randomly selected observations (with replacement) from the original data set, and by only trying a random subset of the input variables at each split, RF achieves much greater predictive accuracy than a single tree.

RF was selected because it stands among the most accurate general-purpose classifiers to date [5], and has shown to be effective in a variety of other fields. To cite only a few examples, the RF algorithm has been used with success to predict patient risk for various diseases [43,46], identify central genes [23], develop automated stock trading strategies

Table 1
Eighty context-free validated injury precursors from Tixier et al. [66]).

Upstream*	*n*
Cable tray	48
Cable	75
Chipping	34
Concrete liquid	58
Concrete	165
Conduit	56
Confined workspace	129
Congested workspace	13
Crane	69
Door	85
Dunnage	29
Electricity	3
Formwork	143
Grinding	133
Grout	18
Guardrail/handrail	91
Heat source	111
Heavy material/tool	79
Heavy vehicle	143
Job trailer	24
Lumber	252
Machinery	189
Manlift	66
Stud	31
Object at height	86
Piping	388
Pontoon	15
Rebar	155
Scaffold	300
Soffit	12
Spool	52
Stairs	137
Steel sections	759
Stripping	114
Tank	85
Unpowered transporter	53
Valve	79
Welding	200
Wire	131
Working at height	268
Working below elevated workspace/material	50
Drill	97
Transitional	
Bolt	186
Cleaning	119
Forklift	39
Hammer	149
Hand size pieces	172
Hazardous substance	156
Hose	95
Insect	105
Ladder	163
Mud	35
Nail	94
Powered tool	239
Screw	37
Slag	75
Spark	9
Slippery surface	142
Small particle	401
Adverse low temperatures	123
Unpowered tool	611
Unstable support/surface	8
Wind	109
Wrench	110
Lifting/pulling/manual handling	553
Light vehicle	133
Exiting/transitioning	132
Sharp edge	47
Splinter/sliver	41
Repetitive motion	66
Working overhead	14
Transitional	
Downstream	
Improper body position	88
Improper procedure/inattention	57
Improper security of materials	87
Improper security of tools	28
No/improper PPE	23
Object on the floor	174
Poor housekeeping	2
Poor visibility	12
Uneven surface	59

* Upstream precursors can be anticipated as soon as during the design phase; transitional precursors are generally not identifiable by designers but can be detected before construction begins based on knowledge of construction means and methods; and downstream precursors are mostly related to human behavior and can only be observed during the construction phase. Note that the original list of attributes is due to Desvignes [22], but minor modifications were made by Tixier et al. [66].

[6], forecast air traffic delays [56], analyze the risk of mortgage prepayment [47], determine the likelihood that a customer will cease doing business with a company [75], predict horse race outcomes [77], and to evaluate the likelihood of being elected to the baseball hall of fame [33].

The tuning parameters of RF are the number *ntree* of trees in the forest, and the number *mtry* of predictors randomly considered as candidates at each split. The "randomForest" package [48] of the R programming language [54] was used in this study to build all the RF models.

3.2. SGTB

Like RF, Boosting is an ensemble approach that combines many base models and let them vote to generate forecasts [34]. Because it can turn an ensemble of weak classifiers (each only slightly better than random guessing) into a strong classifier, Boosting was qualified as being one of the most powerful advances in ML in the last 20 years ([41], p. 337). Like RF, Boosting is often used with decision trees as base models, as it has proven extremely effective in that case ([41], p. 340). However, while RF grows large trees in parallel, Tree Boosting builds a sequence of very small trees, such that each successive tree focuses on capturing the regions of the training set that were missed by the preceding one.

SGTB [35,36] is an improvement of Tree Boosting where the gradient of some differentiable loss function is used to identify the regions missed, and a random subsample of the training set (instead of the full training set) is used to fit and add each new tree to the model.

Table 2
Safety outcomes predicted.

Energy source	Injury type	Body part	Injury severity
Biological	Caught in or compressed	Head	Pain
Chemical	Exposure to harmful substance	Neck	First aid
Electricity	Fall on same level	Trunk	Medical case
Gravity	Fall to lower level	Upper extremities	Lost work time
Mechanical	Overexertion	Lower extremities	Permanent disablement
Motion	Struck by or against		Fatality
Pressure	Transportation accident		
Radiation			
Thermal			

Table 3
Number of observations for each level of the four safety outcomes predicted.

Energy source	*n*	Injury type	*n*	Body part	*n*	Severity	*n*
Biological	108	Caught in or compressed	334	Head	899	Pain/First aid	1521
Chemical	197	Exposure to harmful substance	496	Neck	61	Medical case	206
Gravity	1030	Fall on same level	570	Trunk	354	Lost work time	101
Mechanical	74	Overexertion	594	Upper extremities	1532	Total	1828
Motion	2780	Struck by or against	2401	Lower extremities	710		
Pressure	47	Total	4395	Total	3556		
Thermal	151						
Total	4387						

Some examples of loss functions are the squared error (for regression), or multinomial deviance (used here, for classification). In this study, SGTB models were created with the "gbm" R package [38].

SGTB has five tuning parameters. The first is the number *n.tree* of trees in the sequence. A high number of trees is needed to achieve good learning, but unlike with RF, too many trees can lead to overfitting on noisy data sets [49], so close monitoring of *n.tree* is indispensable. Overfitting describes the instance when a too complex model encodes the peculiarities of the training data (i.e., the noise) as rules rather than its general structure (i.e., the signal). It always deteriorates extrapolation. The second parameter of Boosting is the size of the trees, which is controlled by *interaction.depth*. This parameter is very important, as it defines the order of predictor–predictor interaction that can be captured. For instance, specifying trees with two final nodes (one single split) allows only main effects to be modeled. Trees with three final nodes (two splits) allow first-order (two-variable) interactions to be captured, and so forth ([41], p. 362). The third parameter is the *learning.rate*, which is a factor between 0 and 1 that shrinks the contribution of each new tree added in the series. By delaying the point when overfitting is reached, low values of *learning.rate* (<0.1) allow more trees to be added to the sequence, which dramatically improves performance [35]. The fourth parameter is the minimum number *n.min* of observations allowed per node. Larger values of *n.min* generate smaller trees, which are less sensitive to noise. The proportion of training examples randomly drawn at each round is the fifth and last tuning parameter, called the *bag.fraction*.

3.3. Class imbalance issue

Our data set featured some significantly underrepresented categories, which is a well-known issue in areas like gene profiling, credit card default, or fraud detection [64,78,79]. Learning from such data sets is a challenge for all ML algorithms, including RF and SGTB [21]. Actually, the problem mainly lies in the *absolute rarity* of the minority class training examples [42,72]. For example, *pressure*, the minority class for the safety outcome *energy type*, featured only 47 training examples. This is definitely not a lot of observations in absolute terms, and represents an imbalance of 1 to 60 compared to the majority class, *motion* (2780 observations). Other categories, such as *mechanical* (74) or *biological* (108) were also severely underrepresented. For the safety outcome *body part*, the minority class (*neck*) comprised only 61 observations, as compared to the 1532 training cases of *upper extremities* (imbalance of 1:25).

