



Predicting Serious Injury and Fatality Exposure Using Machine Learning in Construction Projects

Elif Deniz Oguz Erkal, Ph.D.¹; Matthew R. Hallowell, Ph.D.²;
Ayoub Ghriss³; and Siddharth Bhandari, Ph.D.⁴

Abstract: Safety academics and practitioners in construction typically use safety prediction models that employ information associated with past incidents to predict the likelihood of future injury or fatality on site. However, most prevailing models utilize only information related to failure (i.e., incident), so they cannot distinguish effectively between success and failure without well-informed comparison. Furthermore, recordable incidents on construction sites are extremely rare, which results in data that are too sparse to make predictions with high statistical power. This paper empirically reviews different approaches to safety to increase the understanding of conditions associated with safety success and failure. Empirical data about business-, project-, and crew-related factors were collected to predict serious injury and fatality (SIF) exposure conditions. A variety of modeling techniques were tested in a machine learning pipeline to identify the most accurate and stable predictive models. Results showed that the multilayer perceptron (MLP) approach best distinguished SIF exposure conditions from safety success conditions using nonlinear decision boundaries. The most influential factors in the models included the crew experience working together, supervisor experience with the crew, total number of workers under the supervisor's purview, and the maturity of leadership development programs for frontline supervisors. This study showed that data sets with both success and failure information yield more reliable and meaningful predictions than data sets with failure alone. Such an approach to safety data collection, analysis, and prediction could be used by future researchers to generate new insights into the causes of serious incidents and the relationships among causal factors. **DOI: 10.1061/JCEMD4.COENG-13741.** This work is made available under the terms of the Creative Commons Attribution 4.0 International license, <https://creativecommons.org/licenses/by/4.0/>.

Author keywords: Construction safety; Machine learning; Serious injury and fatality (SIF); Prediction.

Introduction

Despite efforts to eliminate serious injury and fatalities (SIFs), fatality rates have plateaued for the last 10 years, and the construction industry still remains as one of the deadliest industries (US Bureau of Labor Statistics 2020a). In 2019 alone there were 1,102 construction worker fatalities, the highest number of deaths in construction since 2011 (Brown et al. 2021; US Bureau of Labor Statistics 2020b). In response to the intolerable societal, emotional, and financial distress associated with SIFs, the construction industry continues to increase its investment in SIF prevention. Among

the many areas of SIF prevention, predictive models always have received considerable attention because they focus on proactive actions, rather than on reactive learnings. In practice, being predictive allows safety professionals to act before a serious injury occurs (Hallowell et al. 2019). In academia, predictive studies help to understand the primary causes of unsafe conditions and the relationships among driving factors. Being able to accurately predict injury potential and take action before an event occurs has become a central mission in modern safety research and practice.

However, the data sets used in such predictive models focused only on failure (e.g., documentation and conditions surrounding injuries), which limited the predictive models to predicting different injury outcomes only if an injury were to occur. Tixier et al. (2016b) used machine learning (ML) to generate a model that predicts injury type, energy type, and body part impacted based on the observable work features (e.g., materials, tools, and equipment) extracted from actual injury reports. Although the model made successful predictions of the characteristics of an injury if one were to occur (e.g., if a worker was to be injured in this work period, it is most likely to be a fall to lower level). Unfortunately, models based only on failure data are incapable of distinguishing success from failure or of estimating the likelihood that an injury will occur. As highlighted by De Finetti et al. (2017), only data sets that capture all outcomes related to the studied consequence can facilitate robust likelihood assessments and differentiate these outcomes from each other given the data space available. Therefore, the present study focused on predicting success and exposure conditions rather than the prediction of incidents. Here, success is defined as a condition in which a high-energy hazard has a corresponding direct control. Conversely, exposure is a condition in which workers are exposed to a high-energy hazard without a

¹Graduate Research Assistant, Dept. of Civil, Environmental, and Architectural Engineering, Univ. of Colorado Boulder, Boulder, CO 80309-0428 (corresponding author). ORCID: <https://orcid.org/0000-0001-9793-8393>. Email: elog7097@colorado.edu

²K. Stanton Lewis Professor of Construction Engineering and Executive Director of the Construction Safety Research Alliance, Dept. of Civil, Environmental and Architectural Engineering, Univ. of Colorado, 1111 Engineering Dr., UCB 428, Boulder, CO 80309-0428. Email: matthew.hallowell@colorado.edu

³Graduate Research Assistant, Dept. of Computer Science, Univ. of Colorado Boulder, Boulder, CO 80309-0428. Email: ayoub.ghriss@colorado.edu

⁴Associate Director of Construction Safety Research Alliance, Research Faculty, Dept. of Civil, Environmental, and Architectural Engineering, Univ. of Colorado, 1111 Engineering Dr., UCB 428, Boulder, CO 80309-0428. Email: siddharth.bhandari@colorado.edu

Note. This manuscript was submitted on March 7, 2023; approved on September 26, 2023; published online on December 26, 2023. Discussion period open until May 26, 2024; separate discussions must be submitted for individual papers. This paper is part of the *Journal of Construction Engineering and Management*, © ASCE, ISSN 0733-9364.

corresponding direct control (Oguz Erkal and Hallowell 2023). Detailed definitions were provided in the “Literature Review” section.

Another limitation of current methods of safety prediction is that factors at different levels of the organization typically are considered independently. For example, leading indicator research focuses on the quality and quantity of injury prevention activities; safety climate research focuses on perceptions of safety and employee satisfaction with the safety system; and precursor analysis focuses on assessment of human factors. However, at any time, all these potential predictors may interact to create conditions of safety success or exposure. Even though they might be evaluated and managed separately in practice, their interaction and interdependency should be considered to increase predictive power and enhance our understanding of how unsafe conditions emerge. This study created a unified predictive model that includes predictors at the business, project, and crew levels.

This study contributes to the body of knowledge by (1) building a new data set that includes both conditions of success and exposure as the target outcome variable, and (2) examining the impact of new potential predictors at different organizational levels on predicting safety success and exposure. By creating a new data set with operational definitions of both success and failure, studying a new set of potential predictors at different levels of the organization, and deploying the latest methods in machine learning to determine predictors that distinguish success and failure on jobsites, this study advances the science of safety prediction and enables construction professionals to better assess safety risk and make evidence-based proactive safety decisions.

Literature Review

This literature review presents the existing knowledge in three primary domains related to safety prediction: (1) safety prediction methods; (2) dependent variables in safety; and (3) analytical and statistical methods used in the models making safety predictions.

Safety Prediction Methods

Safety predictions typically are based on different pieces of available safety information, including but not limited to risk assessments of observable attributes (e.g., materials, tools, and work practices), leading indicators (e.g., prejob safety meetings), safety climate (e.g., perception of and satisfaction with management commitment to safety), and human factors (e.g., distractions, fatigue, and work pressures) (Hallowell et al. 2019).

As the most traditional safety prediction tool, safety risk data often are recorded as text-based reports that describe the circumstances surrounding an injury; categorical information about the injury such as severity, type, and body part affected; and counts of injuries over time (Esmaeili et al. 2015a; Tixier et al. 2017). Risk analysis techniques typically are used to associate observed conditions such as behaviors, work tasks, and environmental conditions with outcomes such as injury type and severity. This approach to safety prediction implicitly assumes that associative relationships (e.g., the correlation between task type and an injury outcome) apply to future conditions as well. Although risk assessments are related closely to operations or tasks, current methods do not model real-time conditions or support proactive decisions (Hallowell et al. 2019; Oguz Erkal et al. 2021).

As a response to traditional risk assessments, safety researchers developed the term “safety leading indicators,” which traditionally refers to the frequency with which a safety activity is performed (Alruqi and Hallowell 2019; Hinze et al. 2013; Salas and Hallowell 2016). This may include safety program elements such

as the frequency of jobsite audits, personal protective equipment (PPE) programs, drug testing, leadership engagements, pretask planning, near-miss reporting, and safety orientations (Salas and Hallowell 2016). There is strong evidence that trends in the frequency of safety activities are highly correlated with the long-term trends in injury rates (Salas and Hallowell 2016; Versteeg et al. 2019). Researchers continue to identify and validate new leading indicators to generate more predictive power (Guo et al. 2017). Despite the benefit of safety leading indicators, there are some limitations that prevent them from being used as perfect indicators of future safety performance. First, the predictive power of safety leading indicators tends to be only moderate because the counts of safety activities do not capture the quality with which the activities are performed (i.e., increasing quantity may result in decreased quality of implementation). Second, a lack of consistent definitions of leading indicators across the industry severely limits the external validity of the modeling results. Third, the directionality of the predictions may be suspect because it is unclear whether the injury rates are responding to trends in the safety activities, or if the implementation of safety activities is responding to trends in injury rates (Lingard et al. 2017).

Another response to reactive safety prediction models is in form of behavior-based safety, which considers safety predictions based on human behaviors such as safety climate and human factors research. Safety climate acts as a proxy measure for safety culture that uses employee perceptions of the safety systems in place, such as management commitment, training, and safety resources to make predictions on future performance (Alruqi et al. 2018; Siu et al. 2004). Recent studies examined many different human factors that may serve as a warning sign before or during a given work period for a SIF. Experiments showed that the factors with the highest predictive capacity were related to human factors such as distraction, fatigue, risk normalization, and schedule pressures (Alexander et al. 2017a, b). Although the predictive capacity of such research has been established by measuring the strength of the associative relationship between metrics and injury rates, the data collected remain highly subjective, are a snapshot in time, and cannot be standardized against a baseline.

In summary, risk assessments can report risks that are related closely to operations and tasks (Baradan and Usmen 2006; Choe and Leite 2016), but current methods do not provide real-time insight that allows practitioners to be truly proactive in their interventions. Leading indicators can capture real-time information about the safety system (Guo and Yiu 2016). However, their relationship with highly consequential results is ambiguous, and their definitions and applications are highly inconsistent in practice (Oguz Erkal et al. 2023). Safety predictions based on human factors could be consequential and measured in real-time, but the collected data are not objective by definition due to dependence on observer or worker perceptions (Alexander et al. 2017b; Schwatka and Rosecrance 2016). As a result, in view of the existing safety prediction methods described previously, this paper identifies of the need for a predictive model that is based on consistent data that can capture real-time information that is objective and closely related to highly consequential risks.

