Journal of Science and Technology in Civil Engineering, HUCE, 2026, 20 (2S): 1–13

# DEEP LEARNING FOR CLASSIFYING AND PREDICTING RISK FACTORS IN RETROSPECTIVE CONSTRUCTION FATALITY REPORTS

Ali E. Esmail[a,*], John Gambatese[a]

[a]*School of Civil and Construction Engineering, Oregon State University, 101 Kearney Hall, Corvallis, USA*

***Article history:***
*Received 06/02/2026, Revised 24/4/2026, Accepted 08/5/2026*

**Abstract**

Despite advances in safety protocols, fatal incidents continue to occur in the construction industry. The National Institute for Occupational Safety and Health (NIOSH), Fatality Assessment and Control Evaluation (FACE) program in the U.S. documents hundreds of fatalities; however, the unstructured narrative nature of these reports constrains systematic analysis. This pilot study applies deep learning to classify and interpret fatal construction incidents from 265 NIOSH FACE reports, primarily using narrative text, with structured attributes supporting annotation and label development. The approach involved curating and labeling incident narratives, fine-tuning transformer-based models for supervised classification, and evaluating performance across four targets: Incident Type, Project Type, Incident Causation, and Temporary Structure Type. Results indicated strong performance for Incident Type (accuracy = 0.981 ± 0.019; macro-F1 = 0.933 ± 0.108), moderate performance for Temporary Structure Type (accuracy = 0.830 ± 0.013), and mixed outcomes for Incident Causation and Project Type due to semantic overlap and class imbalance. Overall, the findings demonstrate the feasibility of converting narrative fatality data into predictive insights and support the development of scalable, data-driven frameworks for improving construction safety research and intervention.

*Keywords:* construction safety; deep learning; fatality analysis; temporary structures; text classification.

https://doi.org/10.31814/stce.huce2026-20(2S)-01 

## 1. Introduction

Construction remains one of the most hazardous industries worldwide, characterized by complex operations, dynamic environments, and persistent safety challenges [1, 2]. Despite decades of progress in safety management systems and regulatory oversight, fatal incidents continue to occur at an alarming rate, reflecting ongoing limitations in risk anticipation, information management, and hazard recognition [3]. The National Institute for Occupational Safety and Health (NIOSH) Fatality Assessment and Control Evaluation (FACE) program in the U.S. has documented hundreds of fatal construction incidents, offering a valuable yet underutilized resource for understanding the mechanisms leading to worker deaths [4]. However, these reports are predominantly narrative and qualitative in nature, creating challenges for systematic analysis and large-scale pattern recognition. Traditional safety research has relied heavily on quantitative injury databases and structured field observations, which, while informative, often lack the contextual richness needed to capture the complex causal pathways underlying fatal incidents [5, 6]. Conversely, qualitative narratives from NIOSH investigations contain detailed descriptions of site conditions, equipment interactions, worker behavior, and procedural failures. These unstructured text reports represent an untapped source of data for computational safety analysis. Nevertheless, their non-standardized format makes it difficult to extract common insights using conventional statistical techniques [7]. As demonstrated in [8], the

*Corresponding author. *E-mail address:* esmaila@oregonstate.edu (Esmail, A. E.)

transition from traditional shallow learning methods to deep learning architectures in construction analytics highlights the need for domain-specific adaptations of NLP models.

Recent advancements in natural language processing (NLP) and deep learning have created opportunities to automate the extraction of meaningful information from unstructured text. Transformer-based architectures such as the Bidirectional Encoder Representations from Transformers (BERT) can learn linguistic representations from textual data, enabling classification and prediction tasks that were previously impractical [9]. Within the construction safety domain, these models can identify recurring patterns in fatality narratives, link descriptive phrases to causal categories, and enhance the interpretability of safety data [10]. Applying these methods to NIOSH reports can transform isolated case descriptions into actionable insights that support data-driven safety interventions. Moreover, the integration of NLP with other computational methods, such as data fusion and automated model checking, has shown promise in advancing construction safety analytics [11].

The present study describes a pilot implementation of a deep learning framework for analyzing fatal construction incident narratives. Using a dataset of 265 NIOSH reports, the study focuses on four key predictive tasks: (1) incident type, (2) project type, (3) incident causation, and (4) temporary structure type. The objectives were to (1) preprocess and label both structured and narrative fields to capture incident characteristics, (2) train supervised deep learning models using NLP techniques, and (3) evaluate model performance in predicting safety-related categories. By integrating computational methods with safety science, this research seeks to demonstrate the feasibility of automated narrative analysis for fatality investigations, offering a replicable foundation for broader applications in construction safety analytics.

## 2. Literature review

Construction safety research has historically focused on identifying the root causes of incidents, emphasizing behavioral factors, site conditions, and management practices [1, 2]. Traditional approaches, such as incident statistics, checklists, safety audits, and Behavior-Based Safety (BBS) frameworks, have provided valuable insights into high-risk activities and recurring hazards [3]. However, these methods rely heavily on structured and quantitative datasets, which, while informative, often fail to capture the complex causal relationships and contextual nuances present in real-world incidents. As a result, recent research has shifted toward more data-driven and computational methodologies, enabling predictive insights to be extracted from larger and more diverse information sources.