Often in such situations, the final ML models do well for the majority classes, but neglect the minorities [1,15,62]. This was a critical issue in this study because accurately predicting the rare categories was at least as important as predicting the majority ones.

To address class imbalance for the RF models, we used stratified oversampling [16,21]. By growing each tree of the forest on a random sample containing more training examples from the minority classes than what would have been obtained by pure chance, oversampling allowed the underrepresented concepts to become more important from the perspective of the learning algorithm, while preserving all the information from the majority categories. This strategy was implemented in R using the *sampsize* argument of the "randomForest" function [48]. For the SGTB models, oversampling was used ahead of model building so that the number of cases from each class matched optimal proportions. This technique produced the same effect as stratified oversampling, by rebalancing the probabilities of randomly drawing examples from each class.

One should note that improvement for the underrepresented categories is always attained at the expense of a decrease in accuracy for the majority classes, regardless of the method used to address class imbalance [16]. Under the severe class imbalance we faced, attaining low error for all categories was impossible. Rather, our goal was to rebalance the overall error between all categories to improve accuracy for the minority classes without losing much accuracy for the majority categories. To achieve best performance, resampling proportions were therefore integrated to the parameter tuning protocols of RF and SGTB, following the recommendation from Sun et al. [62]. We describe these procedures in what follows.

3.4. Parameter optimization

This section describes how the optimal parameter values of the models were found. As was previously explained, one RF model and one SGTB model were created for each of the four safety outcomes that were to be predicted, that is, (1) *energy type* involved, (2) *injury type*, (3) *body part* affected, and (4) *injury severity*. This gave four RF and four SGTB models. Parameter optimization is a fundamental step of statistical learning that seeks to find the optimal level of model complexity, that is, the right tradeoff between training and predictive performance [4]. The overall strategy consists in searching through the parameter space and recording predictive error in terms of an objective function selected by the user. The combination of parameters minimizing the objective function gives the optimal model. The choice of the objective function and of the searching scheme is often dictated by the dimensionality of the parameter space, the computational resources available, and the nature of the ML algorithm [19]. In what follows, we describe the approach we adopted to tackle parameter optimization.

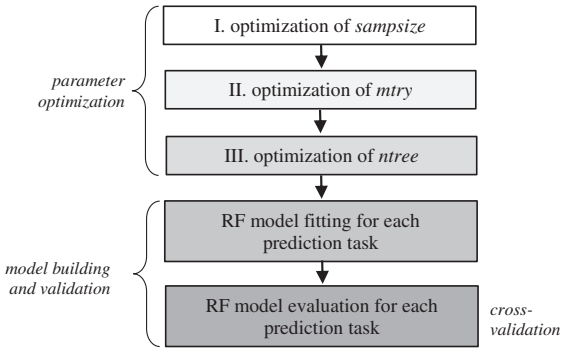

Fig. 3. Overview of the parameter tuning and model evaluation procedure for RF.

3.4.1. Parameter optimization for RF

As already explained, the tuning parameters of RF are the total number of trees *ntree*, and the number *mtry* of predictors randomly tested at each split. As was also already explained, class imbalance was addressed using stratified oversampling. The first step of the optimization procedure involved finding the best stratified bootstrap proportions (*sampsize* parameter). Then, *mtry* and *ntree* were optimized in sequence, as shown in Fig. 3.

3.4.1.1. Step 1: optimization of the sampsize parameter. Fig. 3 shows the procedure followed to determine the best stratified oversampling proportions. Initially, each category was assigned a weight inversely proportional to the number of observations it contained. For instance, as summarized in Table 3, the safety outcome *body part* featured 5 levels: *neck* (61 training examples available), *head* (899), *trunk* (354), *upper extremities* (1532), and *lower extremities* (710). The initial weights for this safety outcome were therefore 1532/61 for *neck*, 1532/899 for *head*, 1532/354 for *trunk*, 1532/1532 for *upper extremities*, and 1532/710 for *lower extremities*.

Randomly drawing with replacement from each class according to these weights generated samples of the original training set where each class was equally represented. Continuing with the body part example, 1532 observations were randomly sampled from each category, making for an initial balanced sample of 7660 observations.

Finally, based on the "out-of-bag" (OOB, [7]) error estimate of the resulting RF model, the classes associated with higher error rates were given more weight, and vice versa. As shown in Fig. 4, this manual trial and error process was repeated until the error was evenly distributed between all classes. We used the OOB error rate estimate as a surrogate for predictive accuracy since it has been proven to be unbiased and at least as accurate as cross-validation [7,74]. Consequently, costly cross-validation procedures could be avoided at this time. Also, because testing many different combinations of weights was usually required before reaching a satisfying between-class error balance, the RF models were at this stage fitted with standard, affordable values of the *mtry* and *ntree* parameters (respectively, 20 and 81). The final weights and *sampsize* values for each model (each prediction task) are given in Table 4.

3.4.1.2. Step 2: optimization of the mtry parameter. The function "tuneRF" from the "randomForest" R package [48] was used to determine the best value of the *mtry* parameter, with arguments *stepFactor* = 1.2, *improve* = 0.01, and *ntreeTry* = 100. This optimization process can be described as: (1) take the initial value of *mtry* to be the largest integer not greater than the default value ($\sqrt{p}$) recommended by Breiman [10]) for classification; (2) fit a RF model with this initial value of *mtry*, and record the OOB error estimate; (3) determine the best search direction by looking to the left (largest integer not greater than $\sqrt{p}/stepFactor$) and to the right (largest integer not greater than $\sqrt{p} \times stepFactor$) of the initial value of *mtry*, fitting a RF model for each direction (each candidate value of *mtry*), and selecting the direction (the value of *mtry*) that maximizes the gain in OOB error reduction; (4a) do not start the search if none of the directions leads to a decrease in OOB error greater than the *improve* parameter (in that case select the initial value $\sqrt{p}$ as the best value of *mtry*); (4b) otherwise, conduct the search in the best direction, by iteratively fitting one RF model for each successive value of *mtry*, and recording the OBB error; (5) stop when iterating (i.e., dividing by *stepFactor*, for searches to the left, or multiplying by *stepFactor*, for searches to the right) does not yield a reduction in OOB error greater than *improve* and return the final value of *mtry* as the best value.

For the four safety outcomes *body part*, *energy source*, *injury type*, and *injury severity*, the best direction was always the right. The best values of *mtry* found are shown in Table 5.