Dependent Variables in Safety

Every predictive model aims to make an educated guess about the likelihood of an event, condition, or trend. Although the importance of safety prediction models cannot be overstated, their quality very rarely is called into question within the occupational health and safety literature, especially as it relates to using injury rate as the data source. Typically, only past safety incidents are used to

predict potential future injuries and fatalities (Raheem and Hinze 2014). The target safety outcome most often is an injury rate, such as the Occupational Safety and Health Administration's (OSHA's) total recordable incident rate (TRIR). However, a truly predictive model would require success- and failure-related data in order to be able to distinguish success- and failure-related characteristics before making likelihood assessments (De Finetti et al. 2017).

Incident rates are recognized as an intuitive metric for predictions due to the ease of data collection, quantitative data representation, and rationalization of the concepts (Lingard et al. 2017). However, injury rate as an outcome variable is statistically problematic to model as an outcome because injuries are relatively rare even over long periods (Hallowell et al. 2021), randomly distributed with very large variation (Hallowell et al. 2021; Hopkins 2009; Lingard et al. 2017), and contextually ambiguous (Raheem and Hinze 2014). Using incident rates as a relative indicator of success or failure inherently is based on the assumption that periods with a lower rate of incidents were definitively safer (i.e., the absence of injuries is the presence of safety).

Modern views of safety suggest that injury rates do not provide information about safety performance because they do not capture the true nature of what safety looks like (Hollnagel 2014). Although safety philosophers have argued that "safety is the presence of safeguards (i.e., capacity)," (Hollnagel 2015) they have not offered an operational way to collect such information (Dekker and Pitzer 2016; Hollnagel 2014; Raheem and Hinze 2014). Further understanding and assessment of safety success is warranted (Choi et al. 2020; De Finetti et al. 2017). Although a few studies have built predictive models using data from incident and nonincident cases (e.g., Alexander et al. 2017b), the chief assumption made is that nonincident cases are always safe. Unfortunately, this is antithetical to the modern understanding of safety as the presence of safeguards rather than the mere absence of injuries (Hollnagel 2015; Lofquist 2010).

Safety prediction techniques require a more robust dependent variable that is statistically stable and in line with the contemporary safety perspectives. For example, safety recently has been reconceptualized as the continuous presence of defenses and controls against safety risks, instead of the mere absence of injuries (Hollnagel 2014, 2015; Lofquist 2010). Building on this new definition of safety, this study focuses on conditions instead of incidents. A new safety performance metric, high energy control assessment (HECA), was used as the dependent variable and the main performance metric in this paper (Oguz Erkal and Hallowell 2023).

HECA is an observation-based variable that measures the proportion of high-energy hazards found on a worksite with a corresponding direct control. Thus, HECA represents two conditions: success, in which all high-energy hazards are observed to have a corresponding direct control; and exposure, in which one or more high-energy hazards do not have a corresponding direct control. As a condition assessment, HECA can be observed any time that work is being performed, and analysts do not need to wait for an incident to occur to measure safety. This allows HECA to be measured at a much greater frequency than incident rates and to yield statistically significant samples relatively quickly, which could allow more-advanced modeling techniques to be used on the data sets. In contrast to the highly reactive injury rates that are based solely on failure, HECA offers an opportunity to capture success and exposure proactively, before an incident occurs. This allows the models to make comparisons between the two classes for more-reliable predictions based on the distinction.

High-Energy Hazard

The hazard assessment component of HECA is built on the understanding that every injury is the result of the contact between a source of energy and the human body (Albert et al. 2014, 2017; Tixier et al. 2018). This theory was extended when researchers found that the magnitude of energy directly predicts the severity of the outcome (Hallowell et al. 2017). Specifically, it was found that when energy exceeds a specific threshold (1,500 J), the hazard is most likely to cause a serious injury or fatality. Thus, when energy sources exceed the 1,500-J threshold, they are labeled high-energy. Because the assessment of some energy sources can be very complex and infeasible in practice, Hallowell (2020) presented the high-energy assessment of 13 high-energy sources that represent approximately 85% of workplace injuries. These include hazards such as heavy mobile equipment, fall from elevation over 4 ft, and suspended loads.

Direct Control

HECA is built on the principle that every high-energy hazard should have an adequate control to protect against a serious injury or fatality. This is aligned with the philosophy that safety is not the absence of injury but the continuous presence of safeguards and controls. A control is considered to be sufficient and is labeled as a direct control if it is (1) targeted at a specific high-energy hazard; (2) mitigates the high-energy hazard when installed verified and used properly either by eliminating the energy or by reducing the energy magnitude to below the 1,500-J threshold; and (3) effective even when someone makes an unintentional human error (Hallowell 2020). This definition aligns with the hierarchy of controls by effectively mitigating risk primarily through elimination, substitution, or engineering controls (NIOSH 2015). Administrative controls such as training, warning signs, rules, and so forth are not considered to be direct controls because they are vulnerable to human error. Although some forms of specialty personal protective equipment such as fall arrest systems or fire-resistant (FR) clothing meet the definition of direct control, many forms of general PPE such as hard hats, gloves, and glasses do not meet this definition because they are not targeted and do not mitigate high-energy.

Predictive Modeling Techniques in Construction Safety

Safety researchers have used many statistical techniques to examine and demonstrate associative relationships between predictor variables to safety outcomes. For example, Sarkar et al. (2020) and Esmaeili et al. (2015b) used generalized linear models to predict the characteristics of injuries. Alexander et al. (2017b) used multiple linear regression to test the predictive validity of human factor precursors based on precursors; and Salas and Hallowell (2016) used a combination of principal components analysis and generalized linear modeling to test the predictive power of safety leading indicators in explaining the variability in lagging indicators.

In addition to utilizing regression analyses, machine learning pipelines have been used for safety prediction, especially to analyze intricate empirical data (Gondia et al. 2020; Liu and Tian 2019; Shin 2019). Machine learning pipelines typically include a comprehensive suite of algorithms to make predictions. These algorithms include simpler models such as linear regressions, as well as more-complex and nonlinear algorithms. ML attempts approach the prediction problems using both trending (regression) and pattern recognition (clustering) in larger data sets.

Researchers have used ML to predict severity of construction accidents, injury type, energy type, and body part injured (Hallowell et al. 2019; Kines 2001; Sarkar et al. 2020; Zhu et al. 2021); to

detect the predictive trends for incidents affecting construction workers (Choi et al. 2020; Zhu et al. 2021); and for risk analyses and simulation (Tixier et al. 2016a). Additionally, several studies have explored how ML may increase predictive power of leading indicators (Costin et al. 2019; Poh et al. 2018; Wang and Razavi 2019). Some common algorithms used in ML pipelines in the safety domain are regressions (Choi et al. 2020; Zhu et al. 2021), decision trees (DTs) (Cheng et al. 2019; Shin 2019; Tixier et al. 2016a), random forests (RFs) (Choi et al. 2020; Guo et al. 2021; He et al. 2021b), support vector machines (SVMs) (Sarkar et al. 2019; Zhang et al. 2015), multilayer perceptron (MLP), and neural networks (NNs) (Sarkar et al. 2020, 2019; Zhu et al. 2021). The quality of such models is assessed using evaluation metrics such as accuracy, precision, recall, and so forth (Goh and Chua 2013; Zhang et al. 2019). Zhu et al. (2021) evaluated the advantages and disadvantages of each ML algorithm for their future effective implementation while developing their model for safety incidents (Zhu et al. 2021). The application of these machine learning methods in safety prediction provided researchers with a robust method to extract actionable information from complex data sets without introducing additional bias. With the growing data collection and processing capacity, tree-based algorithms and unsupervised machine learning have been chosen abundantly over linear models in recent years.

The safety literature has found ML implementations to be useful because ML helps to explore and fine-tune a suite of algorithms, which achieves better predictive performance than individual models. A ML approach was chosen to examine the associations between the predictor and predictand variables because it can facilitate a comprehensive statistical analysis while handling the multi-dimensional, intricate, and noisy empirical data set.

Research Methods

This sections describes the research design, selection of independent and outcome variables, data collection, preprocessing data, model building using various algorithms, evaluation of the resultant

models, and application of the best models to achieve the research objectives.

Research Design

This paper follows the stages of a typical ML pipeline as its overall research design (Fig. 1). The main objective was to predict serious injury and fatality exposure depending on workplace attributes of business, project, and crew levels based on data that includes both safety success and failure. Survey data were collected from construction companies per the data collection strategy. Systematic crew observations assessed SIF exposure and success given the safeguards against high-energy hazards. Survey data were cleaned, coded, and processed in preparation for predictive model building using the chosen ML algorithms.

The priorities considered while selecting the ML algorithms were (1) both linear and nonlinear model utilization, (2) informed feature selection, and (3) ability to handle small-scale data sets without overfitting. Feature importance information was provided because the ML models in this study were created to be utilized in real-time safety management and monitoring systems as decision-making support. In line with the ML algorithm considerations and in alignment with previous research, logistic regression (LR), decision trees, random forests, gradient boosted decision trees (GBDT) and multilayer perceptron models were chosen to be deployed. Area under receiver operating characteristic curve (AUC) rates were used as the evaluation metric for the derived ML models. The following subsections provide in-depth explanations of the methodology.

Independent Variable Selection: Safety Predictors

Independent variables were selected from a preceding study which investigated potential predictors at the crew, project, and business levels (Oguz Erkal et al. 2021). In this study, researchers identified 40 critical potential predictors using the Delphi process with a panel of experts representing the different key sectors of the construction industry. For each potential predictor, multiple questions

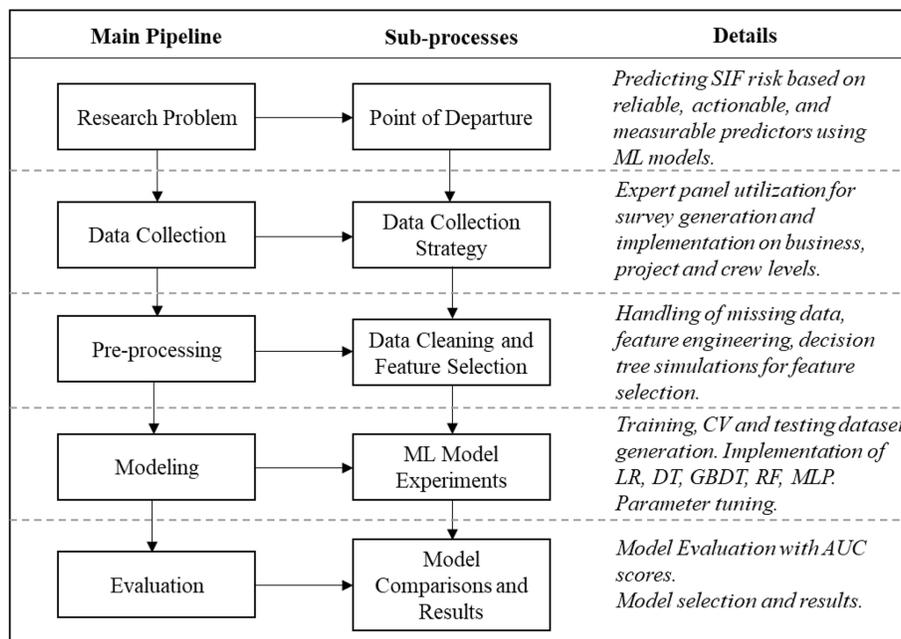


Fig. 1. Overall research design.

were designed to enable a valid collaborative assessment by the supporting industry team. This step was critical for ensuring consistency and internal validity. An independent survey tool was used for data collection.