### *2.1. Natural language processing and deep learning in safety analytics*

The growing availability of unstructured text data, such as incident reports, investigation summaries, and inspection comments, has created new opportunities for applying NLP in safety analytics. Text mining techniques have been employed to classify, cluster, and summarize safety-related documents, allowing researchers to systematically extract risk factors, causal mechanisms, and emerging hazard trends from qualitative data [5, 7]. More recently, transformer-based deep learning architectures such as BERT have revolutionized text classification and contextual understanding [9]. In the construction domain, previous research [8] has shown that deep learning approaches extend beyond shallow models by capturing complex, non-linear relationships within textual and visual safety data, marking a paradigm shift toward intelligent data interpretation. These models capture semantic dependencies between words, enabling more accurate predictions than traditional keyword-based or bag-of-words approaches. In safety research, deep learning has shown potential for identifying injury causes, classifying construction accident narratives, and detecting latent patterns in unstructured data [10, 12–14].

While early applications of NLP in occupational safety focused on general injury types or root-cause extraction, there remains limited work addressing multi-category classification using narrative fatality data. Prior studies in construction safety NLP have primarily focused on single-label or independently modeled tasks (e.g., hazard classification or precursor extraction), with limited work addressing the simultaneous prediction of multiple interdependent outcomes such as incident type, causation, and project characteristics [7, 12–14]. For example, existing NLP-based safety research has primarily focused on extracting specific attributes such as injury precursors, hazard types, or accident categories independently, without capturing the interconnected nature of construction incidents [12–14]. Moreover, narrative data from programs such as NIOSH are often underused due to their qualitative nature and inconsistent linguistic structure [4, 12]. Prior work has advanced text-mining approaches for risk-factor extraction [7] and documented the transition from shallow to deep learning in construction analytics [8]; however, there remains limited evidence on multi-task narrative classification that simultaneously models incident type, causation, project context, and temporary-structure involvement in NIOSH FACE narratives.

### *2.2. Gaps in the literature and research focus*

Despite the promising results of deep learning and NLP applications, several key gaps remain. First, narrative-based fatality reports have not been systematically leveraged for predictive modeling. The FACE database, for instance, contains rich textual detail describing contextual and procedural failures in construction fatalities, yet most prior research has relied on coded or aggregated summaries rather than raw narrative content [4]. Although prior research [7] advanced text-mining approaches to extract safety risk factors from accident reports, these methods remained primarily rule-based and did not utilize deep contextual representations, limiting their ability to model complex linguistic dependencies within narrative data. Second, studies have rarely addressed small-sample and class-imbalance challenges that characterize fatality datasets. Many NLP models assume large, balanced corpora; however, real-world safety data often exhibit skewed class distributions, where some incident types or project categories occur far more frequently than others, reducing model stability and generalizability [5]. Prior work [8] has highlighted that construction datasets often suffer from data sparsity and heterogeneity, making the application of deep learning methods particularly challenging without careful model calibration and validation strategies. Third, few studies have evaluated NLP-based safety models using robust experimental designs, such as repeated stratified validation across multiple random seeds, to confirm reproducibility. Performance variability due to random partitioning remains underexplored, particularly in small safety datasets. Addressing these gaps, the present study introduces a supervised deep learning framework designed to classify fatal construction narratives into four interrelated categories: incident type, project type, incident causation, and temporary structure type. By integrating manual annotation, transformer-based text encoding, and stratified validation, this research contributes a replicable methodological foundation for transforming unstructured fatality narratives into structured, predictive insights. The study thus extends the current body of knowledge by demonstrating the feasibility and interpretability of NLP-driven fatality analysis within the context of construction safety research.

## 3. Research methodology

This study employed a structured five-phase methodological framework to transform unstructured fatality narratives into predictive insights using NLP and supervised deep learning. The framework was designed to ensure a systematic and reproducible workflow, consisting of (1) data collection and curation, (2) data preprocessing and labeling, (3) model development, (4) model evaluation, and (5)

results synthesis and interpretation, as illustrated in Fig. 1. This phased approach is consistent with established machine learning and NLP pipelines, which typically involve data acquisition, preprocessing, model development, evaluation, and interpretation to ensure methodological transparency and robustness [5, 8]. The use of transformer-based architecture, specifically RoBERTa, was motivated by its strong performance in text classification tasks and its ability to capture contextual dependencies through pretrained bidirectional language representations [15]. Pretrained transformer models are particularly advantageous in settings with limited labeled data, as they leverage knowledge learned from large-scale corpora to enhance downstream prediction performance. This is especially relevant for the present study, which utilizes a relatively small dataset of 265 fatality narratives. Compared to traditional approaches such as bag-of-words or shallow learning models, RoBERTa enables richer semantic understanding of domain-specific language, allowing the model to identify nuanced patterns related to incident causation, project context, and safety conditions embedded within narrative descriptions.

