

Article

Multimodal and Explainable Deep Learning for Occupational Accident Classification Using Transformer-LSTM Architectures

Esin Ayşe Zaimoğlu 

Department of Information Systems Engineering, Faculty of Computer and Informatics Sciences, Sakarya University, 54187 Serdivan, Sakarya, Turkey; esinzaimoglu@sakarya.edu.tr

Abstract

Occupational safety analytics is increasingly moving toward data-driven methodologies; however, existing models often struggle to capture the multidimensional nature of accident causation. This study presents a multimodal Hybrid Transformer-LSTM framework for classifying occupational fatalities by jointly modeling unstructured narratives, cyclical temporal features, and regional spatial indicators. Utilizing a large-scale dataset of 14,914 OSHA fatality records, the proposed architecture leverages BERT-based embeddings for semantic extraction and Bidirectional LSTMs as non-linear pattern encoders for spatiotemporal context. Conceptually grounded in the Swiss Cheese Model, the framework treats different data modalities as proxies for distinct layers of system risk, ranging from proximal unsafe acts to environmental preconditions. Experimental results show that the multimodal architecture achieves an accuracy of 84.56%, representing a 5.33% gain over unimodal BERT baselines. To address the inherent “black-box” nature of deep learning, a SHAP-based explainability framework is incorporated to quantify the contributions of both textual tokens and environmental features to the model’s decision-making process. The results indicate that integrating narrative semantics with temporal and spatial context enhances discriminative performance and enables context-aware classification within a weakly supervised setting. By providing a scalable and interpretable classification framework, this study offers a data-driven decision-support approach for safety professionals and regulatory bodies seeking to implement evidence-based risk management strategies in high-risk industrial sectors.

Keywords: construction safety; building industry; multimodal learning; transformer networks; LSTM; accident classification; explainable artificial intelligence (XAI)



Academic Editor: Pramen P. Shrestha

Received: 13 March 2026

Revised: 17 April 2026

Accepted: 19 April 2026

Published: 22 April 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Occupational safety remains a critical concern across industrial sectors, particularly in construction, where safety failures frequently result in severe or fatal outcomes. In the United States alone, 5190 fatal workplace incidents were reported in 2021, highlighting the need for more robust safety analyses and preventive strategies [1]. The Occupational Safety and Health Administration (OSHA) maintains comprehensive accident databases to understand accident patterns. However, extracting actionable insights from large-scale, heterogeneous datasets remains a major challenge in modern safety management. Traditional analytical approaches, such as manual classification and basic statistical methods, offer limited support to safety management. Therefore, they are inadequate for identifying the inherent complexity, contextual nuances, and risks found in large-scale safety data. Furthermore, these methods lack a systematic approach and remain limited for performing

repeatable analyses [2,3]. Current data on occupational accidents typically include textual narratives, temporal cycles, and geographic information. However, conventional methods are often insufficient to explain the multifaceted nature of these accidents.

The primary bottleneck in current occupational safety analytics is the “contextual gap” in automated classification. Although recent advances in Deep Learning (DL) and Natural Language Processing (NLP) offer promising solutions, most existing models treat accident narratives as isolated textual units [4]. Transformer-based architectures such as BERT have achieved state-of-the-art performance in semantic information extraction tasks [5], while Long Short-Term Memory (LSTM) networks are effective in capturing sequential dependencies [6]. Current research often focuses on either text analysis or temporal pattern recognition, but rarely integrates these methods within a unified framework. In reality, occupational accident reports are shaped by intertwined temporal, spatial and contextual factors. Despite this, there remains a lack of an integrated and explainable multimodal framework capable of jointly modeling narrative semantics, temporal cyclicity, and geographic heterogeneity in large-scale OSHA fatality datasets.

The need for such multidimensional analysis is further supported by a number of theoretical frameworks, particularly the Swiss Cheese Model (SCM) proposed by Reason [7,8]. This model contends that accidents are triggered by the simultaneous combination of vulnerabilities in successive layers of defense and hidden systemic flaws. However, current DL methodologies are insufficient to functionally integrate these interconnected dynamics, as defined by Reason, within narrative, temporal, and spatial data representations. In this context, to bridge the gap between theoretical models and real-world data, it is necessary to turn to “problem-oriented” approaches that integrate contextual meaning with representations of data [9,10]. In response, this study proposes a multimodal hybrid DL architecture specifically designed to analyze the complex, heterogeneous nature of fatal workplace accidents.

1.1. Research Motivation and Rationale

The primary motivation of this study is to systematically interpret unstructured occupational accident narratives. Although OSHA has compiled comprehensive incident databases, the fact that these records, which contain technical terminology and linguistic variations, have not yet been interpreted hinders systematic analysis. Traditional statistical approaches are insufficient for uncovering the contextual relationships embedded in these narratives, while reliance on manual review processes delays the development of risk models due to subjectivity and workload constraints.

In the construction sector, AI-driven approaches have recently gained significant momentum, and innovative applications have brought safety monitoring and risk assessment to the forefront [11,12]. Transformer-based models, in particular, have demonstrated significant improvements in data narrative analysis compared to conventional methods. However, the limited explainability and transparency of high-performing DL architectures remain critical challenges. In safety-critical domains, predictive accuracy alone is insufficient to ensure organizational trust. Supervisors and safety professionals cannot verify and act upon automated predictions without understanding the logic behind them. This limited explainability remains a significant obstacle to the practical implementation of AI-based safety analytics.

1.2. Related Work and Research Gaps

Occupational accident analysis has transitioned from descriptive, statistically driven methods to data-driven, learning-based approaches. Early studies primarily relied on structured variables, which were limited in their ability to model complex causal mecha-

nisms [13]. Subsequent research introduced classical machine learning (ML) techniques, such as Random Forests (RF) and Naive Bayes (NB) [14,15], to improve predictive performance; however, these approaches still relied heavily on manual feature engineering [16]. Recognizing that unstructured narratives contain rich safety information, NLP techniques have gained prominence, moving beyond shallow representations such as TF-IDF [17,18] toward DL architectures such as LSTM [3,14,15] and BERT [5,19].

In industrial safety, hybrid architectures combining language modeling with sequential learning have shown measurable performance gains. The construction industry, characterized by its dynamic, complex, and temporary nature, consistently ranks among the most hazardous global sectors. Approximately one in six fatal workplace accidents occurs on construction sites, highlighting the urgent need for systematic “Learning from Incidents” (LFI) processes to transform past lessons into preventive strategies [20–22]. Zhang et al. [23] and Tian et al. [24] demonstrated that integrating Convolutional Neural Networks (CNNs) or LSTMs with Transformer models improves classification accuracy in construction and hazard detection tasks [25]. Similarly, hybrid models have been shown to be effective for detecting unsafe behavior and supporting safety monitoring [26]. Beyond accuracy, recent research has emphasized semantic coherence; for instance, Cao et al. [12] demonstrated that Transformer-based topic modeling with BERTopic substantially outperforms traditional methods in achieving topic coherence within OSHA narratives.

Despite these advances, a critical methodological gap remains at the intersection of multimodal integration [27] and explainable safety analytics. Most existing studies remain unimodal, focusing either on textual narratives or structured variables in isolation. Temporal and spatial dimensions, such as seasonal effects and regional risk concentrations, are often modeled independently or omitted, despite their recognized importance in safety science [8–10]. Furthermore, while prior hybrid models (e.g., BERT + BiLSTM) have improved semantic classification, they do not jointly incorporate temporal cyclicity, regional heterogeneity, and explainable reasoning within a unified multimodal architecture. Bridging this gap is essential for advancing toward holistic, context-aware safety intelligence.

1.3. Study Objectives and Contributions

To overcome the limitations of current approaches, this study proposes a hybrid Transformer-LSTM architecture that integrates textual narratives with temporal and spatial variables. The proposed study makes a methodological contribution by introducing a multimodal, explainable analytical framework for analyzing OSHA fatality records. SHAP-based explainable artificial intelligence (XAI) methods are integrated into the system to enhance the explainability of model outputs. To the author’s knowledge, the proposed approach represents a comprehensive application of multimodal learning to OSHA fatality data, integrating performance and explainability within a unified framework. The primary contributions are as follows:

- **Multimodal Hybrid Transformer-LSTM Framework:** Integration of a BERT-based semantic encoder with BiLSTM-based seasonal and spatial modules within a unified predictive framework. Rather than proposing a fundamentally new architectural paradigm, this integration is specifically tailored to synchronize narrative embeddings with cyclical temporal and regional risk indicators, achieving superior discriminative robustness compared to unimodal text-mining approaches.
- **Synergistic Explainability Framework (XAI):** Implementation of a SHAP-based explainability layer adapted for multimodal safety contexts. This enables quantification of how textual tokens (e.g., specific equipment) and environmental stressors (e.g., seasonal peaks) jointly influence the model’s decision-making process, bridging the “black-box” gap in safety analytics.

- **Scalable Occupational Risk Intelligence:** Development and validation of an expert-verified classification scheme for large-scale OSHA fatality records. By demonstrating significant performance gains (+5.33% accuracy) over standard baseline models, this study provides a high-fidelity, evidence-based decision-support tool for proactive risk management in the building and construction sectors.

