Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

Modeling Amputation-Causing Construction-Related Accidents in the U.S. using Machine Learning

Mahram Rajabi Jorshari
Ph.D. Student, Department of Civil, Environmental, and Construction Engineering Box 41023, Texas Tech University, Lubbock, TX 79409−1023;

Tewodros Ghebrab, Ph.D., P.E.
Assistant Professor, Department of Civil, Environmental, and Construction Engineering Box 41023, Texas Tech University, Lubbock, TX 79409−1023;

Clifford B. Fedler, Ph.D., P.E
Professor, Department of Civil, Environmental, and Construction Engineering Box 41023, Texas Tech University, Lubbock, TX 79409−1023;

Yaser Pakzad Jafarabadi
Master of Engineering, Department of Civil, Environmental, and Construction Engineering Box 41023, Texas Tech University, Lubbock, TX 79409−1023;

Abstract: **In the United States, over 300 construction-related accidents resulting in amputations are reported to the Occupational Safety and Health Administration (OSHA) annually. These types of accidents create significant challenges for employees, their families, and their employers. Efforts must be made to identify ways to reduce such accidents. An analysis using machine learning (ML) to predict amputation has been conducted using OSHA data reported between January 2015 and December 2022, a dataset of 79,000 records. The dataset includes critical details such as event description and event source. This study examines the advantages and limitations of amputation prediction models, considering variables such as accuracy, precision, recall, and F1-score. By analyzing the effectiveness of ML modeling systems in predicting amputations resulting from workplace accidents in various construction industries, the research can assist safety professionals and the OSHA department in taking extra precautionary measures to mitigate such accidents. Utilizing the Python programming language and a range of ML algorithms, including Logistic Regression and Dummy Variables, the research focuses on identifying key predictors such as event type (e.g., slip, fall) and source type (e.g., machinery, tools). Despite challenges like data defects, overfitting, noise, and imbalanced data, the best-performing model achieves a prediction accuracy of 87%.**

1. INTRODUCTION

In the United States, workplace accidents and injuries are reported to the Occupational Safety and Health Administration (OSHA). These accidents are for different industries, such as construction, transportation, manufacturing, retail, food, and healthcare. These reports are critical for maintaining workplace safety, identifying hazards, and developing safety policies to prevent future incidents [1]. Table 1 illustrates the total number of accidents, construction-related accidents, and construction-related amputations reported to Federal OSHA. It indicates a decrease in reported accidents in the recent decade and a slight decrease in construction-related accidents, indicating that the number of reported amputations has almost remained steady between the period of 2015 and 2022.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

Table 1. Number of accidents, construction-related accidents, and construction-related amputations as reported by OSHA by year.

Year	Number of accidents	Number of construction-related accidents	Number of construction-related amputations
2015	9852	1829	281
2016	10091	1929	271
2017	10448	2080	293
2018	11156	2225	325
2019	11075	2271	354
2020	8915	1929	280
2021	8704	1893	288
2022	9110	2012	327
Total	79351	16168	2416
Mean	10544	2021	302
SD	785	139	27

Over the period between 2015 and 2022 (Table 1), the number of construction-related amputations compared to total construction-related accidents as reported by the Federal OSHA was nearly constant. A large percentage of non-fatal accidents reported to OSHA involve amputations, while high-risk tasks, such as operating heavy machinery resulting in amputations continue to pose significant hazards. Additionally, OSHA's safety measures and policies aimed at preventing amputations may not be fully or effectively implemented within the construction industry.

Under the direction of the US Department of Labor, and thus, OSHA is a federal organization. It is OSHA's responsibility to ensure that all workers have safe working conditions by investigating accidents and imposing rules and policies that can reduce the risk of future accidents [2]. While OSHA's efforts are intended to reduce workplace accidents, challenges continue to pose difficulties in ensuring worker safety. These challenges include limited resources for inspections, under-reporting of workplace injuries by employers, and a lengthy process for developing new safety standards [3]. Additionally, high-risk industries like construction present unique hazards, and ensuring compliance can be difficult, especially in temporary or rapidly changing work environments [4].

Some states operate their own OSHA-approved State Plans, enabling them to customize workplace safety regulations to their specific industries and requirements. These plans must be at least as effective as federal standards and can even include stricter or additional rules. While Federal OSHA oversees these plans to ensure compliance and offers funding assistance, the states retain independence in enforcing their safety regulations [5]. For example, the state of California has its own OSHA, which is called Cal-OSHA. Under such circumstances, employers are required to report accidents to the state agency in addition to the Federal OSHA. As a result, states may have different standards, and employers should know the specific OSHA regulations that apply to them. To ensure compliance with reporting requirements, companies and organizations should be aware of all applicable federal and state regulations regarding their location and industry. [6].

Workers in various industries face significant risks of accidents leading to amputation. Amputations in the construction industry are distinct from other types of accidents since they have severe and often life-altering consequences. Unlike other workplace accidents, which may cause temporary injuries or disabilities, amputations typically cause permanent physical and psychological impacts. Amputations often result from accidents involving heavy machinery, power tools, and equipment used in construction. These accidents impact these construction workers' (victims) future physical health, livelihoods, and quality of life [7].

In an effort to reduce the possibility of severe accidents and improve workplace safety, there is a need to integrate data science and safety data. Machine Learning (ML) is an approach that can be used to help analyze construction-based data. ML refers to a system's ability to learn from data, enabling it to assist in making predictions about the outcomes of the data [8]. ML plays a vital role in data science as it provides tools to analyze and extract patterns from large datasets. The efficiency, speed, and adaptability of ML enable the automation of decision-making procedures, the extraction of insightful information from data, and the creation of prediction models with a wide range of uses [9].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

The application of ML in analyzing accident-related data to enhance workplace safety is growing. Recent developments in ML have provided new paths in anticipating and preventing catastrophic accidents [10]. Recent ML tools have several strengths, including the use of predictive analytics, pattern recognition, automation of data processing, real-time monitoring, and tailored training programs. These strengths can be applied to workplace accident analysis by predicting high-risk scenarios for workplace accidents, identifying hidden patterns in accident reports, automated analysis of safety inspection data, and adapting safety protocols based on emerging trends [11]. However, it also has some weaknesses, such as data quality issues interpretability and high costs for data collection and infrastructure. These weaknesses of ML can manifest when analyzing workplace accidents with incomplete or inaccurate incident reporting, making the analysis results inappropriate [12].

Researchers and safety experts have carried out several studies and investigations in this area since they are aware of how these technologies might improve workplace safety [13]. They often focused on more common workplace injuries, overlooking amputation accidents because of their rarity and the complexity of predicting such severe outcomes. For example, Zhu et al. (2023) worked on predictive models of construction fatality characteristics using ML [14]. Chokor et al. (2016) worked on analyzing Arizona OSHA injury reports using ML [15].

Currently, no research has focused on predicting amputations using ML based on accidents reported to the Federal OSHA. This study explores the field of predictive analytics by analyzing the abilities of ML models to predict amputation-causing construction-related accidents.