### *3.1. Phase 1. Data collection and curation*

Fatal construction incident reports were obtained from the NIOSH FACE program website. The dataset included 265 federally published FACE reports describing circumstances, tasks, and causal factors leading to worker fatalities. State-level and engineering investigation reports were excluded to ensure consistency in reporting format and data completeness. Both unstructured textual narratives and structured case attributes were manually extracted, cleaned, and integrated into a unified database. The objective of this phase was to establish a coherent dataset that links qualitative narrative content with quantitative descriptors for each incident.

Although both structured attributes and narrative text were compiled into a unified database, the structured variables (e.g., project type, incident classifications, and causation categories) were used exclusively to support the labeling process and define ground truth targets. The predictive model itself utilized only the unstructured narrative text as input. No feature fusion or multimodal integration was implemented, as the objective of this study was to evaluate the capability of transformer-based models to extract safety-relevant information directly from textual narratives.

### *3.2. Phase 2. Data preprocessing and labeling*

The narrative text from each incident report was tokenized, normalized, and non-informative segments, such as metadata, stop words, and repetitive boilerplate phrases, were removed. Each narrative was then manually annotated into four primary classification targets (“heads”), representing key analytical categories in construction fatalities: (1) incident type—contact with object and equipment, electrocution, falls, slips, and trips, exposure to harmful substances or environments, and fires and explosions; (2) project type—building, residential, industrial, and heavy engineering; (3) incident cause—blunt force or trauma injuries, crush/compression/asphyxia injuries, drowning or non-crush asphyxiation, electrocution or electrical burns, falls from height, and other or medical/toxic causes; and (4) temporary structure type—scaffolding, formwork, excavation support, and non-temporary. To mitigate class imbalance, the frequency of cases within each category was analyzed to determine weighting factors during model training, ensuring that minority categories received appropriate representation and that model learning remained balanced across all fatality types.

To further clarify the operational process, the transformation from raw narrative data to model predictions follows a structured pipeline. Each case begins with a raw NIOSH FACE narrative describing the incident circumstances, worker activities, and environmental conditions. The narrative is first preprocessed through tokenization, normalization, and removal of non-informative content. It

is then annotated into predefined classification categories (incident type, project type, incident causation, and temporary structure type) using the developed codebook. The processed narrative text is subsequently used as input to the transformer-based model, where contextual embeddings are generated and passed to task-specific classification heads. The model outputs predicted class labels along with associated confidence scores for each classification target.

For example, a narrative describing workers being crushed by a collapsing masonry wall due to wind-induced instability is processed and classified by the model as Contact with Object and Equipment (incident type), Heavy Engineering Construction (project type), Blunt Force/Trauma Injuries (causation), and Formwork (temporary structure type). This illustrates how unstructured narrative descriptions are systematically converted into structured, multi-dimensional safety insights across multiple classification tasks.

a. Annotation protocol and quality control

To ensure label quality and reproducibility, a codebook was developed to define class boundaries for each target (incident type, project type, incident causation, temporary structure type), including positive and negative examples as well as edge-case rules (e.g., equipment-assisted falls). The use of a structured codebook is consistent with established qualitative research practices, where clearly defined coding schemes are essential to ensure consistency, transparency, and replicability of thematic classification. Prior to formal annotation, annotators participated in structured calibration sessions using a pilot subset of reports to align interpretation of the codebook, resolve ambiguities, and standardize classification decisions across all targets. Employing multiple annotators and evaluating inter-rater agreement is a widely adopted approach in annotation-based studies to validate labeling reliability. A stratified subset of 15% of the dataset was selected for dual annotation, which aligns with common practices in NLP and qualitative coding studies that balance reliability assessment with practical constraints in manual labeling efforts. Cohen's kappa was used to assess inter-annotator agreement, as it accounts for chance agreement and is widely recommended for categorical classification tasks in both qualitative research and machine learning annotation studies.

Two trained annotators independently labeled a 15% stratified subset of reports; disagreements were resolved via adjudication. Inter-annotator agreement was computed using Cohen's $\kappa$ for each target, achieving $\kappa \geq 0.78$ for incident type and $\kappa \geq 0.70$ for incident causation, indicating substantial agreement. Project type and temporary structure type yielded $\kappa \approx 0.63$–$0.68$ due to semantic overlap and sparse minority classes. The adjudicated subset served as calibration references during the remaining single-annotator pass. The final label distribution informed the class-weighting used in model training.

### *3.3. Phase 3. Model Development*

Table 1 presents the class distribution and corresponding weights used to address class imbalance in model training. Weights were computed using an inverse-frequency approach ($w_i = N/n_i$), where $N$ denotes the total number of samples and $n_i$ represents the number of observations in class $i$. Frequencies shown are based on the full dataset ($N = 265$), while weights applied during model training were computed within each training fold using the same formulation. Extremely rare classes resulted in large weight values; however, these were retained to preserve sensitivity to minority classes while maintaining stable model optimization during training. This approach is commonly used in imbalanced classification settings to improve minority class learning without altering the underlying data distribution.