3.4.1.3. Step 3: optimization of the ntree parameter. Eight different values of *ntree* (101 to 801, by 100) were compared based on 36 runs of "leave-5%-out" cross-validation. The proportion of training examples left out was set to 5% (rather than 10% or 20%) due to class imbalance, in order to avoid discarding too many training observations from the minority classes at each run. Cross-validation ([41], Section 7.10) is a general and standard procedure used to optimize the parameters and objectively estimate the predictive skill of any model [44]. More

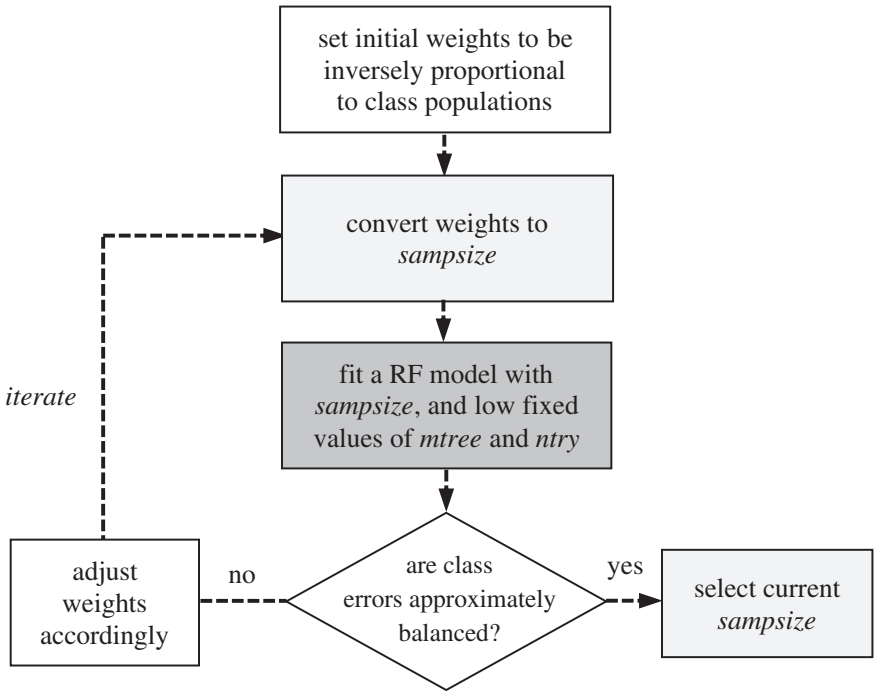

Fig. 4. Class error balancing procedure for the RF models.

Table 4
Optimal weights and values of the *sampsize* parameter for each prediction task (RF).

Body part	**Head**	**Neck**	**Trunk**	**Upper extremities**	**Lower extremities**			**Total**
Size	899	61	354	1532	710			3,556
Weights	1.5	7.26	3.2	1.07	1.45			
Sampsize	1348	443	1133	1633	1030			5,587
Energy source	**Biological**	**Chemical**	**Gravity**	**Mechanical**	**Motion**	**Pressure**	**Thermal**	**Total**
Size	108	197	1030	74	2780	47	151	4,387
Weights	6.5	3.5	3.13	9.5	1.17	14.74	4.5	
Sampsize	702	690	3219	703	3239	693	680	9,926
Injury type	**Caught**	**Exposure**	**Fall**	**Overexertion**	**Struck**			**Total**
Size	334	496	570	594	2401			4,395
Weights	5.25	1	2.25	5.5	1.5			
Sampsize	1753	496	1282	3267	3602			10,400
Severity	**Pain/first aid**		**Medical case**		**Lost work time**			**Total**
Size	1521		206		101			1,828
Weights	1		4.66		6.66			
Sampsize	1521		960		672			3,153

precisely, as shown in Fig. 5, 5% of the observations were randomly put aside (without replacement) of the full data set at each round. This set of observations constituted the testing set. The model was trained on the remaining observations, called the training set. It should be emphasized that the training and the testing sets were mutually exclusive (as is always the case with cross-validation). The model learned the mapping between the input variables (i.e., the predictors) and the target variable (i.e., the safety outcome) from the training set. Then, it was provided with the predictor portion of the testing set and asked to predict the target variable. Predictive skill was then evaluated by comparing the probabilistic forecasts that had been generated by the model to the true known values of the target variable. As will be discussed in a following section, predictive skill was measured in terms of the Rank Probability Skill Score (RPSS, [73]). The optimal parameter values found are shown in Table 5.

3.4.2. Parameter optimization for SGTB

The procedure followed is summarized in Fig. 6. As previously explained, SGTB requires the selection of an appropriate loss function, and the tuning of five parameters: the (1) number of trees in the sequence *n.tree*, the (2) maximal order of interaction that can be captured *interaction.depth*, the (3) minimum number of observations in each leaf *n.min*, the (4) *learning.rate*, and (5) the proportion of observations that are drawn at random from the original data set to grow each tree of the sequence, *bag.fraction*. The loss function appropriate for the multiclass classification problems of this study was the multinomial deviance [57]. As a preliminary step, all parameters except *n.tree* were set to values recommended by the literature. Then, oversampling was used to address class imbalance. Finally, *n.tree* was optimized using cross-validation.

3.4.2.1. Step I: setting bag.fraction, interaction.depth, n.min, and learning rate to standard values. In theory, the value of *interaction.depth* should be chosen to reflect the true order of interaction prevailing in the underlying process studied. However, most of the time, it is unknown ([41], p. 363, [25]), and this research was no exception. However, in practice, low order interactions tend to dominate, and capturing them is generally sufficient to explain most of the interplay between input and output variables ([41] p. 363, [35]). Also, it was empirically shown that values between 4 and 8 give best results, and that all the values in that range can be considered equivalent ([41], p. 363). Therefore, we set *interaction.depth* to a value of 5.

Table 5
Optimal parameter values for each prediction task (RF).

	mtry	*n.tree*
Energy source	44	201
Injury type	37	701
Body part	31	601
Injury severity	26	701