Business factors included leading indicator protocols, frontline supervisor leadership training, investment in safety research and development, intervention protocols, and so forth. Because business factors tend to be stable over a longer period, the data on these factors were collected only once during a 4-month data collection period.

Project factors included frequency and attributes of job safety orientations, presence of project manager, quality of supervisor, weather conditions, and so forth. For some companies that do not complete projects in the traditional sense, an option was added to report in terms of geographical business unit. Although construction companies typically complete work on discrete projects (e.g., a building or a dam), other construction-adjacent companies such as power-generation and delivery companies use complete dispersed work through functional business units that typically correspond to a narrow geographical region (e.g., the Denver metropolitan area). The project survey was designed to be filled out once for each project or business unit during the data collection period.

Crew factors included crew size, duration of time that the crew has worked together, staffing levels, and so forth. The crew survey was designed to be filled out once per field visit of a crew performing a task by the field leader in charge (e.g., superintendent, foreman, and crew lead).

Outcome Variable Selection: HECA

One of the key objectives and intended contributions of this paper was to build a predictive model that incorporates both success- and failure-related data. As noted previously, the authors did not use traditional outputs of prediction models such as fuzzy risk assessments, incident rates, and so forth. An observation-based dependent variable, high-energy controls assessment, was used in this study (Oguz Erkal and Hallowell 2023). The HECA per observation was recorded as binary assessment: success, or failure.

Data Collection Strategy

The data collection undertaken for this study yielded empirical data collected from construction sites representing the different sectors of the industry, including oil and gas, service and utility works, industrial construction, heavy civil construction, and commercial building construction. A total of 28 unique organizations collected information from 74 projects and 698 crews between June and September 2021. This was a particularly large effort because the predictor and outcome variables were new, and therefore were not collected automatically in existing safety management systems. The partnership between the academic researchers and the industry professionals was paramount for success.

To ensure valid and reliable data collection across sites and employers, points of contact from each organization responsible for data collection were required to complete an introductory training course in which the potential predictors were defined, all the survey instruments were reviewed, and rules for collecting HECA observations were discussed. Although an approach with multiple practitioners assisting with data collection is not ideal because it compromises internal validity to some degree, it was a limitation accepted by the authors to generate the voluminous data necessary to build a predictive model. To enhance the internal validity, the training on collecting data was delivered virtually by one individual

from the academic team throughout the process to ensure minimal calibration errors and nonsystemic biases in data collection process.

The HECA assessment was facilitated using the form in Fig. 2. When completing the form, the observer was required to (1) identify all high-energy hazards present in the work task, and (2) identify which high-energy hazards were mitigated by the presence of a direct control. The observer was required to use 16 high-energy icons and the strict definition of direct control (Albert et al. 2017; Hallowell 2020; Tixier et al. 2018). If all high-energy hazards were mitigated, the observation was marked as success, and if one or more hazards did not have a corresponding direct control, the observation was marked as exposure. This binary assessment represented the outcome variable for the safety predictions.

Data Preparation

The data preprocessing process was composed of data cleaning, feature engineering, and dependent variable projection to prepare the data set for further analysis, feature selection, and ML model training.

Data Cleaning and Transformations

To clean the data, a five-step process was followed: (1) handling of missing data, and (2) cleaning no-variance independent variables, (3) one-hot encoding of categorical variables, (4) numerical data transformations, and (5) normalization of variables (Seger 2018; Shehadeh et al. 2021). First, the independent variables that had more than 20% of the data missing were removed from the data set. Second, for variables that had less than 20% of data missing, the missing entries were replaced with median values, and independent variables that had no variance were removed from the data set. Third, all categorical variables were one-hot encoded into separate columns by converting the variables with multiple categorical values without a hierarchy or order of magnitude into separate binary variables. Fourth, ordinal data on Likert scales were coded by numerical means between 0 as a minimum and 1 as a maximum (i.e., 2 on a 5-point Likert scale was coded as 0.25). Finally, all continuous variables were normalized. As a result of the data cleaning and transformations process, all data were encoded into numerical variables that were normalized with no missing values.

Feature Engineering

New features in the independent variables space were created using two strategies to increase predictive performance. Questions with many categorical variables were summed into one new feature to represent the total number of attributes included in the measured predictor. Leading indicators at the business level is an example for which the summation technique was used to present the total number of leading indicators that the company tracks. Such features were created to represent an overall summary of the predictor in addition to individual categories. Second, selected independent variables were divided to calculate ratios. For example, a new feature was designed to present the field management to worker ratio (P2) at the project level by dividing the total number of field managers by the number of workers.

The dependent variable, HECA, was generated using feature engineering to create a binary metric. Accordingly, if the respondent identified high-energy hazards but responded “No” to the existence of direct controls, the instance was recorded as 1, exposure. If the respondent did not identify high-energy hazards or if all the high-energy hazards identified were mitigated by direct controls, the instance was recorded as 0, success.

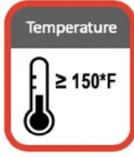
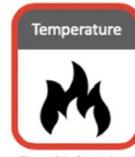
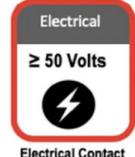
THE OBSERVATION: This section is to be filled by a safety professional ON SITE about the work being performed by this crew. Which of the following high energy sources were observed on the work being performed by this crew? Please check all that apply.				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
 Gravity Suspended Load	 Gravity Fall from Elevation	 Mechanical Heavy Rotating Equipment	 Pressure Excavation or Trench	 Pressure Explosion
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
 Motion Mobile Equipment and Workers on Foot	 Motion Motor vehicle Incident (occupant)	 Temperature High Temperature	 Temperature Fire with Sustained Fuel Source	 Temperature Steam
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
 Electrical Arc Flash	 Electrical Electrical Contact with Source	 Chemical / Radiation High Dose of Toxic Chemical or Radiation	 Other	 No High Energy Hazards
Was a direct control present for every high-energy hazard? A direct control: <ol style="list-style-type: none"> 1. is specifically targeted to the high-energy source 2. effectively mitigates exposure to the high energy source when installed, verified, and used properly 3. is effective even if there is unintentional human error during the work. 			<input type="checkbox"/> YES <input type="checkbox"/> NO	

Fig. 2. Survey section for HECA data collection. [Images reprinted with permission from Hallowell (2020).]

Outcome Variable Projection

Businesses have multiple projects, and projects deploy multiple crews. Therefore, business attributes apply to multiple projects, and project attributes apply to multiple crews. This hierarchical structure was used to connect the business and project attributes in parallel with the outcome variable. Pursuant to data cleaning and feature engineering stages, the outcome variable (i.e., dependent variable) collected on the crew survey level was projected to the independent variables collected in business and project surveys. This was made possible using the consistent coding of businesses, projects, and crews during data collection. After the dependent variable was projected onto business- and project-related independent variables, three unique data sets were created under each functional unit sharing the same dependent variable, HECA. The three data sets were used to create the ML models of business and crew (Business), project and crew (Project), and crew [Crew (C)] data sets. For example, Business A is working on Project AX, with Crews AX-1 and AX-2. Crews AX-1 and AX-2 were observed on site, and HECA data were collected. The resultant Crew data set included only the attributes of Crews AX-1 and AX-2 and HECA results. The Project data set included the attributes of Crews

AX-1 and AX-2 and Project AX, and HECA results. The Business data set included the attributes of Crews AX-1 and AX-2 and Business A, and HECA results.

Predictive Model Development

Understanding the Feature Space

A common first step in machine learning is exploring the latent structure of the feature space. In this study, the feature space is the collection of safety predictors (independent variables) in the business, project, and crew levels for the given data set. This commonly is done via dimensionality reduction or analysis of embedding spaces of self-supervised models. Here, *t*-distributed stochastic neighbor embedding (TSNE) was used for visualizing the data in two-dimensional (2D) space to identify outliers and to gain insight into the linear separability of class labels (Van der Maaten and Hinton 2008).