Table 1. Class distribution and corresponding weights used in model training ($N$ = 265)

| Classification Task | Class | Frequency ($n$) | Weight |
|---|---|---|---|
| Incident type | falls, slips, and trips | 107 | 2.48 |
| | contact with object and equipment | 62 | 4.27 |
| | exposure to harmful substances or environment | 95 | 2.79 |
| | Fires and Explosions | 1 | 265 |
| Project type | building construction | 95 | 2.79 |
| | residential construction | 37 | 7.16 |
| | industrial construction | 63 | 4.21 |
| | heavy engineering construction | 70 | 3.79 |
| Incident causation | blunt force / trauma injuries | 67 | 3.96 |
| | crush / compression / asphyxia injuries | 34 | 7.79 |
| | drowning / asphyxiation (non-crush) | 9 | 29.44 |
| | electrocution / electrical burns | 86 | 3.08 |
| | falls from height | 68 | 3.90 |
| | Other / Medical / Toxic | 1 | 265 |
| Temporary structure type | Falsework | 3 | 88.33 |
| | formwork | 7 | 37.86 |
| | scaffolding | 30 | 8.83 |
| | support of excavation | 10 | 26.50 |
| | non-temporary structure | 215 | 1.23 |

RoBERTa-base transformer architecture was employed as the foundation for model training. The encoder was frozen, and four task-specific classification heads were fine-tuned independently to predict each categorical variable. Training parameters were empirically optimized, including a learning rate of $5 \times 10^{-5}$, batch size = 16, maximum token length = 384, and label smoothing = 0.05. Each head was trained for 12 epochs to capture linguistic patterns embedded in the fatality narratives and to predict structured safety-related categories.

a. Training, reproducibility, leakage controls

Implementation utilized Hugging Face Transformers with RoBERTa-base as the frozen encoder and four task-specific classification heads trained independently. Structured attributes (e.g., project type, weather conditions, and training information) were used during dataset annotation and label construction but were not directly incorporated as input features in the model, which relied solely on narrative text for classification. To prevent data leakage, all preprocessing steps (tokenization, truncation to 384 tokens, and label mapping) were performed after each stratified split; no text from the validation fold was used during training. The repeated 80/20 stratified evaluation employed fixed random seeds {42, 77, 123, 222, 314}. Performance was reported as mean ± standard deviation (SD) across seeds for Accuracy and Macro-F1. Early stopping was not applied to ensure consistent optimization schedules across runs; instead, training was conducted for a fixed 12 epochs with label smoothing set to 0.05. Class weights were computed using training-fold frequencies only. Hyperparameters (learning rate = $5 \times 10^{-5}$, batch size = 16) were selected based on pilot experiments conducted on a held-out subset not included in the five evaluation replicates. Training was performed on a single GPU with mixed precision enabled [15].

