EPiC Series in Built Environment

Volume 6, 2025, Pages 856–865

Proceedings of Associated Schools of Construction 61st Annual International Conference

Classification of Fatal and Non-fatal Construction Incidents in the Southeastern U.S. Using Machine Learning

Mayowa Oladele[1], M. Myung Jeong[1], and Younghan Edwin Jung[2]
[1]Georgia Southern University, [2]Stephen F. Austin State University

Construction sites face persistent safety challenges, with incidents often resulting in severe injuries or fatalities. In the U.S., these safety concerns are heightened due to the high volume of construction projects and complex working conditions. This study conducts a data-driven analysis of construction safety incidents in the Southeastern U.S., utilizing five Machine Learning (ML) techniques to classify fatal or non-fatal incidents. A dataset of 1,963 incidents obtained from the OSHA was analyzed with the ML techniques for their prediction accuracy of classifications. Key findings reveal that random forest and decision trees achieved the highest accuracy and reliability in classifying fatal or non-fatal incidents, with random forest outperforming all models in the classifications. Feature importance analysis highlighted factors such as age, height, occupation, and event type as significant predictors of injury severity. The study's implications are substantial for construction safety management; ML models can provide predictive insights that support proactive safety measures on construction sites. By identifying high-risk factors associated with severe injuries, this research contributes to the development of data-driven safety interventions and policy improvements aimed at reducing incident rates. The findings underscore the potential of ML in advancing construction safety through targeted risk assessment and preventive strategies.

Keywords: Construction Safety, Machine Learning, Data-Driven Analysis, Risk Prediction

Introduction

The construction industry often compromises safety management due to cost-cutting pressures, especially in competitive bidding scenarios, unless required by law or client expectations (Ayhan and Tokdemir, 2019). Construction megaprojects pose significant safety risks with their complex indoor and outdoor work environments yet implementing comprehensive prevention strategies remains difficult (Lin et al., 2024). Recent safety advancements emphasize proactive measures, transitioning from reactive approaches based on historical data to using predictive indicators such as safety controls and risk perception to avert incidents (Ghosh et al., 2023). Analyzing accident precursors (e.g., near-misses) enhances early detection and site safety (Hashmi et al., 2024).

This research fills a critical gap by applying Machine Learning (ML) to understand the severity of safety incidents, which is underexplored in the context of Southeastern U.S. construction, specifically considering thirteen U.S. states in the region. They include Alabama, Arkansas, the Carolinas, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, Tennessee, Virginia, and West Virginia. This study employs five ML techniques, Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Trees, and Random Forests, to build predictive models aimed at informing proactive safety strategies. These models are developed using a comprehensive 10-year (2013-2023) safety dataset obtained from the Occupational Safety and Health Administration (OSHA) (OSHA Enforcement Data, n.d.). This study aims to enhance proactive safety management by applying ML to predict safety incidents, answering two important questions: 1) Which ML technique offers the most accurate prediction in classifying between fatal and non-fatal incidents? 2) How can ML models be used to assess the likelihood of fatal and non-fatal incidents?

This research is motivated by the potential of ML to revolutionize construction safety by reducing incident rates through timely and data-driven interventions. Ultimately, the findings will have the potential to inform policy, enhance on-site safety strategies, and reduce the frequency and severity of construction-related accidents, setting a new standard for safety management practices in the construction sector.

Literature Review

The use of ML techniques has been dramatically increasing in construction safety management nowadays (Kaveh, 2024). ML is categorized, depending upon the type of data to be analyzed, into supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised learning, the most common approach in this domain, uses labeled data to predict safety incidents. For instance, Poh et al. (2018) used these models to forecast hazards based on historical incident reports, aiding early risk prevention (Bugalia et al., 2022). Unsupervised learning, which identifies patterns in unlabeled data, has been leveraged to group similar safety hazards (Kaveh, 2024). Bugalia et al. (2022) employed clustering techniques to discern emerging risks. Semi-supervised models combine labeled and unlabeled data and have proven valuable in scenarios where complete data labeling is impractical (Lin et al., 2024). Reinforcement learning, which adapts safety strategies through feedback mechanisms, is effective for developing optimized safety protocols (Kaveh, 2024). By employing these ML techniques, construction firms can enhance risk management, making construction sites safer and more efficient.