Crew Data set

For the Crew data set, TSNE was used to project the original data of 29 features into a two-dimensional space (Fig. 3). The learning of

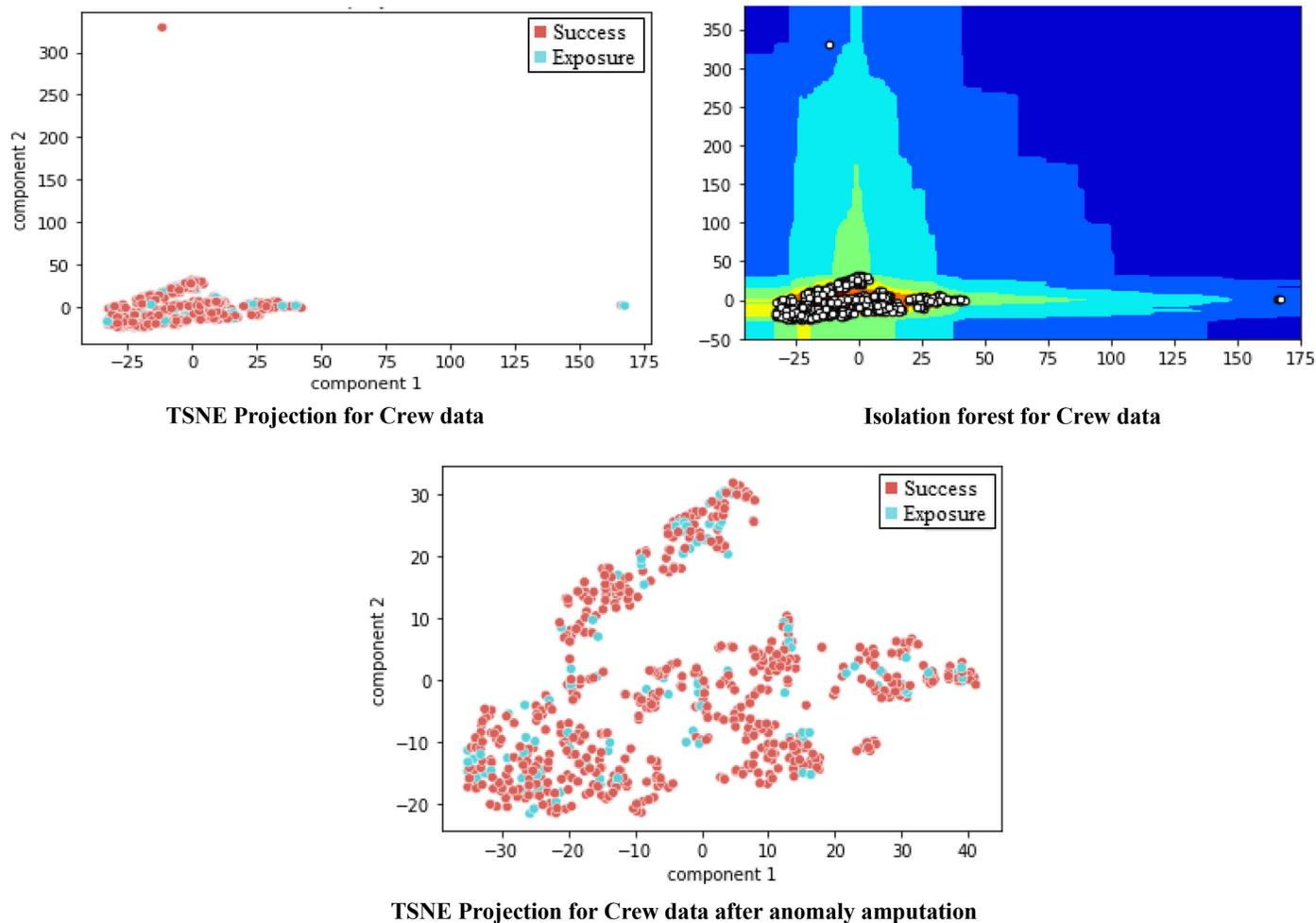


Fig. 3. TSNE projection (before and after anomaly amputation) and isolation forest for Crew data.

the projected 2D variable was performed via minimization of the divergence between the variable and the similarities between samples in the original feature space. This projection demonstrated that there were outliers that might affect the performance of a model. To remove outliers, an isolation forest with 10 estimators was used to compute an anomaly score. This method can be effectively used when there is a large number of variables that might contain irrelevant features (Liu et al. 2012). Specifically, the implementation of isolation forest in scikit-learn (version 1.0.2) was used. To compute the anomaly score, the average length of the path from the root of the trees to the corresponding sample node was calculated. Six samples with scores lower than -0.12 were considered to be anomalous and hence were excluded.

TSNE projection demonstrated that there were no clearly isolated samples clusters with different labels. Thus, it showed that linear models would have modest performance, and classification models that could learn complex boundaries would perform better. A disadvantage of more-complex models is that they require larger data sets; however this drawback could be alleviated using data augmentation.

Business and Project Data

When business or project features were combined with crew features to generate Business and Project data sets, the TSNE projection on 2D space and isolation forest did not yield clear outliers. However, evaluating the TSNE projections in Fig. 7 indicated that

there were clusters of exposure datapoints in both projections, which signified the existence of a predictive pattern. Multiple algorithms with different strengths and weaknesses were tested in parallel to produce models with the best predictive performance for this data set.

Machine Learning Models

Machine learning models were developed to test the relationship between the independent business, project, and crew variables and the binary HECA outcome variable. ML models were developed separately for the crew layer first to analyze the features closest to the work. The most important features chosen using the crew model were merged with the business and project layers separately to identify the additional information that could be provided using the business and project feature space. The use of three models was chosen strategically to provide actionable information to different functional levels of the organization that may or may not have complete information about other hierarchical levels. For example, a project manager could use the final Project model to make predictions without entering data related to business features. The HECA metric results were recorded as exposure ($Y = 1$) and success ($Y = 0$). Logistic regression (Karacasu et al. 2014; Zhu et al. 2021), decision trees (Tixier et al. 2016a; Zhu et al. 2021), gradient boosted decision trees (He et al. 2021b), random forests (Breiman 2001), and multilayer perceptron (He et al. 2021a; Zhu et al. 2021)

models were tested independently for predictive model development. MLP was chosen specifically to better handle the expected noise in the data set while approximating complex nonlinear relationships in the multidimensional feature space (Zhu et al. 2021). All models were developed, tested, and evaluated using the scikit-learn and Keras version 2.11 packages for Python 3.8.

Data Handling

Training, Cross-Validation, and Test Sets

All data sets were split into 80% training samples and 20% test samples. The training samples were subjected to a fivefold cross validation, in which the original data set was split into five different parts, and iterations used one part for validation and the other four parts for training. The performance estimate was taken as the average performance over the five partitions.

Data Augmentation

Whereas neural networks require larger data sets to train, smaller data sets require the utilization of shallow MLP structures with a low number of nodes. Data augmentation to increase the training set is a common practice, and noise was added to the data set to increase the number of samples (Shawky et al. 2020). For the MLP models that required larger data sets, the sample space was augmented by adding randomness (Leung et al. 2010). For each feature f , the sorted unique values of feature were used to calculate the gaps between consecutive values, and the smallest gap was used to offset the values randomly. For example, for a binary variable $f: f_1, f_2, \dots, f_n$, two unique values could be calculated (0,1), and the gap is computed as 1. Because this is the only gap, the smallest gap also is 1, denoted G_f . To augment a sample x_i by adding random values to each feature, the value added to feature f_i is in the interval $[-0.25 \times G_f, 0.25 \times G_f]$. Practically, the model was forced to assign the label to a region of the feature space instead of to a specific point in the space to help construct a clear margin between close samples from different class labels (regularization) (Krogh and Hertz 1991). Data augmentation was applied to the training data set with an augmentation factor of 20 conserving the proportion of class labels across the data splits, and the remaining validation and test data were kept separate.

Feature Selection

Choosing the right features was critical for the resultant performance of the models generated (Poh et al. 2018). It is an essential part of preprocessing to increase the predictive power of resultant ML models (Soibelman and Kim 2002; Son et al. 2012; Witten and

Frank 2002). For feature selection in the crew data set, the augmented data were used to perform a decision tree classification. Decision trees were built by creating splits on features to obtain the highest information gain and the least possible entropy in child nodes. To estimate such features in the data set, 100,000 random DTs were built on the training set. A histogram was plotted with 20 bins showing the total number of features chosen by the DT ensemble. The resultant bins formed clusters showing the number of times each feature was selected for a split. All features that belonged to the first cluster of bins, that were not chosen frequently by the DT simulations as important features, were excluded to reduce the number of features. The feature importance values were calculated by averaging the feature importance values over the DT ensemble.

Fine-Tuning and Testing

Tuning and testing the models involved finding the best parameters for each machine learning model (depth of the tree, activation function, number of trees, and so forth). A grid search technique was used for hyperparameter selection. Using the validation sets, this process selected the parameters that yielded the best performance on the validation set. Random multiple iterations were performed to reduce the exponentially large hyperparameter space and constrain search regions, reducing the search space. When the best performing hyperparameters were chosen, the performance of the model was evaluated using the test data (20% of the original data). The test samples were used mainly to check if the model generalized well to unseen samples. This process is demonstrated in Fig. 4.

For each of the explored ML methods, the hyperparameters that were modified from the default settings of the scikit-learn package are provided in Table 1.

Model Evaluation

Pursuant to the preprocessing, all three data sets had a class imbalance issue in the dependent variable, because there were only 16.18% exposure cases in the crew data set. Such a class imbalance may impair the learning of ML algorithms from underrepresented classes (Poh et al. 2018). The class imbalance could be handled without using any data manipulation techniques to augment the underrepresented class. To do so, the model evaluation metric must be chosen carefully.

Binary classification models generally are evaluated in terms of their predictive power to accurately predict true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). However, reporting the accuracy of a model could be misleading when the class distribution is skewed (Choi et al. 2020).

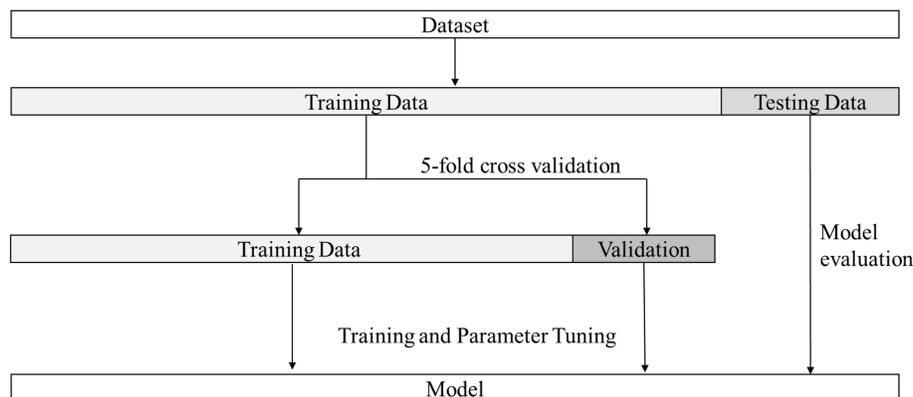


Fig. 4. Parameter tuning and testing process.