2. OBJECTIVE

The goal of this research is to analyze the effectiveness of using Machine Learning (ML) system in predicting instances of amputations resulting from workplace accidents in various construction industries based on categorized events and sources data. These data were extracted from the Occupational Safety and Health Administration (OSHA) database, which concentrated on amputations in the construction industry. To accomplish this goal, the project was divided into four sub-objectives.

The first sub-objective was to collect relevant data that can be used to develop a predictive algorithm. The OSHA database was used as the data source for this study, which concentrated on amputations in the construction industry. The second sub-objective focused on analyzing the collected data to identify relevant features and parameters that are essential for modeling and predicting amputations. Additionally, this sub-objective aimed to categorize the data into distinct parameters suitable for ML modeling. As the third sub-objective, an ML analysis was conducted using ML algorithms to analyze the extracted features and develop a predictive model for identifying potential accidents resulting in amputation. The third sub-objective was to use the categorized data to develop the ML algorithms used to predict the elements that cause amputation from construction-related accidents. The fourth sub-objective was to evaluate the advantages and limitations of the developed predictive models.

3. METHODOLOGY

To complete sub-objective 1, data associated with the incidence of amputation resulting from construction-related accidents as reported to OSHA was gathered. The dataset included details such as event date, employer's information, address, final report, part of body involved in accidents, source of accidents, and events explaining the accidents. The outcome was to ensure a comprehensive dataset supporting reliable model training and practical, actionable insights for construction enterprises to predict workplace amputations. Approximately 79,000 total records were collected from OSHA for the time period between 2015 to 2022. The data consists of accidents from all types of industries and for all types of accidents. The dataset was initially filtered to select only accidents related to the building construction industry. This process resulted in 16,168 construction industry-related accidents: of those, only 2,416 involved amputations.

For sub-objective 2, the data requires consistent terminology, such that the data features were standardized to maintain consistency. Source and Event were chosen as the features. These features are predictive relevance, contextual meaning, and alignment with the research goal. Sources talk about the accident's details, and the Event mentions the kind of accident. The quality of these features dramatically impacts the model's ability to identify meaningful patterns. Analyzing features helps identify high-risk equipment and the most common accident types. Both of these features are directly related to the causes and mechanisms of injuries such as amputations, and they are well-documented in OSHA data, making them reliable and consistent features for analysis. In addition to features, the Categorized Source

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

(CS) and Categorized Event (CE) were added to the features used for modeling. CS is extracted from the source, and CE from the Event. For example, all source titles beginning with the number 357, including 3570, 3571, 3572, 3573, 3574, 3575, 3576, 3577, 3578, and 3579 are from the same category. In other words, sources that mention tools involved in accidents, including sawing machinery and all pieces related to it, are from the same category, and they can be in the same category as number 357. By grouping related Sources or Events into broader categories, the model reduces noise and sparsity in the data, which improves prediction accuracy.

Correlation quantifies the strength and direction of the relationship between two numerical variables, helping to analyze how changes in one variable correspond to changes in another [16]. A positive correlation (+1) indicates that both variables increase together, while a negative correlation (-1) signifies that as one variable increases, the other decreases. A value of zero denotes no relationship between the variables. Correlation is particularly useful in feature selection (to address multicollinearity), data analysis (to understand variable relationships), and model improvement (to identify impactful predictors) [17]. The P-value is checked when working with data to determine the statistical significance of features in an ML model. It helps identify whether a feature has a tangible impact on the target variable or if its effect is due to chance. It ranges from 0 to 1. A low P-value (typically ≤ 0.05) indicates that the feature is statistically significant, suggesting strong evidence that it influences predictability. On the other hand, a high P-value indicates that the feature cannot make a significant contribution to predictions [18]. The Variance Inflation Factor (VIF) measures the degree of multicollinearity in a dataset. Multicollinearity occurs when two or more independent variables are highly correlated, making one variable predictable from the others. Multicollinearity leads to redundancy, which can cause unstable regression coefficients, inflated standard errors, and poor model interpretability (Cheng et al., 2022). VIF quantifies how much the variance of an individual feature's coefficient is inflated due to correlation with other features. High multicollinearity can make model coefficients unstable and affect the model's interpretability. If VIF < 5, there is no multicollinearity, which is ideal. If VIF is between 5 and 10, there is moderate multicollinearity, which needs to be monitored carefully, and if VIF equals 10, it indicates high multicollinearity, meaning a feature is highly correlated with others, which can distort model interpretations (Yu et al., 2015). The correlation between Source and amputations was 0.40, and the P-value was less than 0.001. The correlation between Event and amputations was 0.38, and the P-value was less than 0.001. These numbers indicate a moderate but statistically significant relationship between the variables. Since multicollinearity can impact model stability, further Variance Inflation Factor (VIF) examination was needed to ensure that redundancy among features remains within acceptable limits. VIF = 1.02 suggested that the selected features contributed meaningful information without introducing excessive multicollinearity, supporting the robustness of the model.

As the third sub-objective, the ML algorithms were developed to predict amputations based on features from construction-related accidents. ML plays a critical role in data analysis by leveraging techniques and algorithms that enable computers to detect patterns, make predictions, and perform evaluations without the need for explicit programming [19]. By analyzing large datasets, ML identifies patterns and correlations that might go unnoticed by humans. These patterns and correlations facilitate continuous learning and performance enhancement over time [20]. Various ML algorithms are employed in data science, including linear regression, logistic regression, decision trees, random forests, and artificial neural networks. Selecting the right algorithm is crucial for ensuring accurate predictions based on the unique attributes of construction accident data. The algorithm selection depends on the characteristics of the data and the specific nature of the problem being addressed [21]. For instance, in the context of construction accident data, logistic regression is well-suited for binary outcomes, such as predicting amputations, due to its efficiency and capability to model linear relationships [22]. Decision trees excel at modeling non-linear relationships and identifying significant predictors [23], while random forests improve predictive accuracy by aggregating the results of multiple decision trees [24]. Artificial neural networks are highly effective for capturing intricate, non-linear patterns within large datasets that involve numerous interacting factors, such as machinery usage and environmental conditions [25].

ML tasks can generally be categorized into regression and classification, each serving distinct purposes. Regression tasks focus on predicting continuous outcomes based on input features, such as predicting home prices using attributes like size, location, and amenities. Classification, on the other hand, involves categorizing data into distinct groups. For example, determining whether an email is spam or not is a classification task [26]. Binary classification, a subset of supervised learning, focuses on assigning data points to one of two unique classes, often labeled as “true” and “false,” “yes” and “no,” or similar pairs. [27]; [28]. Logistic regression, a commonly used ML algorithm for binary classification tasks, demonstrates the significant role of statistical models in predictive analytics [29]. It is particularly suitable for binary outcomes and provides insights into the importance of features by analyzing model coefficients. Its simplicity and computational efficiency make it an attractive option for many applications [30].