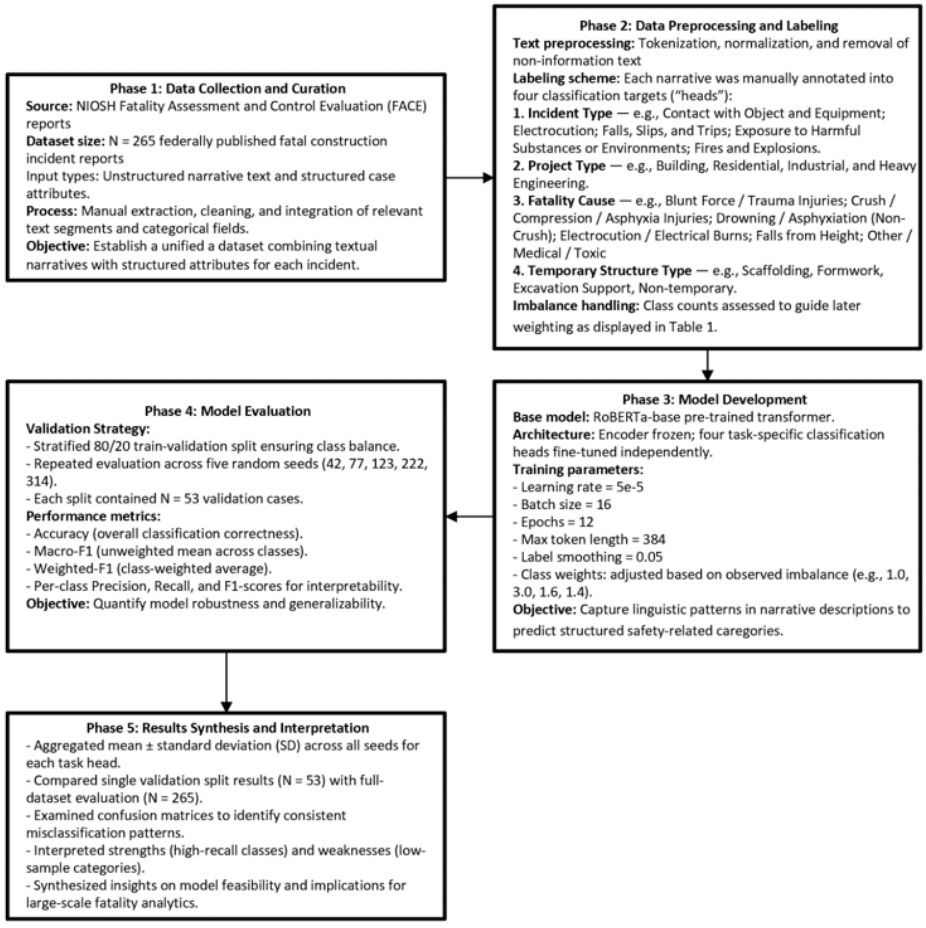


Figure 1. Methodological framework for the study

### *3.4. Phase 4. Model Evaluation*

Model performance was assessed using a repeated stratified 80/20 train–validation procedure to maintain class balance while incorporating randomization. Five randomized seeds (42, 77, 123, 222, 314) yielded distinct yet stratified data partitions, enabling robust generalization estimates. Each validation subset contained 53 incidents. Evaluation metrics included accuracy (overall classification correctness), macro-F1 (unweighted mean across classes), and weighted-F1 (class-weighted average). The F1-score represents the harmonic mean of precision and recall, providing a balanced measure of classification performance, particularly under uneven class distributions. Macro-F1 computes the unweighted average of F1-scores across all classes, assigning equal importance to both majority and minority categories, and is therefore especially appropriate for evaluating performance under class imbalance conditions. Per-class precision, recall, and F1-scores were analyzed to interpret class-level behavior. Mean ± standard deviation (SD) across seeds was reported to quantify robustness and generalizability.

### *3.5. Sample adequacy and data distribution*

Traditional statistical power analyses are not directly applicable to supervised deep learning models trained on archival datasets. Instead, sample adequacy in this study was evaluated based on the distributional balance and representativeness of the available reports. The dataset consisting of 265

NIOSH FACE narratives was sufficient to train and validate lightweight classification heads on a frozen RoBERTa-base encoder. While dominant categories such as Non-temporary Structures and Falls provided ample linguistic variability for model learning, minority classes (e.g., Scaffolding, Formwork, Drowning) remained underrepresented, limiting per-class generalization. Model robustness was therefore assessed through repeated stratified validation across multiple random seeds and reporting of mean ± SD metrics, serving as a practical analog to power assessment in small-sample machine learning studies.

### *3.6. Phase 5. Results synthesis and interpretation*

Results from the five repeated stratified validation runs were aggregated using the mean ± SD for each classification task (head). Model performance on the representative validation subset ($n$ = 53) was compared with full-dataset inference ($N$ = 265) to evaluate stability and scalability. Confusion matrices were analyzed to identify systematic misclassification patterns, most notably, overlap between Industrial and Building Construction categories. High-recall classes demonstrated distinctive linguistic patterns in the fatality narratives, whereas low-frequency categories reflected the effects of limited sample representation. These evaluation outcomes collectively informed the interpretation of model feasibility, predictive reliability, and potential scalability of data-driven fatality analytics for larger occupational safety datasets.

## 4. Results and discussions

The developed deep learning framework was evaluated across four primary classification tasks – incident type, project type, incident causation, and temporary structure type – to assess its ability to extract predictive patterns from unstructured fatality narratives. Model performance was measured using accuracy, macro-F1 (unweighted mean across classes), and weighted-F1 (class-proportional mean). Results were obtained through five independent stratified 80/20 train–validation splits, each preserving class balance and evaluated on validation subsets of 53 incidents from the total dataset of 265 reports. Table 2 presents the aggregated results averaged across the five randomized runs (Seeds: 42, 77, 123, 222, and 314), with mean ± SD reported for both accuracy and macro-F1 to capture generalization stability across partitions.

Table 2. Aggregated classification performance across tasks

| Task (Head) | Accuracy (mean ± SD) | Macro-F1 (unweighted; mean ± SD) |
|---|---|---|
| Incident type | 0.9811 ± 0.0189 | 0.9325 ± 0.1076 |
| Project type | 0.4679 ± 0.1062 | 0.3984 ± 0.1011 |
| Incident causation | 0.7057 ± 0.0510 | 0.5172 ± 0.0739 |
| Temporary structure type | 0.8302 ± 0.0133 | 0.3028 ± 0.0561 |

As shown in Table 2, the model achieved the highest performance in predicting Incident Type (accuracy = 0.9811 ± 0.0189; macro-F1 = 0.9325 ± 0.1076), followed by Temporary Structure Type (accuracy = 0.8302 ± 0.0133) and Incident Causation (accuracy = 0.7057 ± 0.0510). The relatively lower accuracy for Project Type (0.4679 ± 0.1062) reflects greater semantic overlap in textual descriptions across building, residential, and industrial projects, making it more difficult to distinguish from narrative context alone. The minor standard deviations across seeds indicate stable generalization performance across repeated stratified splits. Table 3 presents replicate-level results for all five randomized seeds (42, 77, 123, 222, 314), illustrating variation in performance due to random partitioning.

The observed discrepancy between accuracy (0.8302 ± 0.0133) and Macro-F1 (0.3028 ± 0.0133) for the Temporary Structure Type classification task is primarily attributable to class imbalance in the dataset. The majority of cases fall under the "non-temporary structure" category, resulting in high overall accuracy driven by correct predictions of the dominant class. However, Macro-F1, which equally weights performance across all classes, reveals reduced predictive performance for minority classes representing specific temporary structure types. Despite the use of class weighting during training, the limited representation of these classes constrains the model's ability to generalize across all categories. These findings highlight the challenges of modeling sparse and imbalanced safety data and suggest that additional strategies, such as data augmentation or hierarchical classification, may further improve minority class performance.

Table 3. Replicate-Level Validation Results Across Random Seeds

| Task (Head) | Replicate (Seed) | Accuracy | Macro-F1 |
|---|---|---|---|
| Incident type | 1 (42) | 1.0000 | 1.0000 |
| | 2 (77) | 0.9623 | 0.9633 |
| | 3 (123) | 0.9623 | 0.9556 |
| | 4 (222) | 0.9811 | 0.7436 |
| | 5 (314) | 1.0000 | 1.0000 |
| Project type | 1 (42) | 0.4151 | 0.3321 |
| | 2 (77) | 0.6226 | 0.5461 |
| | 3 (123) | 0.5283 | 0.4529 |
| | 4 (222) | 0.3585 | 0.2952 |
| | 5 (314) | 0.4151 | 0.3659 |
| Incident causation | 1 (42) | 0.7736 | 0.5225 |
| | 2 (77) | 0.6415 | 0.4729 |
| | 3 (123) | 0.7358 | 0.6425 |
| | 4 (222) | 0.6792 | 0.4589 |
| | 5 (314) | 0.6981 | 0.4894 |
| Temporary structure type | 1 (42) | 0.8113 | 0.2240 |
| | 2 (77) | 0.8302 | 0.3131 |
| | 3 (123) | 0.8302 | 0.3002 |
| | 4 (222) | 0.8491 | 0.3816 |
| | 5 (314) | 0.8302 | 0.2949 |

The results show consistent performance across replications, confirming the robustness of the deep learning framework. Incident Type predictions maintained near-perfect scores across all seeds (macro-F1 $\approx$ 1.0), while Temporary Structure Type achieved moderate yet stable performance (macro-F1 $\approx 0.30 \pm 0.06$). Larger variability observed in Project Type and Incident Causation tasks indicates model sensitivity to category representation and narrative diversity. These findings further emphasize the influence of dataset imbalance, particularly when low-frequency classes are underrepresented or absent in validation subsets. To interpret class-level behavior, Table 4 presents precision, recall, and F1-scores for each class in a representative validation run (seed 42; $n = 53$).

For Incident Type, the model demonstrated perfect classification across the three dominant categories—Falls, Slips, and Trips; Contact with Object and Equipment; and Exposure to Harmful Sub-

Table 4. Per-class performance metrics for a representative validation split (seed 42; $n = 53$)

| Task (Head) | Class | Precision | Recall | F1 | Support | Macro-F1 |
|---|---|---|---|---|---|---|
| Incident type | falls, slips, and trips | 1.000 | 1.000 | 1.000 | 15 | 1.000 |
| | contact with object and equipment | 1.000 | 1.000 | 1.000 | 12 | |
| | exposure to harmful substances or environment | 1.000 | 1.000 | 1.000 | 26 | |
| Project type | building construction | 0.533 | 0.421 | 0.471 | 19 | 0.332 |
| | residential construction | 0.000 | 0.000 | 0.000 | 7 | |
| | industrial construction | 0.263 | 0.385 | 0.312 | 13 | |
| | heavy engineering construction | 0.474 | 0.643 | 0.545 | 14 | |
| Incident causation | blunt force / trauma injuries | 0.429 | 0.375 | 0.400 | 8 | 0.523 |
| | crush / compression / asphyxia injuries | 0.400 | 0.333 | 0.364 | 6 | |
| | drowning / asphyxiation (non-crush) | 0.000 | 0.000 | 0.000 | 3 | |
| | electrocution / electrical burns | 0.923 | 1.000 | 0.960 | 24 | |
| | falls from height | 0.800 | 1.000 | 0.889 | 12 | |
| Temporary structure type | formwork | 0.000 | 0.000 | 0.000 | 2 | 0.224 |
| | scaffolding | 0.000 | 0.000 | 0.000 | 5 | |
| | support of excavation | 0.000 | 0.000 | 0.000 | 3 | |
| | non-temporary structure | 0.811 | 1.000 | 0.896 | 43 | |

stances or Environment—indicating that linguistic cues describing fall mechanisms, electrocution, and exposure events were highly distinctive in the narratives. However, Fires and Explosions incidents, though present in the overall dataset ($N = 265$), were not represented in this particular validation subset ($n = 53$) and therefore are omitted from the table. Their absence reflects natural class imbalance in the sampled partition rather than model omission. The model was trained in these cases, but their rarity prevented validation-level evaluation in this specific run. With respect to Project Type, performance was uneven, with moderate recall for Building and Heavy Engineering Construction, but weaker generalization for Residential and Industrial Construction, where shared contextual wording (e.g., "foundation," "crew," "installation") limited differentiation. For Incident Causation, Electrocution and Falls from Height achieved high precision ($\geq 0.80$), while low-frequency classes such as Drowning or Asphyxiation yielded F1 $\approx 0$, reflecting insufficient examples for pattern learning. Similarly, for Temporary Structure Type, the model effectively identified Non-temporary Structure narratives (precision = 0.811; recall = 1.000) but underperformed on Formwork, Scaffolding, and Excavation Support, again due to data scarcity. These results highlight the importance of increasing sample diversity to improve recognition of minority event types and specialized structural categories. To assess scalability and overall model generalization, Table 5 compares performance between the validation subset ($n = 53$) and full-dataset inference ($N = 265$).