Table 1. Fine-tuning results for all models

ML model	Crew data set	Project data set	Business data set
LR	Penalty = ridge $C = 0.001$	Penalty = ridge $C = 0.001$	Penalty = lasso $C = 0.1$
DT	Solver = liblinear max_depth = 17 Criterion = Gini max_features = 9 min_samples_split = 2	Solver = liblinear max_depth = 10 Criterion = Gini max_features = 10 min_samples_split = 2	Solver = saga max_depth = 13 Criterion = entropy max_features = log2 min_samples_split = 15
GBDT	N_estimators = 700 learning_rate = 0.46	n_estimators = 500 learning_rate = 0.46	n_estimators = 400 learning_rate = 0.46
RF	n_estimators = 10 max_depth = none	n_estimators = 25 max_depth = 6 min_samples_split = 3 max_features = 0.1	n_estimators = 20 max_depth = 5 min_samples_split = 4 max_features = log2
MLP	Layers = linear(number of units, activation): linear(64, tanh) dropout (0.1); linear(48, tanh) dropout (0.1); linear(32, tanh) dropout (0.1); linear (1, sigmoid); all layers use L1 (alpha = 0.015) regularization	Layers = linear(number of units, activation): linear(64, tanh) dropout (0.1); linear(48, tanh) dropout (0.1); linear(32, tanh) dropout (0.1); linear (1, sigmoid); all layers use L1 (alpha = 0.015) regularization	Layers = linear(number of units, activation): linear(64, tanh) dropout (0.1); linear(48, tanh) dropout (0.1); linear(32, tanh) dropout (0.1); linear (1, sigmoid); all layers use L1 (alpha = 0.015) regularization

In a skewed data set, such as the ones that were collected in this study, even predictions resulting from a majority vote classifier can produce high accuracy levels without creating any intelligence associated with the independent variables.

A better evaluation metric for ML models using skewed data sets is the area under the receiver operating characteristic curves (AUC). Operating characteristic curves are the visualization of true positive and false positive rates on the vertical and horizontal axis, respectively (Fawcett 2006). The performance of the evaluated model increases as the reported AUC approaches 1.00. AUC can be interpreted as the probability that a classifier correctly determines which sample belongs to which label. As a result, AUC may be evaluated as an accuracy approximation for an imbalanced data set. AUC scores were chosen to fairly evaluate and choose the best performing model. Choi et al. (2020) defined an AUC score of 0.70–0.80 as fair and of 0.80–0.90 as good. The goal of this paper was to achieve at least 0.7 AUC, which is the common minimum threshold to prove that a model can detect some dependency between the features and the labels.

Results and Interpretation

In this section, the outcomes of the designed machine learning pipeline are presented with generated the best model that predicts HECA from the crew, project, and business features. The hyperparameters were determined from the fine-tuning process for replicability. Model evaluation results are provided to compare the models' performance.

As a result of preprocessing process, the resultant data set included 692 binary HECA observations (580 success observations, and 112 exposure observations) as the dependent variable; and 29 crew features, 137 project features, and 111 business features as independent variables that were prepared for further analysis.

Feature Selection and Importance

For all data sets, the feature selection method was based on decision tree classification. As a result, feature importance for Crew, Project, and Business data sets was calculated. Twelve features from the crew feature space were selected for the resultant model. Twenty features each from the project and business feature spaces were

selected for the resultant models. Ten of these features in each model were unique to the Project and Business data sets. Feature selection results are shown in Fig. 5.

Model Evaluation Results

The performance of the average model from each machine learning method was reported using AUC scores (Table 2). The average models were reported to ensure the reliability and generalizability of the resultant models.

MLP models outperformed all other models and had higher stability throughout the batch training. Random forest did not have significantly improved performance over that of decision tree. RF usually reduces overfitting, but for the crew variables, this tendency of reducing variance prevented it from capturing information relevant to HECA.

Although it is difficult to explain the impact of MLP architecture and the overall learning in neural networks, the authors attempt to provide insight into the developed model using step-by-step analysis of model generation. During the training of the MLP model (Fig. 6), the AUC scores at the end of each training epoch was provided. The learning curves were presented for the augmented training set, the nonaugmented validation set, and for the nonaugmented test set. The usual trend in deep learning starts with a noticeable improvement of the objective on all partitions, followed by an overfitting regime in which the training score improves but it decreases on evaluation sets. However, it was noticed that the AUC on the training set started to decrease as well after a certain threshold. Per the authors interpretation, as the training AUC score decreased, the model converged to a state that would yield similar results to the other models (AUC < 0.70).

Postprocessing

For the postprocessing phase, the TSNE projections were used with an overlay of nonlinear decision boundaries determined by the selected MLP and DT models to provide more insight into each model. In other words, the risk probability of exposure was calculated on the data set and demonstrated using linear interpolation to represent the decision surface on the TSNE space (Fig. 7).

The MLP decision boundary was smoother and captured smaller clusters of labels that were defined as exposure (Fig. 7).

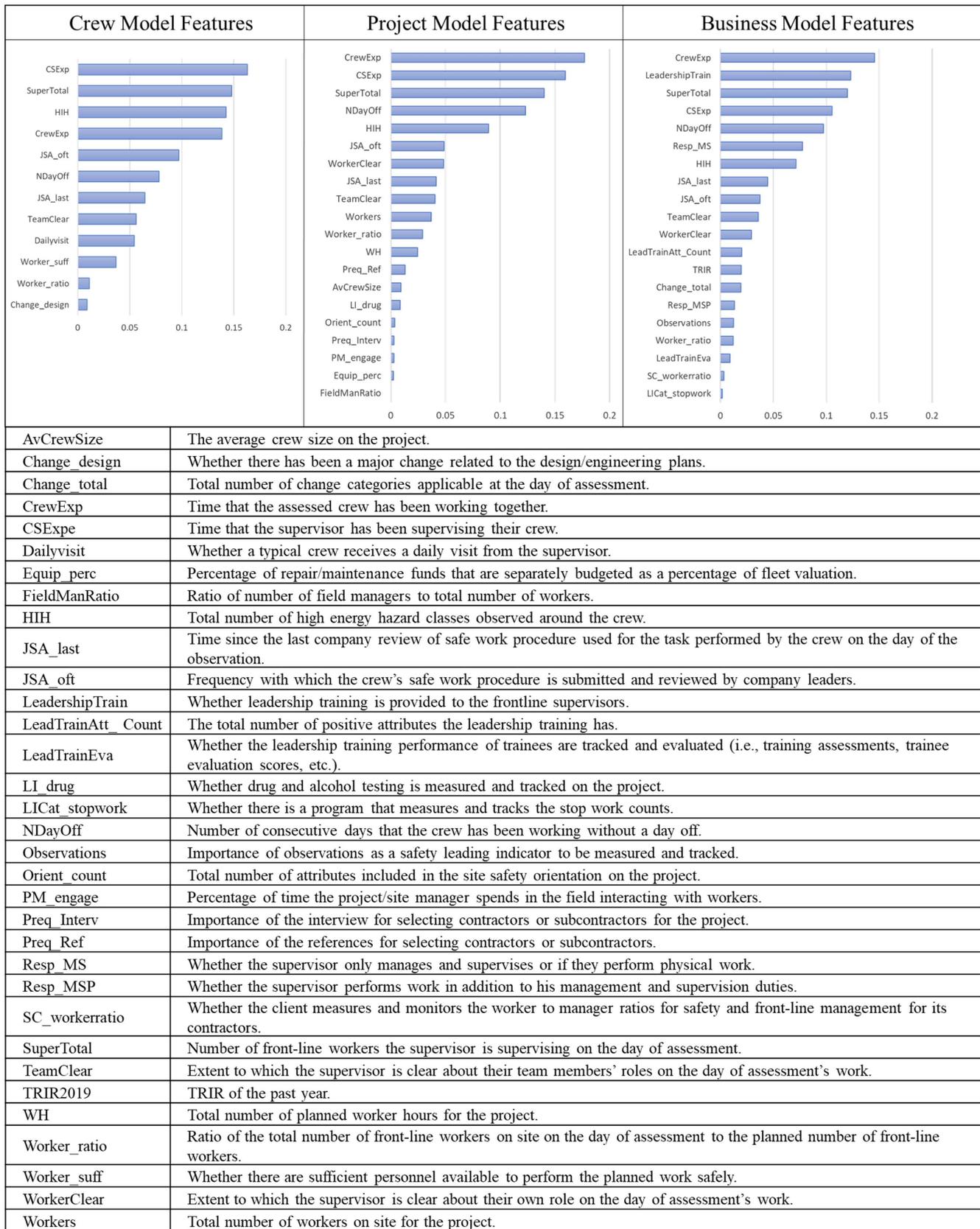


Fig. 5. Feature importance for crew, project, and business variables (in alphabetical order).

Table 2. AUC scores for crew, project, and business models

ML model	Crew data set	Project data set	Business data set
LR	0.55	0.58	0.65
DT	0.62	0.60	0.61
GBDT	0.61	0.67	0.65
RF	0.56	0.61	0.62
MLP	0.69	0.79	0.79

Conversely, DT's decision boundary had discrete separation levels which prevented the learned decision boundary from generalizing to unseen samples. Although MLP also eventually would converge to a similar sharper boundary, as is seen in the learning curves in Fig. 6, the authors managed to choose an optimum model by tracking the validation scores and selecting a model before this convergence was reached. It is expected that a decision surface with an abrupt transition between the different exposure levels might not generalize sufficiently, because TSNE already displayed a low separability between samples from different classes.

Discussion and Key Takeaways

The predictive models presented here offer robust predictions of a new safety variable, HECA. The separation of predictors into crew, project, and business levels was used strategically to inform different functional units within an organization regarding their decision-making processes related to safety. The ML pipeline revealed the key predictors.

- Business- and project-level factors are critical for holistic safety predictions. Crew-level features alone do not yield a high-performing prediction model (the maximum average AUC score was 0.69); however, they create a robust baseline. The models were highly predictive only when business and project features were added, resulting in an AUC score of 79%. By adding business and project predictors to the crew predictors, probability of HECA exposures may be accurately predicted. Practitioners may use this intelligence not only to estimate their risk of exposure, but also to prioritize their improvement efforts using highly ranked factors.
- Crew member experience working together and supervisors experience supervising the crew is crucial. In all models, the time

that the crew has worked together and supervisor experience with the crew were highly predictive and influential. This was followed by the number of frontline workers under the direction of the supervisor and the extent to which the supervisor is clear about their team's role. Collectively, these factors highlight the critical role of supervision and crew connectivity. For example, smaller crews with more experience with each other and with their supervisor tend to perform better than their counterparts. Although this may seem intuitive, the fact that this was the strongest trend among approximately 250 features is surprising.