As most ML algorithms require numerical input, dummy variables are used in ML and statistical modeling to convert categorical data into numerical formats. These binary variables, represented as 0 or 1, indicate the absence or presence

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

of a categorical effect. By incorporating dummy variables, categorical data can be effectively included in regression models and other statistical analyses. This transformation improves model performance and facilitates the interpretation of the effects of categorical variables [31]. By allowing each category to be represented as a binary variable, dummy variables avoid assumptions about ordinal relationships between categories and enable the model to learn distinct patterns for each [31]. This process ensures the model can process categorical data, preserves information, and avoids introducing numerical bias. In the research, dummy variables were used to convert the Event (e.g., "slip", "trip", "fall ") and Source (e.g., "machinery," "construction tools") features into binary numerical columns, indicating the presence or absence of each category.

To achieve acceptable results, ML models must undergo training and testing. Training involves using a dataset to teach the model how to make predictions or classifications by providing it with input data and the corresponding correct outputs. Through this training, the model learns to identify patterns and relationships within the data [28].

In logistic regression, training iterations represent the number of steps the optimization solver takes to find the optimal model parameters. For smaller datasets, a default of 100 iterations is usually adequate, while larger or more complex datasets may require higher iteration counts—often 500 or more—to ensure the solver successfully converges [32].

Moreover, another key concept in ML training is the learning rate, a hyperparameter that determines the size of the model's steps when updating weights during training. The learning rate controls how much the model's parameters are adjusted in response to the error or loss calculated at each training iteration. Fine-tuning this hyperparameter is critical for optimizing the model's performance [33].

The model's performance is typically assessed using a separate subset of the data, known as the test set, to ensure it has effectively learned patterns. The test size represents the fraction of the dataset allocated for testing purposes. For instance, a test size of 0.2 indicates that 20% of the dataset is reserved for testing [34].

To train the models, the dataset was divided into training and testing sets, and each model was trained on a fraction of the data, providing an effective training process. 80 percent of the data was used for training, while 20 percent was used to test the models. The process should be repeated, and in this case, 500 iterations of the process were used. The learning rate was 0.001, which is the default learning rate for many optimizers like Adam, often used in logistic regression. During training, the model uses the loss function to measure its performance and iteratively updates its parameters, such as weights and biases, to minimize the loss. A loss function quantifies the accuracy of a model's predictions by comparing them to the actual data. It evaluates the predicted values against the target values and provides a numerical representation of the "cost" or "error" associated with the model's performance [35]. This process enables the model to learn effectively from the data. Selecting an appropriate loss function is crucial, as it enhances the model's ability to generalize to unseen data [36].

As the loss function for the binary classification problem, Binary Cross-Entropy (BCE) was used. By penalizing predictions that deviate from the actual labels, BCE helps the model make more accurate predictions. For instance, when the actual label is 1 (positive class) but the model predicts a probability closer to 0 (negative class), BCE increases, prompting the model to adjust its parameters to improve predictions. Conversely, as the model's predicted probabilities close to the actual labels, the BCE decreases, reflecting better alignment with the target. Minimizing BCE during training enhances the model's ability to distinguish between classes, thereby improving overall performance [37].

To minimize the model's error, optimization is the process of determining the best parameters or hyperparameters. Several optimization techniques are commonly used in ML, including grid search for tuning hyperparameters and gradient-based optimizers for deep learning models [38]. Gradient-based optimizers leverage the gradient of the loss function with respect to the model's parameters to iteratively update and minimize the loss [39]. After optimization and validation, the model can be deployed in real-world applications, where it continues to learn from new data, further improving its ability to detect patterns and correlations [40].

Precision, recall, accuracy, and F1-score are used to report the model's evaluation. Precision is the ratio of true positive to the sum of true positive and false positive. Precision measures the accuracy of the positive predictions made by the model. It indicates the proportion of the predicted positive cases that were actually positive [41]. Precision is calculated using Equation 1 [42].

Equation 1. $$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

High precision shows a low false positive rate. In the context of amputation prediction, high precision means that the model is likely to be correct when it predicts an amputation [43].

Recall, which is the ratio of true positive predictions to the actual positives, measures how well the model identifies all relevant cases within a dataset [44]. Recall is calculated using Equation 2 [42].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

Equation 2. $$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

High Recall indicates that the model identifies most of the positive cases. For amputation prediction, this means that the model detects most actual amputation cases [45].

In the dataset, amputation cases represented a minority of the total observations (almost 25 percent), introducing a class imbalance issue that could negatively affect model performance, particularly Recall. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE is a method used to address class imbalance by generating synthetic examples of the minority class, such as amputation cases in this study. Instead of simply duplicating existing data, SMOTE creates new, realistic samples by interpolating between existing minority class instances and their nearest neighbors. This technique enables the model to better learn the characteristics of underrepresented cases, thereby improving performance metrics such as Recall and F1-score. SMOTE is typically applied only to training data to prevent data leakage and is especially useful in classification problems where rare but critical outcomes need to be accurately detected.

Accuracy is the ratio of correct predictions after the test to the total observations. Accuracy is calculated using Equation 3 [42].

Equation 3. $$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

The F1 score is a metric used in classification tasks to evaluate a model's accuracy by balancing both precision and recall. It is particularly useful for datasets with imbalanced context. The F1 score provides a single measure that considers both precision and recall which ranges from 0 to 1. The best possible value is 1, which indicates perfect precision and recall [46]. F1 is calculated using Equation 4 [42].

Equation 4. $$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

4. PROGRAMMING LANGUAGE, PYTHON

Programming languages used in ML provide essential tools and libraries for managing data, training models, evaluating performance, and implementing solutions. Among them, Python stands out as a widely popular choice due to its simplicity, readability, versatility, and extensive libraries [47]. Python's rich ecosystem, including libraries like Pandas and NumPy, supports diverse aspects of data analysis and ML, such as processing large construction datasets for data cleaning, filtering, and feature engineering. These capabilities are particularly valuable in construction safety, where Python enables analysis, visualization, and ML applications to predict amputation risks [48]. For instance, ML models like Logistic Regression and Artificial Neural Networks (ANNs) can be employed to evaluate factors such as machinery, tools, and environmental conditions. Additionally, visualization tools such as the confusion matrix were used in the model to help present the outcomes of predictions. A confusion matrix organizes predictions into a table displaying true positives, true negatives, false positives, and false negatives [49].

To address the fourth sub-objective, the study tackled research limitations commonly encountered in ML, such as missing data, biased reporting, noisy data, and overfitting. Managing these defects involved removing duplicates, handling missing values, scaling features, and encoding categorical variables [50]. Missing data happens when some observations lack values for certain features. Solutions include removing rows or columns with missing values—provided the missing data is minimal—or imputing values using the mean, median, mode for numerical data, or the most frequent value for categorical data [51]. This study, however, used a placeholder value (9999) to indicate the extent of missing data. Using 9999 as a placeholder for missing data preserves the dataset's structure and facilitates the identification of gaps. However, it is essential to preprocess the data to ensure algorithms interpret 9999 as a missing value, thereby preventing inaccuracies in analyses or models. Additionally, underreporting or biased reporting of workplace accidents creates incomplete datasets, which limits the model's ability to accurately predict future accidents. Preprocessing techniques such as categorization and oversampling were applied to mitigate these issues, though further refinement and adaptive approaches could enhance prediction accuracy and reliability. Since ML research often deals with imbalanced data, oversampling is an effective method to improve prediction, particularly for metrics such as recall and F1-score, which are crucial when addressing rare but severe outcomes, like amputations. To address class imbalance in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied. This

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

enabled the model to better recognize patterns associated with rare amputation cases, improving evaluation metrics such as recall and F1-score without significantly compromising precision. [52].

