Application of Machine Learning to Construction Injury Prediction

Antoine Tixier, Matthew Hallowell, Balaji Rajagopalan and Dean Bowman

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 ran two state-of-the-art ML models, Random Forest (RF) and Stochastic Gradient Tree Boosting (SGTB), on a data set of validated and carefully featured fundamental construction binary 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 two models 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 skill reached suggests that construction safety features a non-random component and should be studied empirically like other natural phenomena, 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. This kind of data-driven predictions had been absent from the field since its inception.

INTRODUCTION AND MOTIVATION

Construction is one of the largest industries in the United States, but is also one of the deadliest (Bureau of Labor Statistics 2013). Between 1992 and 2010, an average of 730 lives have been claimed each year (CPWR 2013). Despite the numerous efforts that have been motivated by this alarmingly poor performance, injury statistics have not significantly improved in the past decade (BLS 2013). 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 (Esmaeili and Hallowell 2011a). 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 (2012, 2011b) 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 its ability to capture and encode the information of every possible construction situation in a finite, standardized format, regardless of trade or project type. Therefore, as illustrated in Figure 1, extracting attributes and various safety outcomes from injury reports (i.e., objective empirical data) enables the constitution of a structured, consistent multivariate data set ideally suited for data mining, predictive modeling, and, thus, 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 Figure 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.

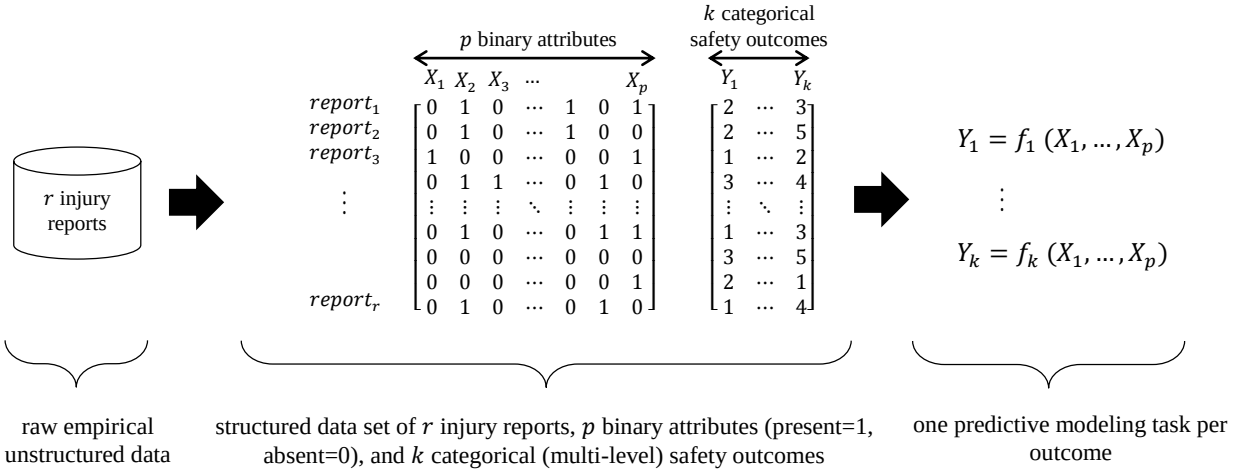

Figure 1. The derivation of predictive models from injury reports is enabled by the attribute-based framework

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., Esmaeili et al. 2015a, Esmaeili et al. 2015b, Prades 2014, Desvignes 2014, Esmaeili and Hallowell 2012, 2011b) is subject to limitations with respect to one or more of the aforementioned parameters.

Building on three recent studies (Prades 2014; Desvignes 2014; and Tixier et al, 2015, in review) that respectively addressed the limitations pertaining to the first three of the aforementioned criteria, 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.

BACKGROUND AND POINT OF DEPARTURE

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

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 (2014) 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 well known that poor hazard recognition skill is a proximal cause of risk misperception and injury in construction (Albert et al. 2014, Carter and Smith 2006). People working upstream of the construction phase, like designers, face an even greater risk of failing to recognize hazards and misestimating risk (Albert et al. 2014, Almén and Larsson 2012).

Furthermore, without even considering the limited experience problem, human judgment and intuition will always be subject to important biases and fallacies (e.g., Kahneman and Tversky 1982). Also, humans have very limited capability of inducing knowledge from large numbers of observations (Skibniewski et al. 1997). 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 (Miller 1956).

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 make safety-related decisions under uncertainty on empirical knowledge, which 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 (2014) trained ML algorithms on large numbers of health records to automatically diagnose new patients, providing physicians with an opportunity to reconsider initial decisions, which eventually increases diagnosis accuracy.

Limitations of previous work on attribute-based construction safety and points of departure

Although Esmaeili and Hallowell (2012, 2011b) 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. (2015a) took the research a step further by using commercial software to automatically extract attributes from a larger amount of reports (1,450). 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. (2015b) used the data set obtained by Esmaeili et al. (2015a) 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. 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 (Hastie et al. 2009, pp. 222-223). Another limitation of Esmaeili et al. (2015b) 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 (Towler et al. 2010, Rajagopalan et al. 2005). Also, because these relationships are unknown, parametric models are not best suited for skillful prediction.

To address the abovementioned limitations, we first used a broadened and more robust list of 80 attributes elaborated and validated by a team of 8 researchers (Prades 2014, Desvignes 2014) and slightly modified by Tixier et al. (2015, in review). This list is provided in Table 2. Second, a rather large database of 5,298 injury reports was used 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 96% accurate natural language processing (NLP) program developed by Tixier et al. (2015 in review), 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.