- Project factors such as project scale, crew size, and leading indicators matter. Although crew features have the highest importance, project features related to project scale are highly influential. For example, the total number of workers, total number of worker hours, average crew size, and size of the project were all significant in the HECA predictions. These project features were followed by the average crew size, the existence of drug tests, and the maturity of safety orientations. Other features at the project level were percentage of the dedicated equipment maintenance budget and importance of references and interview in prequalifying contractors. In light of these results, industry professionals are encouraged to review their prequalification strategies that often depend on injury rates (Lofquist 2010), and consider developing processes to include more contextual factors related to the project demographics and the way that the contractor organizes their work.
- Businesses should consider allocating resources for leadership training to frontline supervisors. Among the business factors, the impact of providing leadership training to the frontline supervisors, which is a business feature, was ranked second; crew's experience working together, a crew feature, was ranked first. The quality indicators related to the leadership training also ranked highly. This finding warrants efforts by academics and industry professionals to develop and implement effective leadership training programs for frontline supervisors.
- Critical business factors include leading and lagging indicators. TRIR has a place in SIF exposure predictions. Supporting the intuition that a balanced approach is warranted to create more robust statistical models for SIF risk prediction, the previous year's TRIR is predictive of future exposure. Other features at the business level were having a leading indicator program

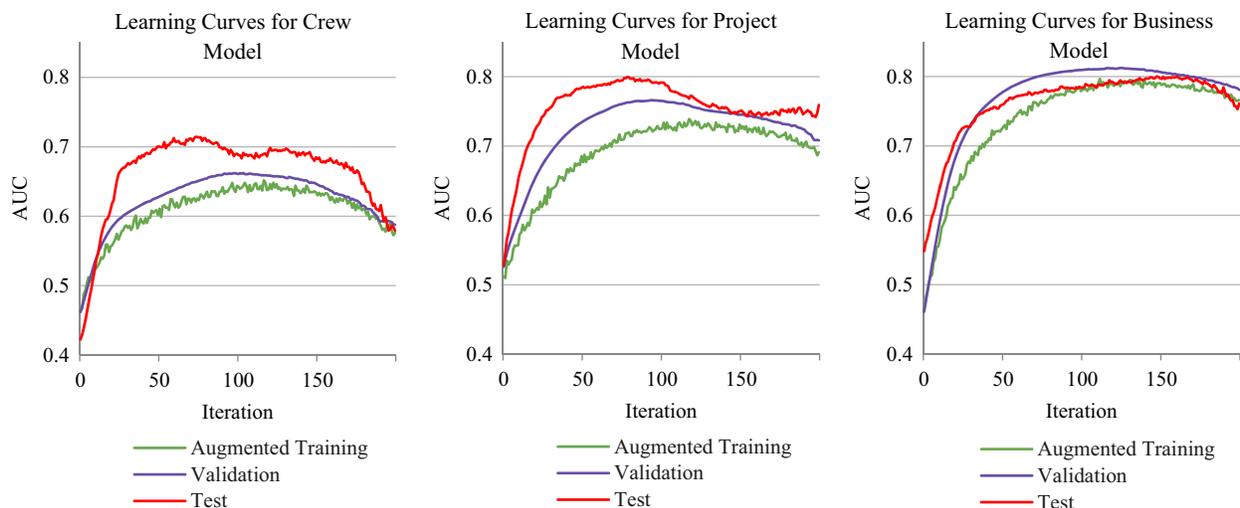


Fig. 6. Learning curves for the MLP models for Crew, Project, and Business data sets.

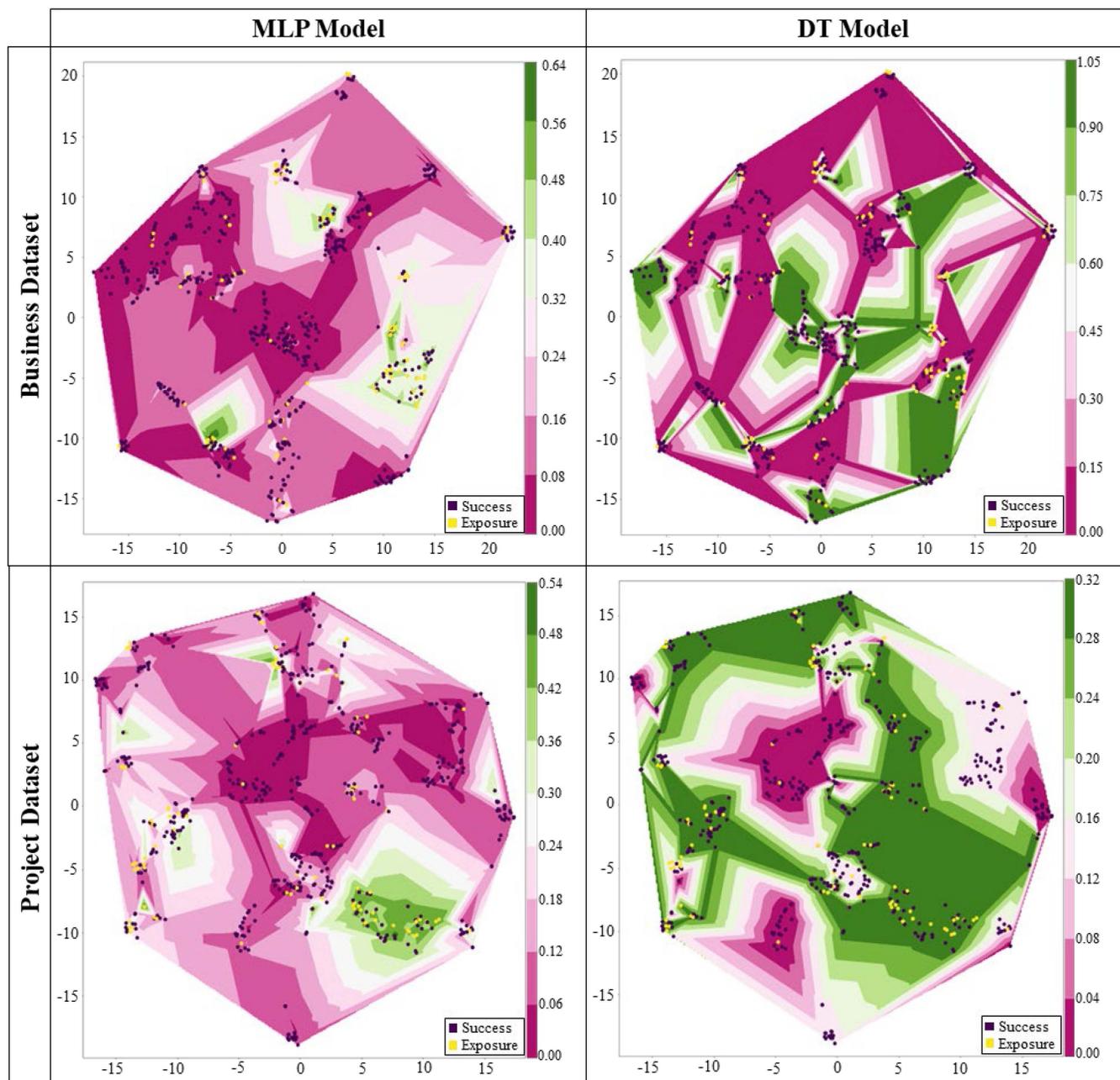


Fig. 7. MLP and DT model decision surface comparisons for Business and Project data.

that monitors stop works and observations, monitoring the supervisor to worker ratio of contractors, and evaluating the performance of trainees after they have received leadership training.

Compared with the previous literature, the findings had significant overlap among important predictors of SIFs such as leading indicators, job safety assessments (JSAs), site supervision, and recordable rates. Despite this overlap, the predictive model uncovered new predictors such as equipment repair funds, leadership training for frontline supervisors, importance given to tracking observations, and the total number of high-energy hazards in the workspace and of production responsibilities of the supervisor. Facilitated by project- and business-level features, predictors related to the prequalification and orientation programs were determined in order to address higher-level risks that have consequences at the crew level.

Conclusions, Limitations, and Future Research

This study made two key advancements in methodology: (1) a focus on predicting success and exposure conditions and adherence to strict operational definitions; and (2) the collection of a novel and empirical data set that included business-, project-, and crew-level factors. State-of-the-art ML algorithms were applied to the data set (i.e., LR, DT, GBDT, RF, and MLP), and MLP performed the best, with AUC scores of 0.79, 0.79, and 0.69 for business, project, and crew levels, respectively. The knowledge gained revealed that, among over 300 factors that were hypothesized to be predictive, a small subset has strong predictive capacity, and therefore deserve most attention.

Construction professionals long have desired robust safety predictions. Despite progress with leading indicators, climate surveys, precursor analysis, and risk assessments, most existing predictive

models rely solely on injury outcomes, which makes the assessment of likelihood of injury almost impossible (De Finetti et al. 2017). By predicting SIF conditions instead of incidents, and including operational definitions of the dependent variable, this study offers a robust predictive model that enables proactive decision making. Two areas of significant improvement were (1) inclusion of safety success along with exposure using a monitoring metric as dependent variable, and (2) incorporation of predictors from different hierarchical levels of an organization.

1. This study suggests a shift from trying to predict very rare and random events to predicting the probability of success or exposure conditions that occur continuously during active work. Using this more-stable prediction approach, researchers may test the effectiveness of their interventions in the short term, and professionals may be more proactive in addressing life-threatening conditions.
2. Specifically, the feature importance of each attribute may be used to guide professionals regarding what to monitor and prioritize. Moreover, the categorization of predictors into business, project, and crew levels helps organizations to assign efforts to different functional groups. For example, although organizations might not be able to make changes in their business structure over a short period, they might be able to implement highly impactful project- or crew-level interventions. Moreover, this structure can be used to explain how business- or project-level decisions impact the crew. The feature importance of each attribute may be used to guide professionals regarding what to monitor and prioritize.

Finally, this study makes important strides toward probabilistic predictions for SIF risk prediction. The ML pipeline and approach could be used by academics to generate and test their own ML models. A similar approach also could be used for planning and testing interventions, creating interdependent network approaches, strategizing change management, and generating innovation adoption strategies for on-site implementations. Moreover, this study demonstrated the implementation of recent ML techniques such as data augmentation, neural networks, and embedding space representations for SIF prevention. The models created were adjusted due to the unique data and analysis challenges and constraints to advance the application of automation methods in construction safety.

Although the sample data set collected in this study was unique and sufficient to facilitate analysis as a proof-of-concept, the major limitation of this study was the sample size. Because the database included only information submitted by 28 construction companies in the United States and Canada, the results cannot be generalized to a larger population. This limits the external validity of the results, and calls for future researchers to collect data at scale. Furthermore, as with any other research that depends on survey data, the collected information is subjected to potential cognitive biases. Even though the authors took measures to mitigate such biases using training and clear guidelines, respondent and research biases still are possible.