Noisy data in ML refers to irrelevant or erroneous information that obscures patterns, often caused by irrelevant features, data entry errors, or random variability. It reduces accuracy, may lead to overfitting, and increases complexity. Fixing noisy data in machine learning (ML) involves cleaning the dataset to remove errors and duplicates, as well as selecting meaningful features to eliminate irrelevant ones [53].

A common issue in ML, which decreases accuracy, is overfitting. Overfitting often arises from factors such as the use of complex models, including deep neural networks, which can detect and memorize insignificant details, patterns, or fluctuations that fail to represent meaningful or generalizable trends. Another common cause is insufficient training data, which forces the model to memorize specific data points rather than learning overarching patterns [54]. To address overfitting, cross-validation is commonly applied. This machine learning technique evaluates and enhances model performance and generalizability by partitioning the dataset into multiple training and validation subsets [55]. Among these techniques, k-fold cross-validation is widely used, where the dataset is divided into k subsets, with the model trained on k-1 subsets and tested on the remaining one. This process is repeated k times to ensure robust model assessment [56].

These limitations reduce the accuracy, generalization, and reliability of ML models in predicting amputations. They also affect interpretability, complicating the application of findings by OSHA and safety professionals.

5. RESULT

Sources and Events in the OSHA database are identified by numerical values, and every Source or Event value is tied specifically to a Source or an Event. There are 625 sources and 165 events in the OSHA database. Figure 1 presents the top 20 identified Sources listed as causing amputations. The difference in counts for the remaining Sources changes minimally, so they are not shown. Table 2 identifies the top 5 Source values represent that resulted in amputation and they are machinery, non-classifiable, forklift, woodworking material, and table saws. Non-classifiable refers to entries where the information is either unclear, missing, or does not align with any established category. These cases are typically assigned the label 9999, and the dataset contains 590 such entries.

Figure 2 illustrates the top 20 identified Categorized Source (CS) values associated with amputations, while Table 3 highlights what the top 5 CS values represent. The data indicates that the top 5 CS causes of amputations are saws, machinery, conveyors, punch presses, and industrial vehicles.

Figure 3 shows the top 20 identified Events leading to amputations, and Table 4 shows what the top 5 Events represent. Table 4 shows that the top 5 Event values that resulted in amputation are caught in running equipment or machinery during regular operation, compressed or pinched by shifting objects or equipment, caught in running equipment or machinery during maintenance, caught in or compressed by equipment or objects, and struck against moving part of machinery or equipment.

Moreover, Figure 4 presents the top 20 identified CE causing amputations, while Table 5 focuses on presenting the top 5 CE values that resulted in amputation. They reveal that the five most frequent CE associated with amputations are the caught in category, compressed or pinched category, struck against (moving object or equipment) category, struck by (falling or dropped) category, and injured by slipping or swinging category.

To show which features have the greatest impact on predictions, the P-value was used. If P-value < 0.05, the feature category is significantly related to amputations. The top 5 Sources that have the greatest impact on predictions are machinery, forklifts, non-classifiable, metal, and table saws. The top 5 CSs that have the greatest impact on predictions are the table saw category, the machinery category, the conveyors category, the punch, press category, and the forklift category. Moreover, the top 5 Events that have the greatest impact on predictions are caught in running equipment or machinery during regular operation, caught in running equipment or machinery during maintenance, compressed or pinched by shifting objects or equipment, struck against a moving part of machinery or equipment, and caught in or compressed by equipment or objects. In addition to Events, the top 5 CEs that have the greatest impact on predictions are the caught in category, the caught in or compressed by equipment or objects category, the struck against category, Struck by falling object or equipment category, and injured by slipping or swinging objects category.

Eight logistic regression models, models 1 to 8 (Table 6), were evaluated using Source, CS, Event, and CE, and a combination of those features. These algorithms used the data to predict the probability of amputation. The prediction capability ranged from about 70% to 74%. Eight more logistic regression models, models 9 to 16, were developed using the data after converting the data into dummy variables. All models 9 to 16 were similar to models 1 to 8, and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

the only difference was the use of dummy variables. The amputation prediction capability after using dummy variables ranged from about 82% to 87%. Table 6 illustrates the accuracy of all tested models, 1 through 16. The model with the highest accuracy is model 13, which utilized dummy variables for Source and Event features and achieved the highest accuracy (87.2 %).

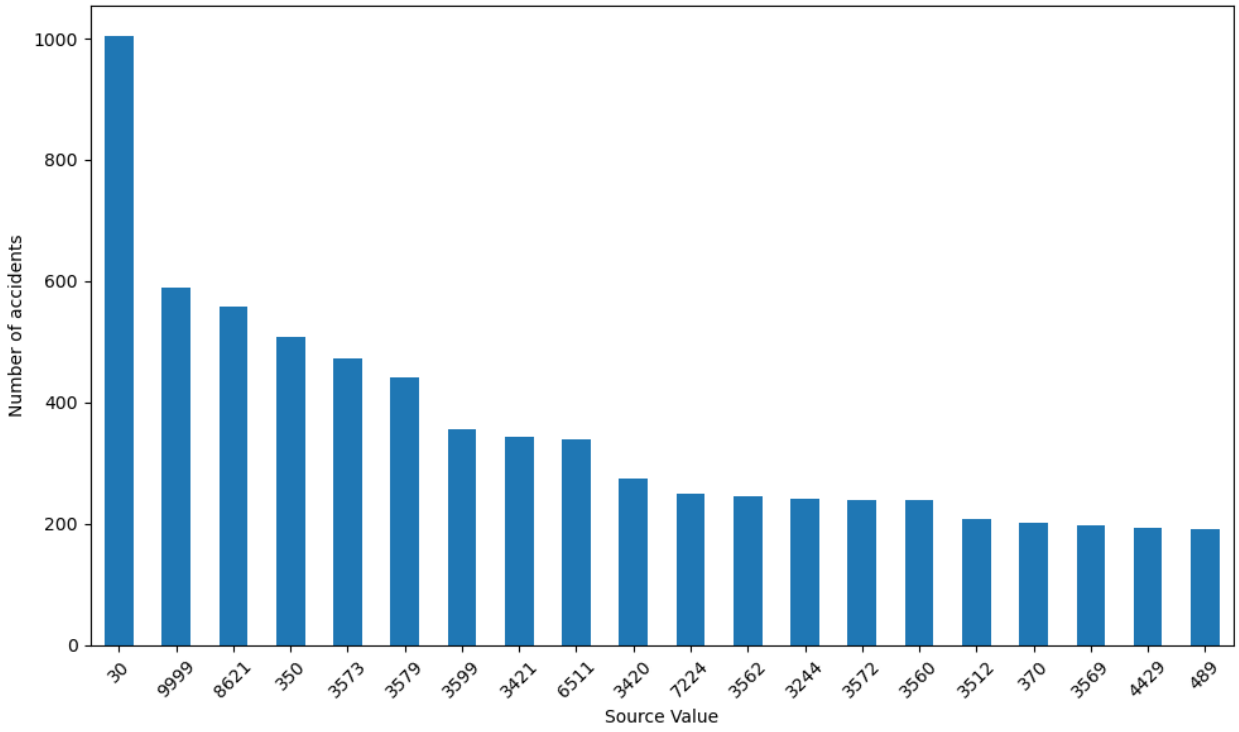

Fig. 1. Top 20 values in the 'Source' category that resulted in amputation. The source number is a numerical identifier of the actual source.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