Construction research has considered ML for more than two decades. Moselhi et al. (1991) first discussed the potential applications of neural networks in construction engineering and management and developed a prototype that provides optimum markup estimates from attributes describing bid situations, such as the number of competitors or the contractor's estimated cost. Later, Skibniewski et al. (1997) 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 (2002) applied decision trees and neural networks to a construction management database to identify the causes of delays.

More recently, Lam et al. (2009) 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. (2010, 2011) 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 and number of rainy days, and to estimate the loss risk associated with a given project given project duration, number of floors, construction season, and geological conditions. Finally, Yang et al. (2010) developed an algorithm to automatically track workers in digital videos; Tsanas and Xifara (2012) 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. (2012) used a support vector machine model to detect concrete structural components in color images from actual construction sites.

Although far from being 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 the first time that ML is used to predict construction safety related outcomes.

Goal of this study

The goal of this study is to use RF and SGTB, two widely used and highly successful ML algorithms, on attribute and outcome data extracted from a large body of injury reports. 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 Figure 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 (2014) is useful, but predicting the most likely categories of various safety outcomes is a complementary and equally valuable strategy.

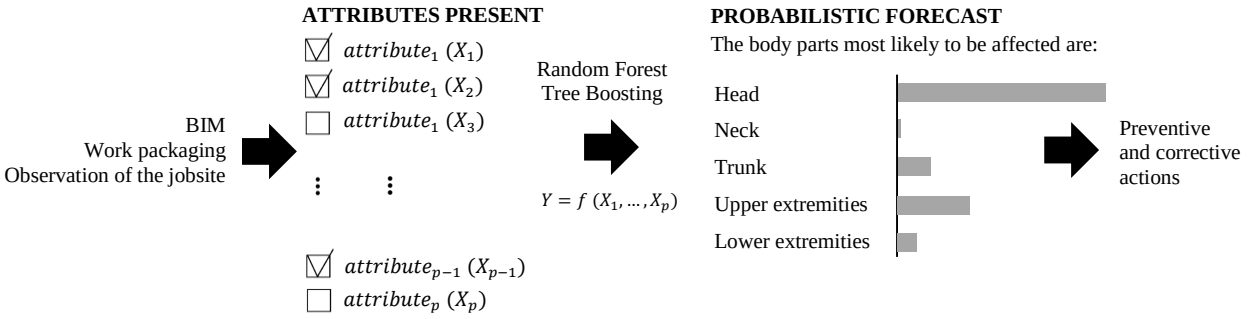

Figure 2. Practical use of the predictive models built in this study

Characteristics of the data set

We had access to a raw database of 5,298 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 (2014), Desvignes (2014), and Tixier et al. (in review). 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 (2015 in review) 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 (Fleming 2009, Haddon 1973). For *injury type*, *body part*, and *injury severity*, the classification scheme is consistent with that of the Bureau of Labor Statistics (BLS) and the Occupational Safety and Health Administration (OSHA) (BLS 2010, Hallowell 2008).

Table 1. Eighty context-free validated injury precursors from Tixier et al. (in review)

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

* 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 (2014), but minor modifications were made by Tixier et al. (in review)

Table 2. Safety outcomes predicted in this study

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			

It should be noted that Prades (2014) and Desvignes (2014) ensured the validity and relevance of the attributes identified 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 machine learning success (Domingos 2012). Also, Tixier et al. (in review) 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 tool. At each round, lessons learned from examining the errors made by the system 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 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 (Tixier et al., 2015 in review).

900 reports out of the 5,298 available were not associated with any attribute, and were therefore removed, making for a final data set of $r = 4{,}398$ observations, $p = 80$ attributes, and $k = 4$ safety outcomes (using the notation from Figure 1). An inspection of these reports showed that they were very short and did not contain any attribute-related information. The number of times each attribute appeared in the final data set is shown in Table 2. 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). The safety outcome *body part affected* could not be inferred for 831 reports, so only 3,556 were available for training for this particular target. Also, because it requires mental projection, Tixier et al.'s (2015 in review) NLP tool cannot extract the safety outcome *injury severity*, so for this prediction task, the 1,829 reports manually analyzed by Prades (2014) and Desvignes (2014) 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* (1,362) since the difference between these two severity levels is very tenuous. The counts of each category of the safety outcomes in the final data sets are presented in Table 3.

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

Energy source	***n***	**Injury type**	***n***	**Body part**	***n***	**Severity**	***n***
Biological	108	Caught in or compressed	334	Head	899	Pain/First aid	1,521
Chemical	197	Exposure to harmful substance	496	Neck	61	Medical case	206
Gravity	1,030	Fall on same level	570	Trunk	354	Lost work time	101
Mechanical	74	Overexertion	594	Upper extremities	1532	TOTAL	1,828
Motion	2,780	Struck by or against	2,401	Lower extremities	710		
Pressure	47	TOTAL	4,395	TOTAL	3556		
Thermal	151						
TOTAL	4,387						

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 Figure 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 $Y_k = f_k(X_1, \dots, X_p)$, where $(X_1, \dots, X_p)$ are the fundamental construction attributes presented in Table 2. The methods used and procedure followed to accomplish these tasks are presented next.