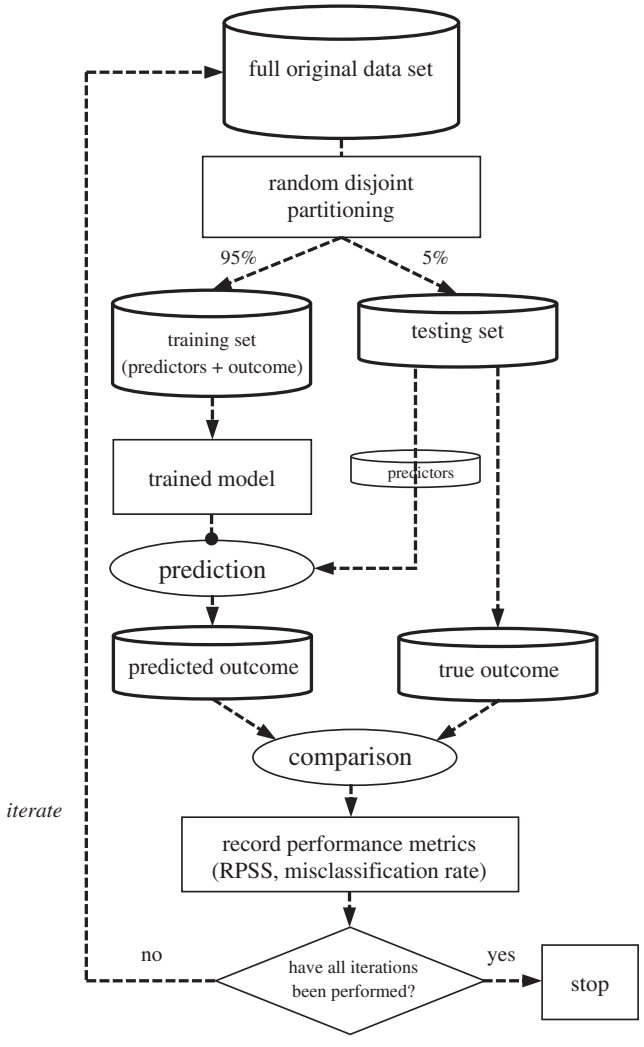

Fig. 5. Leave 5% out cross-validation procedure.

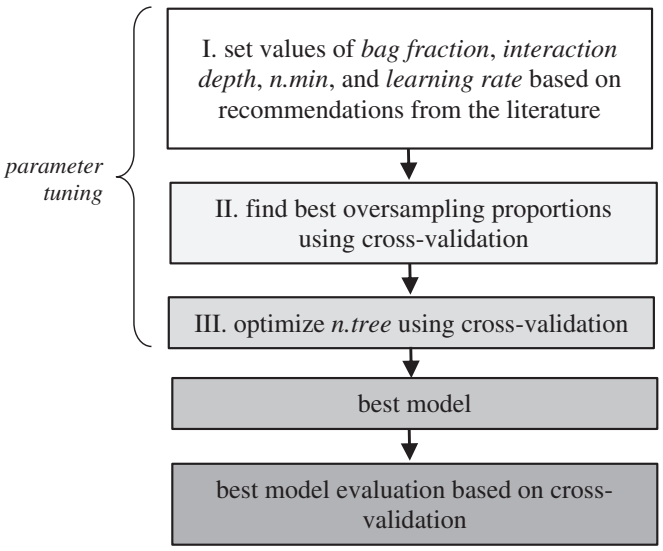

Fig. 6. Parameter optimization and model evaluation procedure used for SGTB.

The *bag.fraction* parameter was set to 0.5 for all prediction tasks since it was found in practice that the best values for this parameter were constantly around 0.5 [25,36,57]. Experiments with neighboring values did not yield any improvement in accuracy, corroborating our choice. Furthermore, following Ridgeway [57], the *learning.rate* parameter was set to 0.005 as this value was reasonably low while still being computationally feasible. For the safety outcome *injury severity*, the procedure with the 0.005 learning rate was aborted after more than twice the training time for the other safety outcomes had passed (a few hours). We then decided to increase the learning rate to 0.01, which is still a decently low value. Note that Ridgeway's [57] advice about "using the lowest learning rate that is still computationally feasible" is subjective, and depends on the machine used and the researchers' position. Ours was to avoid brute-force training since in our experience the gain in predictive skill at the end is not significant. Rather, most of the skill comes from the quality of the features. For already low values of the learning rate, decreasing it even more only returns increasingly marginal gains. Finally, a standard value of 5 was used for *n. min*, the minimum number of observations allowed per leaf.

3.4.2.2. Step II: tackling class imbalance. At step II, we used oversampling to address the class imbalance issue previously explained. Starting with all classes equal in terms of number of observations, oversampling proportions were adjusted (i.e., cases were duplicated) until the misclassification rate was approximately equally shared among all classes. Combinations were compared on the basis of 16 runs of leave 5% cross-validation, with an affordable value of *n.tree* (1200) to ensure fair convergence without risking to overfit. The best sampling proportions found are summarized in Table 6.

3.4.2.3. Step III: optimizing n.tree. Finally, at step III, the *n.tree* parameter was optimized. We followed best practice which consists in finding the optimal number of trees by cross-validation after the value of the *learning.rate* has been set ([41], p. 365; [35,57]). This step was implemented by using the R "gbm" function [38] which offers internal cross-validation (we used 8-fold cross-validation here). A sufficiently large initial value of *n.tree* was prescribed in order to let the "gbm" function find the inflection point when the models began to overfit the data. This stopping value corresponded to the optimal tradeoff between goodness of fit and generalization ability. The optimal parameter values found for each prediction task are reported in Table 7.

3.5. Measuring predictive skill with RPSS

We used the Rank Probability Skill Score (RPSS, [73]) to evaluate the predictive skill of the models. The RPSS is a widely used metric in climatology where probabilistic forecasts are common. Such forecasts, as illustrated in Fig. 2, assign a probability of occurrence to each level of the output variable instead of providing a single "best guess" prediction. Because it strongly penalizes confident forecasts of the wrong categories, the RPSS can be considered to be a stringent test of model performance (e.g., [37]). In this study, using this metric was even more harsh since the RPSS assumes the categories to be ordered (e.g., low, medium, high), and penalizes forecasts more severely when their probabilities are further from the actual outcome [32]. The "rps" function from the "verification" R package [52] was used to compute the RPSS.

$$\mathrm{RPSS} = 1 - \frac{\overline{\mathrm{RPS}_{\mathrm{forecast}}}}{\overline{\mathrm{RPS}_{\mathrm{reference}}}}. \quad (1)$$

Equation 1: Rank Probability Skill Score (RPSS)

As one can see from Eq. (1), the RPSS takes the ratio of the average Rank Probability Score (RPS) of the forecasts generated by the model and the

Table 6
Optimal resampling proportions and final numbers of cases in the resampled data sets for each prediction task (SGTB).

Body part	**Head**	**Neck**	**Trunk**	**Upper extremities**	**Lower extremities**			**Total**
Original proportions	899	61	354	1532	710			3,556
Weights	1.33	8	3.33	1	1.33			
Resampled proportions	1200	488	1180	1532	947			5,346
Energy source	**Biological**	**Chemical**	**Gravity**	**Mechanical**	**Motion**	**Pressure**	**Thermal**	**Total**
Original proportions	108	197	1030	74	2780	47	151	4,387
Weights	1	3	6	15	2	20	2	
Resampled proportions	108	591	6180	1110	5560	940	302	14,791
Injury type	**Caught**	**Exposure**	**Fall**	**Overexertion**	**Struck**			**Total**
Original proportions	334	496	570	594	2401			4,395
Weights	11	1	3	6	2.33			
Resampled proportions	3674	496	1710	3564	6403			15,847
Injury severity	**Pain/first aid**		**Medical case**		**Lost work time**			**Total**
Original proportions	1521		206		101			1,828
Weights	1		6		8			
Resampled proportions	1521		1236		808			3,565

Table 7
Optimal parameter values for each prediction task (SGTB).