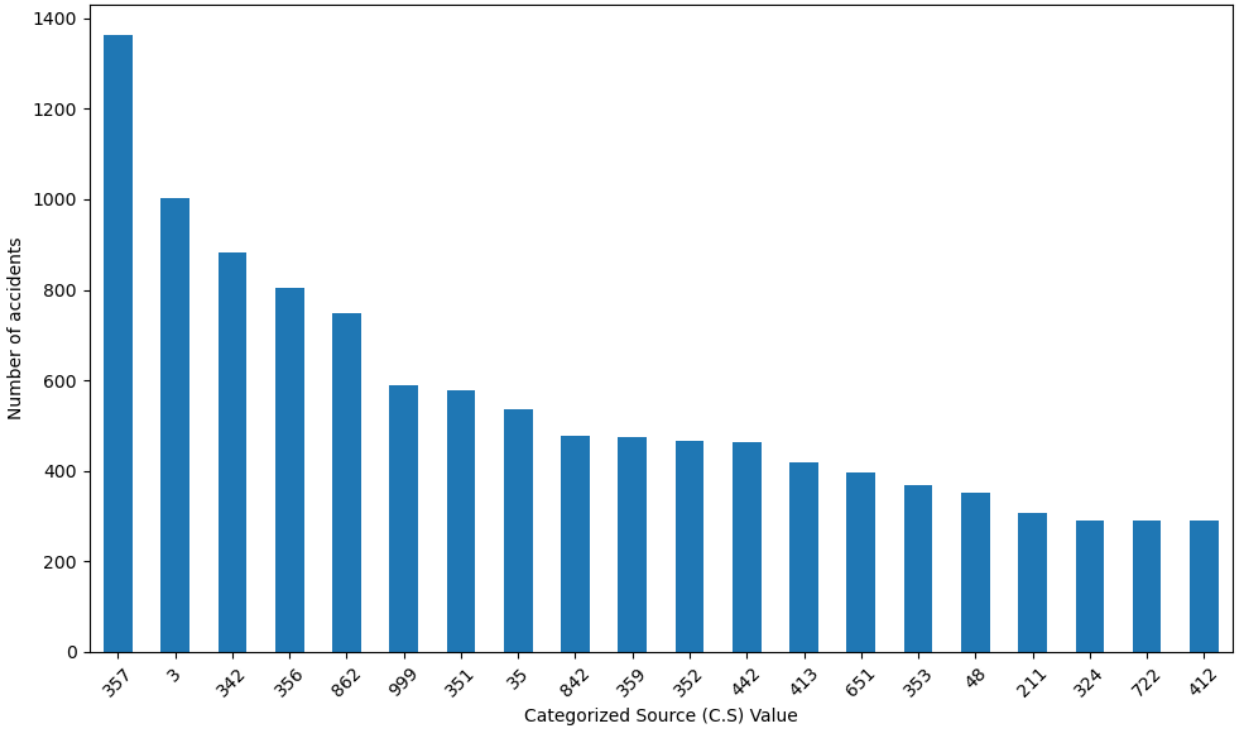

Fig. 2. Top 20 values in the 'Categorized Source (CS)' category that resulted in amputation. The CS number is a numerical identifier of the actual CS.

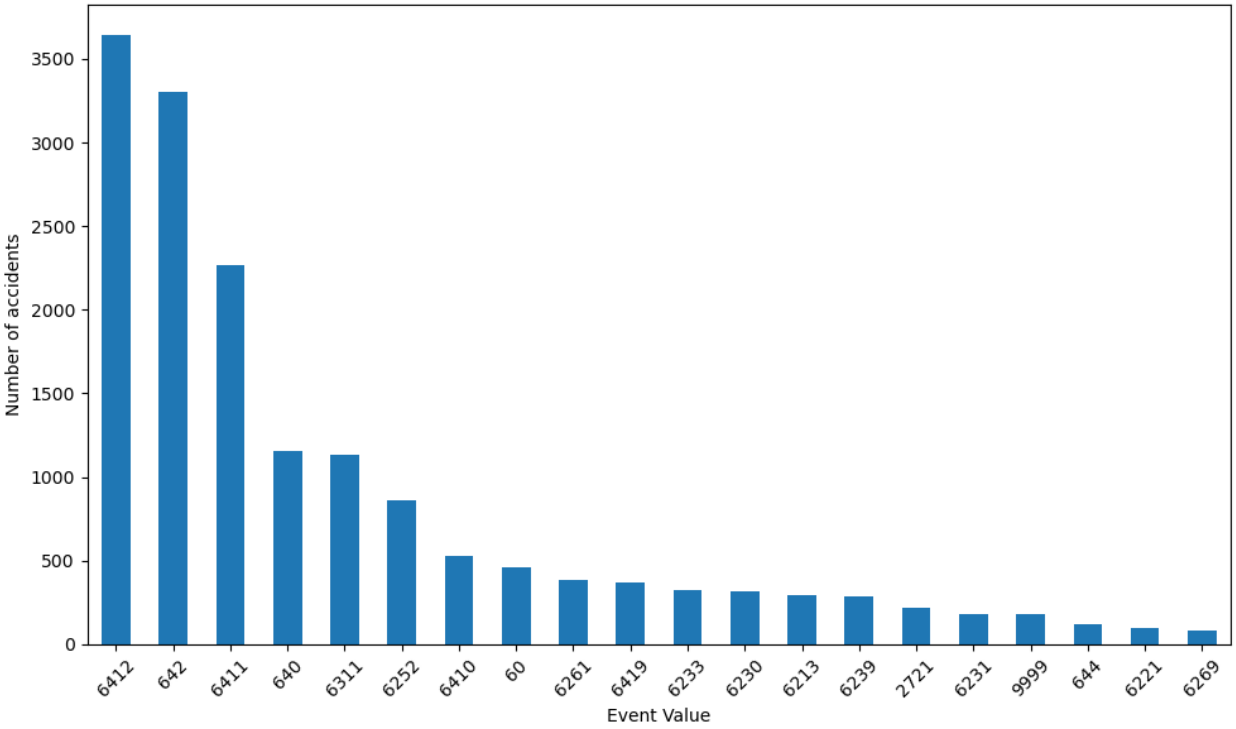

Fig. 3. Top 20 values in the 'Event' category resulted in amputation. The Event number is a numerical identifier of the actual Event.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