APPLICATION OF MACHINE LEARNING (ML)

After a brief general presentation of the Random Forest (RF) and Stochastic Gradient Tree Boosting (SGTB) machine learning algorithms, we present and justify the combination of methodological choices made to address class imbalance and parameter optimization, and discuss the application of the procedures in practice. This results in an unusual level of detail, but is necessary to guarantee transparency and reproducibility.

RF and SGTB were applied to the $r = 4{,}398$ by $p = 80$ structured data set of attributes and outcomes shown in Figure 1. 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 need to compare predictive skill.

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" (Breiman 2001a). On the other hand, ML algorithms do not assume that the data have been generated by any parametric model prescribed *a priori* 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.

Here, the features, or input variables, were the fundamental construction attributes $(X_1, \ldots, X_{80})$ listed in Table 2, such as *welding*, *uneven surface*, or *adverse low temperatures*, and the targets, or output variables, were the four 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, associated a specific combination of attributes to a specific combination of safety outcomes. Based on such training data, ML algorithms could infer rules to map combinations of attributes to safety outcomes, and use these rules later on to predict the most likely outcomes for brand new observations. In what follows, RF and SGTB are briefly introduced.

Random Forest (RF)

The RF algorithm Breiman (2001b) grows many decision trees built via CART (Breiman et al. 1984) and aggregates their output (majority vote in the case of a categorical output variable). Using binary splits, decision trees recursively partition the predictor space by identifying the regions that have the most homogeneous responses to predictors. Then, a constant is locally fit to each final region (or leaf): for a categorical outcome variable, it is the most probable category (Elith et al. 2008). 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. Random Forest 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 robustness to outliers and to the inclusion of irrelevant predictors (Sutton 2005, Timofeev 2004). Furthermore, by growing each tree on randomly selected observations (with replacement) from the original data set, and by only trying a random subset of 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 (Biau 2012), 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 (Lebedev et al. 2014, Khalilia et al. 2011), identify central genes (Díaz-Uriarte and de Andrés 2005), develop automated stock trading strategies (Booth et al. 2014), forecast air traffic delays (Rebollo and Balakrishnan 2014), analyze the risk of mortgage prepayment (Liang and Lin 2014), determine the likelihood that a customer will cease doing business with a company (Xie et al. 2009), predict horse race outcomes (Lessmann et al. 2010), and to evaluate the likelihood of being elected to the baseball hall of fame (Freiman 2010).

The tuning parameters of RF are the number of trees in the forest $ntree$, and the number of predictors randomly considered as candidates at each split $mtry$. The "randomForest" package (Liaw and Wiener 2002) of the R programming language (R Core Team 2015) was used in this study to build all the RF models.

Stochastic Gradient Tree Boosting (SGTB)

Like RF, the Boosting algorithm is an ensemble approach that combines many base models and let them vote to generate forecasts (Freund et al. 1999). 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 machine learning in the last 20 years (Hastie et al. 2009, p. 337). Like RF, Boosting is often used with decision trees as base models, as it has proven extremely effective in that case (Hastie et al. 2009, 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 (Friedman 2002, Friedman 2001) is an improvement of Tree Boosting where each new tree added to the sequence is fitted on a random subsample of the training data (without replacement), and is fitted directly on the gradient of the loss function of the current model, made of all the trees so far in the sequence In this study, SGTB models were created with the "gbm" R package (Ridgeway et al. 2015).

SGTB has five tuning parameters. The first is the number of trees in the sequence, $n.tree$. 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 (Opitz and Maclin 1999), so close monitoring of the $n.tree$ parameter is indispensable. 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 (Hastie et al. 2009, 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 in the sequence, which dramatically improves performance (Friedman 2001). The fourth parameter is $n.min$, the minimum number 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 $bag.fraction$.

Class imbalance issue

Our dataset had some categories that were significantly underrepresented, which is also commonly observed in areas like gene profiling, credit card default, or fraud detection (Tang et al. 2009, Jaehee and Thon 2006, Chawla et al. 2002). Learning from such data sets is a challenge for all ML algorithms, including RF and SGTB (del Rio et al. 2014). Actually, the problem mainly lies in the *absolute rarity* of the minority class training examples (He and Garcia 2009, Weiss 2004). 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* (2,780 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 1,532 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 minority classes (Sun et al. 2007, Chawla 2005, Akbani et al. 2004). This was a critical issue in this study because accurately predicting the rare categories was as important as predicting the majority categories.

To address class imbalance for the RF models, we used stratified oversampling (del Rio et al. 2014, Chen et al. 2004, Chawla 2002). 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, while preserving all the information from the majority categories. This strategy was implemented in R using the *sampsize* argument of the "randomForest" function (Liaw and Wiener 2002). 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 (Chen et al. 2004). Under the severe class imbalance we faced, attaining low error for all categories was not possible. 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 classes. To achieve best performance, resampling proportions were therefore integrated to the parameter tuning protocols of RF and Boosting, following the recommendation from Sun et al. (2007). We describe these procedures in what follows.

Parameter optimization

This section describes how the optimal parameter values of the models were found. As was previously explained, one RF and one Boosting model were fitted 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 which comes down in practice ensuring that the parameter values of the ML algorithms are chosen so as to minimize the predictive error, and not the error on the training set (Bergstra and Bengio 2012).

Parameter optimization for Random Forest (RF)

The parameters $sampsize$, $mtry$, and $ntree$ were optimized in sequence, as shown in Figure 3. The first step of the optimization procedure involved finding the best stratified bootstrap proportions ($sampsize$).