The comparison shows that full-dataset inference ($N = 265$) preserved the same performance trends observed in the validation subset ($n = 53$). Incident Type retained superior accuracy (0.9698),

Table 5. Comparison of Validation-Set and Full-Dataset Performance Across Classification Tasks

| Task (Head) | Accuracy (Validation, $n = 53$) | Accuracy (Full Dataset, $N = 265$) |
|---|---|---|
| Incident Type | 0.9434 | 0.9698 |
| Project Type | 0.5849 | 0.5056 |
| Incident Causation | 0.8113 | 0.6716 |
| Temporary Structure Type | 0.8113 | 0.8528 |

followed by Temporary Structure Type (0.8528), demonstrating model scalability and stability. Minor reductions in Incident Causation (−0.14) and Project Type (−0.08) suggest that limited data diversity and class imbalance remain the primary constraints to generalization. Overall, the findings confirm the feasibility of extracting structured safety intelligence from unstructured fatality narratives. Detailed practical implications and applications of these findings are discussed in Section 5. Remaining challenges related to semantic overlap across project types and the limited representation of rare causal and structural categories highlight the need for dataset expansion, improved annotation granularity, and enhanced feature representation. Future work will focus on enlarging the dataset and applying interpretability methods to better understand the linguistic features driving classification outcomes.

## 5. Practical implications

Full-dataset inference ($N = 265$) preserved the performance hierarchy observed in validation ($n = 53$), with Incident Type highest, followed by Temporary Structure Type. Differences between validation and full-dataset scores were modest, suggesting effective generalization despite class imbalance. Incident Type maintained superior accuracy (0.9698), followed by Temporary Structure Type (0.8528). Minor differences between validation and full-dataset scores indicate that the model generalized effectively despite sample imbalance. Slight reductions in Incident Causation and Project Type accuracies (−0.14 and −0.08, respectively) suggest that expanding the dataset or applying class-weighted fine-tuning could further improve predictive reliability. Overall, the results confirm the feasibility of using deep learning and NLP to extract structured safety insights from fatality narratives. The near-perfect prediction of Incident Type demonstrates strong linguistic regularities in fatal event reporting, while moderate success in Temporary Structure Type classification shows potential for integrating textual and structured data in safety analytics. Nonetheless, challenges in modeling semantically overlapping project descriptions and underrepresented causal factors underscore the need for larger, more balanced datasets. Future work will focus on expanding data coverage, refining annotation granularity, and applying interpretability tools to reveal the narrative features driving each prediction. Practically, three levers are likely to yield the largest gains: (1) dataset expansion with targeted oversampling of rare temporary-structure and causation classes; (2) hierarchical taxonomies (e.g., coarse-to-fine project types) to reduce semantic overlap; and (3) lightweight domain prompts or vocabularies injected via adapters or prompt-tuning to stabilize minority-class representations without unfrozen encoders.

Beyond methodological contributions, the proposed framework has direct implications for construction safety practice. First, the ability to automatically classify fatality narratives can support the development of targeted safety training programs by identifying recurring incident types, causation patterns, and high-risk activities embedded in historical reports. This enables safety managers to design data-driven training interventions tailored to specific hazards, such as falls, electrocution, or temporary structure failures. Second, the model can be integrated into automated risk detection sys-

tems, where new incident reports, inspection notes, or near-miss descriptions are analyzed in real time to flag high-risk conditions and emerging hazard patterns. Such capabilities can enhance proactive safety management by shifting from reactive incident analysis to predictive risk monitoring. Third, the framework can be incorporated into digital safety management systems and knowledge platforms, enabling centralized analysis of safety narratives across projects and organizations. This integration supports continuous learning, improved hazard communication, and more informed decision-making at both the project and organizational levels.

## 6. Conclusions and recommendations

### *6.1. Summary of Key Findings*

This study demonstrated the feasibility of applying deep learning and natural language processing (NLP) to extract structured safety insights from fatality narratives within the NIOSH FACE program. Using a dataset of 265 incident reports, the proposed framework successfully classified four key dimensions of construction fatalities—Incident Type, Project Type, Incident Causation, and Temporary Structure Type—based primarily on narrative text, with structured attributes supporting annotation and label development. The results indicated high predictive performance for Incident Type (mean accuracy = 0.9811 ± 0.0189; macro-F1 = 0.9325 ± 0.1076), suggesting that linguistic patterns describing falls, electrocution, and exposure events are consistent and readily distinguishable within narrative reports. Moderate performance in Temporary Structure Type classification (accuracy = 0.8302 ± 0.0133; macro-F1 = 0.3028 ± 0.0561) reflects the influence of class imbalance and limited representation of specific structural categories, while still demonstrating the potential of identifying structural risk factors from textual descriptions. Lower performance in Project Type (accuracy = 0.4679 ± 0.1062) and Incident Causation (accuracy = 0.7057 ± 0.0510) highlights challenges associated with semantic overlap across categories and variability in narrative expressions. Overall, the findings confirm that deep learning–based NLP models can transform unstructured fatality narratives into quantifiable and analyzable safety information. This capability enables more systematic identification of risk patterns, supports large-scale safety analytics, and provides a scalable foundation for integrating AI-driven methods into construction safety research and management practices.