In this study, the five ML techniques have shown significant potential in predicting safety incidents. Logistic Regression is valued for simplicity and interpretability in binary classification, aiding decision-making in safety contexts (Kuhle, 2018; Zhu et al., 2021). Decision Trees offer transparent decision pathways but risk overfitting, making them suitable for identifying risk factors in construction safety (Charbuty & Abdulazeez, 2021; Ghosh et al., 2014). Random Forest enhances prediction accuracy by aggregating multiple trees, proving effective in managing complex datasets in safety research (Baykal, 2024; Kim et al., 2023). Support Vector Machines (SVM) excel in high-dimensional spaces, efficiently identifying accident-prone sites (Shetty et al., 2024). Lastly, K-Nearest Neighbors (KNN) provides intuitive pattern recognition, though computationally intensive, helping classify hazardous conditions for proactive safety management (Cheng & Hoang, 2016; Zhu et al., 2021). Together, these models contribute to a nuanced approach to predicting and mitigating construction safety incidents.

Research Method

Data Description

This study uses a dataset from the OSHA on construction safety incidents in the Southeastern U.S. from 2013 to 2023, totaling 1,963 incident cases (OSHA Enforcement Data, n.d.). Each case provides very detailed information (e.g., injury type, injury time, inspection outcomes, etc.), allowing each case to have a total of 24 variables at least. The comprehensive list of the variables (or features as an ML term) is presented in Table 1. Feature engineering expanded the dataset from the original 24 to 28 variables by including time-related information such as time of day, day of week, month, and season (Variable No. 25 to 28 in the table). It should be noted that not all 1,963 incident cases included the 28 variables. For example, the cause-of-fatality variable (Variable No. 22) is only applicable for fatality cases and does not apply to non-fatal cases. Additionally, categorical variables were numerically encoded to ensure compatibility with ML models.

Table 1. Data points for each variable for feature engineering application

Variable No.	Variables	No. of Data	Variable No.	Variables	No. of Data
1	Summary NR	1963	15	Source of Injury	263
2	Reporting ID	1963	16	Event Type	1963
3	Event Date	1963	17	Human Factor	263
4	Event Description	1963	18	Occupation	1963
5	Event Keywords	1963	19	Degree of Injury	1963
6	End Use	1454	20	Tasks Assigned	1961
7	Stories in Building	1035	21	Cause of Injury	1372
8	Height for Non-Building	800	22	Cause of Fatality	241
9	Project Cost	327	23	Distance of Fall	425
10	Project Type	1497	24	Height of Person	454
11	Age	1963	25	Month	1963
12	Sex	1963	26	Day of Week	1963
13	Nature of Injury	263	27	Time of Day	1963
14	Part of Body	263	28	Season	1963

Methodology

The *Jupyter* notebook was used to code the five ML models with the Python programming language. The coding steps for all five models were similar but not precisely the same, as indicated in Figure 1. At first, some essential libraries like *pandas*, *numpy*, and *scikit-learn* were imported in the *Jupyter* notebook for data manipulation, visualization, and model training. The original dataset was then loaded for a feature engineering application following by converting categorical variables into a binary format (also known as a one-hot encoding). The dataset was then arbitrarily split into 80% for model training and 20% for model testing. It is worth noting that SVM, KNN, Decision Trees and Random Forests took a data scaling step to standardize feature scales for better optimization prior to being split, while Logistic regression was run without any standardization treatment. The prediction performance of each ML was evaluated based on four metrics (precision, recall, F1-score, and accuracy), which will be explained in the next section. In addition, a confusion matrix is provided to report the performance of ML models, displaying the number of true positives, true negatives, false positives, and false negatives.