The crew-, project-, and business-related features chosen using the ML pipeline were incorporated into the machine learning model in parallel without any built-in or predetermined relationships. Although the data preparation and feature selection methods prevented the inclusion of highly correlated features within the resultant model, this decision might have imposed a limitation due to the exclusion of potential interdependencies or causal relationships between features in the model. Future research could explore such interrelationships to embed related underlying safety knowledge within the predictive models for increased performance.

This paper collected data related to a new outcome variable, HECA. Although HECA offers great promise as a monitoring

metric that can generate robust and large data sets that can facilitate more advanced automation methods in making predictions to prevent SIFs, its association with traditional injury rates are unknown. Future researchers are encouraged to collect data over time for both HECA and injury rates to compare their individual attributes and study their potential correlations. Furthermore, this research piloted HECA data collection in a limited binary form (as discrete conditions of success and exposure). Because this was the first time that HECA has been deployed in an empirical study, the assessments were considered as binary for each observation to ensure that the concept could be applied consistently. As this concept matures and is understood better, future researchers may consider collecting and analyzing HECA data as a continuous variable by recording individual high-energy hazards and their corresponding direct controls. Although this approach would be more complex, it would capture a more detailed representation of safety conditions.

Future research could address these issues by automation of site assessments and data collection. With automated site data collection systems, future researchers will be able to generate larger data sets from a larger number of companies while fully eliminating cognitive biases due to dependence on survey data. Because relevant research already has explored automated hazard recognition systems (Teizer and Cheng 2015), the prospect of developing a comprehensive high-energy and direct controls-based automated risk recognition system is critical in SIF prevention. In parallel with the ever-increasing data processing technology, and inspired by interdisciplinary research in traffic safety (Karacasu et al. 2014) and aviation (Yeoum and Lee 2013), opportunity for automation should continue to be recognized by researchers in construction (Choi et al. 2020; Tixier et al. 2016a). The shift toward using larger data sets will offer more predictive power and will allow the utilization of more-complex machine learning methods that could be used to make construction sites safer for workers.

Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

The authors thank the Construction Safety Research Alliance for providing support for this research, and the experts who participated in the study for their dedication, creativity, and insight.

References

- Albert, A., M. R. Hallowell, and B. M. Kleiner. 2014. "Experimental field testing of a real-time construction hazard identification and transmission technique." *Construct. Manage. Econ.* 32 (10): 1000–1016. <https://doi.org/10.1080/01446193.2014.929721>.
- Albert, A., M. R. Hallowell, M. Skaggs, and B. Kleiner. 2017. "Empirical measurement and improvement of hazard recognition skill." *Saf. Sci.* 93 (Mar): 1–8. <https://doi.org/10.1016/j.ssci.2016.11.007>.
- Alexander, D., M. Hallowell, and J. Gambatese. 2017a. "Precursors of construction fatalities. I: Iterative experiment to test the predictive validity of human judgment." *J. Constr. Eng. Manage.* 143 (7): 1–12. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001304](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001304).
- Alexander, D., M. Hallowell, and J. Gambatese. 2017b. "Precursors of construction fatalities. II: Predictive modeling and empirical validation." *J. Constr. Eng. Manage.* 143 (7): 1–12. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001297](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001297).

- Alruqi, W. M., and M. R. Hallowell. 2019. "Critical success factors for construction safety: Review and meta-analysis of safety leading indicators." *J. Constr. Eng. Manage.* 145 (3): 1547–1556. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001626](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001626).
- Alruqi, W. M., M. R. Hallowell, and U. Techera. 2018. "Safety climate dimensions and their relationship to construction safety performance: A meta-analytic review." *Saf. Sci.* 109 (5): 165–173. <https://doi.org/10.1016/j.ssci.2018.05.019>.
- Baradan, S., and M. A. Usman. 2006. "Comparative injury and fatality risk analysis of building trade." *J. Constr. Eng. Manage.* 132 (5): 533–539. [https://doi.org/10.1061/\(ASCE\)0733-9364\(2006\)132:5\(533\)](https://doi.org/10.1061/(ASCE)0733-9364(2006)132:5(533)).
- Breiman, L. 2001. "Random forests." *Mach. Learn.* 45 (Oct): 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Brown, S., W. Harris, R. D. Brooks, and X. S. Dong. 2021. "Fatal injury trends in the construction industry." Accessed January 15, 2023. <https://www.cpw.com/wp-content/uploads/DataBulletin-February-2021.pdf>.
- Cheng, J., G. Li, and X. Chen. 2019. "Research on travel time prediction model of freeway based on gradient boosting decision tree." *IEEE Access* 7 (Dec): 7466–7480. <https://doi.org/10.1109/ACCESS.2018.2886549>.
- Choe, S., and F. Leite. 2016. "Assessing safety risk among different construction trades: Quantitative approach." *J. Constr. Eng. Manage.* 143 (5): 04016133. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001237](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001237).
- Choi, J., B. Gu, S. Chin, and J. S. Lee. 2020. "Machine learning predictive model based on national data for fatal accidents of construction workers." *Autom. Constr.* 110 (Feb): 102974. <https://doi.org/10.1016/j.autcon.2019.102974>.
- Costin, A., A. Wehle, and A. Adibfar. 2019. "Leading indicators—A conceptual IoT-based framework to produce active leading indicators for construction safety." *Safety* 5 (4): 86. <https://doi.org/10.3390/safety5040086>.
- De Finetti, B., A. Machì, and A. F. M. Smith. 2017. *Theory of probability: A critical introductory treatment*. Chichester, UK: Wiley.
- Dekker, S., and C. Pitzer. 2016. "Examining the asymptote in safety progress: A literature review." *Int. J. Occup. Saf. Ergon.* 22 (1): 57–65. <https://doi.org/10.1080/10803548.2015.1112104>.
- Esmaili, B., M. R. Hallowell, and B. Rajagopalan. 2015a. "Attribute-based safety risk assessment. I: Analysis at the fundamental level." *J. Constr. Eng. Manage.* 141 (8): 04015021. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000980](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000980).
- Esmaili, B., M. R. Hallowell, and B. Rajagopalan. 2015b. "Attribute-based safety risk assessment. II: Predicting safety outcomes using generalized linear models." *J. Constr. Eng. Manage.* 141 (8): 1–11. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000981](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000981).
- Fawcett, T. 2006. "An introduction to ROC analysis." *Pattern Recognit. Lett.* 27 (8): 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>.
- Goh, Y. M., and D. Chua. 2013. "Neural network analysis of construction safety management systems: A case study in Singapore." *Construct. Manage. Econ.* 31 (5): 460–470. <https://doi.org/10.1080/01446193.2013.797095>.
- Gondia, A., A. Siam, W. El-Dakhkhni, and A. H. Nassar. 2020. "Machine learning algorithms for construction projects delay risk prediction." *J. Constr. Eng. Manage.* 146 (1): 04019085. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001736](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001736).
- Guo, B. H. W., and T. W. Yiu. 2016. "Developing leading indicators to monitor the safety conditions of construction projects." *J. Manage. Eng.* 32 (1): 04015016. [https://doi.org/10.1061/\(ASCE\)ME.1943-5479.0000376](https://doi.org/10.1061/(ASCE)ME.1943-5479.0000376).
- Guo, B. H. W., T. W. Yiu, V. A. González, and Y. M. Goh. 2017. "Using a pressure-state-practice model to develop safety leading indicators for construction projects." *J. Constr. Eng. Manage.* 143 (2): 04016092. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001218](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001218).
- Guo, R., D. Fu, and G. Sollazzo. 2021. "An ensemble learning model for asphalt pavement performance prediction based on gradient boosting decision tree." *Int. J. Pavement Eng.* 23 (10): 3633–3646. <https://doi.org/10.1080/10298436.2021.1910825>.
- Hallowell, M. 2020. "Safety classification and learning (SCL) model." *Edison Electric Insitute*. Accessed January 15, 2023. <https://www.eei.org/-/media/Project/EEI/Documents/Issues-and-Policy/Power-to-Prevent-SIF/eeiSCLmodel.pdf?la=en&hash=4E03097C0292F52CB4FA186D0D8CE11876032836>.
- Hallowell, M., S. Bhandari, and W. Alruqi. 2019. "Methods of safety prediction: Analysis and integration of risk assessment, leading indicators, precursor analysis, and safety climate." *Construct. Manage. Econ.* 38 (4): 308–321. <https://doi.org/10.1080/01446193.2019.1598566>.
- Hallowell, M., M. Quashne, R. Salas, B. MacLean, and E. Quinn. 2021. "The statistical invalidity of TRIR as a measure of safety performance." *Prof. Saf.* 66 (4): 28–34.
- Hallowell, M. R., D. Alexander, and J. A. Gambatese. 2017. "Energy-based safety risk assessment: Does magnitude and intensity of energy predict injury severity?" *Construct. Manage. Econ.* 35 (1–2): 64–77. <https://doi.org/10.1080/01446193.2016.1274418>.
- He, X., K. Zhao, and X. Chu. 2021a. "AutoML: A survey of the state-of-the-art." *Knowl.-Based Syst.* 212 (Jan): 106622. <https://doi.org/10.1016/j.knosys.2020.106622>.
- He, Z., D. J. Armaghani, M. Masoumnezhad, M. Khandelwal, J. Zhou, and B. R. Murlidhar. 2021b. "A combination of expert-based system and advanced decision-tree algorithms to predict air-overpressure resulting from quarry blasting." *Nat. Resour. Res.* 30 (2): 1889–1903. <https://doi.org/10.1007/s11053-020-09773-6>.
- Hinze, J., S. Thurman, and A. Wehle. 2013. "Leading indicators of construction safety performance." *Saf. Sci.* 51 (1): 23–28. <https://doi.org/10.1016/j.ssci.2012.05.016>.
- Hollnagel, E. 2014. "Is safety a subject for science?" *Saf. Sci.* 67 (Aug): 21–24. <https://doi.org/10.1016/j.ssci.2013.07.025>.
- Hollnagel, E. 2015. "From safety-I to safety-II: A white paper." Accessed January 15, 2023. <https://www.england.nhs.uk/signuptosafety/wp-content/uploads/sites/16/2015/10/safety-1-safety-2-white-papr.pdf>.
- Hopkins, A. 2009. "Thinking about process safety indicators." *Saf. Sci.* 47 (4): 508–510. <https://doi.org/10.1016/j.ssci.2007.12.006>.
- Karacasu, M., B. Ergül, and A. A. Yavuz. 2014. "Estimating the causes of traffic accidents using logistic regression and discriminant analysis." *Int. J. Inj. Control Saf. Promot.* 21 (4): 305–313. <https://doi.org/10.1080/17457300.2013.815632>.
- Kines, P. 2001. "Occupational injury risk assessment using injury severity odds ratios: Male falls from heights in the Danish construction industry, 1993–1999." *Hum. Ecol. Risk Assess.* 7 (7): 1929–1943. <https://doi.org/10.1080/20018091095492>.
- Krogh, A., and J. A. Hertz. 1991. "A simple weight decay can improve generalization." In Vol. 4 of *Advances in neural information processing systems*. San Mateo, CA: Morgan Kaufmann.
- Leung, C. S., J. Sum, and S. K. Mak. 2010. "Generalization error of faulty MLPs with weight decay regularizer." In *Proc., Int. Conf. on Neural Information Processing*, 160–167. Berlin: Springer.
- Lingard, H., M. Hallowell, R. Salas, and P. Pirzadeh. 2017. "Leading or lagging? Temporal analysis of safety indicators on a large infrastructure construction project." *Saf. Sci.* 91 (Jan): 206–220. <https://doi.org/10.1016/j.ssci.2016.08.020>.
- Liu, F. T., K. M. Ting, and Z. H. Zhou. 2012. "Isolation-based anomaly detection." *ACM Trans. Knowl. Discovery Data* 6 (1): 1–39. <https://doi.org/10.1145/2133360.2133363>.
- Liu, H., and G. Tian. 2019. "Building engineering safety risk assessment and early warning mechanism construction based on distributed machine learning algorithm." *Saf. Sci.* 120 (Dec): 764–771. <https://doi.org/10.1016/j.ssci.2019.08.022>.
- Lofquist, E. A. 2010. "The art of measuring nothing: The paradox of measuring safety in a changing civil aviation industry using traditional safety metrics." *Saf. Sci.* 48 (10): 1520–1529. <https://doi.org/10.1016/j.ssci.2010.05.006>.
- NIOSH (The National Institute for Occupational Safety and Health). 2015. "Hierarchy of controls." Accessed August 1, 2022. <https://www.cdc.gov/niosh/topics/hierarchy/default.html>.
- Oguz Erkal, E. D., and M. R. Hallowell. 2023. "Moving beyond TRIR: Measuring and monitoring safety performance with high-energy control assessments." *Prof. Saf.* 68 (5): 26–35.