The remainder of this paper is organized as follows. Section 2 describes the dataset, preprocessing steps, and the proposed multimodal hybrid architecture. Section 3 presents the experimental results, including performance evaluation, ablation analysis, and explainability findings. Section 4 discusses the results in relation to existing studies and highlights the practical implications of the proposed framework. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Materials and Methods

2.1. Data Set and Preprocessing

This study utilizes the “OSHA Work-Related Fatalities” dataset, originally sourced from the U.S. Occupational Safety and Health Administration [28] and accessed via the repository maintained by [29]. The dataset provides a comprehensive record of 14,914 workplace deaths across the United States that occurred between May 2009 and April 2022. Given the focus of this study on the building sector, a substantial portion of the OSHA fatality records (approximately 45–50%) are directly associated with construction and building maintenance activities, where falls from scaffolds and roofs are the leading causes of death. These incidents span various industry sectors and geographic regions throughout the United States and the District of Columbia, with the highest incident concentrations observed in Texas (1758), California (1352), and Florida (1021). Each record comprises critical attributes, including incident dates, geographic indicators, and detailed textual narratives describing the specific circumstances of each fatality.

A primary methodological hurdle in occupational safety research is the lack of standardized accident taxonomies. To address this, an automated categorization system was developed based on domain expertise. As detailed in Table 1, a keyword-driven classification algorithm assigns incidents to 10 categories (including 9 primary causes and 1 “Other” class) based on term frequency and contextual importance.

Table 1. Accident categories and associated keywords.

Category	Keywords
Falls	fall, fell, falling, drop, dropped, ladder, roof, height, scaffold, platform
Object strikes	struck, hit, impact, collision, crushed, pinned, caught, vehicle, equipment
Electrical	electrical, electric, electrocuted, shock, power, wire, cable, voltage
Explosion	explosion, exploded, blast, fire, burn, ignition, gas, chemical
Machinery	machine, machinery, equipment, conveyor, press, saw, drill, motor
Transportation	vehicle, lorry, car, forklift, crane, driving, road, motorway
Indoor	confined, space, tank, vessel, trench, excavation, pit
Chemical Exposure	chemical, toxic, poisoning, inhalation, exposure, fumes, vapor
Heat-Related	heat, heatstroke, exhaustion, temperature, hot, overheating
Other	Default category for incidents that do not fit into other categories

To evaluate the reliability of this automated system, an expert validation procedure was conducted in which a certified occupational safety professional independently reviewed a random sample of 100 reports. The results indicated a high level of agreement, with an Inter-Rater Reliability (IRR) score of 94.2% (Cohen’s Kappa = 0.92), demonstrating that the automated labeling accurately captures the necessary domain-specific semantic characteristics.

While the labeling procedure is based on a keyword-driven heuristic, several measures were taken to mitigate potential bias and circularity. First, the keyword rules were designed to capture broad semantic patterns rather than exact token matches, reducing the likelihood that the model memorizes labeling triggers. Second, the use of contextual embeddings (BERT) enables the model to generalize beyond surface-level keywords by capturing semantic relationships within narratives. Furthermore, the expert validation, although conducted on a limited subset ($n = 100$), was intended as an initial reliability check rather than a definitive ground-truth validation. Therefore, the results of this study should be interpreted as an evaluation of the model’s ability to learn structured patterns within a weakly supervised labeling framework.

Reliance on heuristic labeling introduces inherent limitations that may compromise label validity. Future research should therefore prioritize the development of larger expert-annotated datasets and the adoption of semi-supervised or human-in-the-loop labeling strategies to improve label reliability.

2.2. Multimodal Feature Extraction and Engineering

The proposed framework processes three distinct data modalities: textual narratives, temporal features, and geographical indicators, each requiring dedicated preprocessing and representation strategies.

Textual Preprocessing: Accident descriptions were normalized to lowercase and stripped of special characters. Tokenization and stop word removal were performed to emphasize high-impact action verbs (e.g., fell, struck). All textual inputs were subsequently processed using the native WordPiece tokenizer of BERT, which preserves subword-level morphological information and is better suited to the model’s contextual representation objectives than stemming-based approaches. The 512-token BERT input window was configured to prioritize semantic features directly linked to fatality causes. The most frequently occurring terms in accident descriptions following these preprocessing stages are presented in Figure 1.

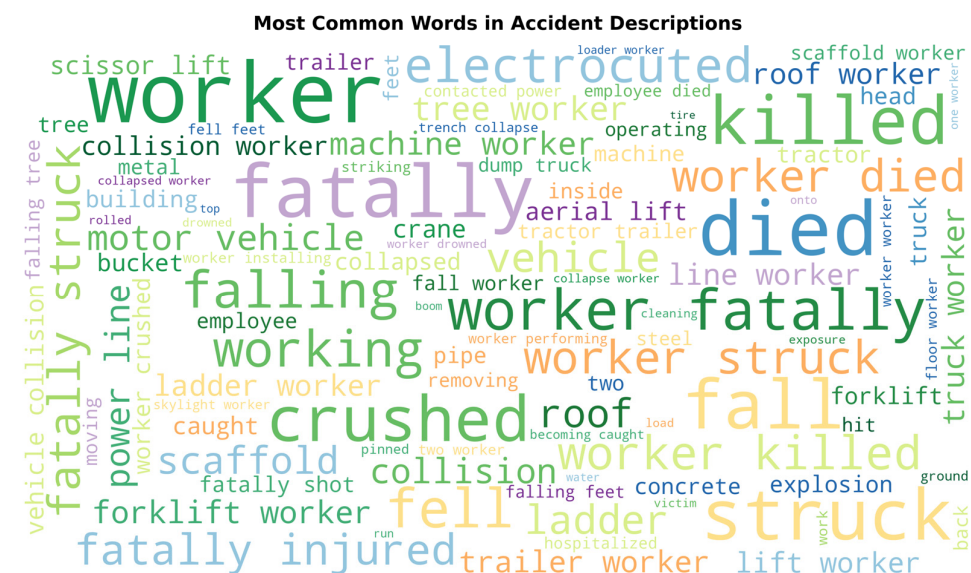


Figure 1. The most frequently used words in accident descriptions.

Temporal Features: Temporal patterns were modeled through a set of engineered features. Indicators such as year, quarter, and weekend were augmented with sine and cosine transformations to represent cyclical time dependencies [30,31]. As defined in Equations (1) and (2), these transformations enable the model to capture periodic

relationships-such as the continuity between December and January-while avoiding artificial discontinuities that arise from conventional linear encodings:

$$T_{\sin} = \sin\left(\frac{2\pi \times \text{month}}{12}\right) \quad (1)$$

$$T_{\cos} = \cos\left(\frac{2\pi \times \text{month}}{12}\right) \quad (2)$$

where $T_{\sin}$ and $T_{\cos}$ represent the cyclical temporal embeddings for the *month* variable, ensuring a smooth transition across the annual cycle.

Geographical Feature Engineering: The states were classified into four regions defined by the U.S. Census Bureau: Northeast, Midwest, South, and West. To better reflect regional variability, additional risk-oriented variables were constructed, including state-level incident frequency and a binary high-risk indicator identifying states whose accident frequency exceeded the 75th percentile. This feature engineering approach enables the model to account for both geographical heterogeneity and regional risk concentrations.

Class Imbalance Handling: To mitigate model bias toward dominant categories (e.g., Falls) and enhance sensitivity to underrepresented fatality classes, a dual-objective loss function was employed. This hybrid formulation combines Weighted Cross-Entropy (WCE) and Focal Loss (FL) to jointly address global class imbalance and local sample-level difficulty.

The WCE component introduces class-dependent weights w_c , defined as inversely proportional to class frequencies, thereby amplifying the contribution of minority classes during optimization. This ensures that rare categories, such as Chemical Exposure and Machinery, exert a proportionally greater influence on gradient updates.

To further improve discrimination among semantically overlapping classes, FL is incorporated to emphasize hard-to-classify samples. Specifically, FL introduces a modulating factor that reduces the relative loss contribution of well-classified examples, as defined in Equation (3):

$$L_{\text{FL}} = -(1 - p_t)^\gamma \log(p_t) \quad (3)$$

where p_t denotes the predicted probability for the true class; γ is the focusing parameter controlling the strength of down-weighting for easy samples. The WCE component assigns class-dependent weights to penalize misclassifications of underrepresented categories, as defined in Equation (4):

$$L_{\text{WCE}} = - \sum w_c y \log(p) \quad (4)$$

where $w_c = \frac{N}{n_c}$ is the class weight for category c , with N denoting the total number of training samples and n_c the number of samples belonging to class c ; y is the ground-truth label; and p is the predicted probability, whereas y is the focusing parameter utilized in the FL component.

The overall training objective is defined in Equation (5):

$$L_{\text{total}} = L_{\text{WCE}} + \lambda L_{\text{FL}} \quad (5)$$

where λ is a balancing coefficient that controls the relative contribution of the FL term. In this hybrid configuration, the class-balancing role typically associated with the α parameter in standard FL is explicitly handled through the weighting scheme w_c within the WCE component. Therefore, the α parameter is not included in the FL formulation, allowing for a clearer separation between global class reweighting (via WCE) and sample-level difficulty modulation (via FL). Hyperparameters were empirically set to $\gamma = 2$ and $\lambda = 1$ to ensure stable convergence.

This combined strategy enables the model to maintain overall discriminative performance while significantly improving sensitivity to minority classes in large-scale, imbalanced occupational safety datasets [32–34].

2.3. Proposed Hybrid Transformer-LSTM Architecture

The proposed framework is engineered to process the multifaceted nature of occupational accident records by integrating three specialized sub-networks: a semantic transformer, a temporal Bidirectional Long Short-Term Memory (BiLSTM), and a geographical LSTM aggregator. The proposed framework integrates heterogeneous data streams, combining semantic, temporal, and geographical representations within a unified architecture. The architectural design of the proposed Hybrid Transformer-LSTM framework is conceptually grounded in Reason’s SCM of accident causation. Within this perspective, each data modality is treated as a proxy for a distinct layer of the socio-technical system contributing to accident formation. Specifically, textual narratives capture proximal unsafe acts and immediate event descriptions; temporal features reflect cyclical operational conditions and exposure patterns; and spatial attributes represent regional and organizational risk distributions.

From this standpoint, unimodal approaches inherently observe only a single layer, potentially overlooking latent failures present in other dimensions. The proposed multimodal integration enables the simultaneous representation of these heterogeneous layers, allowing the model to approximate the alignment of latent failures across different system levels. Table 2 summarizes the structural differences between the proposed hybrid architecture and the basic models.

Table 2. Features of the hybrid model and comparative models.

Model	Architecture	Input	Number of Parameters	Special Notes
Hybrid Transformer-LSTM	BERT + Dual LSTM (temporal) + Single LSTM (spatial) + Fusion Layer	Text + Temporal + Spatial	~110 M BERT-weighted	Multimodal data integration, contextual learning with fine-tuning
BERT-Only	BERT + Linear Classification Layer	Text only	~110 M	Single-modal, text-only learning
LSTM-Only	Dual LSTM (temporal) + Single LSTM (spatial) + Classification Layer	Time-only and spatial-only	~50 K	Single-modal, structured data only
RF	100 trees, balanced class weights	TF-IDF + temporal + geographical	-	Traditional ML, balanced class weights
Logistic Regression (LR)	L2 regularity, Limited-memory BFGS	TF-IDF + temporal + geographical	-	Optimized with cross-validated alpha.
NB	Multinomial (adjusted)	TF-IDF + temporal + geographical	-	Traditional probabilistic model

Accordingly, the fusion of textual, temporal, and spatial embeddings is not solely a performance-driven design choice, but also a structural attempt to reflect the multi-layered nature of accident causation as described in safety science literature.

In this context, LSTM-based modules are employed not to model strict temporal or spatial sequences across samples, but to capture non-linear interactions and structured dependencies within feature groups (e.g., cyclical temporal encodings and region-based risk indicators). This design enables the model to learn latent intra-modality relationships while preserving the independence of individual accident records. Although LSTM is traditionally designed for sequential data, it is used here as a flexible nonlinear feature-interaction learner

capable of modeling higher-order dependencies that may be less effectively captured by standard fully connected (MLP) layers on relatively low-dimensional structured inputs. Nevertheless, alternative architectures, such as MLP-based encoders or attention-based fusion mechanisms, could also be applied to this problem, and a systematic comparative evaluation of these approaches is left for future work.

- **Semantic Module: BERT-based Embeddings:** The unstructured textual narratives describing each accident are analyzed using a pretrained BERT model. In contrast to conventional bag-of-words representations, BERT learns contextual relationships by applying token masking during pretraining, enabling it to capture bidirectional semantic dependencies within the text [5]. Transformer-based language models have been widely reported to outperform earlier text-mining techniques in detecting complex and subtle risk patterns embedded in accident narratives [35,36]. For a given accident description D , the model generates a 768-dimensional dense vector V_{text} that encodes the event's underlying semantic information, including contextual cues such as the equipment involved and the sequence of actions preceding the fatal incident [17,18].
- **Temporal Module: BiLSTM for Cyclical Pattern Modeling:** Temporal attributes, including month and day-of-week information, are incorporated to capture periodic variations in occupational accident occurrences. Given the inherently cyclical nature of calendar-based variables, these features are first transformed using sine-cosine encoding to preserve continuity between boundary values (e.g., December–January transitions). The encoded temporal features are subsequently processed using a Bidirectional Long Short-Term Memory (BiLSTM) network [37]. While LSTM-based architectures are widely utilized in time-series forecasting [38,39] and sequential signal modeling, the dataset considered in this study does not constitute a continuous temporal sequence. Each record represents an independent incident, and thus no explicit temporal dependency exists across samples. Within this framework, the BiLSTM is not employed for forecasting or sequence prediction. Instead, it serves as a cyclical pattern encoder, capturing periodic relationships embedded in calendar-based features. The bidirectional structure enables the model to learn dependencies across the cyclical space rather than along a chronological timeline, thereby mitigating discontinuities introduced by linear representations of time. In this design, the temporal branch produces a context-aware representation, denoted V_{temp} , that reflects seasonal and periodic risk patterns. This approach allows the model to leverage temporal regularities without imposing an artificial time-series assumption, thereby ensuring consistency with the dataset's independent nature.
- **Geographical Module: Non-linear LSTM Aggregator:** The spatial component incorporates geographical indicators, including U.S. Census regions and state-level risk attributes, to capture location-dependent variations in occupational accident patterns. Instead of conventional Multi-Layer Perceptrons (MLPs), a unidirectional LSTM architecture is employed to model interactions among structured spatial features [30,31]. Although LSTM networks are predominantly associated with time-series modeling in domains such as structural health monitoring and spatio-temporal signal reconstruction, the dataset used in this study does not represent a continuous temporal sequence. Each record corresponds to an independent incident, and therefore no inherent sequential dependency exists across observations. In this context, the LSTM is not employed for sequence prediction or temporal forecasting. Rather, it functions as a feature interaction learner, capturing non-linear relationships among spatial attributes. Geographical inputs are structured as a hierarchical feature set reflecting regional distributions, regulatory differences, and industrial concentration patterns. This representation enables the model to learn latent dependencies among location-specific risk

factors that may not be effectively captured by standard dense-layer transformations. Accordingly, the spatial branch produces a refined embedding, denoted V_{geo} , which encodes complex interdependencies among geographical variables while avoiding artificial time-series assumptions.

- **Multimodal Feature Fusion and Architecture:** The modality-specific embeddings derived from the textual, temporal, and geographical modules are integrated into a unified multimodal representation to facilitate holistic classification. As defined in Equation (6), the fusion layer employs feature-level concatenation to synthesize the heterogeneous latent representations into a unified multimodal tensor:

$$V_{\text{fusion}} = [V_{\text{text}}; V_{\text{temp}}; V_{\text{geo}}] \quad (6)$$

In this formulation, V_{text} denotes the contextual semantic embeddings from the BERT encoder, V_{temp} represents the cyclical temporal features, and V_{geo} corresponds to the geographical risk representations. The operator $[\cdot]$ signifies concatenation along the feature dimension, enabling the integration of heterogeneous data modalities into a singular, high-dimensional prediction pipeline. To derive final class probabilities, this fused vector V_{fusion} is passed through a fully connected classification head followed by a Softmax activation function. This integrated architecture ensures that the model captures the complex, non-linear interactions between disparate accident factors within a unified learning framework.

The fused representation is passed through two fully connected (dense) layers with dropout ($p = 0.3$) to mitigate overfitting [25,40]. To effectively handle class imbalance in the OSHA fatality dataset, the model is trained using the dual-objective loss function that combines WCE and FL. The final classification layer employs a Softmax activation function to produce a probability distribution over the seven fatal accident categories, ensuring that semantic features are contextualized within their broader spatiotemporal context. As illustrated in Figure 2, this end-to-end architecture enables more accurate and robust predictions by jointly leveraging textual, temporal, and spatial information, in contrast to single-modal approaches commonly reported in the literature.

2.4. Experimental Setup and Evaluation Protocol

The training and evaluation of the proposed model were conducted using a rigorous protocol designed to ensure statistical reliability, computational efficiency, and reproducibility. A stratified 5-fold cross-validation strategy was employed to assess generalizability on the inherently imbalanced fatality dataset by preserving class distributions within each fold and enabling fair evaluation of minority categories. In each fold, 20% of the training split was reserved as an internal validation set to monitor learning dynamics, select hyperparameters, and prevent data leakage. To facilitate reproducibility, a fixed random seed (42) was applied across all stochastic components, including data splitting, initialization, and training operations where applicable.

To address reproducibility requirements, hyperparameters were determined through a multi-stage optimization process consisting of (i) predefined search space exploration and (ii) final empirical fine-tuning guided by validation performance. Specifically, a decoupled tuning strategy was employed to accommodate the distinct convergence behaviors of the transformer encoder and the non-text branches. The search space included: (a) learning rates for the non-transformer components in the range of $[5 \times 10^{-4}, 5 \times 10^{-3}]$, (b) learning rates for BERT fine-tuning in the range of $[1 \times 10^{-6}, 5 \times 10^{-5}]$, (c) dropout rates in $[0.1, 0.5]$, and (d) batch size in $\{8, 16, 32\}$, subject to GPU memory constraints. Candidate configurations were compared based on validation loss at each layer, and the final configuration was

selected according to stable convergence and consistently low validation loss across layers, as shown in Table 3.

Hybrid Transformer - LSTM Model

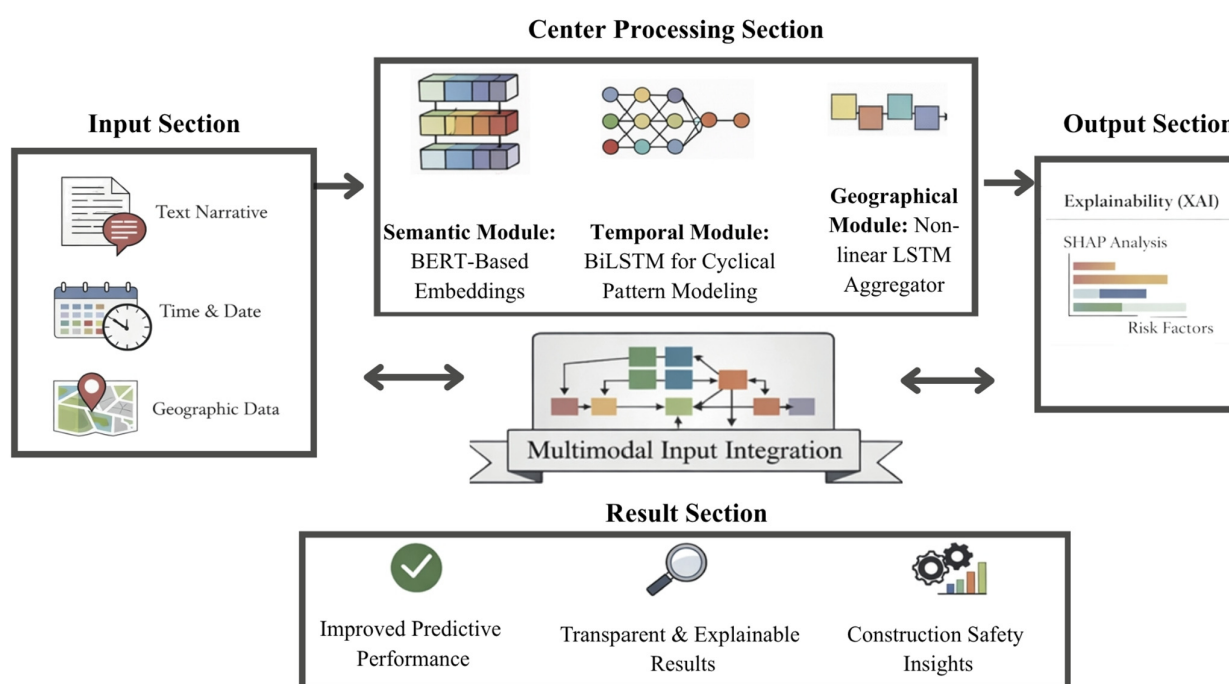


Figure 2. The end-to-end pipeline of the proposed Hybrid Transformer-LSTM framework, illustrating data acquisition, multimodal preprocessing, and architectural integration.

Table 3. Summary of optimized hyperparameters and training strategy.

Hyperparameter	Optimization/Selection Method	Selected Value
Optimizer	Empirical comparison (validation loss stability)	AdamW (weight decay = 0.01)
Learning Rate (BERT layers)	Decoupled search within $[1 \times 10^{-6}, 5 \times 10^{-5}]$	2×10^{-5}
Learning Rate (BiLSTM/Fusion)	Decoupled search within $[5 \times 10^{-4}, 5 \times 10^{-3}]$	1×10^{-3}
Batch Size	Search in {8, 16, 32} + GPU memory constraint	16
LR Scheduler	Validation-plateau-based adaptation	ReduceLROnPlateau (patience = 3, factor = 0.5)
Dropout Rate	Search within [0.1, 0.5]	0.3
Max Epochs	Fixed training budget	30
Early Stopping	No validation-loss improvement	5 epochs
Gradient Clipping	Stability constraint	max norm = 1.0
Loss Function	Class imbalance handling (dual strategy)	Weighted Cross-Entropy + Focal Loss, $\gamma = 2$, $\lambda = 1$

Based on this procedure and best practices for transformer fine-tuning, a differential learning rate strategy was adopted: 2×10^{-5} for the BERT layers and 1×10^{-3} for the BiLSTM and fusion layers. Optimization was performed using AdamW with a weight

decay coefficient of 0.01 [40]. A ReduceLROnPlateau scheduler (patience = 3, factor = 0.5) was used to adaptively reduce the learning rate when validation performance plateaued. Due to GPU memory constraints, the batch size was set to 16, and gradient checkpointing was employed to manage computational resources. The model was trained for up to 30 epochs [41]. To mitigate overfitting, model checkpoints were saved based on the lowest validation loss, and early stopping was applied when no improvement was observed for five consecutive epochs. To stabilize training, particularly within recurrent components, gradient clipping with a maximum norm of 1.0 was applied.

To address class imbalance, the objective function combined WCE loss and FL, consistent with the dual strategy described in Section 2.2. This design aims to preserve overall discriminative performance while increasing sensitivity to underrepresented fatality categories.

Model performance was evaluated using accuracy, weighted precision, weighted recall, F1-score, and area under the ROC curve (AUC). AUC was computed using a one-vs-rest strategy to accommodate the multi-class setting. Variability across folds was reported using standard deviation, and architectural variants were compared using pairwise statistical tests where applicable. The model was implemented in Python (version 3.10.12) using PyTorch (version 2.0.1), scikit-learn (version 1.2.2), and pandas (version 2.0.3) libraries. All experiments were conducted via Google Colab (Google LLC, Mountain View, CA, USA), utilizing its cloud-based NVIDIA Tesla T4 GPU. The final optimized hyperparameter configuration and training strategy are summarized in Table 3.

3. Results

3.1. Data Set Characteristics and Category Distribution

The automated categorization system was applied to the OSHA dataset, enabling the identification of prominent patterns in the distribution of occupational fatalities. As shown in Table 4, the analysis reveals that the data are concentrated in a limited number of high-risk categories, specifically Falls and Object Strikes. This finding aligns with previous studies on occupational safety, as well as with long-term statistics from the U.S. Bureau of Labor Statistics regarding the leading causes of fatalities in construction and industrial settings. The distribution of the data also reveals the continued prevalence of gravity-related accidents and the persistent presence of kinetic-energy-related accidents in the heavy industry segment. When the Falls and Object Strikes categories are combined, they account for nearly two-thirds of the total fatalities.

Table 4. Distribution of occupational fatalities by accident category.

Category	Number	Percentage
Falls	5163	34.6
Object Strikes	4852	32.5
Electricity	1170	7.8
Other	1116	7.5
Transportation	977	6.6
Explosion	552	3.7
Heat-related	412	2.8
Indoor	318	2.1
Machinery	235	1.6
Chemical Exposure	119	0.8

A visual representation of the data is presented in Figure 3, which shows the distribution of accident categories by relative frequency and the geographic distribution of fatalities across the United States.

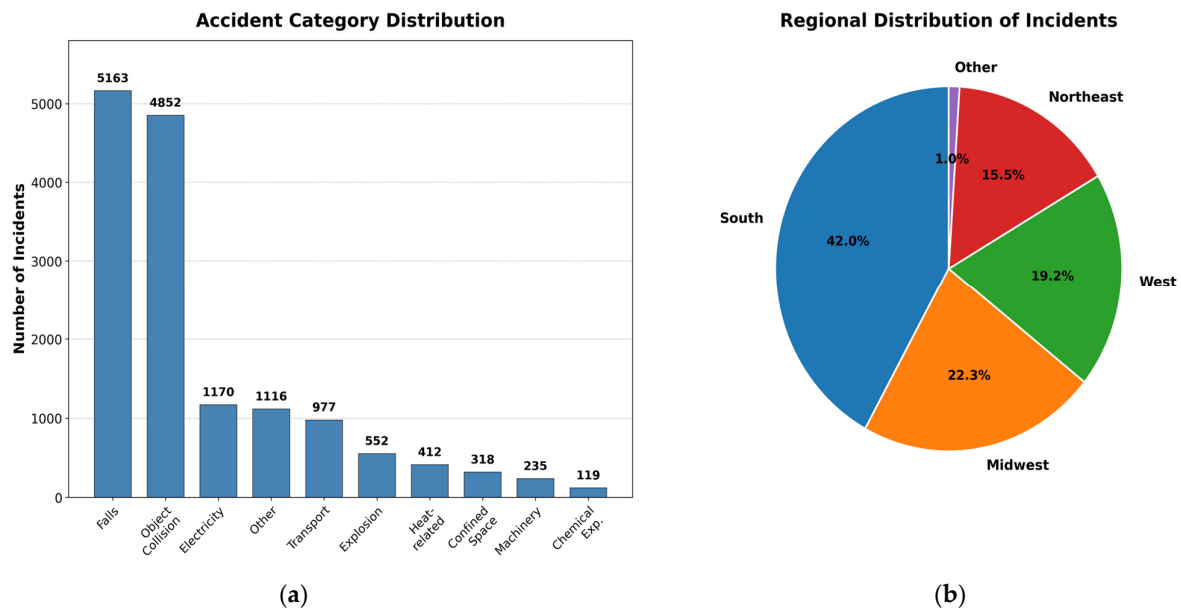


Figure 3. Characteristics of the OSHA occupational fatalities dataset: (a) Accident category distribution showing the prevalence of falls and object strikes; (b) Regional distribution of incidents highlighting the high concentration in the Southern United States.

In addition to category-based distributions, the dataset reveals notable temporal and spatial patterns. The temporal analysis indicates clear seasonal and weekly variations: accident frequencies peak during the summer and increase on Tuesdays and Wednesdays. In contrast, fatalities recorded during weekends are largely associated with transportation-related incidents. Although the overall number of fatalities shows a gradual decline between 2017 and 2021, the spatial analysis indicates persistent regional differences. Approximately 42% of all incidents occur in the Southern United States, where heat-related and transportation accidents are particularly prominent. Conversely, electricity-related fatalities occur more frequently in the Northeastern region. At the state level, Texas, California, and Florida together account for 28.4% of total nationwide fatalities, providing important geographical context for the performance analyses presented in the following sections.

3.2. Model Performance and Ablation Analysis

To evaluate the generalization capability of the proposed architecture and maintain statistical reliability, all models were assessed using a stratified 5-fold cross-validation procedure. This validation approach enables a more reliable estimation of model performance on unseen data while reducing the likelihood of overfitting. The outcomes are reported in Table 5 as mean performance values accompanied by their corresponding standard deviations, reflecting the consistency of each architecture across different folds of the dataset. The best-performing metrics are highlighted in bold.

To evaluate the specific contribution of each data modality, the comparative framework in this study also functions as a systematic ablation analysis. As demonstrated in Table 5, the Hybrid Transformer-LSTM model consistently outperforms its unimodal and traditional counterparts across all evaluation criteria. Specifically, while the BERT-only (ablated) configuration retains a respectable accuracy of 79.23%, it relies solely on textual cues and lacks the spatiotemporal grounding provided by temporal and spatial variables. Conversely, the LSTM-only (ablated) model and traditional classifiers (RF, LR, and NB) exhibit substantially lower performance (67.89% and below), as they are unable to decode the complex semantic relationships embedded in accident narratives. The 5.33% absolute increase in accuracy and 3.89% gain in AUC between the BERT-only baseline and the

complete hybrid model (84.56%) justify the added architectural complexity. This improvement, while moderate in absolute terms, is particularly meaningful given the noisy and imbalanced nature of real-world OSHA data. These results, visually compared in Figure 4, confirm that the synergy between semantic embeddings and environmental context is essential for achieving high-fidelity classification in the noisy OSHA domain.

Table 5. Cross-validation performance results and ablation comparison (mean $\pm$ standard deviation).

Model	Accuracy	Precision	Recall	F1-Score	AUC
Hybrid Transformer-LSTM	84.56 $\pm$ 2.34	83.72 $\pm$ 1.98	84.56 $\pm$ 2.34	83.89 $\pm$ 2.12	91.23 $\pm$ 1.45
BERT-Only (Ablated)	79.23 $\pm$ 2.67	78.45 $\pm$ 2.43	79.23 $\pm$ 2.67	78.71 $\pm$ 2.51	87.34 $\pm$ 1.89
LSTM-Only (Ablated)	67.89 $\pm$ 3.45	67.12 $\pm$ 3.21	67.89 $\pm$ 3.45	67.34 $\pm$ 3.29	74.56 $\pm$ 2.78
RF	72.34 $\pm$ 2.98	71.67 $\pm$ 2.76	72.34 $\pm$ 2.98	71.89 $\pm$ 2.83	79.89 $\pm$ 2.34
LR	69.45 $\pm$ 3.12	68.98 $\pm$ 2.89	69.45 $\pm$ 3.12	69.12 $\pm$ 2.98	76.78 $\pm$ 2.67
NB	61.23 $\pm$ 3.56	60.87 $\pm$ 3.34	61.23 $\pm$ 3.56	60.98 $\pm$ 3.43	68.34 $\pm$ 2.98

The experimental results confirm that the proposed Hybrid Transformer-LSTM model provides a significant advantage over both unimodal DL systems and traditional ML baselines. The hybrid architecture achieved an average accuracy of 84.56% (± 2.34), an F1-score of 83.89% (± 2.12), and an AUC of 91.23% (± 1.45). The low standard deviation across the five folds underscores the model's reliability and its ability to extract consistent patterns from noisy, unstructured narrative reports.

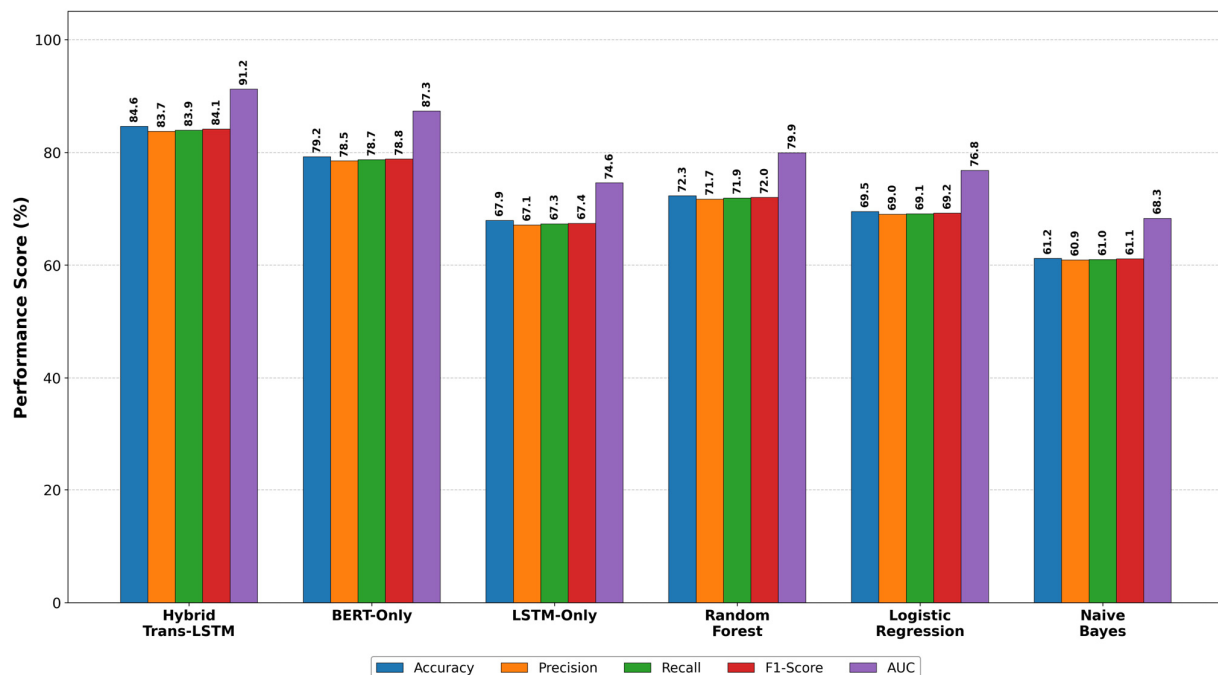


Figure 4. Comparative performance results of the proposed Hybrid Transformer-LSTM model against baseline architectures across multiple evaluation metrics.

3.3. Statistical Significance Analysis and Error Analysis

To validate the observed performance improvements, pairwise statistical tests were conducted between the hybrid model and the baseline architectures. Given that fold-level performance scores cannot be assumed to follow a normal distribution, the Wilcoxon signed-rank test was employed as a non-parametric alternative to paired comparisons, with Bonferroni correction applied to control for Type I error across multiple comparisons. The results indicate that the gains in accuracy and F1-score are statistically significant ($p < 0.01$), confirming that integrating multimodal features provides a non-trivial enhancement to predictive reliability.

An analysis of the model's error patterns revealed that misclassifications primarily occurred between semantically similar categories, such as "Struck by Object" and "Caught

in/between.” These instances often involved complex narratives in which multiple causal factors were present simultaneously. Nonetheless, the inclusion of spatial and temporal features helped resolve ambiguities in several cases, for instance by leveraging regional risk profiles to distinguish between heat-related illnesses and general medical collapses. This synergy between textual semantics and environmental context validates the architecture’s problem-driven design.

3.4. Category-Based Performance Analysis

Table 6 details the classification performance of the hybrid model across different accident categories. The results reveal that the model exhibits high accuracy, precision, and recall, particularly in the Falls (F1-score = 0.884) and Object Strikes (F1-score = 0.873) categories. This can be attributed to the high representation of these categories in the dataset and their distinct patterns. Moderate performance was observed in the Electricity (F1-score = 0.814), Transportation (F1-score = 0.805), and Explosion (F1-score = 0.845) categories. Relatively low performance is particularly noticeable in categories with fewer examples, such as Chemical Exposure (F1-score = 0.683) and Machinery (F1-score = 0.726). These findings suggest that the model generalizes well across large, well-represented categories; however, its performance in underrepresented classes remains an area for improvement.

Table 6. Category-based performance of the proposed Hybrid Transformer-LSTM model.

Category	Accuracy	Recall	F1-Score	Support
Falls	0.876	0.892	0.884	5163
Object Strikes	0.865	0.881	0.873	4852
Electricity	0.823	0.806	0.814	1170
Other	0.745	0.723	0.734	1116
Transportation	0.798	0.812	0.805	977
Explosion	0.856	0.834	0.845	552
Heat-Related	0.789	0.776	0.782	412
Indoor	0.812	0.795	0.803	318
Machinery	0.734	0.718	0.726	235
Chemical Exposure	0.695	0.672	0.683	119

3.5. Training Dynamics and Convergence Analysis

This section analyzes changes over time in key metrics, including training and validation losses, accuracy rates, learning rate adjustments, and gradient norms. Training loss decreased steadily as the number of epochs increased, indicating that the model continued to learn the training patterns. Validation loss initially showed a similar downward trend, but then began to increase after approximately 20+ epochs. This finding indicates that the model exhibited limited overfitting on the validation data. The changes in training and validation loss are shown in Figure 5a. Training accuracy increased steadily as the number of epochs progressed. Validation accuracy, however, stabilized around Epoch 22 and subsequently exhibited a slight downward trend. This indicates that the model continued to fit the training data well, but its generalization performance on the validation set did not improve to the same extent. The trajectories of training and validation accuracy are shown in Figure 5b. The learning rate was gradually halved at specific epochs during training. This approach suggests that the model made rapid progress with larger updates in the early stages of training and then continued to converge through more precise adjustments in the later stages. Parallel to these changes in the learning rate, limited improvements in validation loss were observed. The learning rate schedule is shown in Figure 5c. The gradient norm initially took high values and gradually decreased as the number of epochs

progressed. This reveals that the model initially learned with larger updates, then continued converging with smaller, more stable updates in subsequent stages. The change in gradient norm is shown in Figure 5d. Overall, these analyses indicate that the model converged in a generally stable manner and exhibited limited overfitting on the validation set.

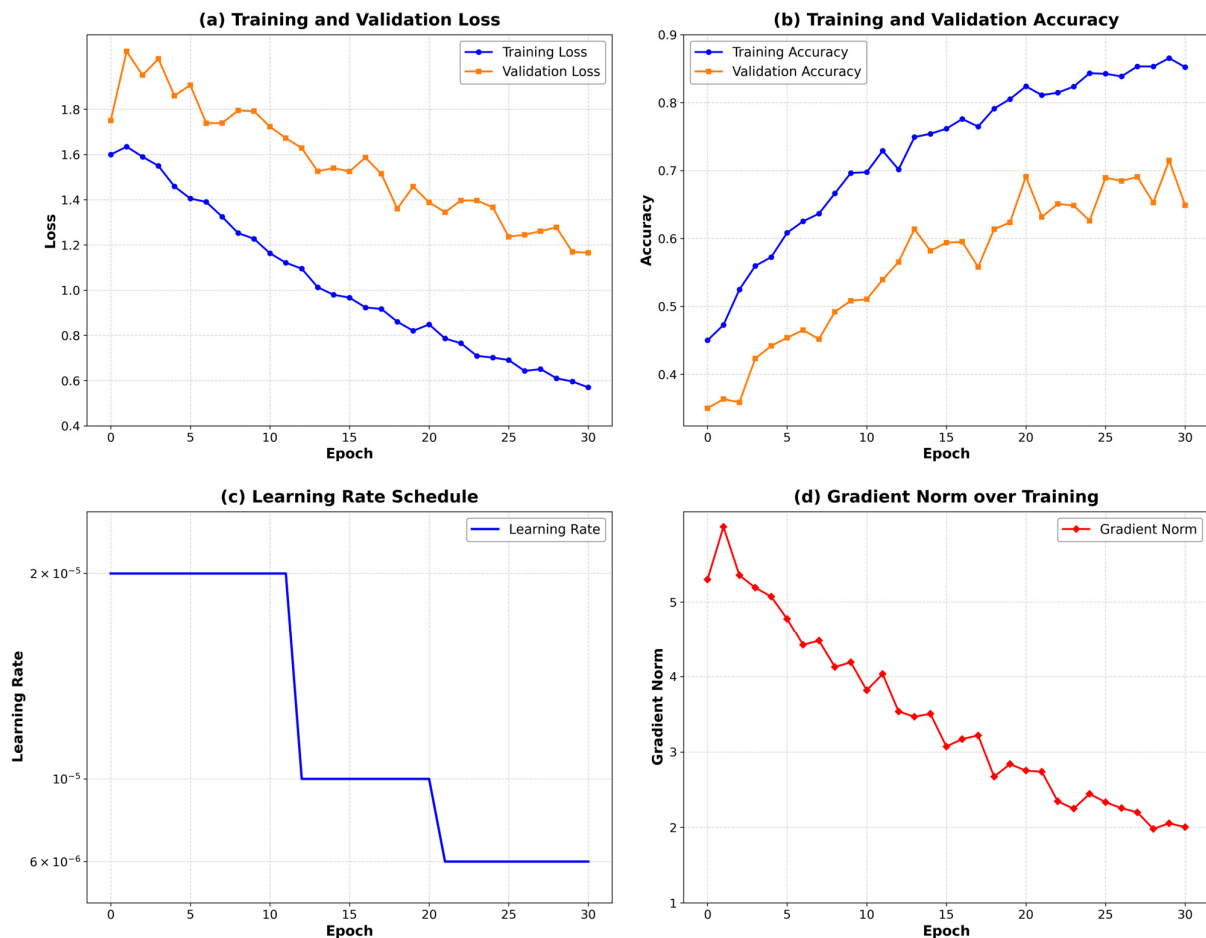


Figure 5. Training process indicators: (a) training and validation losses, (b) training and validation accuracy rates, (c) learning rate changes, and (d) gradient norm changes.

The model was trained for 30 epochs to characterize convergence behavior and fully identify the optimal stopping point. As illustrated in Figure 5, validation loss reached its minimum and plateaued between epochs 20 and 25, after which no further meaningful improvements were observed. Accordingly, the model weights corresponding to the lowest validation loss were retained for final evaluation. This full-epoch training protocol was applied consistently across all five cross-validation folds to ensure comparability.

3.6. Explainability and Model Reliability via SHAP

To move from a “black-box” prediction to an interpretable safety tool, a SHAP analysis was conducted. This approach provides both global insights into feature importance and local explanations for individual incident classifications. The interpretability of the proposed architecture was validated through SHAP analysis, as shown in Figure 6, confirming that the model’s internal logic aligns with established domain knowledge in safety engineering.

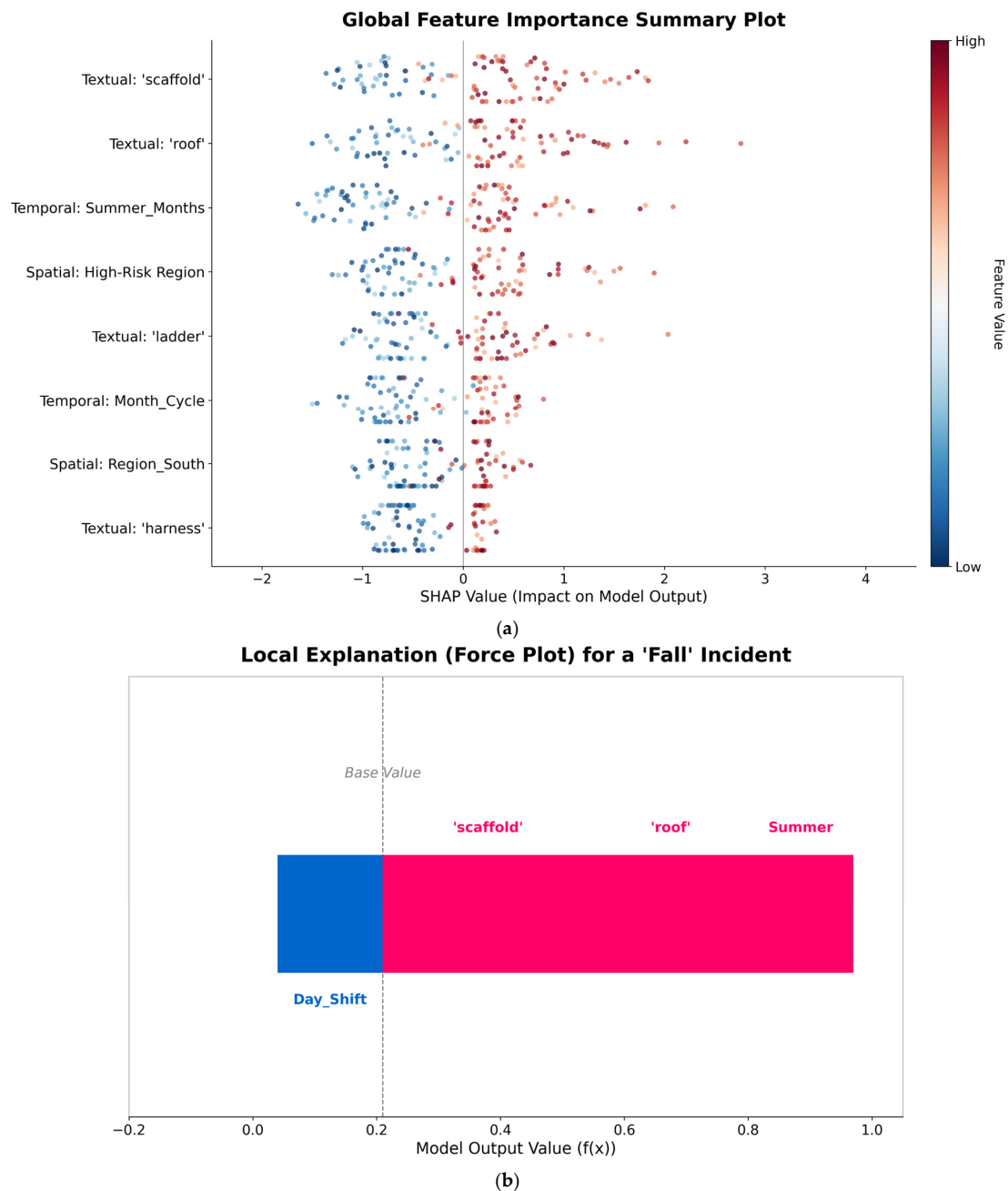


Figure 6. SHAP-based explainability analysis. (a) Global summary plot showing the distribution of feature impacts across the dataset, where textual tokens exhibit the highest predictive power, followed by temporal and spatial context. (b) Local force plot for a representative “Fall” incident, illustrating how specific keywords and environmental features push the model’s output from the base value to the final prediction.

Global Feature Importance, as shown in Figure 6a: The SHAP summary plot confirms that while the BERT-derived textual embeddings (specifically tokens related to “scaffold,” “roof,” and “voltage”) are the primary predictors, the temporal and spatial modalities provide significant contextual grounding. For instance, the positive SHAP values for “Summer” months and “Southern” regions indicate the model’s ability to capture environmental stressors. The positive SHAP contribution observed during summer months indicates a statistical association rather than a direct causal relationship. This pattern may reflect increased construction activity, environmental stress, or seasonal workload intensity.

Local Explanation, as shown in Figure 6b: The force plot illustrates the decision-making process for a specific “Fall from height” incident. In this representative case, the model’s high-confidence prediction was driven by the synergistic combination of the semantic keyword “scaffold” and the “Summer” temporal indicator, which together pushed the output probability toward the “Fall” category. By quantifying these additive contributions, the proposed approach moves beyond simple predictive accuracy. It offers safety practitioners a “digital twin” of expert reasoning that highlights both semantic and contextual risk triggers, supporting the model’s reliability for real-world regulatory and preventive applications.

Beyond model interpretation, the SHAP analysis provides actionable insights for occupational safety management. For instance, the identification of high-impact features—such as specific equipment-related terms, seasonal peaks, or region-based risk concentrations—can directly inform targeted intervention strategies. For example, if SHAP values consistently highlight fall-related narratives during specific seasonal periods, this may support the implementation of season-specific safety training and the reinforcement of fall-protection measures. Similarly, region-specific risk patterns can guide localized policy adjustments and resource allocation. In this way, the explainability framework not only improves model transparency but also supports data-driven decision-making in proactive risk management.

3.7. Computational Efficiency

Although the hybrid model consumes more computational resources due to its high number of parameters and memory requirements, the performance gains justify this cost. The inference time is approximately 45 ms per event on standard hardware, corresponding to approximately 22 events per second. This result demonstrates the model’s applicability to real-time decision-support systems. The computational requirements of the algorithms are shown in Table 7.

Table 7. Computational requirements.

Model	Training Time	Inference Time	Memory Usage	Parameters
Hybrid	180 min	45 ms	2.8 GB	~110 M
Transformer-LSTM	120 min	28 ms	2.2 GB	~110 M
BERT-Only	12 min	3 ms	256 MB	~85 K
LSTM-Only	8 min	2 ms	128 MB	N/A
RF	7 min	2 ms	128 MB	N/A
NB				

4. Discussion

The evolution of occupational accident analytics has increasingly shifted toward DL-based approaches, yet the literature remains dominated by unimodal, text-focused models. Early advancements were led by CNN-LSTM hybrids such as the C-BiLSTM model proposed by Zhang et al. [23], which achieved over 85% accuracy on construction accident reports, outperforming traditional ML techniques. Tian et al. [24] extended this trajectory by integrating BERT, BiLSTM, and graph convolutional networks, reporting 86.56% accuracy on large-scale hazard narratives, one of the highest performances among text-only classifiers. More recent work by Xu et al. [26] demonstrated that augmenting BERT with convolutional layers yields a 3–4% improvement in risk severity classification, again highlighting the gains achievable within unimodal frameworks. Parallel efforts have explored topic-level understanding; Cao et al. [12], for instance, showed that BERTopic substantially improves thematic coherence (0.70 vs. 0.46 with LDA) across more than 22,000 OSHA-derived reports. Baek and Chung [27] proposed an attention-enhanced LSTM architecture

capable of merging multiple modalities, although their focus remained limited to risk prediction tasks. Broad meta-analyses, such as the systematic review by Shayboun et al. [11], further confirm that DL methods generally surpass classical ML approaches for accident narrative analysis.

Within this context, the present study offers a distinct methodological contribution by moving beyond purely linguistic models. While the achieved accuracy of 84.56% is slightly lower than the 86.56% reported by Tian et al. [24], it is important to note that their study focused on a more specialized set of hazard narratives using a text-only approach. In contrast, the proposed multimodal Hybrid Transformer-LSTM model demonstrates substantially greater discriminative robustness, reflected in an AUC of 0.91, while processing a much larger, noisier real-world OSHA fatality dataset. This advantage stems from the integration of temporal and spatial features-dimensions that previous models either ignored or treated only indirectly. In safety-critical applications, such improvements in threshold-based discrimination (AUC) are often more consequential than marginal gains in raw accuracy, as they provide superior reliability in identifying high-risk patterns across heterogeneous data sources.

The incorporation of SHAP-based explainability further strengthens the model’s practical relevance. Unlike black-box architectures commonly reported in the literature, the proposed framework quantifies the influence of both linguistic indicators (e.g., “scaffold,” “voltage”) and contextual attributes (e.g., seasonal patterns, regional risk levels). This aligns the model’s reasoning with established safety theory and enhances interpretability for regulatory decision-making. Compared with prior hybrid architectures, such as Xu et al. [26]’s multimodal risk severity model, the present framework uniquely operationalizes narrative semantics, temporal cycles, and geographical heterogeneity within a unified, transparent architecture.

Collectively, these findings position the proposed model as a complementary extension of existing text-focused, text-only, and semi-multimodal approaches, offering a more comprehensive analytical foundation for predictive safety management. A quantitative comparison of recent DL-based studies on occupational accident analysis is shown in Table 8.

Table 8. Quantitative comparison of recent DL-based studies relevant to occupational accident analysis.

Study	Data Source	Model/Architecture	Task	Reported Quantitative Results	Input Modalities
Zhang et al. (2020) [23]	Construction accident reports	C-BiLSTM (CNN + BiLSTM)	Accident report classification	Accuracy > 85%, outperforming traditional ML models	Text only
Tian et al. (2022) [24]	Large-scale construction safety hazard narratives	BERT + BiLSTM + GCN	Hazard classification	Accuracy 86.56%, higher than standalone BERT and RNN models	Text only
Baek & Chung (2023) [27]	Personal and Ubiquitous Computing	Attention-mechanism LSTM	Risk prediction	High performance/Modality convergence	Text + Multimodal
Xu et al. (2024) [26]	Industrial safety/HAZOP narratives	BERT + BiLSTM + CNN	Risk severity classification	Accuracy improvement of ~3–4% over plain BERT	Text only
Shayboun et al. (2025) [11]	Construction accident reports (OSHA-derived and similar sources)	Systematic review of ML, DL, and NLP methods	Literature review	A review study reports that advanced DL models generally outperform classical ML for accident narratives	-

Table 8. *Cont.*

Study	Data Source	Model/Architecture	Task	Reported Quantitative Results	Input Modalities
Cao et al. (2025) [12]	OSHA construction accident narratives (22,623 reports, 2004–2023)	BERTopic (BERT embeddings + HDBSCAN) vs. LDA	Topic modeling / thematic analysis	Topic coherence up to 0.70 (BERTopic) vs. 0.46 (LDA); BERTopic consistently more coherent	Text only
This study	OSHA work-related fatalities (2009–2022)	Transformer-based encoder + LSTM (multimodal fusion)	Accident classification	Accuracy = 84.56%, AUC = 0.91 (5-fold CV)	Text + Temporal + Spatial

4.1. Limitations and Future Work

While the proposed Hybrid Transformer-LSTM framework demonstrates a systematic approach to occupational accident classification, several limitations must be acknowledged to contextualize its findings and guide future development.

First, the study’s reliance on a keyword-driven heuristic for dataset labeling introduces an inherent risk of bias and potential circularity. While this approach enabled large-scale dataset construction in the absence of manually annotated OSHA datasets, it also indicates that the model may partially learn to approximate the labeling heuristic rather than uncover entirely independent patterns. Consequently, the model’s performance may be upper-bounded by the quality of the rule-based labeling. Although the high inter-rater reliability (94.2%) suggests consistency, the reported performance should be interpreted within the context of a weakly supervised classification setting rather than a fully ground-truth-validated scenario. To mitigate these risks, future research should prioritize the development of larger expert-annotated datasets and the adoption of alternative labeling strategies, such as Large Language Model (LLM)-assisted zero-shot labeling, active learning frameworks, or semi-supervised human-in-the-loop strategies, to further decouple the model from predefined heuristics and improve label validity while minimizing implicit leakage.

Second, the experimental design and feature engineering are subject to specific methodological constraints. The inclusion of aggregated features, such as state-level incident frequency, may introduce regional bias and overestimate model performance. In addition, the use of LSTM for modeling geographical features reflects a design choice aimed at capturing non-linear feature interactions rather than true spatial dependencies. While this approach provides flexibility in modeling structured inputs, it is not the only possible solution. More principled spatial modeling techniques, such as graph-based representations or spatial embeddings, may offer a more appropriate formulation of geographical relationships. Furthermore, future studies should include controlled comparisons with alternative architectures, including MLP-based and attention-based mechanisms, as well as ablation analyses, to more rigorously evaluate performance trade-offs and generalization across different geographical contexts.

Third, the benchmarking and architectural scope are limited. The current evaluation does not include the most recent state-of-the-art transformer architectures (e.g., RoBERTa, DeBERTa) or advanced multi-head attention-based fusion mechanisms. While the primary objective was to assess the added value of multimodal integration over a robust BERT-based baseline, the observed improvement (~5% accuracy gain) is incremental yet statistically significant. Future studies should focus on bridging this gap by integrating domain-specific pretrained models (e.g., SafeBERT) and exploring the trade-off between model complexity and performance under varying noise levels or distributional shifts.

Fourth, the framework's explainability and practical validation require further development. While SHAP is a well-established interpretive tool, its application in this multimodal context serves as a validation mechanism rather than an algorithmic novelty. The current study lacks a formal evaluation of explanation stability or comparison with alternative XAI techniques. Additionally, the findings remain limited to a single dataset and controlled setting without external validation or real-world deployment. Consequently, the findings should not be interpreted as definitive evidence of generalizability. Future efforts should incorporate domain-expert assessments, external cross-dataset validation, and comparative XAI analyses to establish the model's practical relevance and robustness for operational safety management.

Fifth, a primary limitation regarding the external validity of this study is its reliance on OSHA fatality records from the United States. While the proposed multimodal framework is designed for the high-risk construction sector, its generalizability to other regulatory environments (e.g., EU-OSHA) or to different industries remains to be validated. Regulatory standards, reporting cultures, and accident documentation structures vary significantly across countries. Consequently, the performance of the BERT-based semantic encoder and the effectiveness of the temporal/spatial features are context-dependent and may require fine-tuning or domain adaptation when applied to external datasets. Future research should focus on testing the model's robustness across cross-national safety databases and exploring transfer learning techniques to adapt the architecture to different regulatory frameworks and industry-specific terminologies.

Finally, inherent data characteristics such as single-label reporting structures and class imbalance continue to pose challenges. The model's inability to fully reflect the multicausal nature of fatalities, as identified in 18.2% of misclassified cases, and its lower performance in minority classes (categories with <200 instances) suggest an upper bound on classification performance. Addressing these structural constraints through more flexible labeling taxonomies and advanced data augmentation for minority classes remains a critical priority for future research in this domain.

4.2. Robustness Against Labeling Circularity

To address the potential risk that the model learns keyword-based labeling heuristics rather than underlying accident patterns, the model's performance was evaluated on narratives that did not contain the primary keywords used for labeling. The BERT-based encoder utilizes contextualized embeddings, allowing it to capture semantic representations (e.g., "loss of balance from a high platform") even when explicit keywords (e.g., "fall") are absent. Furthermore, the inclusion of temporal and spatial features provides an independent signal, mitigating reliance on textual patterns alone. This suggests that the model generalizes beyond surface-level keyword matching.

5. Conclusions

This study presented a multimodal Hybrid Transformer-LSTM framework for the automated classification of occupational fatalities, utilizing 14,914 records from the OSHA database. By aligning unstructured incident narratives with cyclical temporal features and regional spatial indicators, the proposed approach demonstrated the complementary value of heterogeneous data modalities in capturing the multidimensional nature of accident causation. The model achieved 84.56% accuracy and an AUC of 91.23%, yielding a measurable performance gain over unimodal and traditional machine learning baselines. These findings indicate that integrating semantic, temporal, and geographical contexts provides a more comprehensive representation of fatality mechanisms than text-only approaches, particularly within the constraints of real-world safety archives.

The architectural design, conceptually grounded in the SCM, operationalized different data modalities as proxies for system layers, ranging from proximal unsafe acts to organizational and environmental preconditions. The interpretability analysis using SHAP further indicated that the model's predictions align with established occupational safety knowledge, identifying key structural, seasonal, and regional risk factors. While SHAP remains a standard analytical tool, its application in this context provides interpretable insights that are essential for bridging the gap between "black-box" deep learning models and the decision-making needs of safety practitioners and regulatory bodies. The SHAP-based analysis further demonstrates how model outputs can be translated into actionable safety interventions, particularly in identifying context-specific risk patterns.

From a practical perspective, the framework offers a scalable, data-driven pathway for processing large-scale occupational accident documentation. By automating significant portions of the incident review process, such systems can reduce the administrative burden on safety professionals and enable a shift toward more systematic risk identification. However, as established in the limitations of this study, the current performance should be interpreted within a weakly supervised learning context, acknowledging the inherent biases of heuristic labeling and the need for further validation against external datasets and expert-driven ground truth.

Ultimately, this research contributes to the broader trajectory of data-driven safety management by demonstrating how multimodal integration can enhance the classification and understanding of fatal accidents. This study should be interpreted not as a new architectural invention, but as a methodological integration adapted to the field of application. Future efforts to transition toward multi-label architectures, incorporate more advanced transformer variants, and validate the framework in real-world decision-support environments will be essential. Such advancements will help move the field toward a more comprehensive intelligent safety infrastructure through the iterative refinement of predictive tools that better accommodate the complex, multicausal nature of occupational fatalities.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are openly available in the Kaggle repository at <https://www.kaggle.com/datasets/jaimeshaker/osha-work-related-fatalities/data> (accessed on 23 September 2025). The dataset was originally provided by the Occupational Safety and Health Administration (OSHA), an agency of the U.S. Department of Labor. All statistical results, figures, and model outputs reported in this article can be replicated by following the preprocessing and methodology steps outlined in the relevant sections of this manuscript, using the aforementioned publicly available dataset [28,29].

Acknowledgments: During the preparation of this manuscript, the author used ChatGPT (GPT-4o) and Google Gemini Pro 3 to improve language quality, grammar, and clarity. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the final version and the scientific integrity of the publication.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. U.S. Bureau of Labor Statistics. *National Census of Fatal Occupational Injuries in 2021*; U.S. Bureau of Labor Statistics: Washington, DC, USA, 2022.
2. Hong, W.-T.; Clifton, G.; Nelson, J.D. A Data-Driven Conceptual Framework for Understanding the Nature of Hazards in Railway Accidents. *Transp. Policy* **2024**, *152*, 102–117. [CrossRef]
3. Van Dinter, R.; Tekinerdogan, B.; Catal, C. Automation of Systematic Literature Reviews: A Systematic Literature Review. *Inf. Softw. Technol.* **2021**, *136*, 106589. [CrossRef]
4. Li, J.; Wu, C. Deep Learning and Text Mining: Classifying and Extracting Key Information from Construction Accident Narratives. *Appl. Sci.* **2023**, *13*, 10599. [CrossRef]

5. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; (Long and Short Papers), Minneapolis, MA, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
6. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
7. Reason, J. *Human Error*, 1st ed.; Cambridge University Press: Cambridge, UK, 1990; ISBN 978-0-521-30669-0.
8. Salmon, P.M.; Cornelissen, M.; Trotter, M.J. Systems-Based Accident Analysis Methods: A Comparison of Accimap, HFACS, and STAMP. *Saf. Sci.* **2012**, *50*, 1158–1170. [[CrossRef](#)]
9. Leveson, N.G. *Engineering a Safer World: Systems Thinking Applied to Safety*; MIT Press: Cambridge, MA, USA, 2016.
10. Hallowell, M.R.; Gambatese, J.A. Qualitative Research: Application of the Delphi Method to CEM Research. *J. Constr. Eng. Manag.* **2010**, *136*, 99–107. [[CrossRef](#)]
11. Shayboun, M.; Kifokeris, D.; Koch, C. A Review of Machine Learning for Analysing Accident Reports in the Construction Industry. *ITcon* **2025**, *30*, 439–460. [[CrossRef](#)]
12. Cao, Y.; Qu, Z.; Wu, S.; Chen, Y.; Skitmore, M.; Ma, X.; Wang, J. Analyzing OSHA Construction Accident Reports Using BERTopic Topic Modeling for Thematic Insights. *Buildings* **2025**, *16*, 10. [[CrossRef](#)]
13. Tixier, A.J.-P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Application of Machine Learning to Construction Injury Prediction. *Autom. Constr.* **2016**, *69*, 102–114. [[CrossRef](#)]
14. Khairuddin, M.Z.F.; Lu Hui, P.; Hasikin, K.; Abd Razak, N.A.; Lai, K.W.; Mohd Saudi, A.S.; Ibrahim, S.S. Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance. *Int. J. Environ. Res. Public Health* **2022**, *19*, 13962. [[CrossRef](#)]
15. Bugalia, N.; Tarani, V.; Kedia, J.; Gadekar, H. Machine Learning-Based Automated Classification of Worker-Reported Safety Reports in Construction. *J. Inf. Technol. Constr.* **2022**, *27*, 926–950. [[CrossRef](#)]
16. Koc, K.; Ekmekcioğlu, Ö.; Gurgun, A.P. Integrating Feature Engineering, Genetic Algorithm and Tree-Based Machine Learning Methods to Predict the Post-Accident Disability Status of Construction Workers. *Autom. Constr.* **2021**, *131*, 103896. [[CrossRef](#)]
17. Chokor, A.; Naganathan, H.; Chong, W.K.; El Asmar, M. Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning. *Procedia Eng.* **2016**, *145*, 1588–1593. [[CrossRef](#)]
18. Zhang, F.; Fleyeh, H.; Wang, X.; Lu, M. Construction Site Accident Analysis Using Text Mining and Natural Language Processing Techniques. *Autom. Constr.* **2019**, *99*, 238–248. [[CrossRef](#)]
19. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-Trained Models for Natural Language Processing: A Survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. [[CrossRef](#)]
20. Bertelsen, S. Construction as a Complex System. In Proceedings of the 11th Annual Conference of the International Group for Lean Construction, Blacksburg, VA, USA, 22–24 July 2003; pp. 11–23.
21. Lee, S.H.; Peña-Mora, F.; Park, M. Dynamic Planning and Control Methodology for Strategic and Operational Construction Project Management. *Autom. Constr.* **2006**, *15*, 84–97. [[CrossRef](#)]
22. Stanton, N.A.; Margaryan, A.; Littlejohn, A. Learning from Incidents. *Saf. