Journal of Asian Architecture and Building Engineering

ISSN: 1346-7581 (Print) 1347-2852 (Online) Journal homepage: www.tandfonline.com/journals/tabe20

Comparative evaluation of machine learning and LLM-based approaches for construction accident prediction

SeungJu Park, KiHyun Jeong, Jinwoo Kim, HyunSoo Kim, SungHwan Yoon, Jaeyoon Kim & JungHo Jeon

To cite this article: SeungJu Park, KiHyun Jeong, Jinwoo Kim, HyunSoo Kim, SungHwan Yoon, Jaeyoon Kim & JungHo Jeon (28 Jan 2026): Comparative evaluation of machine learning and LLM-based approaches for construction accident prediction, Journal of Asian Architecture and Building Engineering, DOI: 10.1080/13467581.2026.2622236

To link to this article: https://doi.org/10.1080/13467581.2026.2622236

Published online: 28 Jan 2026.

JOURNAL OF ASIAN ARCHITECTURE AND BUILDING ENGINEERING
https://doi.org/10.1080/13467581.2026.2622236

Taylor & Francis
Taylor & Francis Group

CONSTRUCTION MANAGEMENT

OPEN ACCESS Check for updates

Comparative evaluation of machine learning and LLM-based approaches for construction accident prediction

SeungJu Park[a], KiHyun Jeong[b], Jinwoo Kim[c], HyunSoo Kim[d], SungHwan Yoon[e], Jaeyoon Kim[f] and JungHo Jeon[b]

[a]Department of Architecture and Architectural Engineering, Seoul National University, Seoul, South Korea; [b]Department of Architectural Engineering, Pusan National University, Busan, South Korea; [c]Department of Architectural Engineering, Kumoh National Institute of Technology, Gyeongsangbuk-Do, South Korea; [d]Department of Architectural Engineering, Dankook University, Yongin, South Korea; [e]Department of Architecture, Pusan National University, Busan, South Korea; [f]Department of Civil Engineering, The University of Texas at El Paso, El Paso, TX, USA

ABSTRACT
High injury and fatality rates in the construction industry continue to result in significant human and economic losses. To improve construction safety, construction accident prediction has become a critical component of safety management by leveraging machine learning (ML) and deep learning (DL) approaches based on historical datasets. However, they are limited to capturing contextual causality within datasets and performing well in new contexts. A promising alternative is to use large language models (LLMs) that enable data-based causal reasoning and capture contextual pre-incident factors. Unfortunately, the feasibility of using LLMs to predict construction accidents has not been sufficiently investigated, leaving several research gaps (e.g. prompting strategies, performance relative to ML methods, and validation) unresolved. Therefore, this study aims to demonstrate feasibility by developing two prediction models (ML and LLM), comparing their performance, and validating them. Fourteen different ML algorithms were used, and a separate LLM classifier was developed using Chain-of-Thought prompting combined with the Issue, Rule, Application, and Conclusion (IRAC) approach, which is effective in shaping logical reasoning processes. Results showed that XGBoost and LLM achieved accuracies of 97.9% and 94.5%, and recalls of 97.5% and 98.9%, respectively. Also, validation with 1,206 hold-out datasets confirmed temporal robustness for both approaches.

ARTICLE HISTORY
Received 22 August 2025
Accepted 22 January 2026

KEYWORDS
Construction safety; construction accident severity prediction; machine learning; large language models

1. Introduction

Construction industry records a large number of fatalities and injuries every year. According to the recent statistics published by the U.S. Bureau of Labor Statistics (BLS), in 2022, the construction industry recorded about 23% of all fatal injuries within the private sector (BLS 2023). Direct and indirect economic costs resulting from such annual fatalities and injuries are estimated to be $11.5 billion (Waehrer et al. 2007). In South Korea, the construction industry accounted for approximately 53% of fatalities across all industry sectors and 24% of all occupational accidents in 2022, resulting in direct and indirect economic losses of $26.2 billion (MOEL 2023).

In order to prevent workers' accidents on construction sites and improve safety and health records, efforts have been made by developing systematic safety programs and employing advanced technologies, such as virtual reality (VR), computer vision, wearable sensing, and robotics (D. Choi and Koo 2023; Jeon and Cai 2021, 2023; Kim, Lee, and Jeon 2024; Kulinan et al. 2024). For instance, a VR module was developed for construction safety training, and its effectiveness in training performance has been demonstrated (Rokooei et al. 2023). Computer vision approach has also been often adopted to remotely monitor workforce safety risks on construction sites (Kulinan et al. 2024). Among various prior research efforts, construction accident prediction has become a critical component within the construction safety management domain, attributed to its capability of systematically forecasting potential incidents on specific construction sites and the potential to be used as a tool for safety managers to allocate limited resources and design corresponding proactive measures (Andolfo and Sadeghpour 2015; Koc, Ekmekcioğlu, and Gurgun 2023a). The underlying rationale is that predictive models developed using well-established methods and historical datasets allow for quantitatively measuring the probability of potential accidents (Go et al. 2022; Jeong, Chae, and Kang 2023).

CONTACT JungHo Jeon junghojeon@pusan.ac.kr Department of Architectural Engineering, Pusan National University, Busandaehak-Ro 63beon-Gil, Busan 46241, South Korea

Varying methods, such as statistical analysis, machine learning (ML), and deep learning (DL) algorithms, and historical datasets have been utilized to improve the reliability of the prediction models and resulting performance (e.g., accuracy) (Ayhan and Tokdemir 2019; Koc, Ekmekcioğlu, and Gurgun 2022; Tixier et al. 2016; Q. Xu and Xu 2021). Unfortunately, however, despite their proven performance and effectiveness, existing approaches requiring users' expertise in data analytics are challenging to implement in actual construction workplaces and often involve demanding model development tasks (e.g., fine-tuning and model construction) (Akinosho et al. 2020; Y. Xu et al. 2021). In addition, construction accident prediction models built upon ML and DL algorithms, which were used as main methodological components in many previous studies, are difficult to reproduce, directly implement in a new construction environment, and achieve consistent performance due to the dynamic and constantly changing nature of construction jobsites (Gondia, Ezzeldin, and El-Dakhakhni 2022; Koc, Ekmekcioğlu, and Gurgun 2021).

A promising alternative is to adopt a large language model (LLM) that can efficiently generate quality outputs (e.g., text) in response to users' queries (K. Zhang, Kwon, and Xiong 2023). With the recent advancement of deep learning, generative models, and LLM architectures, LLM has started to be widely used for content generation, personalization, and problem-solving (Fui-Hoon Nah et al. 2023; J. Lee and Ahn 2025). At the same time, it has started to be used by researchers to address long-standing challenges in various domains (e.g., psychology, linguistics, and construction) (Hansen and Hebart 2022; Liu, Zhang, et al. 2023; Yoo et al. 2024). Some studies have suggested LLM-centered approaches to predict dynamic trends in the financial market, medical outcomes, and the type of natural disaster propagation (Amacher et al. 2024; Che et al. 2024; Khanmohammadi et al. 2024).

However, despite the unique potential of leveraging LLM in prediction tasks, several critical knowledge gaps remain to be addressed to achieve comprehensive and precise construction accident prediction. First, given the nature of construction datasets, which comprise a mix of structured and unstructured variables, the feasibility of using an LLM for forecasting construction accidents has not been sufficiently investigated. Also, whether the LLM, which allows for causal reasoning and capturing contextual patterns within the data, could offer advantages or disadvantages over existing approaches (e.g., ML and DL) remains unclear. Second, developing effective prompting strategies remains a key challenge because LLM performance is highly sensitive to prompt structure, yet few prior studies have examined which prompting frameworks – such as zero-shot, few-shot, or structured prompting – are suitable for guiding domain-specific LLMs to reason over dynamic construction data for accident prediction. Third, prior work lacks validation on new hold-out accident datasets, making it unclear whether the developed models generalize beyond historical data or remain reliable when applied to newly occurring incidents.

To bridge the aforementioned gaps, this paper aims to examine the feasibility of an LLM-powered construction accident prediction approach and compare the performance with that achieved from well-established ML algorithms that were adopted by a majority of previous studies. The methodology consists of the following five steps. First, a dataset is constructed based on publicly available databases containing historical accident cases that occurred on the jobsite. Second, data points are augmented due to class imbalance issues observed in the dataset. Third, a total of fourteen classifiers forecasting the potential construction accidents in two classes (death and injury) are developed, each of which is separately built upon an individual machine learning algorithm. Fourth, an LLM (i.e., ChatGPT)-centered classifier is developed through prompt engineering based on a theoretical foundation. Lastly, the prediction performance is compared using widely adopted metrics (e.g., accuracy and precision), validation on the new datasets that were not used for training and testing ML and LLM classifiers is conducted, and critical observations are discussed. The findings of this paper are expected not only to promote the potential application of LLM in construction accident prediction and classification but also to ultimately contribute to enhancing safety performance on construction sites.

2. Background and related studies

2.1. Construction accident prediction in safety management

In the construction safety management domain that aims to identify, assess, and control risks in construction environments to prevent workers' injuries and fatalities, construction accident prediction has received a significant amount of attention as it serves as a risk-forecasting foundation, shifting safety management from reactive responses to preventive strategies (e.g., reallocating limited resources to high-risk operations and implementing task-specific controls) (Andolfo and Sadeghpour 2015; Koc, Ekmekcioğlu, and Gurgun 2023a). One of the methodologies used for effective construction accident prediction is statistical techniques (e.g., regression modeling and Bayesian inference) that use varying forms of historical datasets (e.g., national accident data and accident investigation reports) to quantify risk probabilities for specific construction activities, site conditions, etc., before incidents occur (Akboğa Kale and Baradan 2020; J. Choi et al. 2020; K. Kang and Ryu 2019). In particular, Generalized Linear Models (GLMs) have been extensively utilized to handle target variables with non-normal distributions. Also, the Firth's logistic regression has been recognized as an effective alternative to mitigate estimation bias caused by imbalanced data distributions (Toptanci, ErgiNel, and Acar 2023). More recently, advanced statistical approaches, including Hierarchical Generalized Additive Models (HGAM) and Explainable Boosting Machines (EBM), have been proposed, offering new analytical possibilities by automatically detecting complex interactions and visualizing non-linear relationships within high-dimensional construction accident data. However, despite their proven effectiveness and potential, a review of the existing literature has revealed the following challenges and observations.