	Interaction depth	*Bag fraction*	*Learning rate*	*n.min*	*n.tree*
Energy source	5	0.5	0.005	5	1200
Injury code	5	0.5	0.005	5	1550
Body part	5	0.5	0.005	5	900
Injury severity	5	0.5	0.01	5	4000

average RPS of some reference. We chose the reference to match the frequencies observed in the data. As shown in Eq. (2), the Rank Probability Score (RPS, [71]) measures the squared error between the cumulative probability mass function of a given forecast and that of a given observation. It takes on positive values, zero indicating a perfect prediction. As a result, the RPSS takes on values from -∞ to 1, where 1 indicates a perfect forecast, and 0 indicates that the model's skill is equivalent to that of the reference. Negative values mean that the model does worse than the reference. Typically, for three-class classification tasks, modest predictive skill is associated with RPSS in the [0.05,0.20] range [37]. Note that the more categories to be predicted, the harder it gets for a model to obtain high RPSS values.

$$\mathrm{RPS} = \sum_{k=1}^{K} (Y_k - O_k)^2 \tag{2}$$

Equation 2: Rank Probability Score (RPS)

Where K is the number of categories of the output variable, $Y_k = \sum_{i=1}^{k} y_i$ is the cumulative vector of forecasted values, and $O_k = \sum_{i=1}^{k} o_i$ is the cumulative vector of the observations. y_i is the probabilistic forecast for the event to happen in category i, and $o_i = 1$ if the observation is in category i, else 0.

4. Results and interpretation

The performance of each of the four RF and four SGTB models was evaluated by recording RPSS for 36 runs of "leave-5%-out" cross-validation (see Fig. 5). At each iteration, (1) 5% of the observations were randomly put aside without replacement from the original data set, (2) the models with the optimal parameter values determined previously were trained on the remaining 95% of observations, and finally (3) the models were tested on the 5% of left-out observations. The numbers of observations in the testing set at each round were 178, 220, 220, and 92 for the safety outcomes *body part*, *energy type*, *injury type*, and *injury severity*, respectively. These steps were repeated 36 times. The RPSS values reported in this study can, therefore, be considered highly reliable as they were computed for each model from several thousands of predictions for brand new, never seen observations.

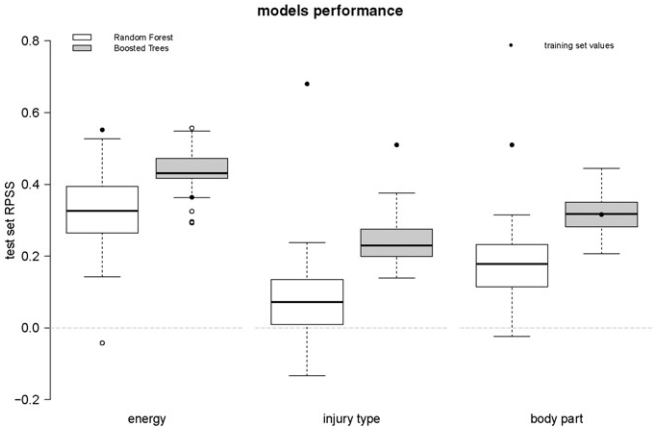

Fig. 7. Predictive skill for the first three prediction tasks, as measured by RPSS recorded in 36 runs of cross-validation.

Fig. 7 represents the distributions (as boxplots) of the RPSS values of the RF and SGTB models for the safety outcomes *energy type*, *injury type*, and *body part*. The thick black bars represent the median values, and the circles filled in black the values on the full original data sets. The dotted horizontal line passing through the origin indicates a RPSS of zero (same skill as the reference). The mean and median RPSS values are reported in Table 9.

It can be clearly seen from Fig. 7 that skill is significantly better than that of random predictions. More precisely, the median RPSS values are included between 0.172 and 0.319 for the RF models, and between 0.236 and 0.436 for the Boosting models. This indicates medium-high to very high skill, especially considering the large number of classes to be predicted (at least 5 for each prediction task). Indeed, according to Goddard et al. [37], modest skill is associated with RPSS values in the [0.05,0.20] range, and very high skill is associated with RPSS values of 0.4 and above. The best performance (mean RPSS of 0.436) was attained by a SGTB model, for the prediction of the outcome *energy type*. This represents a relative improvement of 276% over the 0.116 RPSS of the best model proposed by Esmaeili et al. [30], possibly justifying the choice of ML over parametric modeling. The high predictive skill of our models can also be viewed as a proof of the validity and promising potential of the attribute-based framework. Indeed, these results show that fundamental construction attributes do carry predictive power and thus make good ML features, and that skillful and useful multi-categorical forecasts can be issued for various safety outcomes.

One can also notice from Fig. 7 that all SGTB models consistently outperform their RF counterparts. This is in accordance with Caruana and Niculescu-Mizil [14] who compared boosted trees and RF for 11 binary classification problems and found that boosted trees have a slight edge over RF (performance was evaluated based on 8 different metrics). This superiority can be partly explained by the fact that RF can only reduce error through decreasing variance (where *error* = *bias* + *variance*), while Boosting reduces error on both fronts ([41], p. 588).

An example of probabilistic forecasts generated by the SGTB model for the safety outcome *injury type* (median RPSS 0.230) is provided in Table 8, along with the true response. Despite the model being the least skillful of the three SGTB models, the most likely class differs from the true response only once (marked in bold in the last column). The shades of grey indicate the magnitude of the probabilities assigned (the greater the probability, the darker).

It is interesting to note that while the reasoning behind certain predictions shown in Table 8 is clear (e.g., *heat source* ➔ *exposure to harmful substance*, *small particle* ➔ *struck*), in some other cases, the combinations of attributes are more complex and the most likely outcome is not as obvious or intuitive. It is in those very situations that our predictive models prove the most useful, by leveraging empirical data to guide decision-making under uncertainty.

Interestingly, as shown in Fig. 8, both the RF and the SGTB model performed worse than the reference for the prediction of the fourth and last safety outcome, *injury severity*. One explanation for this absence of predictive skill is that injury severity may not be predictable from combinations of fundamental attributes alone. Additional predictive layers may be required, such as the amount of energy present in the environment [3]. Also, it should be noted that a random component obviously plays a role in dictating injury severity. For instance, a worker slipping on ice may feel discomfort (pain), twist their ankle (first aid, medical case, or lost work time), or even badly fall backwards and sustain a head trauma (permanent disablement or fatality). Thus, in the same situation, injuries of radically different severity levels can occur based on pure chance only. This injects a lot of noise in the data. Finally, description of injury severity is impacted by reporting practices. The same injury can be classified as

Table 8
Example of probabilistic forecasts issued by the SGTB model for *injury type*.

Attributes	Caught in or compressed	Exposure to harmful sub.	Fall on same level	Overexertion	Struck by or against	Truth
Hose, object on the floor	0.026	0.002	0.702	0.187	0.083	Fall
Ladder	0.212	0.006	0.049	0.274	0.459	**Caught**
Grinding, small particle	0.010	0.001	0.004	0.015	0.969	Struck
Concrete, formwork, heavy mat. tool, rebar, exiting/transitioning	0.109	0.009	0.224	0.447	0.210	Overexertion
Insect	0.017	0.926	0.003	0.02	0.033	Exposure
Small particle	0.0194	0.001	0.005	0.0186	0.956	Struck
Rebar, wire, lifting pulling manual handling	0.107	0.003	0.027	0.200	0.663	Struck
Heat source, piping	0.055	0.863	0.005	0.031	0.047	Exposure

pain, first aid or medical case based only on whether the injured worker chose to seek medical attention, and whether they were evaluated directly onsite or transported to some external medical facility. Again, this injects a lot of noise in the data. Nevertheless, despite the low skill observed, the probabilistic forecast for *injury severity* could serve as a measure of *potential severity* or *potential risk of severe injury*, which can be of significant use in risk-based safety decision-making.

5. Conclusions, limitations and recommendations

Traditional construction safety research is limited as it was built on the assumption of independence of tasks and is primarily based upon expert opinion or subjective, aggregated, or secondary data. The attribute-based framework introduced by Esmaeili and Hallowell [26,28] provided the basis for addressing both limitations, by showing possible the extraction of universal and structured safety information from raw, unstructured injury reports. However, the framework had yet to be used to its full potential due to the high cost of manual content analysis and the limitations of the statistical tools previously used for prediction. The recourse to an extended list of attributes validated by past research [22,70]and to a highly accurate NLP system [66] allowed a large data set of 4400 attributes and safety outcomes to be constituted. Then, we applied two state-of-the art machine learning (ML) algorithms, Random Forest (RF) and Stochastic Gradient Tree Boosting (SGTB), to this structured data set. Using binary fundamental construction attributes as input, the resulting models predict three safety outcomes out of four with high skill ($0.236 < RPSS < 0.436$), namely *injury type*, *energy type*, and *body part*. This clearly outperforms the models developed in past research in terms of skill (276% relative improvement over [30]) but also in terms of variety of outcomes predicted. It is also to be noted that the SGTB models systematically reached higher predictive skill than their RF counterparts.

5.1. Contributions to theory

The high predictive skill reached shows that construction injuries do not occur in a chaotic fashion, but rather that underlying patterns and trends exist and can be uncovered and captured via statistical learning when applied to sufficiently large data sets. This finding suggests that construction safety should be studied empirically and scientifically like natural phenomena rather than strictly being qualitatively approached through the analysis of subjective, aggregated, or secondary data; expert-opinion; and with a regulatory and managerial perspective. Thus, this line of inquiry opens the gate to a new research field, where construction safety is considered an empirically grounded quantitative science. The high predictive skill reached also acts as evidence that the attribute-based framework is viable, as it produces valuable structured data from unstructured injury reports. Especially, it shows that the feature engineering of Prades Villanova [70] and Desvignes [22] was successful. It also justifies the choice of algorithmic modeling over parametric modeling.

The absence of skill for the output variable *injury severity* suggests that unlike other safety outcomes, *injury severity* is mainly random, or that additional layers of predictive information should be taken into account in making predictions, such as the amount of energy in the environment [3]. Future research should try to incorporate such energy-based data into the predictive models to test whether skill can be improved for *injury severity*. However, large-scale gathering of this information remains a challenge as it does not seem to be readily accessible from text. Nonetheless, current predictions *for injury severity* can be used as an estimate of *potential injury severity risk*.

Also, it should be noted that current predictions are conditional on the occurrence of an accident. Indeed, all that can be learned from attribute and outcome data extracted from injury reports is what happens when an accident occurs. Making unconditional predictions would necessitate the recording of "non-accident" cases. Such data, currently unavailable, could be gathered by making random attribute-based observations of onsite conditions.

Other suggestions for future research include extracting attributes and outcomes from larger amounts of injury reports, in order to overcome the absolute rarity issue faced in this research for certain levels

Table 9
Mean and median RPSS for each prediction task.

Prediction task:	Body part		Energy source		Injury type		Injury severity	
Model	RF	SGTB	RF	SGTB	RF	SGTB	RF	SGTB
Mean RPSS	0.172	0.324	0.319	0.436	0.068	0.236	−0.1	−0.650
Median RPSS	0.170	0.318	0.326	0.432	0.0725	0.230	−0.89	−0.522

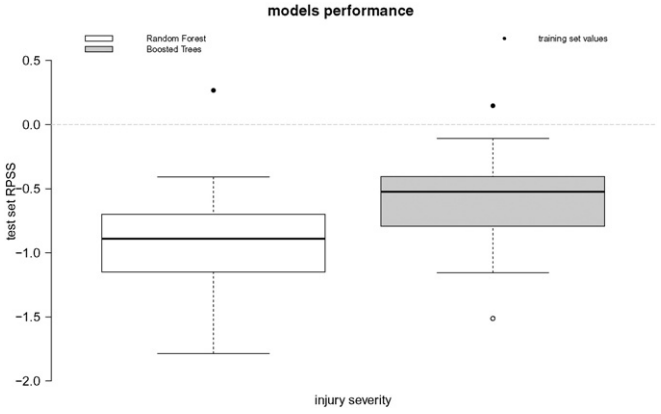

Fig. 8. Predictive skill for the last prediction task, as measured by RPSS recorded in 36 runs of cross-validation.

of the target variables. This should yield improvement in predictive skill for all prediction tasks. Also, using training data extracted from injury reports originating from other sectors than the industrial, energy, infrastructure, and mining ones would widen the range of application of the models. Another way to improve the current predictions would be to train a learning algorithm that combines the predictions of various models: RF, SGTB, but also others, such as support vector machines or artificial neural networks. This approach, known in the ML field as model *stacking*, has proven highly successful [24].

To sum up, this study makes important strides in that the final models can provide reliable probabilistic forecasts of the likely outcomes should an accident occur, for a given construction situation. This kind of predictions had been absent from the field since its inception. Safety analysts in the broader context may also find important methodological advancements in the extraction of structured data from unstructured text via NLP and the attribute-based framework, and from subsequent prediction made via ML. This combination opens the field to automated safety analysis from massive data sets (i.e., "big data").

5.2. Contributions to practice

Professionals have long aimed to add prediction to safety. The field of construction safety research has recently grown to include risk analysis, leading indicators, and precursor analysis. To achieve the goal of being predictive, practitioners have turned to expert input, particularly from knowledgeable safety professionals. However, as human beings, even the most experimented safety experts have limited personal history with injuries (thousands of worker hours), and a plethora of cognitive biases alter their judgment under uncertainty. On the other hand, the ML algorithms used in this study can learn lessons from large volumes of objective, empirical data corresponding to millions of worker hours.