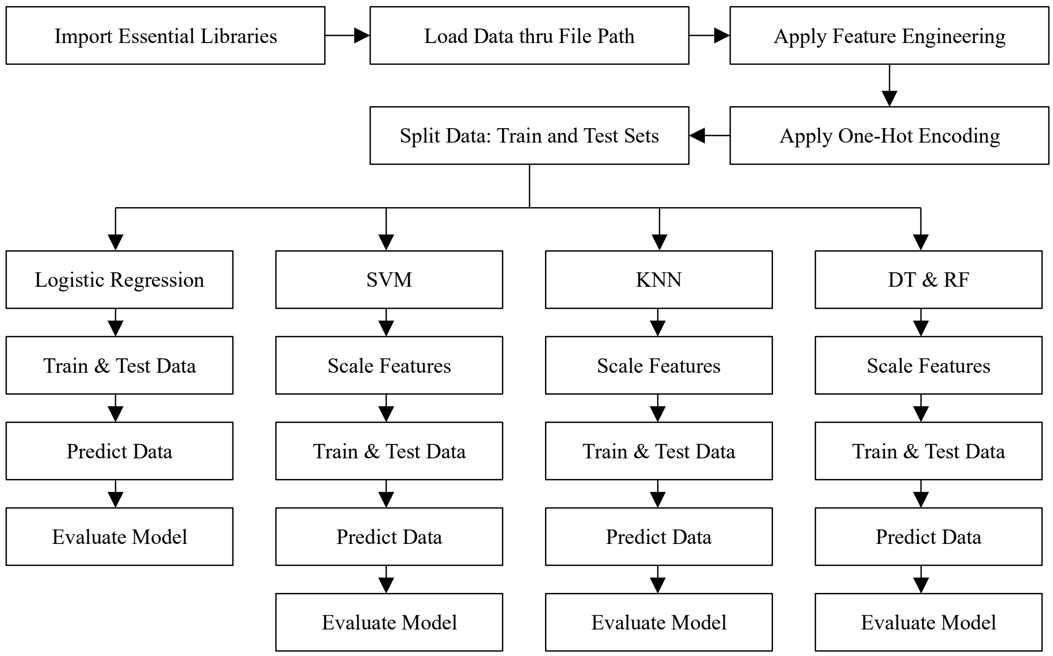


Figure 1. Coding Steps for Machine Learning Techniques

Evaluation Metrics

In evaluating ML models, especially for classification tasks, precision, recall, F1-score, and accuracy are the fundamental metrics. Each of these metrics provides unique insights into the model's performance and can be used to assess its strengths and weaknesses in different aspects. Precision measures the accuracy of positive predictions, representing the proportion of true positive predictions (correctly identified positives) among all positive predictions made by the model. Recall, also known as sensitivity or true positive rate, assesses the model's ability to identify all relevant instances in a dataset. It is the proportion of true positives out of the total actual positives. A high recall indicates that the model captures most of the true positives, which is crucial in applications like fraud detection, where missing a positive case could be risky. The F1-score is the harmonic average of precision and recall, providing a single measure that balances both. It's particularly useful when there is an uneven class distribution, as it penalizes extreme values in either precision or recall. A high F1-score indicates that the model achieves a good balance between precision and recall, making it useful in applications where both false positives and false negatives need to be minimized. Accuracy is the ratio of correctly predicted instances to the total number of instances. While accuracy can be helpful, it may be misleading when dealing with imbalanced datasets. For example, in a dataset where one class vastly outnumbers another, a model can achieve high accuracy simply by predicting the majority class. Each of these metrics are, respectively calculated for each ML model, using the following equations:

- Precision = TP/(TP+FP) where TP = True Positives, and FP = False Positives
- Recall = TP/(TP+FN) where FN = FN= False Negatives
- F1-Score = 2*(Precision*Recall)/(Precision+Recall)
- Accuracy=(TP+TN)/(TP+TN+FP+FN) where TN= True Negatives