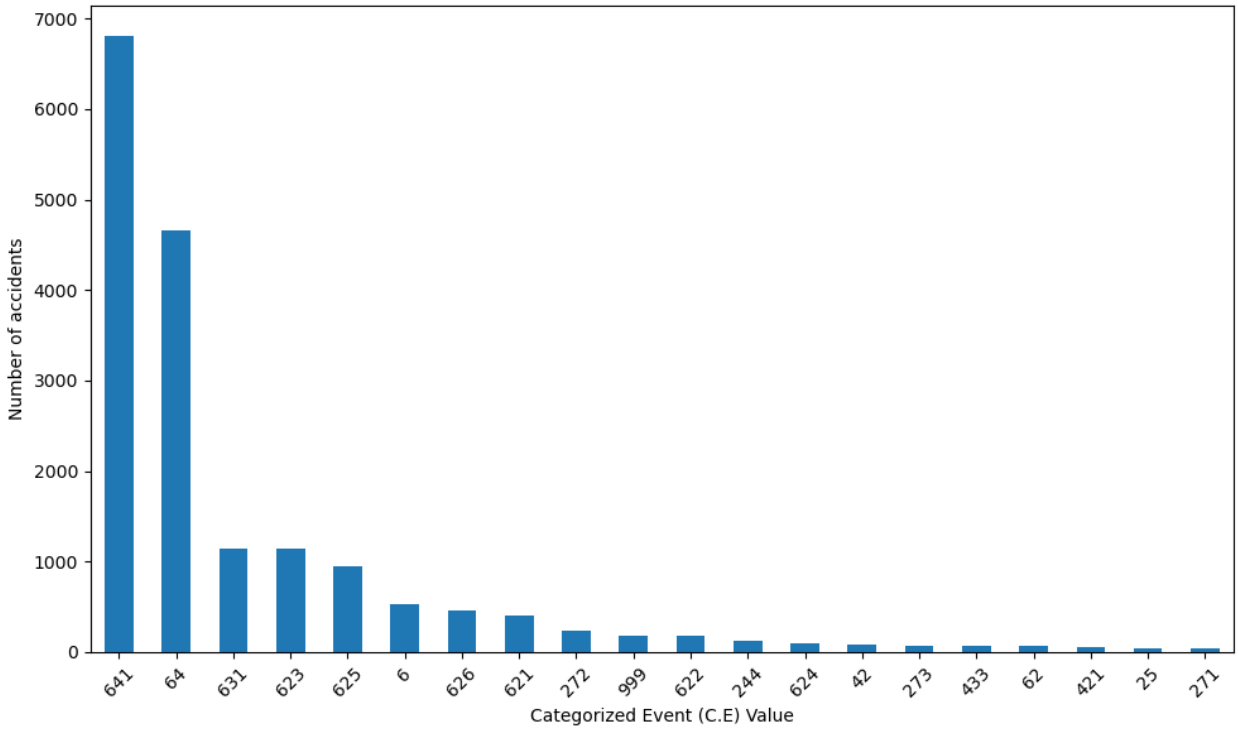

Fig. 4. Top 20 values in the 'Categorized Event (CE)' category that resulted in amputation. The CE number is a numerical identifier of the actual CE.

Table 2. Top 5 Sources identified by OSHA by their source value and associated title that resulted in amputation.

Source Number	Source Title
30	Machinery
9999	Non-classifiable
8621	Forklift, order picker, platform truck-powered
350	Metal, woodworking, and special material
3573	Table saws

9999 means non-classifiable, which refers to entries where the information is either unclear, missing, or does not align with any established category.

Table 3. Top 5 Categorized Sources (CS) identified by OSHA by their CS value and associated title that resulted in amputation.

categorized sources (CS) Number	categorized sources (CS) Title
357	Saws category
3	Machinery
342	Conveyors category
356	Punch, press category
862	Industrial vehicle, material hauling, and transport-powered

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

Table 4. Top 5 Events identified by OSHA by their Event value and associated title that resulted in amputation.

Event Number	Event Title
6412	Caught in running equipment or machinery during regular operation
642	Compressed or pinched by shifting objects or equipment
6411	Caught in running equipment or machinery during maintenance, cleaning
640	Caught in or compressed by equipment or objects, unspecified
6311	Struck against moving part of machinery or equipment

Table 5. Top 5 Categorized Events (CE) identified by OSHA by their CE value and associated title that resulted in amputation.

Categorized Event (CE) number	Categorized Event (CE) title
641	Caught in category
6	Compressed or pinched category
631	Struck against (moving object or equipment) category
623	Struck by (falling or dropped) category
625	Injured by slipping or swinging category

Table 6. Details of Feature Engineering, Dummy Encoding, Accuracy Metrics, and Statistical Evaluation

Model number	Features used	Dummy variables	Accuracy (%)
Original data			
1	Source	No	74.0
2	Categorized Source (CS)	No	73.7
3	Event	No	73.4
4	Categorized Event (CE)	No	73.3
5	Source and Event	No	68.8
6	Categorized Source (CS) and Event	No	73.6
7	Source and Categorized Event (C.E)	No	73.6
8	Categorized Source (CS) and Categorized Event (CE)	No	73.6
t-statistic	1.43		
p-value	0.19		
Standard Deviation	1.88		
Dummy variable tests			
9	Source	Yes	83.1
10	Categorized Source (CS)	Yes	82.6
11	Event	Yes	86.8
12	Categorized Event (CE)	Yes	86.5
13	Source and Event	Yes	87.2
14	Categorized Source (CS) and Event	Yes	87.0
15	Source and Categorized Event (CE)	Yes	86.8
16	Categorized Source (CS) and Categorized Event (CE)	Yes	86.9
t-statistics	1.30		
p-value	0.24		
Standard Deviation	1.88		

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

The t-statistic, p-value, and standard deviation were calculated to assess whether models 1 to 8 differ significantly. The t-statistic assesses whether the mean of a sample significantly differs from a known value or another sample, making it a key tool in hypothesis testing [57]. The p-value quantifies the likelihood that observed results occurred by chance; a small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, warranting its rejection [58]. Standard deviation measures the dispersion of data points relative to the mean, reflecting variability within a dataset, smaller values signify closely clustered data, while larger values indicate greater spread [59]. These statistical concepts enable analysts to evaluate significance, interpret patterns, and confidently make data-driven decisions. For models 1 to 8, the t-statistic is approximately 1.43, with a p-value of 0.19 and a standard deviation of 1.73. Since p-value > α (0.05), the differences among models 1 to 8 are not statistically significant, and they all have the same predictability.

Similarly, the t-statistic, p-value, and standard deviation were evaluated for models 9 to 16. The t-statistic for this group is approximately 1.30, with a p-value of 0.24 and a standard deviation of 1.88. As with models 1 to 8, the p-value > α (0.05), indicating that the differences among models 9 to 16 are also not statistically significant, and they all have the same predictability.

When comparing models 1 to 8 with models 9 to 16, the t-statistic is approximately -14.36, and the p-value is extremely small, approaching 0.000, and the standard deviation is 1.79. Since the p-value is less than $\alpha = 0.05$, the difference between models 1 to 8 and models 9 to 16 is statistically significant.