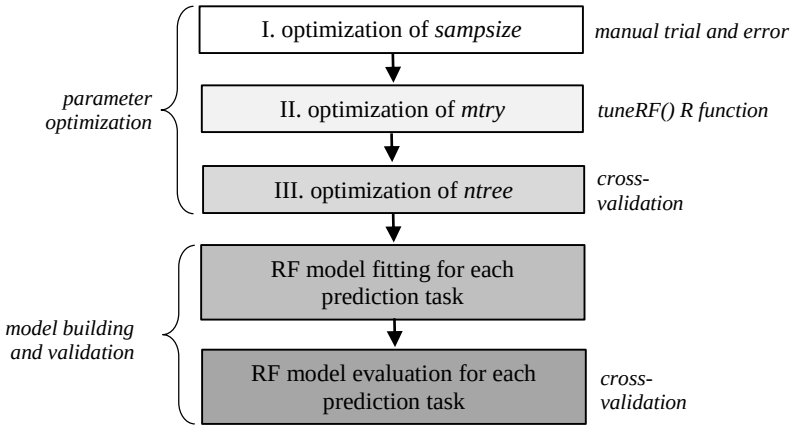

Figure 3. Overview of the parameter tuning and model evaluation procedure for RF

Step 1: Optimization of the sampsize parameter

Figure 4 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). Rounded to the nearest integer, the initial weights for this safety outcome were therefore $1532/61 = 25$ for *neck*, $1532/899 = 2$ for *head*, $1532/354 = 4$ for *trunk*, $1532/1532 = 1$ for upper extremities, and $1532/710 = 2$ 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 approximately equally represented. Continuing with the *body*

part example, the numbers of observations sampled from each category were: $25 * 61 = 1{,}525$ for *head*, $899 * 2 = 1{,}798$ for *neck*, $354 * 4 = 1{,}416$ for *trunk*, $1532 * 1 = 1{,}532$ for *upper extremities*, and $710 * 2 = 1{,}420$ for *lower extremities*, making for initial samples of 7,691 observations, where classes were represented with equal proportions (one fifth each).

Finally, based on the "out-of-bag" (OOB, Breiman 1996b) error estimate of the resulting RF model, the classes associated with higher error rates were given more weight, and vice versa. As shown in Figure 4, this manual trial and error process was repeated until the error was evenly distributed between all classes. The OOB error rate estimate was used as a surrogate for predictive accuracy since it was proven to be unbiased and as accurate (or more), than cross-validation (Wolpert and Macready 1999, Breiman 1996b). 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.

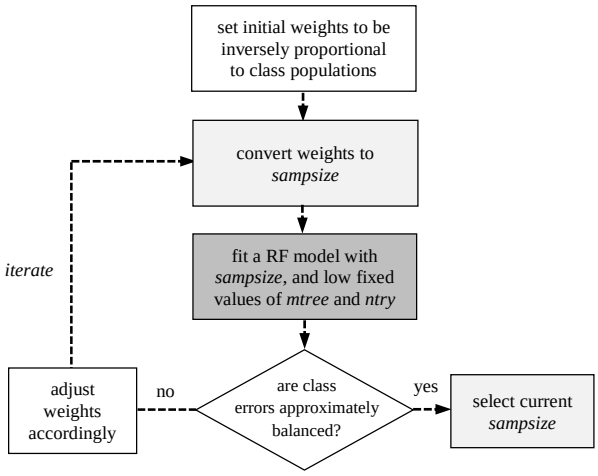

Figure 4. Class error balancing procedure for the RF models

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

Step 2: Optimization of the mtry parameter

The function "tuneRF" from the "randomForest" R package (Liaw and Wiener 2002) 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 (2001b) for classification; (2) fit a RF model with this initial value of $mtry$, and record the out-of-bag (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 new 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 the $improve$ parameter (in that case, 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 via this procedure were 31, 44, 37, and 26, respectively.

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.

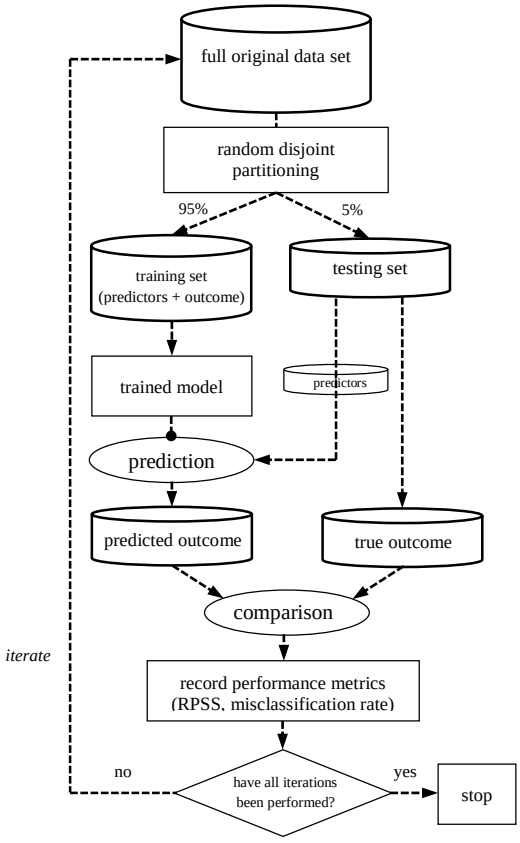

Figure 5. Leave 5% out cross-validation procedure

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

The proportion of training examples left out was set to 5% (rather than 10% or 20%) in order to avoid discarding too many training observations from the minority classes at each run. Cross-validation (Hastie et al., section 7.10) is a general and standard procedure used to optimize the parameters and objectively estimate the predictive skill of any model (Kohavi 1995). More precisely, as shown in Figure 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, the model 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 known true 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, Wilks 1995). The optimal parameter values found are shown in Table 5.

Parameter optimization for Stochastic Gradient Tree Boosting (SGTB)

As previously explained, SGTB required 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 dataset to grow each tree of the sequence, called $bag.fraction$. The loss function appropriate for the multiclass classification problems of this study was the multinomial deviance (Ridgeway 2012).

Step I

As shown in Figure 6, all parameters except $n.tree$ and the oversampling proportions were first set to values recommended by the literature. 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 (Hastie et al. 2009, p. 363, Elith et al. 2008), and this research was no exception. Because in practice, low order interactions tend to dominate, capturing them is generally sufficient to explain most of the interplay between input and output variables (Hastie et al. 2009 p. 363, Friedman 2001). 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 (Hastie et al. 2009, p. 363). Therefore, $interaction.depth$ was set to a value of 5.

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 (Ridgeway 2007, Elith et al. 2008, Friedman 2002). Experiments with neighboring values did not yield any improvement in accuracy, corroborating this choice. Following Ridgeway (2007), the $learning.rate$ parameter was set as 0.005 as this value was reasonably low while still being computationally feasible. For the safety outcome *injury severity*, 0.005 was too slow, so the $learning.rate$ was set to 0.01 for this prediction task. Finally, a standard value of 5 was used for $n.min$, the minimum number of observations allowed per leaf.

Step II

At step II, oversampling was used 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. Models were compared on the basis of 16 runs of leave 5% cross-validation, with an affordable value of $n.tree$ (1200) that ensured approximate convergence without risking to overfit. The best sampling proportions found are summarized in Table 6.

Step III

Finally, at step III, the $n.tree$ parameter was optimized. Following Ridgeway (2007) and Hastie et al. (2009, p. 365), once a value of the $learning.rate$ was selected, the optimal number of trees was found by cross-validation while keeping all the other parameters fixed. This step was implemented by using the "gbm" R function (Ridgeway et al. 2015), which offers an internal cross-validation functionality (8-fold cross-validation was used in this study). A sufficiently large initial value of $n.tree$ was prescribed in order to let the "gbm" function find the inflexion 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 tasks are summarized in Table 7.

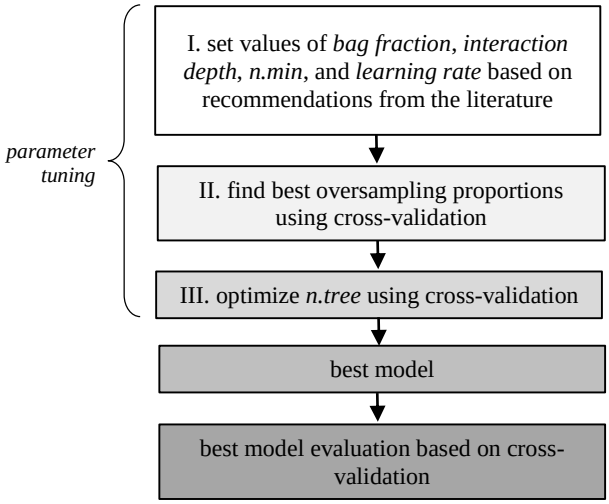

Figure 6. Parameter optimization and model evaluation procedure used for Boosting.

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

Measuring predictive skill with RPSS

We used the Rank Probability Skill Score (RPSS, Wilks 1995) to evaluate the predictive skill of the models. The RPSS is a metric widely used in climatology where probabilistic forecasts are common. Such forecasts, as illustrated in Figure 2, assigns 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., Goddard et al. 2003). In this study, using this metric was even more harsh because it 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 (Franz and Sorooshian 2002). The "rps" function from the "verification" R package (NCAR - Research Applications Laboratory 2015) was used to compute the RPSS.

$$\mathrm{RPSS} = 1 - \frac{\overline{\mathrm{RPS}_{\mathrm{forecast}}}}{\overline{\mathrm{RPS}_{\mathrm{reference}}}}$$

Equation 1: Rank Probability Skill Score (RPSS)

As one case see from equation 1, the RPSS takes the ratio of the average Rank Probability Score (RPS) of the forecasts generated by the model and the average RPS of some reference. We chose the reference to match the frequencies observed in the data. As shown in Equation 2, the Rank Probability Score (RPS, Weigel et al. 2007) 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 is equivalent to 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 range

$[0.05, 0.20]$ (Goddard et al. 2003). Note that the more categories to be predicted, the harder it gets for a model to obtain high RPSS values.

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

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.

RESULTS AND INTERPRETATION

The performance of each of the four RF and four Boosting models was evaluated by recording the RPSS for 36 runs of "leave-5%-out" cross-validation (see Figure 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 observations left-out. 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.

Figure 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 means, 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 Figure 7 that all SGTB models outperform their RF counterparts. The mean RPSS values are comprised 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. (2003), modest skill is associated with RPSS values in the range $[0.05, 0.20]$, 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 an important improvement over the 0.116 RPSS of the best model proposed by Esmaeili et al. (2015b), possibly justifying the choice of machine learning over parametric modeling. The high predictive skill of the models obtained can also be viewed as a proof of the validity and promising potential of the attribute-based framework. Indeed, these results show that attributes carry predictive power, and that skillful and useful multi-categorical forecasts can be issued for various safety outcomes.

An example of probabilistic forecasts issued 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).

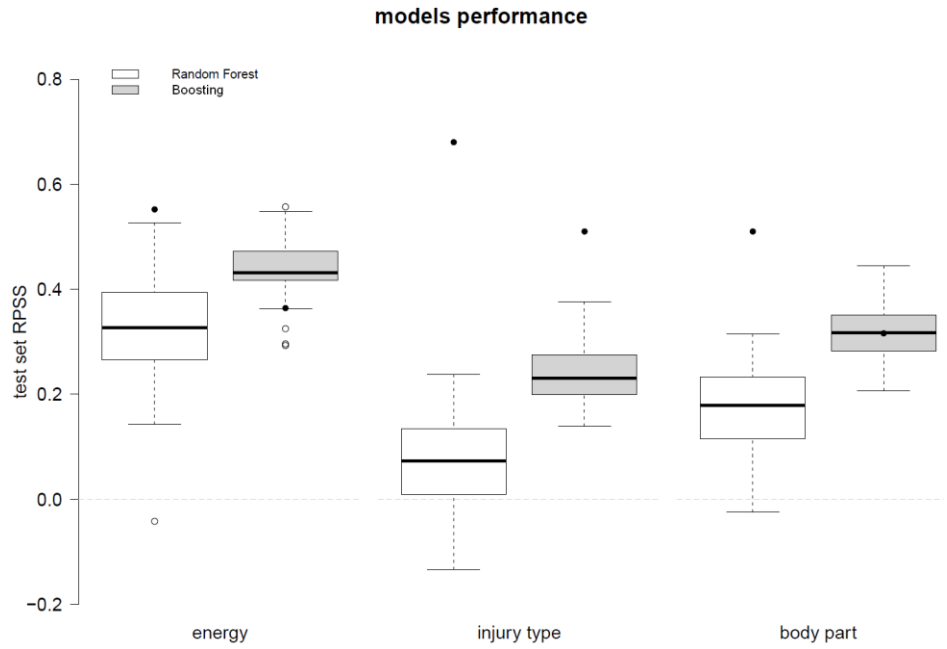

Figure 7. Predictive skill for the first three prediction tasks, as measured by RPSS recorded in 36 runs of cross-validation

It is interesting to note in Table 8 that while the reasoning behind certain predictions 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 these very situations that the predictive models developed in this research prove the most useful, by leveraging empirical data to guide decision-making under uncertainty.

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

As shown in Figure 8, both the RF model 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 simply from combinations of fundamental attributes alone. Additional predictive layers may be required, such as the amount of energy present in the environment (Alexander et al. 2015). 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 simply feel discomfort in their legs (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. Finally, injury severity as reported in accident

reports is impacted by reporting practices. The same injury can be classified as pain, first aid or medical case only based on whether the injured worker chose to seek medical attention, and whether they were evaluated directly onsite or transported to some external medical facility. While the skill of injury severity is low, the probabilistic forecast of this category could serve as a measure of *potential severity* or *potential risk of severe injury,* that can be of significant use in risk-based safety decision.

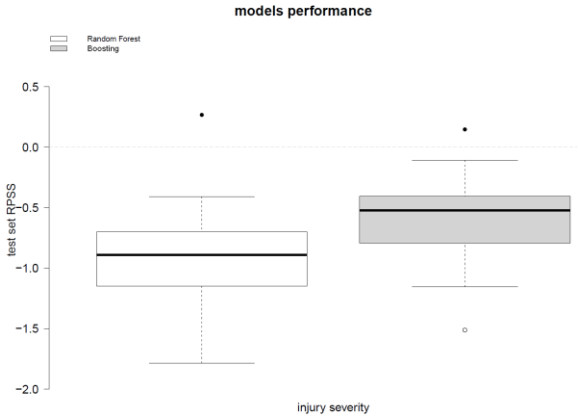

Figure 8. Predictive skill for the last prediction task, as measured by RPSS recorded in 36 runs of cross-validation

Table 9. Mean and median RPSS for the RF and SGTB models on 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

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 (2012, 2011b) 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 to build predictive models. The recourse to an extended list of attributes validated by past research (Desvignes 2014, Prades 2014) and to a highly accurate NLP system (Tixier et al., in review) allowed a large data set of 4,400 attributes and safety outcomes to be constituted. This study 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 (RPSS of 0.116, Esmaeili et al. 2015b) 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.

Contributions to theory

The high predictive skill reached by the models for three safety outcomes out of four 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 like other natural phenomena rather than strictly being 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 (2014) and Desvignes (2014) 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. Examples of such information may include the energy level in the environment (Alexander et al. 2015). Future research should try to incorporate such energy-based data into the predictive models to test whether predictive skill can be improved for *injury severity*. Furthermore, this can be used as an estimate of *potential injury severity risk.*

Also, it should be noted that the predictions made by the models 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 currently unavailable data could be gathered by recording random observations of the conditions onsite in terms of attributes.

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 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 (Domingos 2012).

To sum up, within the context of safety analysis, this study makes important strides in that the results provide reliable probabilistic forecasts of the most likely outcomes should an accident occur. 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 very large datasets (i.e., "big data").