This objective knowledge can be used to complement potentially biased individual opinions, leading to better-informed, safer decision-making. For example, a user simply needs to identify the attributes expected for a work package and the new models can predict, with good accuracy, the most likely type of energy, type of injury, and body part to be involved should an accident occur. Such actionable feedback can be used to better plan a worksite by removing (in time and/or space), replacing, or communicating attributes before exposure. Also, the predictions can be used to better target pre-job safety meetings. For example, a forecasted high probability of hand injury can be used to spur focused discussions about proper gloves for the task, or the prediction of a high probability for the pressure type of energy can encourage focusing hazard recognition programs on sources of pressure energy.

Finally, these predictions have great potential for integration with advanced work packing and building information modeling software as the models use binary attributes as input variables. Before construction work begins, designers, engineers, and planners can be provided with predictions of the most likely outcomes should an accident occur. Also, new configurations can be considered and objectively balanced against time, cost, and quality as a competing criterion. Safety professionals have long languished the fact that safety is considered as a fragmented function. The attribute-based framework of Esmaeili and Hallowell [26,28], coupled with the NLP tool of Tixier et al. [66] and with the methodology proposed in this study may take strides toward true, objective integration of empirical safety data within construction planning and design.

Acknowledgments

We would like to thank the National Science Foundation for supporting this research through an Early Career Award (CAREER) Program. This material is based upon work supported by the National Science Foundation under grant no. 1253179. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We would also like to recognize Bentley Systems for their financial support.

References

[1] R. Akbani, S. Kwek, N. Japkowicz, Applying Support Vector Machines to Imbalanced Datasets, Machine Learning: ECML 2004, Springer, Berlin Heidelberg 2004, pp. 39–50.
[2] A. Albert, M.R. Hallowell, B. Kleiner, A. Chen, M. Golparvar-Fard, Enhancing construction hazard recognition with high-fidelity augmented virtuality, J. Constr. Eng. Manag. 140 (7) (2014) 04014024.
[3] D. Alexander, M. Hallowell, J. Gambatese, Energy-Based Safety Risk Management: Using Hazard Energy to Predict Injury Severity, *Proceedings of ICSC15: The Canadian Society for Civil Engineering 5th International/11th Construction Specialty Conference*, University of British Columbia, Vancouver, Canada, 2015 (June 7–10).
[4] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (1) (2012) 281–305.
[5] G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (1) (2012) 1063–1095.
[6] A. Booth, E. Gerding, F. McGroarty, Automated trading with performance weighted random forests and seasonality, Expert Syst. Appl. 41 (8) (2014) 3651–3661.
[7] L. Breiman, Out-of-bag Estimation, Technical Report, Statistics Department, University of California Berkeley, Berkeley CA 94708 1996, pp. 1–13 (1996b. 33, 34).
[8] L. Almén, T.J. Larsson, Design measures for construction site safety, Conference of the Nordic Ergonomics Society, 2012.
[9] L. Breiman, Statistical modeling: the two cultures (with comments and a rejoinder by the author), Stat. Sci. 16 (3) (2001) 199–231.
[10] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[11] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC Press, 1984.
[12] Bureau of Labor Statistics (BLS), National Census of fatal occupational injuries in 2013 (preliminary results) - Cfoi.pdfAccessed August 21, 2015 http://www.bls.gov/news.release/pdf/cfoi.pdf 2013.
[13] G. Carter, S.D. Smith, Safety hazard identification on construction projects, J. Constr. Eng. Manag. 132 (2) (2006) 197–205.
[14] R. Caruana, A. Niculescu-Mizil, An Empirical Comparison of Supervised Learning Algorithms, Proceedings of the 23rd International Conference on Machine Learning, ACM June 2006, pp. 161–168.
[15] N.V. Chawla, Data Mining for Imbalanced Datasets: An Overview, Data Mining and Knowledge Discovery Handbook, Springer, US 2005, pp. 853–867.
[16] C. Chen, A. Liaw, L. Breiman, Using Random Forest to Learn Imbalanced Data, University of California, Berkeley, 2004.
[17] M.Y. Cheng, H.S. Peng, Y.W. Wu, T.L. Chen, Estimate at completion for construction projects using evolutionary support vector machine inference model, Autom. Constr. 19 (5) (2010) 619–629.
[18] M.Y. Cheng, H.S. Peng, Y.W. Wu, Y.H. Liao, Decision making for contractor insurance deductible using the evolutionary support vector machines inference model, Expert Syst. Appl. 38 (6) (2011) 6547–6555.
[19] M. Claesen, B. De Moor, Hyperparameter Search in Machine Learning, 2015 (*arXiv preprint arXiv:1502.02127*).
[20] CPWR, The Center for Construction Research and Training, Produced with Support from the National Institute for Occupational Safety and Health Grant Number OH009762, The Construction Chart Book | CPWR, 2013 (Accessed August 21, 2015. http://www.cpwr.com/publications/construction-chart-book).
[21] S. del Río, V. López, J.M. Benítez, F. Herrera, On the use of MapReduce for imbalanced big data using Random Forest, Inf. Sci. 285 (2014) 112–137.
[22] M. Desvignes, Requisite Empirical Risk Data for Integration of Safety with Advanced Technologies and Intelligent Systems(Master Thesis) University of Colorado at Boulder, 2014.
[23] R. Diaz-Uriarte, S.A. de Andrés, Variable Selection from Random Forests: Application to Gene Expression Data, 2005 (*arXiv preprint q-bio/0503025*).
[24] P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (10) (2012) 78–87.
[25] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J. Anim. Ecol. 77 (4) (2008) 802–813.
[26] B. Esmaeili, M. Hallowell, Attribute-Based Risk Model for Measuring Safety Risk of Struck-by Accidents, Construction Research Congress May 2012, pp. 289–298.
[27] B. Esmaeili, M.R. Hallowell, Diffusion of safety innovations in the construction industry, J. Constr. Eng. Manag. 138 (8) (2011) 955–963.
[28] B. Esmaeili, M.R. Hallowell, Using Network Analysis to Model Fall Hazards on Construction Projects, *Safety and Health in Construction, CIB W*, 992011 24–26.
[29] B. Esmaeili, M.R. Hallowell, B. Rajagopalan, Attribute-based safety risk assessment. I: analysis at the fundamental level, J. Constr. Eng. Manag. (2015) 04015021.
[30] B. Esmaeili, M.R. Hallowell, B. Rajagopalan, Attribute-based safety risk assessment. II: predicting safety outcomes using generalized linear models, J. Constr. Eng. Manag. (2015) 04015022.
[31] M.A. Fleming, Hazard Recognition, By Design, ASSE 2009, pp. 11–15.
[32] K.J. Franz, S. Sorooshian, Verification of National Weather Service Probabilistic Hydrologic Forecasts, *University of Arizona, report prepared for the National Weather Service*, 2002.