Results and Discussion

Comparison of Machine Learning Techniques

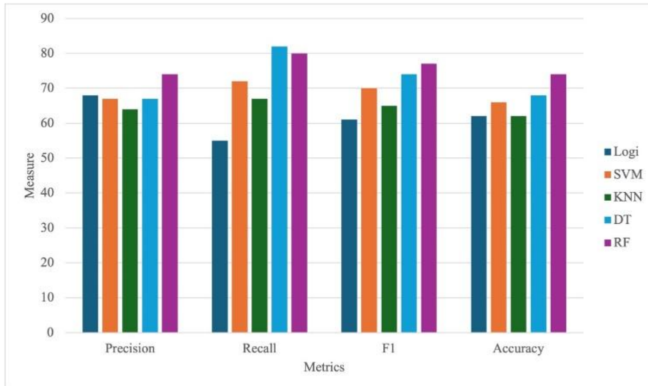
Table 2 summarizes the performance metrics for the five ML techniques used. It is important to recognize that the evaluation scores provided are based exclusively on a subset of the original 1,963 cases, specifically 20% (approximately 393 cases). In other words, these scores do not include any data points used in model training. The analysis revealed varying scores among the techniques. In the table, note that since the accuracy metric mixes both precision and recall components, it accounts for both fatal and non-fatal injury classifications, while the other three metrics (precision, recall, and F1-score) are presented separately for each of fatal and non-fatal classifications. All these scores are also presented for a visual comparison in a bar chart for each of the classifications in Figure 2.

For considering non-fatal classifications only, decision trees and random forest achieved the best scores for the first three metrics (i.e., precision, recall, and F1-score), while Logistic regression and KNN showed the weakest performance for the same metrics. Despite the decision trees having a higher recall rate (0.82), random forest maintained superior precision, successfully predicting 74% of non-fatal incidents. This result reinforces its efficacy in handling the non-fatal case classification task. For a fatal injury classification, random forest recorded the best prediction performance for the first three metrics as the non-fatal classification, establishing its superiority over the other techniques. Logistic regression and KNN delivered subpar results, deeming them unsuitable for this classification.

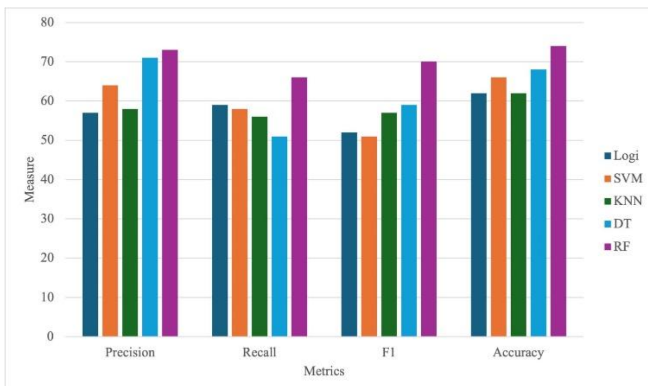
Table 2. Comparison of Prediction of Machine Learning Techniques

ML Type	Degree of Injury	Precision	Recall	F1-Score	Accuracy
Logistic Regression	Non-Fatal	0.68	0.55	0.61	0.62
	Fatal	0.57	0.59	0.52	
SVM	Non-Fatal	0.67	0.72	0.70	0.66
	Fatal	0.64	0.58	0.51	
KNN	Non-Fatal	0.64	0.67	0.65	0.62
	Fatal	0.58	0.56	0.57	
Decision Trees	Non-Fatal	0.67	0.82	0.74	0.68
	Fatal	0.71	0.51	0.59	
Random Forest	Non-Fatal	0.74	0.80	0.77	0.74
	Fatal	0.73	0.66	0.70	

As for the accuracy metric that represents the prediction performance for both non-fatal and fatal cases, Logistic regression and KNN demonstrated the lowest accuracy, indicating their inefficacy in correctly predicting fatality classifications. SVM and decision trees slightly outperformed logistic regression and KNN but lagged behind random forest. Random forest stood out as the top-performing algorithm, boasting a 6% higher accuracy. Overall, random forest proved the most effective, excelling across all evaluation metrics (precision, recall, F1-score, and accuracy). The high accuracy of random forest highlights its reliability in classifying injury severity.



(i)



(ii)

Figure 2. Comparison of different machine learning models for (i) non-fatal classification and (ii) fatal classification

Confusion Matrix

The confusion matrix, a vital tool for evaluating classification models, compares predicted labels with actual outcomes, categorizing results as true positives, true negatives, false positives, or false negatives. The matrix establishes the accuracy of a classification model by presenting the number of accurate and inaccurate predictions for each category. It comprises four categories: true positives, false positives, true negatives, and false negatives. The true positives are the correctly predicted positive class, while the true negatives are the correctly predicted negative class. On the other hand, the false positives are the incorrectly predicted positive for a true negative outcome, and the false negatives are the incorrectly predicted negative for a true positive outcome.