Table 7 reports all the various prediction models considered for predicting amputations, which shows that models using dummy variables generally outperformed those without, and categorizing features did not significantly improve the models in the research.

Table 7. Models' Accuracy, Precision, Recall, and F1 Score report

Model Number	Accuracy (%)	Precision	Recall	F1 Score
Original data				
1	74.0	0.55	0.74	0.63
2	73.7	0.54	0.74	0.63
3	73.4	0.60	0.73	0.63
4	73.3	0.55	0.73	0.63
5	68.8	0.52	0.69	0.64
6	73.6	0.68	0.74	0.65
7	73.6	0.68	0.74	0.65
8	73.6	0.68	0.74	0.65
t-statistic	1.43	0	21.5	11.06
p-value	0.19	1	0	0
Standard Deviation	1.88	0.07	0.02	0.01
Dummy variable tests				
9	83.1	0.82	0.83	0.83
10	82.6	0.82	0.83	0.82
11	86.8	0.88	0.87	0.87
12	86.5	0.87	0.86	0.87
13	87.2	0.88	0.87	0.87
14	87.0	0.88	0.84	0.87
15	86.8	0.87	0.87	0.87
16	86.9	0.88	0.87	0.87
t-statistic	1.30	27.92	38.95	34.85
p-value	0.24	0	0	0
Standard Deviation	1.88	0.03	0.02	0.02

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

To choose the most appropriate and accurate model, all models with dummy variables were evaluated. Models 9 to 16 had high F1 Scores, high accuracy, precision, and recall. While model 13 achieved the highest accuracy of 87.2%, model 11, with an accuracy of 86.8%, was chosen as the best model because of several important factors that make it easier for practical application in real-world scenarios. The first reason is its simplicity and convenience in practical application. Model 11 uses the Event feature, which simplifies both the analysis and prediction processes. In construction, where safety personnel and OSHA officials need to make quick decisions, having a model with a transparent and interpretable feature is crucial. The Event feature is straightforward, enabling practitioners to quickly understand and apply the model's results, significantly enhancing its utility in day-to-day operations. The second reason is its stability and robustness. Model 11 demonstrated excellent stability across key performance metrics like precision, recall, and F1 score, which makes the model a dependable tool for amputation prediction by ensuring it performs well across various data subsets and under various conditions. The third reason is computational efficiency. By using a single feature, Model 11 minimizes the complexity of the model compared to those that use multiple features. This complexity reduction makes computation times faster, which is especially important when the model needs to be deployed in environments where computational resources may be limited. Additionally, by focusing on a single feature (Event), the model is less likely to overfit the data, which contributes to its long-term reliability and robustness when applied to new, unseen data.

6. CONCLUSIONS

This research aimed to evaluate the effectiveness of a Machine Learning (ML) system in predicting workplace-related amputations across various construction sectors, using categorized data on event types and sources. The research investigated the application of ML techniques, especially logistic regression and dummy variables. Federal OSHA data collected between January 2015 and December 2022 were used for the study. Multiple predictive models were developed by analyzing 79,000 records and utilizing various features such as Source, Categorized Source (CS), Event, and Categorized Event (CE). Figures 1 to 4 and Tables 2 to 5 show the most common Source, CS, Event and CE respectfully.

Addressing data defects, missing values, feature selection, encoding categorical data, and balancing the dataset using dummy variables were used to improve the overall model performance.

The accuracy of the models varied depending on the features used and the usage of variables. The logistic regression models using original data without dummy variables (Models 1 to 8) achieved accuracy between 68.8% and 74%. To compare these models, the p-value > α (0.05) showed that the differences among models 1 to 8 are not statistically significant, and they all have the same predictability. The introduction of dummy variables to models 9 to 16 significantly enhanced model performance, resulting in accuracy ranging from 82.6% to 87.2%. As with models 1 to 8, the p-value > α (0.05), indicated that the differences among models 9 to 16 are also not statistically significant, and they all have the same predictability.

The highest accuracy of 87.2% was for model 13, which was achieved using a combination of Source and Event with the usage of dummy variables. Since models 9 to 16's accuracies are not significantly different, model number 11, which used the Event as the feature with an accuracy of 86.8%, precision of 0.88, recall of 0.87, and F1 Score of 0.87, was the most appropriate model. It could be appropriate since the accuracy is high, and there is just one feature to use, which makes it easier and faster.

This study focused on logistic regression due to its simplicity, efficiency, and suitability for binary classification problems such as predicting the occurrence of amputations. Logistic regression is a well-established model that performs well when the relationship between input features and the target variable is approximately linear. It is fast to train, requires fewer computational resources, and is less prone to overfitting when regularized properly. However, its main limitation is that it may not capture complex, non-linear relationships within the data, which can result in reduced predictive performance in certain scenarios. Although this model offers practical advantages, we acknowledge that comparing its performance to other machine learning models would provide a more comprehensive view of its strengths and limitations. This comparison is recommended for future work to validate the robustness of logistic regression in handling imbalanced and complex datasets.

The models developed in this study can assist safety professionals and OSHA experts in predicting instances of amputations resulting from workplace accidents in various construction industries based on Categorized Event (CE) and Source data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