[33] M.H. Freiman, Using random forests and simulated annealing to predict probabilities of election to the baseball hall of fame, J. Quant. Anal. Sports 6 (2) (2010).
[34] Y. Freund, R. Schapire, N. Abe, A short introduction to boosting, J. Jpn. Soc. Artif. Intell. 14 (771–780) (1999) 1612.
[35] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.
[36] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (4) (2002) 367–378.
[37] L. Goddard, A.G. Barnston, S.J. Mason, Evaluation of the IRI's "net assessment" seasonal climate forecasts: 1997–2001, Bull. Am. Meteorol. Soc. 84 (12) (2003) 1761–1781.
[38] Greg Ridgeway with contributions from others, gbm: generalized boosted regression models. R package version 2.1.1, http://CRAN.R-project.org/package=gbm 2015.
[39] W. Haddon, Energy damage and the ten countermeasure strategies, Hum. Factors 15 (4) (1973) 355–366.
[40] M.R. Hallowell, A Formal Model for Construction Safety and Health Risk Management(Doctoral dissertation) Oregon State University, 2008.
[41] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, vol. 2, no. 1Springer, New York, 2009.
[42] H. He, E. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[43] M. Khalilia, S. Chakraborty, M. Popescu, Predicting disease risks from highly imbalanced data using random forest, BMC Med. Inform. Decis. Mak. 11 (1) (2011) 51.
[44] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, Ijcai, vol. 14, no. 2 August 1995, pp. 1137–1145.
[45] K.C. Lam, E. Palaneeswaran, C.Y. Yu, A support vector machine model for contractor prequalification, Autom. Constr. 18 (3) (2009) 321–329.
[46] A.V. Lebedev, E. Westman, G.J.P. Van Westen, M.G. Kramberger, A. Lundervold, D. Aarsland, AddNeuroMed consortium, Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness, NeuroImage 6 (2014) 115–125.
[47] T.H. Liang, J.B. Lin, A two-stage segment and prediction model for mortgage prepayment prediction and management, Int. J. Forecast. 30 (2) (2014) 328–343.
[48] Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (3) (2002) 18–22.
[49] D. Opitz, R. Maclin, Popular ensemble methods: an empirical study, J. Artif. Intell. Res. (1999) 169–198.
[50] G.A. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information, Psychol. Rev. 63 (2) (1956) 81–97.
[51] O. Moselhi, T. Hegazy, P. Fazio, Neural networks as tools in construction, J. Constr. Eng. Manag. (1991).
[52] NCAR - Research Applications Laboratory, Verification: weather forecast verification utilities. R package version 1.42, http://CRAN.R-project.org/package=verification 2015.
[53] Bureau of Labor Statistics (Ed.), Occupational Injury and Illness Classification Manual Version 2.0, U.S. Department of Labor, September 2010 (http://www.bls.gov/iif/oiics_manual_2010.pdf).
[54] R. Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015 (URL http://www.R-project.org/).
[55] B. Rajagopalan, K. Grantz, S. Regonda, M. Clark, E. Zagona, Ensemble streamflow forecasting: methods and applications, Adv. Water Sci. Methodol. (2005) 97–116.
[56] J.J. Rebollo, H. Balakrishnan, Characterization and prediction of air traffic delays, Transp. Res. C 44 (2014) 231–241.
[57] G. Ridgeway, Generalized boosted models: a guide to the gbm package, Update 1 (1) (2007).
[58] M. Seera, C.P. Lim, A hybrid intelligent system for medical data classification, Expert Syst. Appl. 41 (5) (2014) 2239–2249.
[59] M. Skibniewski, T. Arciszewski, K. Lueprasert, Constructability analysis: machine learning approach, J. Comput. Civ. Eng. 11 (1) (1997) 8–16.
[60] L. Soibelman, H. Kim, Data preparation process for construction knowledge generation through knowledge discovery in databases, J. Comput. Civ. Eng. 16 (1) (2002) 39–48.
[61] H. Son, C. Kim, C. Kim, Automated color model-based concrete detection in construction-site images by using machine learning algorithms, J. Comput. Civ. Eng. 26 (3) (2011) 421–433.
[62] Y. Sun, M.S. Kamel, A.K. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recogn. 40 (12) (2007) 3358–3378.
[63] C.D. Sutton, Classification and Regression Trees, Bagging, and Boosting, Handbook of Statistics, 242005 303–329.
[64] Y. Tang, Y.Q. Zhang, N.V. Chawla, S. Krasser, SVMs modeling for highly imbalanced classification, IEEE Trans. Syst. Man Cybern. B Cybern. 39 (1) (2009) 281–288.
[65] R. Timofeev, *Classification and regression trees (CART) theory and applications*, (Master thesis), Humboldt University, Berlin, 2004.
[66] A.J.P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Automated content analysis for construction safety: a natural language processing system to extract precursors and outcomes from unstructured injury reports, Autom. Constr. 62 (2016) 45–56.
[67] E. Towler, B. Rajagopalan, R.S. Summers, D. Yates, An approach for probabilistic forecasting of seasonal turbidity threshold exceedance, Water Resour. Res. 46 (6) (2010).
[68] A. Tsanas, A. Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, Energy Build. 49 (2012) 560–567.
[69] A. Tversky, D. Kahneman, The framing of decisions and the psychology of choice, Science 211 (4481) (1981) 453–458.
[70] M. Prades Villanova, Attribute-based Risk Model for Assessing Risk to Industrial Construction Tasks (Master Thesis) University of Colorado at Boulder, 2014.
[71] A.P. Weigel, M.A. Liniger, C. Appenzeller, The discrete Brier and ranked probability skill scores, Mon. Weather Rev. 135 (1) (2007) 118–124.
[72] G.M. Weiss, Mining with rarity: a unifying framework, ACM SIGKDD Explor. Newsl. 6 (1) (2004) 7–19.
[73] D.S. Wilks, Statistical Methods in the Atmospheric Sciences, Elsevier, New York, 1995.
[74] D.H. Wolpert, W.G. Macready, An efficient method to estimate bagging's generalization error, Mach. Learn. 35 (1) (1999) 41–55.
[75] Y. Xie, X. Li, E.W.T. Ngai, W. Ying, Customer churn prediction using improved balanced random forests, Expert Syst. Appl. 36 (3) (2009) 5445–5449.
[76] J. Yang, O. Arif, P.A. Vela, J. Teizer, Z. Shi, Tracking multiple workers on construction sites using video cameras, Adv. Eng. Inform. 24 (4) (2010) 428–434.
[77] S. Lessmann, M.C. Sung, J.E. Johnson, Alternative methods of predicting competitive events: An application in horserace betting markets, International Journal of Forecasting 26 (3) (2010) 518–536.
[78] J. Jung, M.R. Thon, Automatic annotation of protein functional class from sparse and imbalanced data sets, Data Mining and Bioinformatics, Springer, Berlin Heidelberg 2006, pp. 65–77.
[79] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research (2002) 321–357.