Figure 3 presents the confusion matrix for the most accurate ML technique for this study: random forest. The model correctly identified 170 true positives (non-fatal cases labeled as Class 0 in the matrix) and 119 true negatives (fatal cases labeled as Class 1 in the matrix). However, it misclassified 43 instances as false positives and 61 as false negatives, suggesting that the model has room for improvement.

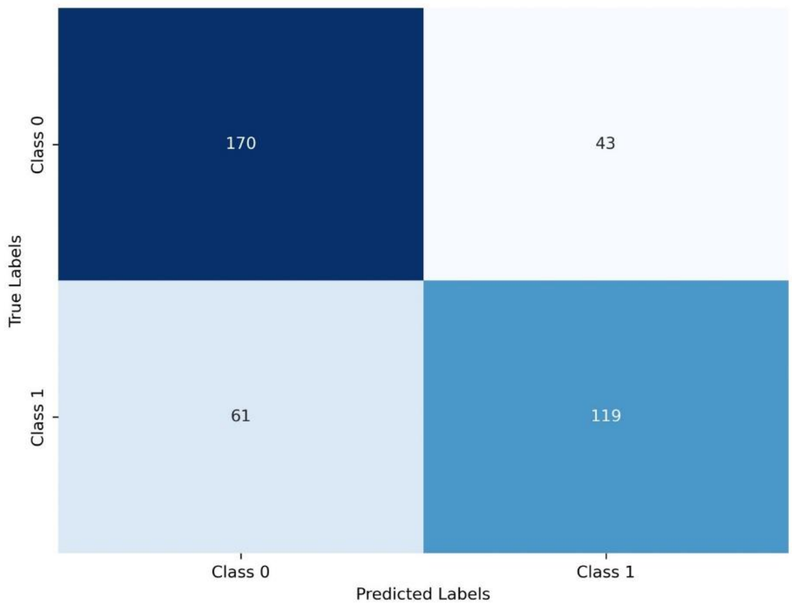


Figure 3. Confusion matrix heatmap for the random forest model

Feature Importance

In ML, feature importance assigns scores to independent variables (or features as a ML term) according to their performance in predicting a target dependent variable. It facilitates the interpretation of models by outlining the contribution of each feature in predicting the target variable. In this study, feature importance was extracted from the random forest model to understand how each factor influences the degree of injury for the fatal and non-fatal classification. The feature importance scores were used to rank the input variables that influence the target variable and enhance the model's accuracy.

Figure 4 displays the top ten influential features that affected the model's performance. The analysis identified age as the most influential predictor (0.085) for determining whether an incident would be fatal or non-fatal, with all other variables having lesser predictive value. Following age, the next most significant features were height for non-building structures (0.04), unreported occupations (0.031), the number of stories in a building (0.028), the individual's height (0.027), fall distance (0.027), time of day in hours (0.016), incidents occurring on Wednesday (0.015), a struck-by event type (0.013), and a fall event type (0.013). These features played a significant role in explaining the current model's accuracy and interpretability among the 28 variables used.

Practicality of ML in Construction Safety and Future Work

The predictions from ML techniques, as shown in Table 2 and Figures 2 and 3, have practical implications for the construction industry, particularly in improving safety protocols and injury prevention strategies. The relatively high precision and recall for non-fatal injuries using random forest (precision: 0.74, recall: 0.80, F1-score: 0.77) suggest that this model can effectively identify scenarios that may lead to non-fatal injuries. Similarly, its performance for fatal injuries (precision: 0.73, recall: 0.66, F1-score: 0.70) indicates its potential for identifying high-risk situations. These insights could help prioritize areas where safety measures need to be strengthened, such as better equipment maintenance or enhanced worker training in high-risk environments.

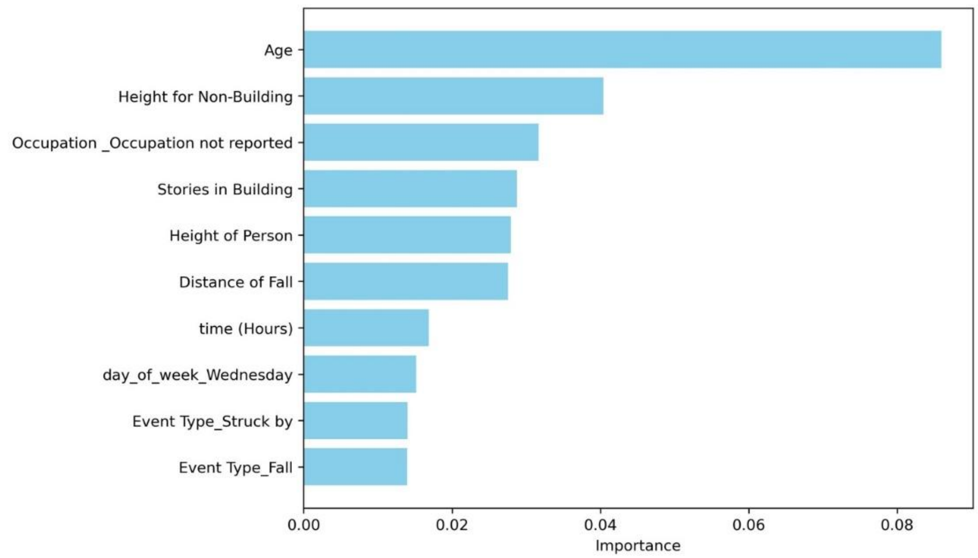


Figure 4. Feature importance for the random forest model

However, the varying performance across models highlights the importance of selecting the most suitable ML technique for specific applications. For instance, decision trees and random forest outperform others in predicting fatal injuries, with random forest offering a better balance between precision and recall. This could guide construction companies in adopting tailored predictive analytics tools to monitor worksite conditions and proactively address hazards. Moreover, integrating these models into real-time monitoring systems could help identify and mitigate risks dynamically, potentially reducing both fatal and non-fatal injuries.

In practical terms, leveraging these predictive models could lead to improved resource allocation for safety interventions, better compliance with regulatory standards, and a reduction in injury-related costs. By adopting ML-driven safety measures, the construction industry can foster a safer working environment, enhancing worker confidence and overall productivity.

Future research could focus on integrating more diverse and comprehensive datasets to improve the predictive accuracy of ML models for construction injury outcomes. Expanding datasets to include variables like environmental conditions, worker fatigue, training levels, and real-time sensor data could provide a more nuanced understanding of injury risks. Additionally, employing advanced feature engineering techniques to identify critical predictors of fatal and non-fatal injuries could enhance the models' performance, particularly in improving recall for fatal injuries, which remains relatively low in some models.

Another consideration could be the exploration of ensemble learning methods and hybrid ML models. Combining the strengths of multiple models, such as random forest and SVM, could yield better predictive performance by leveraging their complementary capabilities. Furthermore, incorporating deep learning techniques, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), may enhance the ability to process complex, time-sequenced, or image-based data, such as video footage from construction sites. These advancements could significantly enhance the practical utility of ML in reducing construction injuries.

Summary and Conclusion

The aim of this study was to evaluate and compare the effectiveness of five ML techniques (Logistic regression, SVM, KNN, decision trees, and random forest) in classifying the severity of construction-related injuries as fatal or non-fatal. Using a dataset of construction safety incidents (a total of 1,963 cases), the prediction performance of each model was analyzed to assess their performance with the classification metrics. The results indicate that random forest was the top-performing technique showing the highest overall accuracy and outperforming the other methods across all evaluation metrics. Random forest accurately classified 74% of non-fatal incidents and 73% of fatal incidents, making it the most effective model for this classification task. Logistic regression and KNN demonstrated the lowest accuracy, suggesting they are less suitable for fatality classification in this context. The feature importance analysis highlighted age as the most influential factor in the model's classification performance.

This study concludes that ML techniques, particularly random forest, can serve as reliable tools for predicting injury severity in construction settings. Random forest's superior performance in both fatal and non-fatal classifications underscores its value for enhancing workplace safety by identifying high-risk factors and tailoring safety measures. Future work could explore incorporating additional features and fine-tuning algorithms to improve predictive accuracy further. Overall, the findings provide a valuable foundation for applying advanced ML in construction safety management, contributing to proactive injury prevention strategies.