- Oguz Erkal, E. D., M. R. Hallowell, and S. Bhandari. 2021. "Practical assessment of potential predictors of serious injuries and fatalities in construction." *J. Constr. Eng. Manage.* 147 (10): 04021129. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0002146](https://doi.org/10.1061/(ASCE)CO.1943-7862.0002146).
- Oguz Erkal, E. D., M. R. Hallowell, and S. Bhandari. 2023. "Formal evaluation of construction safety performance metrics and a case for a balanced approach." *J. Saf. Res.* 85 (Jun): 380–390. <https://doi.org/10.1016/j.jsr.2023.04.005>.
- Poh, C. Q. X., C. U. Ubeynarayana, and Y. M. Goh. 2018. "Safety leading indicators for construction sites: A machine learning approach." *Autom. Constr.* 93 (Sep): 375–386. <https://doi.org/10.1016/j.autcon.2018.03.022>.
- Raheem, A. A., and J. W. Hinze. 2014. "Disparity between construction safety standards: A global analysis." *Saf. Sci.* 70 (Dec): 276–287. <https://doi.org/10.1016/j.ssci.2014.06.012>.
- Salas, R., and M. Hallowell. 2016. "Predictive validity of safety leading indicators: Empirical assessment in the oil and gas sector." *J. Constr. Eng. Manage.* 142 (10): 1–11. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001167](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001167).
- Sarkar, S., A. Pramanik, J. Maiti, and G. Reniers. 2020. "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data." *Saf. Sci.* 125 (May): 104616. <https://doi.org/10.1016/j.ssci.2020.104616>.
- Sarkar, S., S. Vinay, R. Raj, J. Maiti, and P. Mitra. 2019. "Application of optimized machine learning techniques for prediction of occupational accidents." *Comput. Oper. Res.* 106 (Jun): 210–224. <https://doi.org/10.1016/j.cor.2018.02.021>.
- Schwatka, N. V., and J. C. Rosecrance. 2016. "Safety climate and safety behaviors in the construction industry: The importance of co-workers commitment to safety." *Work* 54 (2): 401–413. <https://doi.org/10.3233/WOR-162341>.
- Seger, C. 2018. "An investigation of categorical variable encoding techniques in machine learning: Binary versus one-hot and feature hashing." Accessed August 1, 2022. <https://www.diva-portal.org/smash/get/diva2:1259073/FULLTEXT01.pdf>.
- Shawky, O. A., A. Hagag, E. S. A. El-Dahshan, and M. A. Ismail. 2020. "Remote sensing image scene classification using CNN-MLP with data augmentation." *Optik* 221 (Nov): 165356. <https://doi.org/10.1016/j.ijleo.2020.165356>.
- Shehadeh, A., O. Alshboul, R. E. Al Mamlook, and O. Hamedat. 2021. "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression." *Autom. Constr.* 129 (Sep): 103827. <https://doi.org/10.1016/j.autcon.2021.103827>.
- Shin, Y. 2019. "Application of stochastic gradient boosting approach to early prediction of safety accidents at construction site." *Adv. Civ. Eng.* 2019 (Dec): 1–9. <https://doi.org/10.1155/2019/1574297>.
- Siu, O. L., D. R. Phillips, and T. W. Leung. 2004. "Safety climate and safety performance among construction workers in Hong Kong: The role of psychological strains as mediators." *Accid. Anal. Prev.* 36 (3): 359–366. [https://doi.org/10.1016/S0001-4575\(03\)00016-2](https://doi.org/10.1016/S0001-4575(03)00016-2).
- Soibelman, L., and H. Kim. 2002. "Data preparation process for construction knowledge generation through knowledge discovery in databases." *J. Comput. Civ. Eng.* 16 (1): 39–48. [https://doi.org/10.1061/\(ASCE\)0887-3801\(2002\)16:1\(39\)](https://doi.org/10.1061/(ASCE)0887-3801(2002)16:1(39)).
- Son, H., C. Kim, and C. Kim. 2012. "Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables." *Autom. Constr.* 27 (Nov): 60–66. <https://doi.org/10.1016/j.autcon.2012.05.013>.
- Teizer, J., and T. Cheng. 2015. "Proximity hazard indicator for workers-on-foot near miss interactions with construction equipment and geo-referenced hazard areas." *Autom. Constr.* 60 (Dec): 58–73. <https://doi.org/10.1016/j.autcon.2015.09.003>.
- Tixier, A. J.-P., A. Albert, and M. R. Hallowell. 2018. "Proposing and validating a new way of construction hazard recognition training in academia: Mixed-method approach." *Pract. Period. Struct. Des. Constr.* 23 (1): 04017027. [https://doi.org/10.1061/\(ASCE\)SC.1943-5576.0000347](https://doi.org/10.1061/(ASCE)SC.1943-5576.0000347).
- Tixier, A. J.-P., M. R. Hallowell, and B. Rajagopalan. 2017. "Construction safety risk modeling and simulation." *Risk Anal.* 37 (10): 1917–1935. <https://doi.org/10.1111/risa.12772>.
- Tixier, A. J.-P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2016a. "Application of machine learning to construction injury prediction." *Autom. Constr.* 69 (Sep): 102–114. <https://doi.org/10.1016/j.autcon.2016.05.016>.
- Tixier, A. J.-P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2016b. "Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports." *Autom. Constr.* 62 (Feb): 45–56. <https://doi.org/10.1016/j.autcon.2015.11.001>.
- US Bureau of Labor Statistics. 2020a. "Injuries, illnesses, and fatalities." Accessed October 2, 2021. https://www.bls.gov/iif/nonfatal-injuries-and-illnesses-tables/soii-summary-historical.htm#20Summary_Tables.
- US Bureau of Labor Statistics. 2020b. "National census of fatal occupational injuries in 2019." Accessed December 16, 2020. https://www.bls.gov/news.release/archives/cfoi_12162020.pdf.
- Van der Maaten, L. J., and G. E. Hinton. 2008. "Visualizing data using t-SNE." *J. Mach. Learn. Res.* 9 (11): 2579–2605.
- Versteeg, K., P. Bigelow, A. M. Dale, and A. Chaurasia. 2019. "Utilizing construction safety leading and lagging indicators to measure project safety performance: A case study." *Saf. Sci.* 120 (Dec): 411–421. <https://doi.org/10.1016/j.ssci.2019.06.035>.
- Wang, J., and S. Razavi. 2019. "Network-based safety leading indicators for safety risk analysis in construction." In Vol. 63 of *Proc., Human Factors and Ergonomics Society Annual Meeting 2019*, 1787–1791. Los Angeles: SAGE.
- Witten, I. H., and E. Frank. 2002. "Data mining: Practical machine learning tools and techniques with Java implementations." *ACM Sigmod Rec.* 31 (1): 76–77. <https://doi.org/10.1145/507338.507355>.
- Yeoum, S. J., and Y. H. Lee. 2013. "A study on prediction modeling of Korea Military aircraft accident occurrence." *Int. J. Ind. Eng.* 20 (Sep): 562–573.
- Zhang, H., F. Yang, Y. Li, and H. Li. 2015. "Predicting profitability of listed construction companies based on principal component analysis and support vector machine—Evidence from China." *Autom. Constr.* 53 (May): 22–28. <https://doi.org/10.1016/j.autcon.2015.03.001>.
- Zhang, M., T. Cao, and X. Zhao. 2019. "Using smartphones to detect and identify construction workers' near-miss falls based on ANN." *J. Constr. Eng. Manage.* 145 (1): 04018120. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001582](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001582).
- Zhu, R., X. Hu, J. Hou, and X. Li. 2021. "Application of machine learning techniques for predicting the consequences of construction accidents in China." *Process Saf. Environ. Prot.* 145 (Jan): 293–302. <https://doi.org/10.1016/j.psep.2020.08.006>.