Firstly, since many studies attempted to minimize input variables for the development of statistical models for avoiding the multicollinearity issues, they often sacrificed latent variables that could have contributed to enhanced construction accident prediction performance (Alin 2010). Also, to mitigate increased computational complexity and the loss of intuitive interpretability, advanced models often impose structural constraints or engage in selective variable identification. Secondly, construction accident data inherently exhibit a highly imbalanced distribution between low- and high-severity accident classes (e.g., injuries vs. fatalities) (J. Choi et al. 2020; K.-S. Kang, Koo, and Ryu 2022; K. Kang and Ryu 2019; Toptanci, ErgiNel, and Acar 2023). To address this issue, research efforts have been made by utilizing data sampling (Koc, Ekmekcioğlu, and Gurgun 2023b) and augmentation techniques (Elamrani Abou Elassad, Mousannif, and Al Moatassime 2020), often combined with machine learning (ML) algorithms (Baker, Hallowell, and Tixier 2020; Koc and Gurgun 2022). Lastly, although the focus of construction accident prediction is on using pre-incident variables (e.g., activities workers engaged with and site conditions) for performing statistical analysis, many post-incident factors (e.g., injured body parts) were used as predictors, representing a limited applicability for proactive safety management (J. Y. Lee et al. 2020; Tixier et al. 2016).

2.2. Machine learning and deep learning approaches for accident prediction

In order to address the limitations of using statistical approaches for forecasting potential construction accidents, ML and DL algorithms have started to be actively adopted. The underlying rationale is that ML and DL approaches allow for capturing latent features in the data to extract risk-relevant patterns, modeling high-dimensional relationships, and improving the resulting prediction performance (J. Choi et al. 2020; K. Kang and Ryu 2019; Mistikoglu et al. 2015).

Among several comparative characteristics (e.g., modeling approach, ability to capture nonlinearity, and data compatibility) of using ML or DL algorithms for accident prediction, the literature shows that DL tends to provide higher predictive performance than ML because DL algorithm's layered architectures enable the effective extraction of latent interactions among data (e.g., activity type, environmental conditions, and situational hazards) and it performs well with unstructured data (Ayhan and Tokdemir 2019; Koc, Ekmekcioğlu, and Gurgun 2022). However, despite the DL's superior performance, it struggles to offer interpretability due to multiple hidden layers, a large number of hyperparameters, and complex transformations, while ML algorithms (e.g., logistic regressions and decision trees) allow for identification of how each feature contributes to the outcome based on transparent mathematical structures (Gondia, Ezzeldin, and El-Dakhakhni 2022; K.-S. Kang, Koo, and Ryu 2022; Koc, Ekmekcioğlu, and Gurgun 2023a). Also, DL demands complex modeling efforts, which require expert intervention, and it tends to focus on accident outcomes rather than contextual precursors.

In contrast, ML algorithms, such as Decision Tree (DT), are most suitable when transparent decision-making is prioritized, as their structure mimics human logic and presents accident causal factors in a visually interpretable form (Mistikoglu et al. 2015). However, DT is prone to capturing noise in accident data, leading to poor generalization. To ensure higher predictive robustness, Random Forest (RF) is often preferred, as its bagging technique – aggregating multiple decorrelated trees – effectively reduces variance and prevents overfitting amidst the high heterogeneity of construction site conditions (J. Choi et al. 2020; K.-S. Kang, Koo, and Ryu 2022).

Although many studies have attempted to utilize ML algorithms for construction accident forecasting, given their unique advantages (e.g., model interpretability, predictive robustness, and mitigation of overfitting), they are limited to holistically comparing the performance of varying algorithms, as individual studies relied on a few algorithms, which restricts the ability to identify insightful relationships between the structure/nature of different algorithms and their corresponding performance. In addition, these approaches are limited to quantifying accident data and identifying underlying patterns, making it challenging to infer contextual causality and resulting in a lack of adaptive response capability to new settings (Requeima et al. 2024). As a result, the introduction of an LLM capable of simultaneously performing data-based causal reasoning has been proposed as an alternative to enhance interpretability and adaptability in novel contexts. More theoretical background on using

LLM for construction accident prediction and comparing the results with those achieved from ML is elaborated in the following subsection.

2.3. Application of large language model in the construction domain

The LLM is a subfield of artificial intelligence that can generate contextually coherent output based on user input (prompt) (K. Zhang, Kwon, and Xiong 2023). LLM supports various modalities (e.g., text-to-text, text-to-image, and text-to-video) and is capable of performing contextual understanding and adaptive reasoning, making it a promising alternative to tackle challenges that were difficult to address with traditional approaches (S.-W. Lee and Choi 2023; Liu, Yang, et al. 2023; Singer et al. 2022; Topal, Bas, and van Heerden 2021).

With the LLM's potential and advantages, it has been widely leveraged in the construction safety domain for automated safety report generation (Nyqvist, Peltokorpi, and Seppänen 2024), multimodal hazard identification (Wang et al. 2024), safety guideline summarization (Smetana et al. 2024), simulation-based risk scenario generation (Samsami 2024), and accident type classification (e.g., falls, struck-by, and slips/trips) (Yoo et al. 2024). Although these studies mainly demonstrated that LLM allows for efficient data summarization, extraction, and information generation, its capability of directly processing textual data (e.g., categorical variables), reasoning, and synthesizing heterogeneous data, in predicting construction accidents, which was difficult to achieve with traditional approaches, has been minimally investigated.

In addition, conventional statistical, ML, and DL techniques often struggle to fully leverage construction datasets that contain a mix of structured and unstructured data points (Samsami 2024). They may overlook latent patterns, fail to infer deeper causal relationships, and inadequately capture the contextual factors essential for improving the performance of construction accident prediction. Nevertheless, studies on demonstrating the LLM's construction accident prediction ability are still in an early stage, leaving significant gaps between prompt engineering strategies, generative mechanisms, and quantitative comparison with existing approaches unexplored.

3. Methodology

This study developed a comprehensive framework, as illustrated in Figure 1, to predict construction accident outcomes as either death or injury, using both ML and LLM approaches. A large dataset of historical

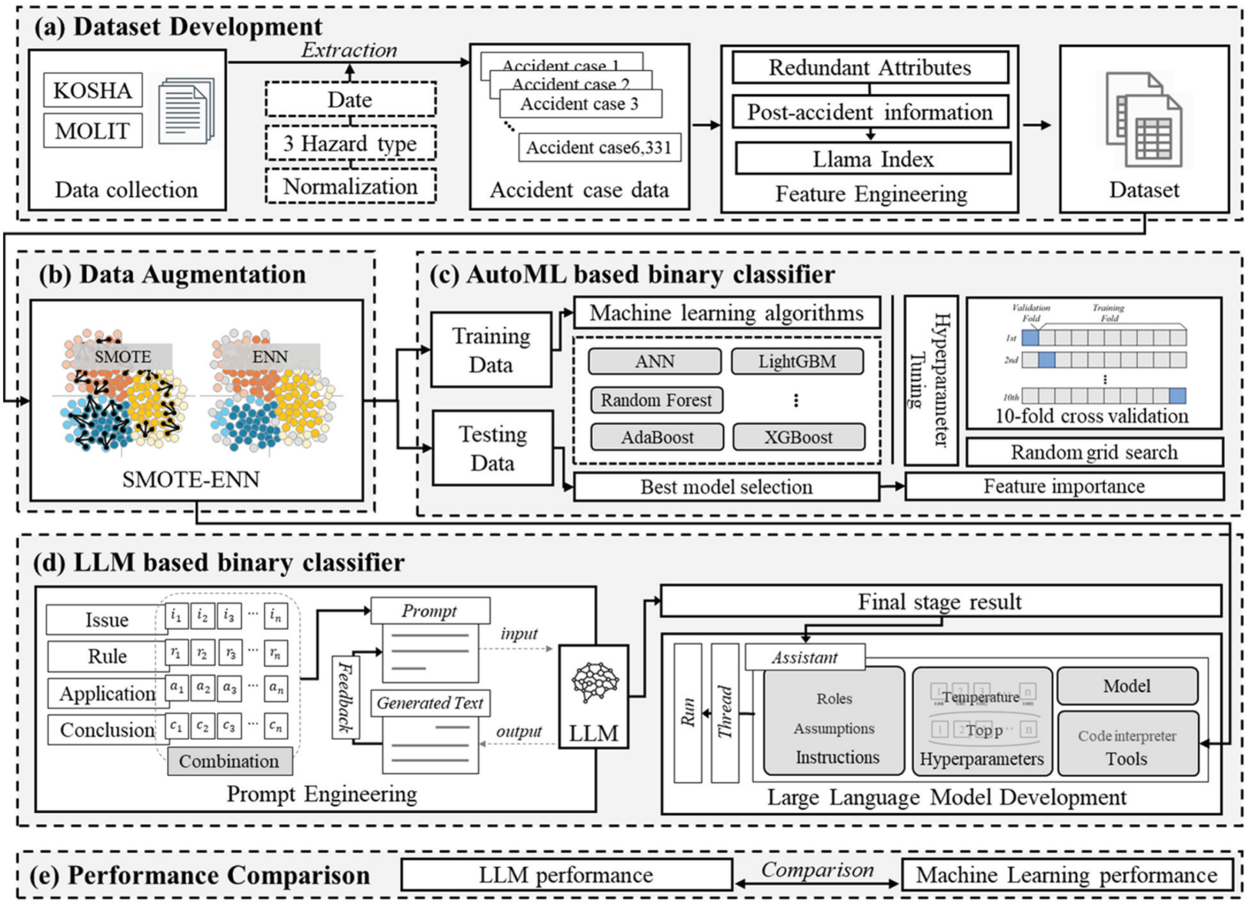

Figure 1. Research framework.

construction accidents was curated, preprocessed, and refined to ensure data quality and relevance. Feature selection techniques were applied to reduce complexity and focus on critical attributes, while data augmentation methods addressed class imbalances, resulting in a balanced dataset for model training. Automated machine learning (AutoML) technique was adopted to simplify processes (e.g., model selection and hyperparameter tuning) and develop binary classification models. For the LLM, a structured CoT prompting framework was introduced to construct a separate classification model. Lastly, the prediction and classification performance were compared based on the above two approaches.

3.1. Dataset development

In order to develop the dataset that serves as a foundation to predict the occurrence of accidents on construction sites, a large amount of publicly available historical database published by the Korea Occupational Safety and Health Agency (KOSHA) and the Ministry of Land, Infrastructure and Transport (MOLIT) was utilized (MOLIT 2023). KOSHA and MOLIT conduct national-level investigations into construction accidents and post all information (e.g., accident name, date/time, hazard type, accident description, root cause, weather/season, and project details) associated with each accident to the database. The rationale for focusing on these data sources was attributed to their comprehensive coverage across regions, projects, and seasons, as well as the completeness of the data, collected via standardized and reliable reporting protocols, all of which are similar to those in OSHA and the Bureau of Labor Statistics (BLS) in the U.S.

The initial dataset was constructed by downloading 23,573 accident case data that occurred at construction sites in South Korea from January 2019 to February 2024; individual data points consisted of 51 features. Among these construction accident cases, three hazard types of fall (18%), slip/trip (24%), and struck-by (15%) were identified as the leading causes of injuries and deaths at construction sites. They were filtered out and used to construct the dataset consisting of 13,559 accident cases. Then, preprocessing was conducted by handling null values, removing outliers (e.g., 90°C temperature), and normalizing the data, resulting in a refined dataset of 6,331 data points.

In the following step, feature selection was performed to remove unnecessary attributes and minimize the complexity of the resulting prediction model. Feature selection is an essential technique in data analytics because it efficiently improves the resulting performance (e.g., accuracy) by eliminating irrelevant/redundant features (predictors) and enhancing computational efficiency (Gupta et al. 2021). Interdependent components were reviewed, redundant attributes were removed, and features associated with post-accident information (e.g., damage description, accident recurrence prevention measure, and injured body type) were excluded, reducing the number of features from 51 to 17. Then, the Llama Index framework, a widely adopted approach for identifying critical attribute components (Jin et al. 2024), was adopted to further refine 17 features through scoring and prioritization. It was used for its retrieval-based feature-scoring mechanism, which evaluates attribute relevance not only through numerical correlations but also through contextual semantic patterns embedded in the data. This aligns with recent literature that emphasizes the value of context-aware feature selection in accident-causation modeling. By identifying features that frequently serve as discriminative cues across similar cases, the Llama Index framework provided a theoretically justified and computationally efficient method for refining mixed-type safety datasets, ensuring that the final feature set maintains strong causal relevance and is compatible with construction accident prediction. As a result, the top 10 critical features (e.g., work type, project process rate, and weather condition) were identified in this process, and they were expected to maximize the computational efficiency when implementing various types of algorithms, as partially demonstrated by previous studies (Suzuki et al. 2021).

As a result, the final dataset comprised a total of 6331 data points (construction accident cases) and 11 corresponding attributes [10 independent variables (features) and one dependent variable (target class)], as summarized in Table 1.

3.2. Dataset augmentation

Data augmentation is an approach used to expand and reshape the size and quality of data while keeping its internal information (J. Zhang et al. 2021). Since it allows for the generation of new synthetic data through transformations of the original data, it can be used to expand the dataset from a limited sample size and mitigate overfitting issues (Khosla and Saini 2020; Maharana, Mondal, and Nemade 2022).

Since the dataset (Table 1) developed in this study was severely imbalanced, with 131 cases of deaths and 6200 cases of injuries, the authors attempted to leverage the data augmentation approach for oversampling by comparing four well-established algorithms. Random Oversampling balances the class distribution by randomly replicating instances of the minority class (Mohammed, Rawashdeh, and Abdullah 2020). Synthetic Minority Oversampling Technique (SMOTE) Tomek combines the SMOTE and Tomek links algorithms (Jonathan, Putra, and Ruldeviyani 2020). It balances the class distribution by generating synthetic instances using SMOTE and enhances class separation

Table 1. Illustration of the constructed dataset.

Feature		Type	Description
Dependent	Occupational Incident Type	Categorical (2 categories)	Injury, Death
Independent	Accident Time	Categorical (7 categories)	Regular work, Holiday work, etc.
	Public/Private sector	Categorical (2 categories)	Private, Public
	Weather Condition	Categorical (6 categories)	Clear, Cloudy, Fog, etc.
	Work type – major	Categorical (7 categories)	Architecture, Civil Engineering, etc.
	Work type – subcategory	Categorical (37 categories)	Bridge, Carpentry, Demolition, etc.
	Work Process	Categorical (40 categories)	Installation, Moving, Pouring, etc.
	Construction Cost	Categorical (17 categories)	More than 1 trillion, 300 billion ~ 1ess than 500 billion won, etc.
	Tender Ratio	Categorical (8 categories)	More than 90%, 70~74%, etc.
	Construction Period (days)	Numerical	–
	Progress	Categorical (10 categories)	More than 90%, 50~59%, etc.

by removing neighboring and conflicting instances using Tomek links. This approach simultaneously addresses class imbalance and improves class separation to enhance model performance (e.g., accuracy, precision, and recall). However, it often increases the complexity of the data preprocessing step and slightly reduces the precision of the majority class. SMOTE-Edited Nearest Neighbors (ENN) is a method devised to combine SMOTE with ENN (Batista, Prati, and Monard 2004; Muntasir Nishat et al. 2022). Specifically, SMOTE generates synthetic instances through linear interpolation between existing minority class samples and their neighboring points, followed by ENN refining datasets by removing instances that introduce redundancy. Adaptive Synthetic Sampling (ADASYN) generates synthetic samples centered around densely populated regions of the minority class (H. He, Yang Bai, and Li 2008). However, it often excessively generates unnecessary samples, lacks consistency in terms of resulting performance, and involves complex computations (Dey and Pratap 2023). Among the aforementioned four oversampling algorithms, SMOTE-ENN was adopted in this study due to its proven performance, computing efficiency, and similar experimental settings (e.g., distribution of imbalance classes) observed in the previous research (Chawla et al. 2002; Choirunnisa and Lianto 2018; Lin et al. 2021; Noviandy et al. 2023).

Oversampling using SMOTE-ENN was performed via the following steps. For each minority class sample of death instances $d_m \in D(m = 1, \ldots, M)$, the algorithm identified k nearest neighbors of the same class K, and selected n samples from K according to the sampling rate. Here, n represents the number of synthetic samples to be generated from each original sample. This is represented as $k_{m,i} \in K_{d_m}(i = 1, \ldots, n, m = 1, \ldots, M)$. Finally, new samples were generated by performing random linear interpolation between d_m and $k_{m,1}, \ldots, k_{m,n}$. The new samples can be described as Eq. (1).

$$D_{new,m} = d_m + \lambda \cdot \left(k_{m,i} - d\right), i = 1, \ldots, n, m = 1, \ldots, M \quad (1)$$

where λ is a random coefficient between 0 and 1, and $D_{new,m}$ is a vector containing n new samples generated by d_m. In total, $M \cdot n$ new samples are generated.

Then, the ENN process was followed by removing a sample in case two out of its three nearest neighbors belonged to different samples. Specifically, if the majority of the k nearest neighbors of a sample d_m belonged to the majority class, the sample was removed. As a result, by generating 6069 synthetic instances of death, the original dataset was balanced (D_{new}) with a distribution of 6200 instances of deaths and 6200 instances of injuries. To evaluate whether the SMOTE-ENN augmentation process introduced bias or unrealistic samples, the Classifier Two-Sample Test (C2ST), which compares the distributional similarity between the original and augmented datasets using a discriminative algorithm, was adopted (Lopez-Paz and Oquab 2016). The C2ST score of 0.51 was achieved, indicating that the original and augmented data points are statistically indistinguishable and that minimal bias and artifacts exist. In addition, the authors performed a conditional consistency check as a complementary method to evaluate whether synthetic records satisfy domain-specific conditions by assessing hierarchical coherence and logical dependencies among input features (e.g., “work type-major” & “progress” and “construction cost” & “construction period” in Table 1). When comparing 6069 synthetic datapoints with the original 131 records of death, no record presenting contradictory or improbable feature combinations was observed.

3.3. AutoML-based binary classifier

AutoML approach that allows for training and assessing the performance of various algorithms by automatically configuring models and streamlining repetitive tasks (e.g., model selection and hyperparameter optimization) has emerged as a promising approach (Waring, Lindvall, and Umeton 2020). AutoML comprises several stages, including data preparation, feature engineering, model generation, and model evaluation. For example, in the model generation step, design principles are established for models, hyperparameters are optimized, and model architectures are adjusted based on several techniques (e.g., random grid search and evolutionary algorithms) (Truong et al. 2019).

Since the performance of many research tasks (e.g., prediction and classification) significantly varies depending on the type of input datasets and selected algorithms, AutoML has started to be widely adopted across many domains, including healthcare, finance, marketing, and construction (Agrapetidou et al. 2021, X. He, Zhao, and Chu 2021, Hutter et al. 2019, Schmitt 2023, Zhuhadar and Lytras 2023). In our study, due to its proven performance (e.g., accuracy) in previous studies and scalability (Toğan et al. 2022, Zhu et al. 2021), AutoML was adopted to develop varying binary classifiers that could predict the occurrence of construction accidents in two classes (death or injury).

Each of the fourteen algorithms [Random Forest (RF), Support Vector Machine (SVM), Light Gradient Boosting Machine (LGBM), Extremely Randomized Trees (ET), Adaptive Boosting (AdaBoost), Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GBC), Ridge Classifier (RC), Linear Discriminant Analysis (LDA), Decision Tree Classifier (DT), Naive Bayes (NB), K Neighbors Classifier (KNN), and Quadratic Discriminant Analysis (QDA)] was used to create the corresponding classifier. Collectively, these algorithms span a broad spectrum of learning paradigms, including linear and non-linear models, tree-based and ensemble approaches, as well as probabilistic algorithms. The dataset, consisting of 12,400 data points, was randomly split into 70% for training and 30% for testing. This distribution ratio was set based on proven predictive performance in prior studies (Sihombing et al. 2022). To optimize hyperparameters and fine-tune models, 10-fold cross-validation was applied with over 20 iterations using a random grid search approach based on established benchmark studies (Kohavi 1995). In each iteration, random combinations of hyperparameters were tested, and the optimal set was derived based on accuracy and loss function. This systematic procedure mitigated the risk of overfitting and subjective bias associated with manual tuning. PyCaret, an end-to-end and open-source Python library, was used for the implementation.

3.4. LLM-based binary classifier

3.4.1. Prompt engineering

In the field of LLM, prompts are defined as queries designed by users that are used as input for producing task-specific responses (Z. Zhang et al. 2022). Since the output quality is highly associated with the type, format, and structure of the input query, prompt engineering that focuses on strategically designing and optimizing prompts has started to be extensively studied by researchers (Ekin 2023).

In the construction domain, four widely used prompting approaches are zero-shot, few-shot, structured, and CoT (Chain-of-Thought) prompting (Prieto, Mengiste, and García De Soto 2023, Samsami 2024). Zero-shot prompting does not provide guidance on the input prompts, whereas few-shot prompting provides several input-output examples, which help LLMs infer task structure, formatting requirements, and reasoning patterns (Liang, Kalaleh, and Mei 2025). Structured prompting pre-defines the format of the desired output and functionalities, and aims to enhance reasoning ability by making the model comply with constraint-based guidelines (Sadowski and Chudziak 2025). CoT is an approach that induces the model to perform step-by-step reasoning, mimicking human thought processes (Li et al. 2023, Luo et al. 2024). It is particularly effective for problems involving causal relationships or requiring complex judgment. In the construction safety research domain, CoT was reported to achieve high performance by analyzing accident progression, compared to other prompting methods (Chaudhary et al. 2025, Chen 2022, Sammour et al. 2024). In addition, a preliminary experiment was conducted to investigate how the four prompting strategies affect prediction performance, and CoT achieved the highest accuracy of 84.74% in predicting construction accidents as either injury or fatality, as illustrated in Table 2.

Thus, CoT was adopted as a primary prompting strategy by further advancing it with a well-established reasoning approach of Issue, Rule, Application, and Conclusion (IRAC), which is effective in logically organizing thought processes and arguments (Lu and Kao 2024). Several reasoning steps

Table 2. Prediction performance by four prompting approaches.

	Zero-shot prompting	Few-shot prompting	Structured prompting	CoT prompting
Accuracy	50%	58.44%	67.50%	84.74

were outlined based on a sequence of sentences, and reasoning chains were established through the sequence to help a LLM-based classifier accurately classify the construction incident as either death or injury. As illustrated in Table 3, the overall process consisted of three steps (initial, intermediate, and final stages), and the final prompt was gradually developed by elaborating and expanding individual components of IRAC (e.g., Issue and Rule) in accordance with its principles.

3.4.2. LLM classifier development

In this study, the OpenAI API-Assistants, which can be used to perform various LLM-driven tasks through tool support, was utilized. The Assistants feature enables the efficient creation of AI assistants tailored to specific research purposes, allowing them to respond to user queries (OpenAI 2025). The configuration and development process consisted of the following steps.

First, instructions need to be written to ensure the *assistant* is used appropriately for its intended purpose. These instructions include *roles* and *basic assumptions*. In this study, the role of predicting human accidents on construction sites was assigned, and a prompt incorporating the IRAC-centered CoT approach, along with the output format, was specified. Second, a specific LLM was identified as a component of the *assistant*. For this study, the most recent model at the time of the experiment, "GPT-4-turbo", was adopted because of its computational efficiency and minimal resource consumption (Son et al. 2024). Furthermore, models in the GPT-4 family have been shown to provide reproducible outputs with low variance across repeated trials in research tasks (Tao et al. 2024). Third, available tools that serve as the input to the *assistant* were designated. To input the construction dataset, the "Code Interpreter" tool was utilized. Lastly, two critical *assistant*'s hyperparameters (*temperature* and *top p*) were identified and set. *Temperature* controls the variability and volatility of outputs (e.g., prediction results), and *top-p* defines the cumulative probability distribution for selecting the next token (Arora et al. 2024). To determine the appropriate configuration for the above two hyperparameters, a preliminary experiment was conducted based on a widely adopted exploratory tuning process (Bommarito and Katz 2022, Renze 2024, Savelka and Ashley 2023). *Top-p* was fixed at 1.0, and *temperature* was varied within a range of 0.0 and 2.0 with an interval of 0.5. As a result, the optimal

Table 3. Development process of the IRAC-based CoT prompt.

Stage	Section	Description
Initial stage	Issue	Predict the type of human accidents ("Deaths" or "Injuries") at the given construction sites based on training data and target data.
	Rule	-Training Data Analysis: Analyze the provided training data (file-*xxx*) to identify factors influencing accident type. - Accident Type Prediction: Use these factors to predict the likely type of human accident ("Injured" or "Deaths") for each construction site in the target file (file- *xxx*).
	Application	- Data Factor Analysis: Identify and analyze factors (e.g., weather, safety equipment use, work conditions) from the training data, and correlate them with accident types. - Similarity Analysis: Compare target site data to training data to predict the most likely accident type for each site.
	Conclusion	Predict the type of human accident for each construction site and present the results in the required format (e.g., '1. Deaths', '2. Injuries').
Intermediate stage	Issue	The task is to predict the type of human accidents ("Deaths" or "Injuries") likely to occur at each construction site. The prediction is based on the analysis of the correlation between independent and dependent variables in the training data.
	Rule	- Dependent Variable: The type of human accident ("Deaths" or "Injuries") must be predicted. - Independent Variables: All other factors except the accident type will be treated as independent variables. -Correlation Analysis: Identify the relationship between independent and dependent variables in the training data (file-*xxx*). -Prediction of Target Site Data: Based on the identified correlations, predict the type of human accidents in the target file (file- *xxx*).
	Application	- Data Analysis: Analyze the correlation between independent and dependent variables in the training data to identify key factors that predict "Deaths" or "Injuries". - Similar Site Prediction: Compare the data from the target sites with the patterns identified in the training data to predict which type of accident is more likely to occur at each site.
	Conclusion	Predict the type of human accidents for each construction site and present the results in the specified format. For example, if a fatal accident is more likely at the first site, output '1. Deaths'. If a non-fatal injury is more likely at the second site, output '2. Injured'
Final stage	Issue	The problem is to predict the type of human accidents ("Deaths" or "Injuries") that are likely to occur at each construction site. Based on the training data, the task is to analyze the correlation between independent and dependent variables and predict the accident type under specific environmental conditions.
	Rule	- Dependent Variable: The type of human accident ("Deaths" or "Injuries"). -Independent Variables: All factors except the accident type, such as weather, humidity, and construction processes, are treated as independent variables. - Correlation Analysis: Analyze the correlation between independent and dependent variables in the training data (file-*xxx*). - Prediction Model: Build a prediction model by identifying important independent variables. - Target Data Prediction: Based on the correlation analysis, predict the type of human accidents for each construction site in the target file (file- *xxx*).
	Application	- Prediction Execution: Use the analysis results to predict the type of accident for each construction site in the target data.
	Conclusion	Predict the type of human accident for each construction site and present the results in the specified format. For example, if a fatal accident is more likely at the first site, output '1. Deaths'. If a non-fatal injury is more likely at the second site, output '2. Injuries'.

Table 4. LLM-based classifier development.

Instructions	Role	You are a construction site human accident prediction model. The file xxx is the training data, and the file xxx is the list of construction sites where you need to predict the types of human accidents. We aim to analyze the impact of various environmental and working conditions, such as weather, humidity, and construction processes, on the types of human accidents at construction sites. By developing a model that predicts the likelihood of specific types of human accidents, 'Injured' or 'Deaths,' under certain conditions, we intend to establish an efficient safety management system that can detect and prevent accident risks in advance.
	Assumption	
	CoT Prompting	To predict human accidents, follow these steps: (1) Based on the site data from the file xxx, predict the dependent variable, the type of human accident (Deaths, Injured). Independent variables are items other than the type of human accident. (2) Identify the correlation between independent variables and the dependent variable. Through analysis, identify important independent variables and use them to build the prediction model. (3) Based on the correlation analysis from step 2, predict the type of human accident for the construction sites in the file xxx. The types are 'Injured' and 'Deaths'.
	Output format	The output format should be displayed as 'ConstructionSiteOrder. AccidentType'. For example, if there is a higher likelihood of a death occurring rather than an injury at the first construction site, it should be displayed as '1. Deaths'. Similarly, if there is a higher likelihood of an injury occurring rather than a death at the second site, it should be displayed as '2. Injured'. Please output the results according to this format.
Model		gpt-4-turbo
Tools		Code interpreter
Hyperparameter	Temperature	1.0
	Top P	1.0

configuration was found to be 1.0 for both hyperparameters. Table 4 illustrates the development process.

4. Results and discussions

4.1. AutoML-based binary classification performance

In order to assess and compare binary classification performance, four well-known metrics (accuracy, precision, recall, and F1-score) were utilized (Chinchor and Sundheim 1993). Accuracy that measures the overall correctness of the model is the ratio of correctly predicted instances to the total number of instances (Eq. (2)). Precision is defined as the ratio of correctly predicted positive instances to the total predicted positive instances (Eq.(3)). It focuses on the accuracy of positive predictions, indicating the proportion of true positives (TP) among all positive predictions, including false positives (FP). Recall measures the proportion of total confidence scores correctly identified as instances of a specific class (Eq.(4)). F1-score is the harmonic mean of precision and recall, offering a balanced measure of a model's performance (Eq.(5)). In this study, the "Death" class was designated as the positive class when calculating the evaluation metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

Table 5 presents the performance of 14 ML classifiers, evaluated using four key metrics: accuracy, precision, recall, and F1-score. The classifiers are ranked in descending order of accuracy, with XGBoost achieving the highest performance (accuracy: 97.9%, precision: 98.2%, recall: 97.5%, and F1-score: 97.9%), followed by LightGBM, RF, and ET. In contrast, QDA showed the lowest performance. Figure 2 illustrates the performance of classifiers with an accuracy above 0.94.

The superior performance of the top classifiers can be attributed to their tree-based ensemble mechanisms, which effectively capture complex nonlinear relationships in the data. Considering that the dataset predominantly comprises nonlinear variables (e.g., work type, construction cost, and construction period), it is likely that tree-based methods like XGBoost and LightGBM emerge as robust alternatives. On the other hand, linear models such as LR and LDA performed relatively well but were less effective in handling nonlinear interactions.

In the next step, the feature importance of the best-performing XGBoost classifier was analyzed to explore the relationship between data features and their impact on classification performance, as shown in Figure 3.

The analysis revealed that "Work type-Major" contributed the most to classification performance, indicating that the type of construction work (e.g., architecture, civil, mechanical, and electrical work) plays a critical role in predicting potential accident outcomes. This result aligns with prior studies

Table 5. Performance of AutoML-based classifiers.

No.	classifiers	Accuracy	Precision	Recall	F1-score
1	XGBoost (Extreme Gradient Boosting)	0.979	0.982	0.975	0.979
2	LightGBM (Light Gradient Boosting Machine)	0.976	0.976	0.975	0.976
3	RF (Random Forest Classifier)	0.975	0.976	0.974	0.975
4	ET (Extra Trees Classifier)	0.971	0.973	0.97	0.971
5	GBC (Gradient Boosting)	0.965	0.969	0.962	0.965
6	LR (Logistic Regression)	0.962	0.968	0.955	0.961
7	RC (Ridge Classifier)	0.959	0.971	0.946	0.958
8	LDA (Linear Discriminant Analysis)	0.959	0.971	0.946	0.958
9	AdaBoost (Adaptive Boosting Classifier)	0.954	0.957	0.951	0.954
10	SVM (SVM - Linear Kernel)	0.953	0.954	0.951	0.953
11	DT (Decision Tree Classifier)	0.943	0.944	0.941	0.942
12	NB (Naive Bayes)	0.904	0.920	0.886	0.903
13	KNN (K Neighbors Classifier)	0.887	0.981	0.788	0.874
14	QDA (Quadratic Discriminant Analysis)	0.747	0.768	0.747	0.744

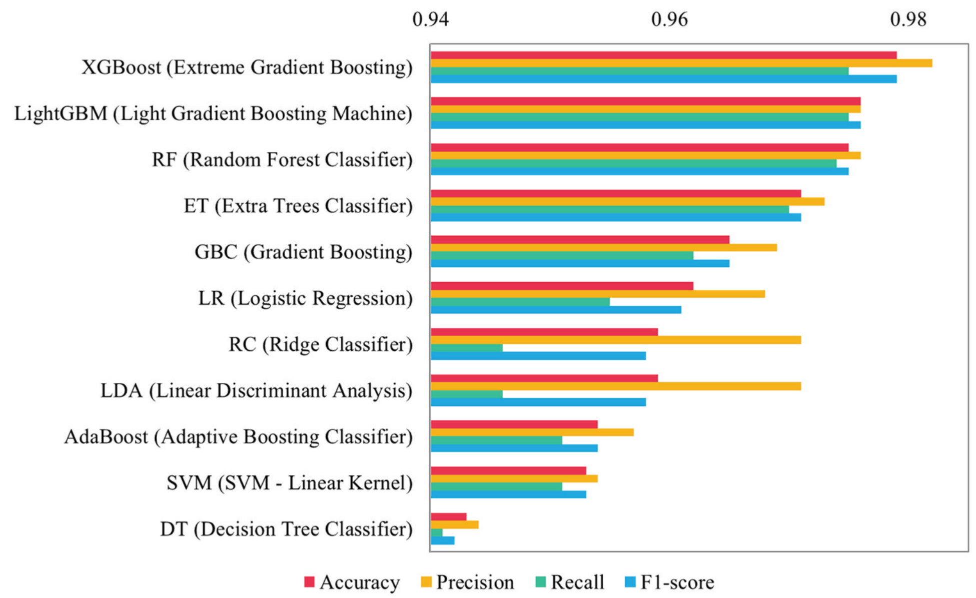

Figure 2. Performance of AutoML-based classifiers.

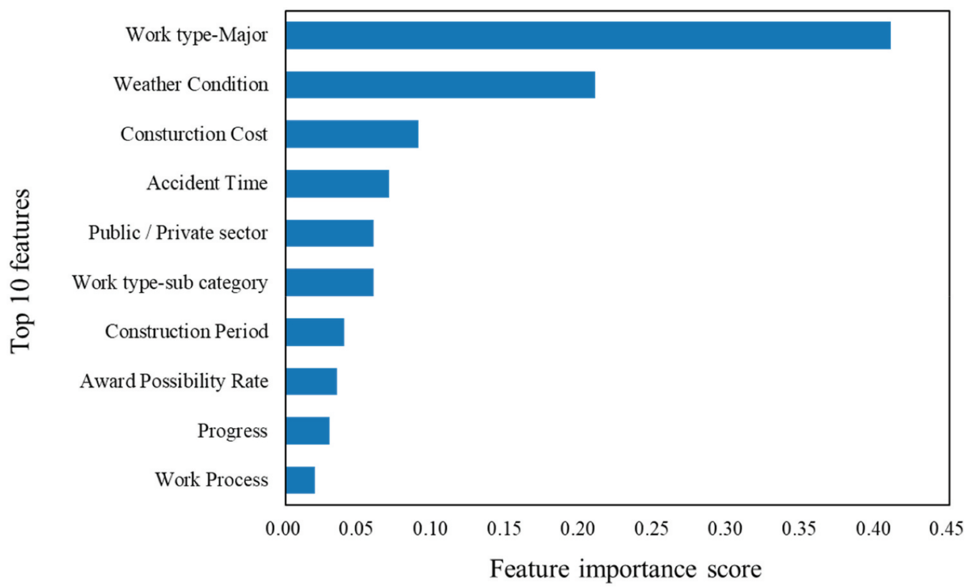

Figure 3. Feature importance computed by XGBoost classifier.

suggesting that specific construction tasks expose workers to unique risks and hazards that collectively influence the severity and type of accidents (Frank and Bacao 2023).

A notable association was also observed between "Weather Condition" and construction accidents. Previous research highlights that adverse weather conditions can significantly impact workers both physically and psychologically, thereby increasing their susceptibility to safety risks (Ha et al. 2023, Trillo Cabello et al. 2021). Similarly, "Construction Cost" emerged as the third most important feature, suggesting that projects with limited budgets may allocate fewer resources (e.g., safety equipment) to safety management, thereby elevating the risk of incidents on jobsites (Tsang et al. 2017).

While the top three features presented the highest importance scores, other variables, such as "Construction Period" and "Accident Reporting Time," also contributed to the model's performance, albeit to a lesser extent. These features may capture subtler influences on accident risk, such as extended project durations or delays in incident reporting. Conversely, variables such as "Tender Ratio" had minimal impact, implying they are less directly connected to safety outcomes in the construction domain.

These findings suggest that prioritizing data related to work type, weather conditions, and construction costs as pre-incident indicators could enhance safety management on construction sites by enabling effective and efficient construction accident prediction. Additionally, while lower-ranked features may seem less critical, they could still offer valuable contextual insights for more holistic accident prediction models. Researchers and practitioners are encouraged to focus on these high-impact features to develop efficient and streamlined classification models, potentially omitting less relevant variables to simplify analysis and implementation.

4.2. Comparative classification performance of ML versus LLM

This section investigates the application of LLM for binary classification aimed at predicting accident types (death or injury). To evaluate the feasibility and performance of an LLM classifier, ten iterative training and testing experiments were conducted. The classifier was trained and tested using the same dataset consisting of 12,400 data points, divided into training (8,680) and testing (3,720) subsets at a 7:3 ratio. This consistent splitting ensured fair and reliable evaluation across all iterations, allowing for a robust assessment of the classifier's predictive capabilities.

The results demonstrated that the LLM classifier achieved an average accuracy of 94.5%, precision of 90.9%, recall of 98.9%, and F1-score of 94.7%. These results highlight the model's ability to capture nearly all true positive cases with very high recall (98.9%), while the comparatively lower precision (90.9%) indicates that some injury instances were incorrectly classified as death. Figure 4 provides a visual representation of the confusion matrix, offering deeper insights into the classifier's strengths and areas for improvement.

When compared with the top-performing ML classifier (XGBoost) the LLM classifier showed a 3.4% lower accuracy and a 7.3% lower precision, but a 1.4% higher recall. This difference can be explained in several aspects. First, when LLMs are pre-trained, they often develop a safety-oriented judgment tendency, leading them to perform conservative reasoning that overestimates potential risks. This can result in an increase in the false positive rate by over-predicting high-risk

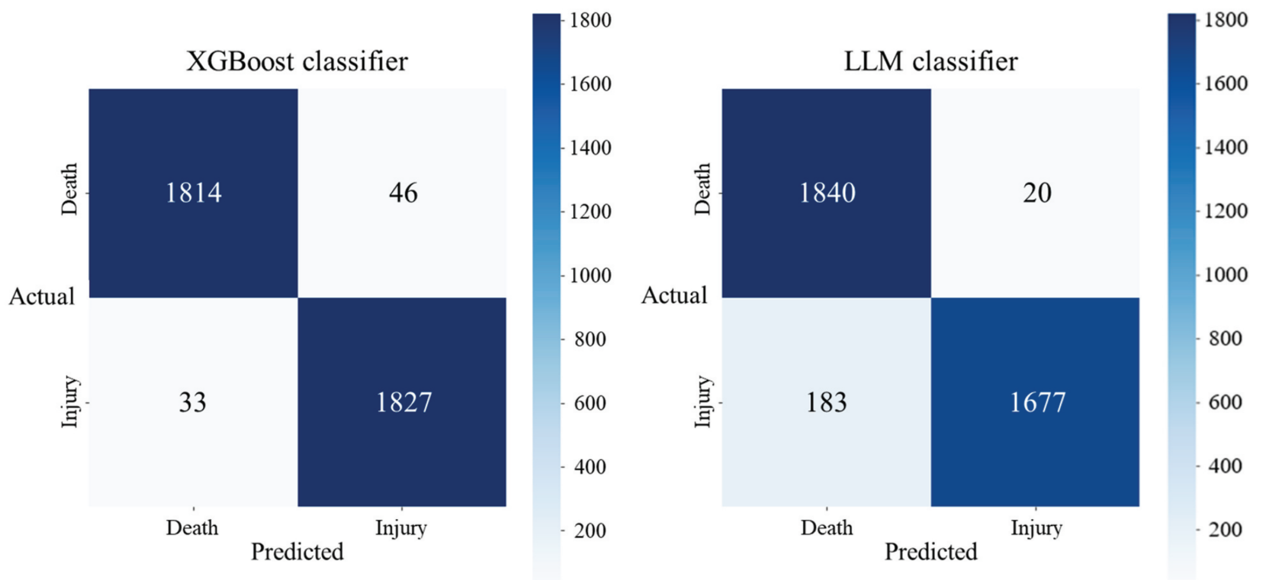

Figure 4. Confusion matrix comparison (XGBoost vs. LLM).

classes such as death. Second, the process of grouping continuous values into intervals was applied equally to both the ML and LLM during data curation. However, when the LLM interpreted these intervals linguistically (e.g., progress between 70% and 79%), it introduced semantic ambiguity within the same bucket. As a result, while the ML model classified cases based on explicit numerical thresholds, the LLM tended to interpret the bucketed expressions contextually, leading to a higher number of false positives.

LLM classifier demonstrated a notable ability to capture nonlinear relationships in the data, as evidenced by its very high recall (98.9%). This indicates its potential to identify nearly all true positive accident cases. Also, the LLM demonstrated a notable aptitude for analyzing accident-related data compared to typical structured datasets. This is likely because the LLM was pre-trained on a massive corpus of unstructured accident reports (e.g., from OSHA and KOSHA), allowing it to leverage deep, domain-specific contextual knowledge during the analysis of structured features.

While LLM offers promising advantages in understanding data complexity and enhancing feature robustness, ML-based models remain highly effective for tasks requiring precise optimization of classification boundaries. Future research could focus on combining the strengths of both approaches to develop hybrid models that leverage the generative capabilities of LLMs while maintaining the classification precision of traditional ML algorithms.

4.3. Validation on hold-out dataset

To assess the temporal generalization capability of the developed models and address potential concerns regarding overfitting to historical data, an additional validation was conducted using a hold-out dataset comprising recent accident cases. Specifically, 1,206 accident records that occurred between March 2024 and August 2025 were collected from the KOSHA and MOLIT databases, representing data that were not included in the training or testing phases described in Section 3. The same preprocessing procedures were applied to maintain consistency with the original dataset structure.

The trained XGBoost and IRAC-based LLM classifiers were directly applied to this hold-out dataset without retraining or parameter adjustments, to strictly evaluate their generalization performance. The prediction accuracy was assessed using the same four evaluation metrics (accuracy, precision, recall, and F1-score), and the results are summarized in Table 6.

The findings indicate that both models maintained stable, high performance when applied to the most recent data, confirming that their learned representations were not dependent on time-specific patterns in the original data set. The XGBoost classifier showed only a minor decrease (approximately 1.1%) in accuracy compared to the original test results, while the LLM maintained a very high recall (96.8%), demonstrating its ability to detect nearly all death cases in the new dataset.

These results suggest that both approaches exhibit strong temporal robustness and that their predictive mechanisms are based on generalizable relationships rather than spurious correlations derived from past data. Although slight performance degradation was observed, it remained within acceptable bounds, validating the long-term applicability of both models to ongoing research on construction accident prediction.

5. Conclusion

The objective of this study was to investigate the feasibility of using LLM for construction accident prediction and to determine whether it could outperform a widely adopted method (i.e., ML) through comparative evaluation. For this, the dataset was constructed by collecting 6331 data points, which were then augmented using the SMOTE-ENN technique. An AutoML approach was adopted for developing 14 separate ML models, and an LLM-centered classifier was designed via IRAC-based CoT prompting. The resulting performance was then compared, and additional validation was performed using 1206 unseen datasets.

The research outcome led to the following findings. First, of 14 ML algorithms, XGBoost was found to be the best-performing classifier, achieving 97.9% accuracy and 97.5% recall. At this stage, feature importance analysis revealed three critical features (work type, weather condition, and construction cost) that significantly contributed to the performance. Second, the LLM also demonstrated strong predictive capability, achieving 94.5% accuracy and 98.9% recall, suggesting it is more effective at capturing fatal cases, likely due to its conservative bias toward safety-related outcomes. Third, both ML and LLM approaches maintained robust performance (accuracies of 96.7% and 92.6%, respectively) when validated on a hold-out dataset, with only a marginal decrease.

Table 6. Validation of XGBoost and IRAC-based LLM classifiers.

Model	Accuracy	Precision	Recall	F1-score
XGBoost classifier	0.967	0.972	0.962	0.967
LLM classifier	0.926	0.893	0.968	0.929

Despite the findings, there remain several limitations of this study that deserve future research efforts. First, since the dataset was developed from Korean construction accident records, it is necessary to apply the proposed framework to datasets from other countries. Second, LLM experiments in this research relied solely on a single LLM (GPT-4-Turbo), which could raise reproducibility concerns and leave the performance of alternative LLM architectures unexplored. The authors will compare varying LLMs (e.g., LLaMA-2) and include open-source alternatives that allow for fine-tuning and transparent benchmarking in future research. Third, although the reliability of SMOTE-ENN for addressing class imbalance was demonstrated via C2ST, further validation can be pursued by applying advanced techniques (e.g., class-weighted loss functions) and analyzing their impact through ablation studies. Fourth, future studies should explore various prompting strategies against the proposed IRAC-based CoT to determine the most effective strategy for accident prediction tasks. Finally, the interpretability of the LLM's decisions remains limited. Unlike the ML model, which benefited from feature importance analysis, the LLM's reasoning process remains an open challenge. Developing methods to extract and present the LLM's decision rationale (e.g., through natural language explanations or user-interactive queries) is an important next step to increase trust in LLM predictors among safety professionals.

Author contributions

SeungJu Park: analysis and interpretation of the data; the drafting of the paper. **KiHyun Jeong**: analysis and interpretation of the data; the drafting of the paper. **Jinwoo Kim**: analysis and interpretation of the data; revising the paper. **HyunSoo Kim**: revising the paper. **SungHwan Yoon**: analysis and interpretation of the data. **Jaeyoon Kim**: revising the paper. **JungHo Jeon**: conception and design; final approval of the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (no. RS-2023-00218875).

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.

References

Agrapetidou, A., P. Charonyktakis, P. Gogas, T. Papadimitriou, and I. Tsamardinos. 2021. "An AutoML Application to Forecasting Bank Failures." *Applied Economics Letters* 28 (1): 5–9. https://doi.org/10.1080/13504851.2020.1725230.

Akboğa Kale, Ö., and S. Baradan. 2020. "Identifying Factors That Contribute to Severity of Construction Injuries Using Logistic Regression Model." *Teknik Dergi* 31 (2): 9919–9940. https://doi.org/10.18400/tekderg.470633.

Akinosho, T. D., L. O. Oyedele, M. Bilal, A. O. Ajayi, M. D. Delgado, O. O. Akinade, and A. A. Ahmed. 2020. "Deep Learning in the Construction Industry: A Review of Present Status and Future Innovations." *Journal of Building Engineering* 32:101827. https://doi.org/10.1016/j.jobe.2020.101827.

Alin, A. 2010. "Multicollinearity." *WIREs Computational Statistics* 2 (3): 370–374. https://doi.org/10.1002/wics.84.

Amacher, S. A., A. Arpagaus, C. Sahmer, C. Becker, S. Gross, T. Urben, K. Tisljar, R. Sutter, S. Marsch, and S. Hunziker. 2024. "Prediction of Outcomes After Cardiac Arrest by a Generative Artificial Intelligence Model." *Resuscitation Plus* 18:100587. https://doi.org/10.1016/j.resplu.2024.100587.

Andolfo, C., and F. Sadeghpour. 2015. "A Probabilistic Accident Prediction Model for Construction Sites." *Procedia Engineering* 123:15–23. https://doi.org/10.1016/j.proeng.2015.10.052.

Arora, C., A. I. Sayeed, S. Licorish, F. Wang, and C. Treude. 2024. "Optimizing Large Language Model Hyperparameters for Code Generation." arXiv.

Ayhan, B. U., and O. B. Tokdemir. 2019. "Predicting the Outcome of Construction Incidents." *Safety Science* 113:91–104. https://doi.org/10.1016/j.ssci.2018.11.001.

Baker, H., M. R. Hallowell, and A. J.-P. Tixier. 2020. "AI-Based Prediction of Independent Construction Safety Outcomes from Universal Attributes." *Automation in Construction* 118:103146. https://doi.org/10.1016/j.autcon.2020.103146.

Batista, G. E. A. P. A., R. C. Prati, and M. C. Monard. 2004. "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data." *ACM SIGKDD Explorations Newsletter* 6 (1): 20–29. https://doi.org/10.1145/1007730.1007735.

BLS (U.S. Bureau of Labor Statistics). 2023. Number and Rate of Fatal Work Injuries, by Private Industry Sector.

Bommarito, M., and D. M. Katz. 2022. "GPT Takes the Bar Exam." arXiv.

Chaudhary, N., S. M. J. Uddin, S. S. Chandra, A. Ovid, and A. Albert. 2025. "Prompt to Protection: A Comparative Study of Multimodal LLMs in Construction Hazard Recognition." arXiv.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16:321–357. https://doi.org/10.1613/jair.953.

Che, C., Z. Huang, C. Li, H. Zheng, and X. Tian. 2024. "Integrating Generative AI into Financial Market Prediction for Improved Decision Making." arXiv.

Chen, W. 2022. "Large Language Models Are Few(1)-Shot Table Reasoners." arXiv.

Chinchor, N., and B. M. Sundheim. 1993. "MUC-5 Evaluation Metrics." In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore*, 69–78. Maryland.

Choi, D., and C. Koo. 2023. "Impact of Personal Characteristics on Learning Performance in Virtual Reality-Based Construction Safety Training - Using Machine Learning and SHAP -." *Korean Journal of Construction Engineering and Management* 24 (6): 3–11. https://doi.org/10.6106/KJCEM.2023.24.6.003.

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." *Automation in Construction* 110:102974. https://doi.org/10.1016/j.autcon.2019.102974.

Choirunnisa, S., and J. Lianto. 2018. "Hybrid Method of Undersampling and Oversampling for Handling Imbalanced Data." In *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 276–280. Yogyakarta, Indonesia: IEEE.

Dey, I., and V. Pratap. 2023. "A Comparative Study of SMOTE, Borderline-SMOTE, and ADASYN Oversampling Techniques Using Different Classifiers." In *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, 294–302. Trichy, India: IEEE.

Ekin, S. 2023. "Prompt Engineering for ChatGPT: A Quick Guide to Techniques, Tips, and Best Practices."

Elamrani Abou Elassad, Z., H. Mousannif, and H. Al Moatassime. 2020. "Class-Imbalanced Crash Prediction Based on Real-Time Traffic and Weather Data: A Driving Simulator Study." *Traffic Injury Prevention* 21 (3): 201–208. https://doi.org/10.1080/15389588.2020.1723794.

Frank, F., and F. Bacao. 2023. "Advanced Genetic Programming vs. State-of-the-Art AutoML in Imbalanced Binary Classification." *Emerging Science Journal* 7 (4): 1349–1363. https://doi.org/10.28991/ESJ-2023-07-04-021.

Fui-Hoon Nah, F., R. Zheng, J. Cai, K. Siau, and L. Chen. 2023. "Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration." *Journal of Information Technology Case and Application Research* 25 (3): 277–304. https://doi.org/10.1080/15228053.2023.2233814.

Go, H., J. Hyun, J. Lee, and J. Ahn. 2022. "Development of a Quantitative Risk Assessment Model by BIM-Based Risk Factor Extraction - Focusing on Falling Accidents -." *Korean Journal of Construction Engineering and Management* 23 (4): 15–25. https://doi.org/10.6106/KJCEM.2022.23.4.015.

Gondia, A., M. Ezzeldin, and W. El-Dakhakhni. 2022. "Machine Learning–Based Decision Support Framework for Construction Injury Severity Prediction and Risk Mitigation." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 8 (3): 04022024. https://doi.org/10.1061/AJRUA6.0001239.

GPT-4 Turbo. 2024. "openAi."

Gupta, V. K., A. Gupta, D. Kumar, and A. Sardana. 2021. "Prediction of COVID-19 Confirmed, Death, and Cured Cases in India Using Random Forest Model." *Big Data Mining and Analytics* 4 (2): 116–123. https://doi.org/10.26599/BDMA.2020.9020016.

Ha, D., Y. Yu, J. Choi, S. Kim, and B. Koo. 2023. "Integrating a Machine Learning-Based Space Classification Model with an Automated Interior Finishing System in BIM Models." *Korean Journal of Construction Engineering and Management* 24 (4): 60–73. https://doi.org/10.6106/KJCEM.2023.24.4.060.

Hansen, H., and M. N. Hebart. 2022. "Semantic Features of Object Concepts Generated with GPT-3." arXiv.

He, H., E. A. G. Yang Bai, and S. Li. 2008. "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning." In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. Hong Kong, China: IEEE.

He, X., K. Zhao, and X. Chu. 2021. "AutoML: A Survey of the State-of-the-Art." *Knowledge-Based Systems* 212:106622. https://doi.org/10.1016/j.knosys.2020.106622.

2019. "Automated Machine Learning: Methods, Systems, Challenges." In *The Springer Series on Challenges in Machine Learning*, edited by Hutter, F., L. Kotthoff, and J. Vanschoren, 219. Cham: Springer International Publishing.

Jeon, J., and H. Cai. 2021. "Classification of Construction Hazard-Related Perceptions Using: Wearable Electroencephalogram and Virtual Reality." *Automation in Construction* 132:103975. https://doi.org/10.1016/j.autcon.2021.103975.

Jeon, J., and H. Cai. 2023. "Wearable EEG-Based Construction Hazard Identification in Virtual and Real Environments: A Comparative Study." *Safety Science* 165:106213. https://doi.org/10.1016/j.ssci.2023.106213.

Jeong, H., J. Chae, and Y. Kang. 2023. "Distribution of Occupational Safety and Health Management Costs (OSHMC) by Project Size and Activity Type with the Consideration of Accident Rates." *Korean Journal of Construction Engineering and Management* 24 (4): 44–51. https://doi.org/10.6106/KJCEM.2023.24.4.044.

Jin, M., Q. Yu, D. Shu, C. Zhang, L. Fan, W. Hua, S. Zhu, et al. 2024. "Health-LLM: Personalized Retrieval-Augmented Disease Prediction System." arXiv.

Jonathan, B., P. H. Putra, and Y. Ruldeviyani. 2020. "Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek." In *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 81–85. Bali, Indonesia: IEEE.

Kang, K., and H. Ryu. 2019. "Predicting Types of Occupational Accidents at Construction Sites in Korea Using Random Forest Model." *Safety Science* 120:226–236. https://doi.org/10.1016/j.ssci.2019.06.034.

Kang, K.-S., C. Koo, and H.-G. Ryu. 2022. "An Interpretable Machine Learning Approach for Evaluating the Feature Importance Affecting Lost Workdays at Construction Sites." *Journal of Building Engineering* 53:104534. https://doi.org/10.1016/j.jobe.2022.104534.

Khanmohammadi, S., M. G. Cruz, D. D. B. Perrakis, M. E. Alexander, and M. Arashpour. 2024. "Using AutoML and Generative AI to Predict the Type of Wildfire Propagation in Canadian Conifer Forests." *Ecological Informatics* 82:102711. https://doi.org/10.1016/j.ecoinf.2024.102711.

Khosla, C., and B. S. Saini. 2020. "Enhancing Performance of Deep Learning Models with Different Data Augmentation Techniques: A Survey." In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, 79–85. London, United Kingdom: IEEE.

Kim, J., K. Lee, and J. Jeon. 2024. "Systematic Literature Review of Wearable Devices and Data Analytics for Construction Safety and Health." *Expert Systems With Applications* 257:125038. https://doi.org/10.1016/j.eswa.2024.125038.

Koc, K., Ö. Ekmekcioğlu, and A. P. Gurgun. 2021. "Integrating Feature Engineering, Genetic Algorithm and Tree-Based Machine Learning Methods to Predict the Post-Accident Disability Status of Construction Workers." *Automation in Construction* 131:103896. https://doi.org/10.1016/j.autcon.2021.103896.

Koc, K., Ö. Ekmekcioğlu, and A. P. Gurgun. 2022. "Accident Prediction in Construction Using Hybrid Wavelet-Machine Learning." *Automation in Construction* 133:103987. https://doi.org/10.1016/j.autcon.2021.103987.

Koc, K., Ö. Ekmekcioğlu, and A. P. Gurgun. 2023a. "Developing a National Data-Driven Construction Safety Management Framework with Interpretable Fatal

Accident Prediction." *Journal of Construction Engineering and Management* 149 (4): 04023010. https://doi.org/10.1061/JCEMD4.COENG-12848.

Koc, K., Ö. Ekmekcioğlu, and A. P. Gurgun. 2023b. "Prediction of Construction Accident Outcomes Based on an Imbalanced Dataset Through Integrated Resampling Techniques and Machine Learning Methods." *Engineering, Construction and Architectural Management* 30 (9): 4486–4517. https://doi.org/10.1108/ECAM-04-2022-0305.

Koc, K., and A. P. Gurgun. 2022. "Scenario-Based Automated Data Preprocessing to Predict Severity of Construction Accidents." *Automation in Construction* 140:104351. https://doi.org/10.1016/j.autcon.2022.104351.

Kohavi, R. 1995. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. ACM.

Kulinan, A. S., M. Park, P. P. W. Aung, G. Cha, and S. Park. 2024. "Advancing Construction Site Workforce Safety Monitoring Through BIM and Computer Vision Integration." *Automation in Construction* 158:105227. https://doi.org/10.1016/j.autcon.2023.105227.

Lee, J., and S. Ahn. 2025. "Development of AI Prototype for Generating Construction Safety Guidelines Through Fine-Tuning of Large-Scale Language Model." *Korean Journal of Construction Engineering and Management* 26 (2): 20–31. https://doi.org/10.6106/KJCEM.2025.26.2.020.

Lee, J. Y., Y. G. Yoon, T. K. Oh, S. Park, and S. I. Ryu. 2020. "A Study on Data Pre-Processing and Accident Prediction Modelling for Occupational Accident Analysis in the Construction Industry." *Applied Sciences* 10 (21): 7949. https://doi.org/10.3390/app10217949.

Lee, S.-W., and W.-J. Choi. 2023. "Utilizing ChatGPT in Clinical Research Related to Anesthesiology: A Comprehensive Review of Opportunities and Limitations." *Anesthesia and Pain Medicine* 18 (3): 244–251. https://doi.org/10.17085/apm.23056.

Li, J., G. Li, Y. Li, and Z. Jin. 2023. "Structured Chain-of-Thought Prompting for Code Generation." arXiv.

Liang, H., M. T. Kalaleh, and Q. Mei. 2025. "Integrating Large Language Models for Automated Structural Analysis." arXiv.

Lin, M., X. Zhu, T. Hua, X. Tang, G. Tu, and X. Chen. 2021. "Detection of Ionospheric Scintillation Based on XGBoost Model Improved by SMOTE-ENN Technique." *Remote Sensing* 13 (13): 2577. https://doi.org/10.3390/rs13132577.

Liu, Y., Z. Yang, Z. Yu, Z. Liu, D. Liu, H. Lin, M. Li, S. Ma, M. Avdeev, and S. Shi. 2023. "Generative Artificial Intelligence and Its Applications in Materials Science: Current Situation and Future Perspectives." *Journal of Materiomics* 9 (4): 798–816. https://doi.org/10.1016/j.jmat.2023.05.001.

Liu, Y., Z. Zhang, W. Zhang, S. Yue, X. Zhao, X. Cheng, Y. Zhang, and H. Hu. 2023. "ArguGPT: Evaluating, Understanding and Identifying Argumentative Essays Generated by GPT Models." arXiv.

Lopez-Paz, D., and M. Oquab. 2016. "Revisiting Classifier Two-Sample Tests." arXiv.

Lu, Y., and H. Kao. 2024. "0x.Yuan at SemEval-2024 Task 5: Enhancing Legal Argument Reasoning with Structured Prompts." In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval*-2024), edited by O. Atul Kr. A. S. Doğruöz, H. Madabushi, D. S. Tayyar, G. Martino, S. Rosenthal and A. Rosá, 385–390. Mexico City, Mexico: Association for Computational Linguistics.

Luo, X., A. Rechardt, G. Sun, K. K. Nejad, F. Yáñez, B. Yilmaz, K. Lee, et al. 2024. "Large Language Models Surpass Human Experts in Predicting Neuroscience Results." *Nature Human Behaviour* 9 (2): 305–315. https://doi.org/10.1038/s41562-024-02046-9.

Maharana, K., S. Mondal, and B. Nemade. 2022. "A Review: Data Pre-Processing and Data Augmentation Techniques." *Global Transitions Proceedings* 3 (1): 91–99. https://doi.org/10.1016/j.gltp.2022.04.020.

Mistikoglu, G., I. H. Gerek, E. Erdis, P. E. Mumtaz Usmen, H. Cakan, and E. E. Kazan. 2015. "Decision Tree Analysis of Construction Fall Accidents Involving Roofers." *Expert Systems With Applications* 42 (4): 2256–2263. https://doi.org/10.1016/j.eswa.2014.10.009.

MOEL. 2023. "Analysis of the Status of Industrial Accidents in 2022."

Mohammed, R., J. Rawashdeh, and M. Abdullah. 2020. "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results." In *2020 11th International Conference on Information and Communication Systems (ICICS)*, 243–248. Irbid, Jordan: IEEE.

MOLIT. 2023. "Construction Accidents Historical Data." Ministry of Land, Infrastructure and Transport (MOLIT).

Muntasir Nishat, M., F. Faisal, I. Jahan Ratul, A. Al-Monsur, A. M. Ar-Rafi, S. M. Nasrullah, M. T. Reza, and M. R. H. Khan. 2022. "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset." *Scientific Programming* 2022:1–17. https://doi.org/10.1155/2022/3649406.

Noviandy, T. R., G. M. Idroes, A. Maulana, I. Hardi, E. S. Ringga, and R. Idroes. 2023. "Credit Card Fraud Detection for Contemporary Financial Management Using XGBoost-Driven Machine Learning and Data Augmentation Techniques." *Indatu Journal of Management and Accounting* 1 (1): 29–35. https://doi.org/10.60084/ijma.v1i1.78.

Nyqvist, R., A. Peltokorpi, and O. Seppänen. 2024. "Can ChatGPT Exceed Humans in Construction Project Risk Management?" *Engineering, Construction and Architectural Management* 31 (13): 223–243. https://doi.org/10.1108/ECAM-08-2023-0819.

OpenAI. 2025. "Assistants API Overview."

Prieto, S. A., E. T. Mengiste, and B. García De Soto. 2023. "Investigating the Use of ChatGPT for the Scheduling of Construction Projects." *Buildings* 13 (4): 857. https://doi.org/10.3390/buildings13040857.

Renze, M. 2024. "The Effect of Sampling Temperature on Problem Solving in Large Language Models." In *Findings of the Association for Computational Linguistics: EMNLP 2024*, edited by Y. Al-Onaizan, M. Bansal and Y-.N. Chen, 7346–7356. Miami, Florida, USA: Association for Computational Linguistics.

Requeima, J., J. Bronskill, D. Choi, R. E. Turner, and D. Duvenaud. 2024. "LLM Processes: Numerical Predictive Distributions Conditioned on Natural Language." *ArXiv*. https://doi.org/10.48550/ARXIV.2405.12856.

Rokooei, S., A. Shojaei, A. Alvanchi, R. Azad, and N. Didehvar. 2023. "Virtual Reality Application for Construction Safety Training." *Safety Science* 157:105925. https://doi.org/10.1016/j.ssci.2022.105925.

Sadowski, A., and J. A. Chudziak. 2025. "Explainable Rule Application via Structured Prompting: A Neural-Symbolic Approach." arXiv.

Sammour, F., J. Xu, X. Wang, M. Hu, and Z. Zhang. 2024. "Responsible AI in Construction Safety: Systematic

Evaluation of Large Language Models and Prompt Engineering." arXiv.
Samsami, R. 2024. "Optimizing the Utilization of Generative Artificial Intelligence (AI) in the AEC Industry: ChatGPT Prompt Engineering and Design." *CivilEng* 5 (4): 971–1010. https://doi.org/10.3390/civileng5040049.
Savelka, J., and K. D. Ashley. 2023. "The Unreasonable Effectiveness of Large Language Models in Zero-Shot Semantic Annotation of Legal Texts." *Frontiers in Artificial Intelligence* 6:1279794. https://doi.org/10.3389/frai.2023.1279794.
Schmitt, M. 2023. "Automated Machine Learning: AI-Driven Decision Making in Business Analytics." *Intelligent Systems with Applications* 18:200188. https://doi.org/10.1016/j.iswa.2023.200188.
Sihombing, D. J. C., J. U. Dexius, J. Manurung, M. Aritonang, and H. S. Adinata. 2022. "Design and Analysis of Automated Machine Learning (AutoML) in PowerBI Application Using PyCaret." In *2022 International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*, 89–94. Denpasar, Bali, Indonesia: IEEE.
Singer, U., A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, et al. 2022. "Make-a-Video: Text-to-Video Generation Without Text-Video Data." arXiv.
Smetana, M., L. Salles De Salles, I. Sukharev, and L. Khazanovich. 2024. "Highway Construction Safety Analysis Using Large Language Models." *Applied Sciences* 14 (4): 1352. https://doi.org/10.3390/app14041352.
Son, G., H. Lee, S. Kim, S. Kim, N. Muennighoff, T. Choi, C. Park, K. M. Yoo, and S. Biderman. 2024. "KMMLU: Measuring Massive Multitask Language Understanding in Korean." arXiv.
Suzuki, K., T. Laohakangvalvit, R. Matsubara, and M. Sugaya. 2021. "Constructing an Emotion Estimation Model Based on EEG/HRV Indexes Using Feature Extraction and Feature Selection Algorithms." *Sensors* 21 (9): 2910. https://doi.org/10.3390/s21092910.
Tao, K., Z. A. Osman, P. L. Tzou, S.-Y. Rhee, V. Ahluwalia, and R. W. Shafer. 2024. "GPT-4 Performance on Querying Scientific Publications: Reproducibility, Accuracy, and Impact of an Instruction Sheet." *BMC Medical Research Methodology* 24 (1): 139. https://doi.org/10.1186/s12874-024-02253-y.
Tixier, A. J.-P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2016. "Application of Machine Learning to Construction Injury Prediction." *Automation in Construction* 69:102–114. https://doi.org/10.1016/j.autcon.2016.05.016.
Toğan, V., F. Mostofi, Y. E. Ayözen, and O. Behzat Tokdemir. 2022. "Customized AutoML: An Automated Machine Learning System for Predicting Severity of Construction Accidents." *Buildings* 12 (11): 1933. https://doi.org/10.3390/buildings12111933.
Topal, M. O., A. Bas, and I. van Heerden. 2021. "Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet." arXiv.
Toptanci, Ş., N. ErgiNel, and I. P. Acar. 2023. *Predicting the Severity of Occupational Accidents in the Construction Industry Using Standard and Regularized Logistic Regression Models*. Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi. https://doi.org/10.28948/ngumuh.1212385.
Trillo Cabello, A., M. Martínez-Rojas, J. A. Carrillo-Castrillo, and J. C. Rubio-Romero. 2021. "Occupational Accident Analysis According to Professionals of Different Construction Phases Using Association Rules." *Safety Science* 144:105457. https://doi.org/10.1016/j.ssci.2021.105457.
Truong, A., A. Walters, J. Goodsitt, K. Hines, C. B. Bruss, and R. Farivar. 2019. "Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools." In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 1471–1479. Portland, OR, USA: IEEE.
Tsang, Y. T., I. W. H. Fung, V. W. Y. Tam, C. P. Sing, and C. T. Lu. 2017. "Development of an Accident Modelling in the Hong Kong Construction Industry." *International Journal of Construction Management* 17 (2): 124–131. https://doi.org/10.1080/15623599.2016.1222664.
Waehrer, G. M., X. S. Dong, T. Miller, E. Haile, and Y. Men. 2007. "Costs of Occupational Injuries in Construction in the United States." *Accident Analysis & Prevention* 39 (6): 1258–1266. https://doi.org/10.1016/j.aap.2007.03.012.
Wang, Q., T. Geng, Z. Wang, T. Wang, B. Fu, and F. Zheng. 2024. "Sample Then Identify: A General Framework for Risk Control and Assessment in Multimodal Large Language Models." arXiv.
Waring, J., C. Lindvall, and R. Umeton. 2020. "Automated Machine Learning: Review of the State-of-the-Art and Opportunities for Healthcare." *Artificial Intelligence in Medicine* 104:101822. https://doi.org/10.1016/j.artmed.2020.101822.
Xu, Q., and K. Xu. 2021. "Analysis of the Characteristics of Fatal Accidents in the Construction Industry in China Based on Statistical Data." *International Journal of Environmental Research and Public Health* 18 (4): 2162. https://doi.org/10.3390/ijerph18042162.
Xu, Y., Y. Zhou, P. Sekula, and L. Ding. 2021. "Machine Learning in Construction: From Shallow to Deep Learning." *Developments in the Built Environment* 6:100045. https://doi.org/10.1016/j.dibe.2021.100045.
Yoo, B., J. Kim, S. Park, C. R. Ahn, and T. Oh. 2024. "Harnessing Generative Pre-Trained Transformers for Construction Accident Prediction with Saliency Visualization." *Applied Sciences* 14 (2): 664. https://doi.org/10.3390/app14020664.
Zhang, J., K. Yu, Z. Wen, X. Qi, and A. Kumar Paul. 2021. "3D Reconstruction for Motion Blurred Images Using Deep Learning-Based Intelligent Systems." *Computers, Materials & Continua* 66 (2): 2087–2104. https://doi.org/10.32604/cmc.2020.014220.
Zhang, K., O. Kwon, and H. Xiong. 2023. "The Impact of Generative Artificial Intelligence." arXiv.
Zhang, Z., A. Zhang, M. Li, and A. Smola. 2022. "Automatic Chain of Thought Prompting in Large Language Models." arXiv.
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 Safety and Environmental Protection* 145:293–302. https://doi.org/10.1016/j.psep.2020.08.006.
Zhuhadar, L. P., and M. D. Lytras. 2023. "The Application of AutoML Techniques in Diabetes Diagnosis: Current Approaches, Performance, and Future Directions." *Sustainability* 15 (18): 13484. https://doi.org/10.3390/su151813484.