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 learned 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 as input variables and the new models can predict, with good accuracy, the type of energy, type of injury, and body part 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 (2012, 2011b), coupled with the NLP tool of Tixier et al. (2015 in review) and with the methodology and the predictive proposed in this study; may take strides toward true, objective integration of empirical safety data within construction planning and design.

ACKNOWLEDGEMENTS

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. The authors would also like to recognize Bentley Systems for their financial support for this research and intellectual contributions.

REFERENCES

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In *Machine Learning: ECML 2004* (pp. 39-50). Springer Berlin Heidelberg.

Albert, A., Hallowell, M. R., Kleiner, B., Chen, A., & Golparvar-Fard, M. (2014). Enhancing construction hazard recognition with high-fidelity augmented virtuality. *Journal of Construction Engineering and Management, 140*(7), 04014024.

Alexander, D., Hallowell, M., & Gambatese, J. (2015). Energy-based safety risk management: using hazard energy to predict injury severity.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys, 4*, 40-79.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research, 13*(1), 281-305.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research, 13*(1), 1063-1095.

Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research, 9*, 2015-2033.

Booth, A., Gerding, E., & McGroarty, F. (2014). *Automated trading with performance weighted random forests and seasonality. Expert Systems with Applications, 41*(8), 3651-3661.

Breiman, L. (1996a). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (1996b). *Out-of-bag estimation* (pp. 1-13). Technical report, Statistics Department, University of California Berkeley, Berkeley CA 94708, 1996b. 33, 34.

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*, *26*(3), 801-849.

Breiman, L. (2001a). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199-231.

Breiman, L. (2001b). Random forests. *Machine learning*, *45*(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brown, G. (2010). Ensemble learning. In *Encyclopedia of Machine Learning* (pp. 312-320). Springer US.

Bureau of Labor Statistics (BLS) (2013). "Census of Fatal Occupational Injuries (CFOI) - Current and Revised Data." Accessed August 21, 2015. http://www.bls.gov/iif/oshcfoi1.htm.

Bureau of Labor Statistics (BLS) (2013). "National Census of Fatal Occupational Injuries in 2013 (Preliminary Results) - Cfoi.pdf." Accessed August 21, 2015. http://www.bls.gov/news.release/pdf/cfoi.pdf.

Carter, G., & Smith, S. D. (2006). Safety hazard identification on construction projects. *Journal of Construction Engineering and Management*, *132*(2), 197-205.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook* (pp. 853-867). Springer US.

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*.

Cheng, M. Y., & Wu, Y. W. (2009). Evolutionary support vector machine inference system for construction management. *Automation in Construction*, *18*(5), 597-604.

Cheng, M. Y., Peng, H. S., Wu, Y. W., & Chen, T. L. (2010). Estimate at completion for construction projects using evolutionary support vector machine inference model. *Automation in Construction*, *19*(5), 619-629.

Cheng, M. Y., Peng, H. S., Wu, Y. W., & Liao, Y. H. (2011). Decision making for contractor insurance deductible using the evolutionary support vector machines inference model. *Expert Systems with Applications*, *38*(6), 6547-6555.

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." Accessed August 21, 2015. http://www.cpwr.com/publications/construction-chart-book.

del Río, S., López, V., Benítez, J. M., & Herrera, F. (2014). On the use of MapReduce for imbalanced big data using Random Forest. *Information Sciences, 285*, 112-137.

Desvignes, M. (2014). *Requisite empirical risk data for integration of safety with advanced technologies and intelligent systems* (Master Thesis, University of Colorado at Boulder).

Diaz-Uriarte, R., & de Andrés, S. A. (2005). Variable selection from random forests: application to gene expression data. *arXiv preprint q-bio/0503025*.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*(4), 802-813.

Esmaeili, B. (2012). *Identifying and quantifying construction safety risks at the attribute level* (Doctoral dissertation, University of Colorado at Boulder).

Esmaeili, B., & Hallowell, M. (2012, May). Attribute-based risk model for measuring safety risk of struck-by accidents. *In Construction Research Congress* (pp. 289-298).

Esmaeili, B., & Hallowell, M. R. (2011a). Diffusion of safety innovations in the construction industry. *Journal of Construction Engineering and Management*, *138*(8), 955-963.

Esmaeili, B., & Hallowell, M. R. (2011). Using network analysis to model fall hazards on construction projects. *Safety and Health in Construction, CIB W, 99*, 24-26.

Esmaeili, B., Hallowell, M. R., & Rajagopalan, B. (2015a). Attribute-Based Safety Risk Assessment. I: Analysis at the Fundamental Level. *Journal of Construction Engineering and Management*, 04015021.

Esmaeili, B., Hallowell, M. R., & Rajagopalan, B. (2015b). Attribute-Based Safety Risk Assessment. II: Predicting Safety Outcomes Using Generalized Linear Models. *Journal of Construction Engineering and Management*, 04015022.

Fleming, M. A. (2009). Hazard recognition. *By Design, ASSE*, 11-15.

Franz, K. J., & Sorooshian, S. (2002). Verification of National Weather Service Probabilistic Hydrologic Forecasts. *University of Arizona, report prepared for the National Weather Service*.

Freiman, M. H. (2010). Using random forests and simulated annealing to predict probabilities of election to the baseball hall of fame. *Journal of Quantitative Analysis in Sports, 6*(2).