REFERENCE

1. Rosenman, K.D., OSHA, well past its infancy, but still learning how to count injuries and illnesses. American journal of industrial medicine, 2016. **59**(8): p. 595-599.
2. Rosner, D. and G. Markowitz, A short history of occupational safety and health in the United States. American journal of public health, 2020. **110**(5): p. 622-628.
3. Felsen, M., Addressing Worker Safety and Health Through the Lens of Strategic Enforcement—Part One. New solutions: a journal of environmental and occupational health policy, 2024. **34**(2): p. 133-146.
4. Nygren, M., et al., Safety and multi-employer worksites in high-risk industries: an overview. Relations industrielles, 2017. **72**(2): p. 223-245.
5. Grueskin, C., At Least as Effective: OSHA, the State Plans, and Divergent Worker Protections from COVID-19. Yale J. Health Pol'y L. & Ethics, 2022. **21**: p. 228.
6. Bennet, E.D. and T.D. Parkin, Fifteen risk factors associated with sudden death in Thoroughbred racehorses in North America (2009–2021). Journal of the American Veterinary Medical Association, 2022. **260**(15): p. 1956-1962.
7. Villanueva, G., et al., Self-efficacy, disability level and physical strength in the injured workers: findings from a major factory disaster in Bangladesh. Disability and rehabilitation, 2017. **39**(7): p. 677-683.
8. Kurani, A., et al., A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science, 2023. **10**(1): p. 183-208.
9. Tien, J.M., Internet of things, real-time decision making, and artificial intelligence. Annals of Data Science, 2017. **4**: p. 149-178.
10. Shayboun, M., Toward Accident Prevention Through Machine Learning Analysis of Accident Reports. 2022, Universidade Tecnica de Lisboa (Portugal).
11. Nithya, B. and V. Ilango. Predictive analytics in health care using machine learning tools and techniques. in 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). 2017. IEEE.
12. Vallmuur, K., Machine learning approaches to analysing textual injury surveillance data: a systematic review. Accident Analysis & Prevention, 2015. **79**: p. 41-49.
13. Tabatabaee, S., et al., Investigating the barriers to applying the internet-of-things-based technologies to construction site safety management. International journal of environmental research and public health, 2022. **19**(2): p. 868.
14. Zhu, J., et al., Developing predictive models of construction fatality characteristics using machine learning. Safety science, 2023. **164**: p. 106149.
15. Chokor, A., et al., Analyzing Arizona OSHA injury reports using unsupervised machine learning. Procedia engineering, 2016. **145**: p. 1588-1593.
16. Gogtay, N.J. and U.M. Thatte, Principles of correlation analysis. Journal of the Association of Physicians of India, 2017. **65**(3): p. 78-81.
17. Mukaka, M.M., A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal, 2012. **24**(3): p. 69-71.
18. Di Leo, G. and F. Sardanelli, Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. European radiology experimental, 2020. **4**: p. 1-8.
19. Alzubi, J., A. Nayyar, and A. Kumar. Machine learning from theory to algorithms: an overview. in Journal of physics: conference series. 2018. IOP Publishing.
20. Balal, A.T., et al., Forecasting solar power generation utilizing machine learning models in Lubbock. 2023.
21. Rodriguez-Galiano, V., et al., Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore geology reviews, 2015. **71**: p. 804-818.
22. Ashqar, H.I., et al. Impact of risk factors on work zone crashes using logistic models and Random Forest. in 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). 2021. IEEE.
23. Chong, M.M., A. Abraham, and M. Paprzycki, Traffic accident analysis using decision trees and neural networks. arXiv preprint cs/0405050, 2004.
24. Umer, M., et al., Comparison analysis of tree based and ensembled regression algorithms for traffic accident severity prediction. arXiv preprint arXiv:2010.14921, 2020.
25. Almeida, J.S., Predictive non-linear modeling of complex data by artificial neural networks. Current opinion in biotechnology, 2002. **13**(1): p. 72-76.
26. Chio, C. and D. Freeman, Machine learning and security: Protecting systems with data and algorithms. 2018: " O'Reilly Media, Inc.".
27. López, V., et al., An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information sciences, 2013. **250**: p. 113-141.
28. Kotsiantis, S.B., I. Zaharakis, and P. Pintelas, Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 2007. **160**(1): p. 3-24.
29. Khandezamin, Z., M. Naderan, and M.J. Rashti, Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. Journal of Biomedical Informatics, 2020. **111**: p. 103591.
30. Ahmed, A., A. Jalal, and K. Kim, A novel statistical method for scene classification based on multi-object categorization and logistic regression. Sensors, 2020. **20**(14): p. 3871.
31. Alkharusi, H., Categorical variables in regression analysis: A comparison of dummy and effect coding. International Journal of Education, 2012. **4**(2): p. 202.
32. Chen, Y., et al. Evaluating iterative optimization across 1000 datasets. in Proceedings of the 31st ACM SIGPLAN Conference on Programming Language Design and Implementation. 2010.
33. Subramanian, M., K. Shanmugavadivel, and P. Nandhini, On fine-tuning deep learning models using transfer learning and hyper-parameters optimization for disease identification in maize leaves. Neural Computing and Applications, 2022. **34**(16): p. 13951-13968.
34. Rácz, A., D. Bajusz, and K. Héberger, Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. Molecules, 2021. **26**(4): p. 1111.
35. Schorfheide, F., Loss function-based evaluation of DSGE models. Journal of Applied Econometrics, 2000. **15**(6): p. 645-670.
36. Wang, Q., et al., A comprehensive survey of loss functions in machine learning. Annals of Data Science, 2022. **9**(2): p. 187-212.
37. Terven, J., et al., Loss functions and metrics in deep learning. A review. arXiv 2023. arXiv preprint arXiv:2307.02694.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

38. Andonie, R., Hyperparameter optimization in learning systems. Journal of Membrane Computing, 2019. **1**(4): p. 279-291.
39. Daoud, M.S., et al., Gradient-based optimizer (GBO): a review, theory, variants, and applications. Archives of Computational Methods in Engineering, 2023. **30**(4): p. 2431-2449.
40. Brink, H., J. Richards, and M. Fetherolf, Real-world machine learning. 2016: Simon and Schuster.
41. Moons, K.G., et al., Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine, 2015. **162**(1): p. W1-W73.
42. Hand, D.J., P. Christen, and N. Kirielle, F*: an interpretable transformation of the F-measure. Machine Learning, 2021. **110**(3): p. 451-456.
43. Monteiro-Soares, M., et al., Lower-limb amputation following foot ulcers in patients with diabetes: classification systems, external validation and comparative analysis. Diabetes/metabolism research and reviews, 2015. **31**(5): p. 515-529.
44. Saito, T. and M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 2015. **10**(3): p. e0118432.
45. Wang, S., et al., Machine learning for the prediction of minor amputation in University of Texas grade 3 diabetic foot ulcers. Plos one, 2022. **17**(12): p. e0278445.
46. Yacouby, R. and D. Axman. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. in Proceedings of the first workshop on evaluation and comparison of NLP systems. 2020.
47. Miller, B.N., D.L. Ranum, and J. Anderson, Python programming in context. 2019: Jones & Bartlett Learning.
48. Ranjan, M.K., et al., Python: Empowering Data Science Applications and Research. Journal of Operating Systems Development & Trends, 2023. **10**(1): p. 27-33.
49. Amin, F. and M. Mahmoud, Confusion matrix in binary classification problems: A step-by-step tutorial. Journal of Engineering Research, 2022. **6**(5): p. 0-0.
50. García, S., J. Luengo, and F. Herrera, Data preprocessing in data mining. Vol. 72. 2015: Springer.
51. Pigott, T.D., A review of methods for missing data. Educational research and evaluation, 2001. **7**(4): p. 353-383.
52. Bazzoli, A., Analyzing workplace accident underreporting: a systematic review, a Monte Carlo simulation, and a real data application. 2023: Washington State University.
53. Gupta, S. and A. Gupta, Dealing with noise problem in machine learning data-sets: A systematic review. Procedia Computer Science, 2019. **161**: p. 466-474.
54. Bejani, M.M. and M. Ghatee, A systematic review on overfitting control in shallow and deep neural networks. Artificial Intelligence Review, 2021. **54**(8): p. 6391-6438.
55. Ghojogh, B. and M. Crowley, The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. arXiv preprint arXiv:1905.12787, 2019.
56. Nti, I.K., O. Nyarko-Boateng, and J. Aning, Performance of machine learning algorithms with different K values in K-fold cross-validation. International Journal of Information Technology and Computer Science, 2021. **13**(6): p. 61-71.
57. Park, H.M., Hypothesis testing and statistical power of a test. 2015.
58. Greenland, S., et al., Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology, 2016. **31**(4): p. 337-350.
59. Rayat, C.S. and C.S. Rayat, Measures of dispersion. Statistical Methods in Medical Research, 2018: p. 47-60.