### *6.2. Limitations and Recommendations*

Despite promising results, several limitations must be acknowledged. First, the dataset was limited to 265 reports drawn exclusively from the federal NIOSH database, excluding state-level or engineering investigation reports. This limitation constrained data diversity and introduced class imbalance, particularly in rare categories such as Fires and Explosions or Scaffolding-related incidents. Second, while the study employed manual annotation and stratified randomization, future applications should explore automated labeling and active learning techniques to enhance scalability and consistency. Third, the NLP model analyzed narrative text but did not incorporate temporal, environmental, or human-factor variables that often influence incident causation. To improve predictive reliability, future implementations should (1) integrate broader datasets across agencies and jurisdictions, (2) apply data augmentation or resampling to balance minority classes, and (3) fine-tune model architectures (e.g., transformer-based contextual embeddings) to capture subtle linguistic differences among overlapping categories.

### *6.3. Future Research Directions*

Building on this pilot study, several research extensions are recommended. First, future work should expand the dataset to include state-level FACE reports, OSHA fatality records, and engineering investigation narratives, enabling multi-source model training and cross-validation. Second, integrating quantitative descriptors, such as project duration, crew size, or equipment type, can support

multimodal learning that fuses textual and numerical predictors. Third, incorporating explainable Artificial Intelligence (XAI) methods (e.g., SHAP or LIME) would allow interpretation of linguistic patterns driving model decisions, promoting trust and transparency in practical applications. Finally, embedding the trained models into automated safety management systems could enable real-time fatality classification, early warning generation, and targeted training interventions, transforming how safety data is utilized for prevention and policy design. Collectively, these directions aim to advance the integration of computational analytics with construction safety research, supporting a transition from retrospective investigation to proactive and data-informed risk mitigation across the construction industry.

## References


[1] Hallowell, M. R., Gambatese, J. A. (2009). Activity-Based Safety Risk Quantification for Concrete Formwork Construction. *Journal of Construction Engineering and Management*, 135(10):990–998.

[2] Zhou, Z., Goh, Y. M., Li, Q. (2015). Overview and analysis of safety management studies in the construction industry. *Safety Science*, 72:337–350.

[3] Carra, S., Bottani, E., Vignali, G., Madonna, M., Monica, L. (2024). Implementation of Behavior-Based Safety in the Workplace: A Review of Conceptual and Empirical Literature. *Sustainability*, 16(23):10195.

[4] Accessed October 12, 2025 (2024). *Fatality Assessment and Control Evaluation (FACE) Program*. National Institute for Occupational Safety and Health (NIOSH).

[5] Khairuddin, M. Z. F., Hasikin, K., Abd Razak, N. A., Lai, K. W., Osman, M. Z., Aslan, M. F., Sabanci, K., Azizan, M. M., Satapathy, S. C., Wu, X. (2022). Predicting occupational injury causal factors using text-based analytics: A systematic review. *Frontiers in Public Health*, 10:984099.

[6] Sunindijo, R. Y., Zou, P. X. W. (2012). Political Skill for Developing Construction Safety Climate. *Journal of Construction Engineering and Management*, 138(5):605–612.

[7] Xu, N., Ma, L., Liu, Q., Wang, L., Deng, Y. (2021). An improved text mining approach to extract safety risk factors from construction accident reports. *Safety Science*, 138:105216.

[8] Xu, Y., Zhou, Y., Sekula, P., Ding, L. (2021). Machine learning in construction: From shallow to deep learning. *Developments in the Built Environment*, 6:100045.

[9] Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Association for Computational Linguistics, 4171–4186.

[10] Wang, Y., Huang, Y., Gu, B., Cao, S., Fang, D. (2023). Identifying mental fatigue of construction workers using EEG and deep learning. *Automation in Construction*, 151:104887.

[11] Cheng, T., Migliaccio, G. C., Teizer, J., Gatti, U. C. (2013). Data Fusion of Real-Time Location Sensing and Physiological Status Monitoring for Ergonomics Analysis of Construction Workers. *Journal of Computing in Civil Engineering*, 27(3):320–335.

[12] Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., Bowman, D. (2016). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, 62:45–56.

[13] Zhang, F., Fleyeh, H., Wang, X., Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, 99:238–248.

[14] Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B. (2017). Construction Safety Risk Modeling and Simulation. *Risk Analysis*, 37(10):1917–1935.

[15] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.