Freund, Y., & Schapire, R. E. (1995, January). A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory* (pp. 23-37). Springer Berlin Heidelberg.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In *ICML* (Vol. 96, pp. 148-156).

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal of Japanese Society For Artificial Intelligence, 14*(771-780), 1612.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), 367-378.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, *28*(2), 337-407.

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225-2236.

Goddard, L., Barnston, A. G., & Mason, S. J. (2003). Evaluation of the IRI's "Net assessment" seasonal climate forecasts: 1997-2001. *Bulletin of the American Meteorological Society*, *84*(12), 1761-1781.

Greg Ridgeway with contributions from others (2015). gbm: Generalized Boosted Regression Models. R package version 2.1.1. http://CRAN.R-project.org/package=gbm

Haddon, W. (1973). Energy damage and the ten countermeasure strategies. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *15*(4), 355-366.

Hallowell, M. R. (2008). *A formal model for construction safety and health risk management*. (Doctoral dissertation, Oregon State University)

Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning (Vol. 2, No. 1). New York: springer.

He, H., & Garcia, E. (2009). Learning from imbalanced data., *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284.

Jung, J., & Thon, M. R. (2006). Automatic annotation of protein functional class from sparse and imbalanced data sets. In *Data Mining and Bioinformatics* (pp. 65-77). Springer Berlin Heidelberg.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, *11*(1), 51.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

Lam, K. C., Palaneeswaran, E., & Yu, C. Y. (2009). A support vector machine model for contractor prequalification. *Automation in Construction*, *18*(3), 321-329.

Lebedev, A. V., Westman, E., Van Westen, G. J. P., Kramberger, M. G., Lundervold, A., Aarsland, D., ... & AddNeuroMed consortium. (2014). Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical*, *6*, 115-125.

Liang, T. H., & Lin, J. B. (2014). A two-stage segment and prediction model for mortgage prepayment prediction and management. *International Journal of Forecasting*, *30*(2), 328-343.

Liaw and M. Wiener (2002). Classification and Regression by randomForest. *R News 2*(3), 18—22

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 169-198.

Miller, George A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Moselhi, O., Hegazy, T., & Fazio, P. (1991). Neural networks as tools in construction. Journal of Construction Engineering and Management.

NCAR - Research Applications Laboratory (2015). verification: Weather Forecast Verification Utilities. R package version 1.42. http://CRAN.R-project.org/package=verification

Occupational Injury and Illness Classification Manual Version 2.0, U.S. Department of Labor, Bureau of Labor Statistics, September 2010 (http://www.bls.gov/iif/oiics_manual_2010.pdf)

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence, 41*(1), 77-93.

Rajagopalan, B., Grantz, K., Regonda, S., Clark, M., & Zagona, E. (2005). Ensemble streamflow forecasting: Methods and applications. *Advances in Water Science Methodologies*, 97-116.

Rebollo, J. J., & Balakrishnan, H. (2014). Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies, 44*, 231-241.

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. Update, 1(1).

Robnik-Šikonja, M. (2004). Improving random forests. In *Machine Learning: ECML 2004* (pp. 359-370). Springer Berlin Heidelberg.

Sandri, M., & Zuccolotto, P. (2006). Variable selection using random forests. In *Data analysis, classification and the forward search* (pp. 263-270). Springer Berlin Heidelberg.

Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. *Expert Systems with Applications, 41*(5), 2239-2249.

Skibniewski, M., Arciszewski, T., & Lueprasert, K. (1997). Constructability analysis: machine learning approach. *Journal of computing in civil engineering, 11*(1), 8-16.

Soibelman, L., & Kim, H. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering*, *16*(1), 39–48.

Son, H., Kim, C., & Kim, C. (2011). Automated color model–based concrete detection in construction-site images by using machine learning algorithms. *Journal of Computing in Civil Engineering, 26*(3), 421-433.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods, 14*(4), 323.

Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition, 40*(12), 3358-3378.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of statistics*, *24*, 303-329.

Tang, Y., Zhang, Y. Q., Chawla, N. V., & Krasser, S. (2009). SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 39*(1), 281-288.

Timofeev, R. (2004). *Classification and regression trees (CART) theory and applications*. (Master thesis, Humboldt University, Berlin).

Towler, E., Rajagopalan, B., Summers, R. S., & Yates, D. (2010). An approach for probabilistic forecasting of seasonal turbidity threshold exceedance. *Water Resources Research, 46*(6).

Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings, 49*, 560-567.

Tserng, H. P., Lin, G. F., Tsai, L. K., & Chen, P. C. (2011). An enforced support vector machine model for construction contractor default prediction. *Automation in Construction, 20*(8), 1242-1249.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

Villanova, M. P. (2014). *Attribute-based Risk Model for Assessing Risk to Industrial Construction Tasks*. (Master Thesis, University of Colorado at Boulder).

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2007). The discrete Brier and ranked probability skill scores. *Monthly Weather Review, 135*(1), 118-124.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter, 6*(1), 7-19.

Wilks, D. S. (1995), Statistical Methods in the Atmospheric Sciences, Elsevier, New York.

Wolpert, D. H., & Macready, W. G. (1999). An efficient method to estimate bagging's generalization error. *Machine Learning, 35*(1), 41-55.

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*(3), 5445-5449.

Yang, J., Arif, O., Vela, P. A., Teizer, J., & Shi, Z. (2010). Tracking multiple workers on construction sites using video cameras. *Advanced Engineering Informatics, 24*(4), 428-434.