References

Al-Aubaidy, N. A., Caldas, C. H., & Mulva, S. P. (2022). Assessment of underreporting factors on construction safety incidents in US construction projects. International Journal of Construction Management, 22(1), 103–120. https://doi.org/ 10.1080/15623599.2019.1613211

Ayhan, B. U., & Tokdemir, O. B. (2019). Predicting the outcome of construction incidents. Safety Science, 113, 91–104. https://doi.org/10.1016/j.ssci.2018.11.001

Baykal, T., Ergezer, F., Eriskin, E., & Terzi, S. (2024). Using ensemble machine learning to estimate international roughness index of asphalt pavements. Iranian Journal of Science and Technology, Transactions of Civil Engineering, 48(4), 2773–2784. https://doi.org/10.1007/s40996-023-01320-6

Bugalia, N., Tarani, V., Kedia, J., & Gadekar, H. (2022). Machine learning-based automated classification of worker-reported safety reports in construction. Journal of Information Technology in Construction, 27. https://doi.org/10.36680/j.itcon.2022.045

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(1), 20–28. https://doi.org/10.38094/jastt20165.

Cheng, M. Y., & Hoang, N. D. (2016). Slope collapse prediction using Bayesian framework with k-nearest neighbor density estimation: Case study in Taiwan. Journal of Computing in Civil Engineering, 30(1). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000456

Occupational Safety and Health Administration (OSHA). (n.d.). Enforcement data. U.S. Department of Labor. https://enforcedata.dol.gov/views/data_summary.php

Ghosh, S., Nourihemadani, M., & Reyes, M. (2024). Effect of Previous Accidents and Near-Miss Incidents on Risk Perceptions of Construction Workers. International Journal of Construction Education and Research, 1–17. https://doi.org/10.1080/15578771.2024.2365206

Hashmi, F., Hassan, M. U., Zubair, M. U., Ahmed, K., Aziz, T., & Choudhry, R. M. (2024). Near-Miss Detection Metrics: An Approach to Enable Sensing Technologies for Proactive Construction Safety Management. Buildings, 14(4), 1005. https://doi.org/10.3390/buildings14041005

Kaveh, A. (2024). Applications of artificial neural networks and machine learning in civil engineering. Studies in Computational Intelligence, 1168.

Kim, H. S., Seong, J., & Jung, H. J. (2023). Real-time struck-by hazards detection system for small- and medium-sized construction sites based on computer vision using far-field surveillance videos. Journal of Computing in Civil Engineering, 37(6). https://doi.org/10.1061/JCCEE5.CPENG-5238

Kuhle, S., Maguire, B., Zhang, H., Hamilton, D., Allen, A. C., Joseph, K. S., & Allen, V. M. (2018). Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: A retrospective cohort study. BMC Pregnancy and Childbirth, 18, 1–9. https://doi.org/10.1186/s12884-018-1971-2

Lin, N. W., Ramirez-Cardenas, A., Wingate, K. C., King, B. S., Scott, K., & Hagan-Haynes, K. (2023). Risk factors for heat-related illness resulting in death or hospitalization in the oil and gas extraction industry. Journal of Occupational and Environmental Hygiene, 21(1), 58–67. https://doi.org/10.1080/15459624.2023.2268142

Love, P. E., Teo, P., Carey, B., Sing, C. P., & Ackermann, F. (2015). The symbiotic nature of safety and quality in construction: Incidents and rework non-conformances. Safety Science, 79, 55–62. https://doi.org/10.1016/j.ssci.2015.05.009

Poh, C. Q., Ubeynarayana, C. U., & Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach. Automation in Construction, 93, 375–386. https://doi.org/10.1016/j.autcon.2018.03.022

Shetty, N. P., Himakar, J., Gnanchandan, P., Prajwal, V., & Jamadagni, S. (2024, March). Enhancing construction site safety: A tripartite analysis of safety violations. In 2024 3rd International Conference for Innovation in Technology (INOCON) (pp. 1–6). IEEE. https://doi.org/10.1109/INOCON60754.2024.10511598.

Zhu, R., Hu, X., Hou, J., & Li, X. (2021). Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Safety and Environmental Protection, 145, 293–302. https://doi.org/10.1016/j.psep.2020.08.006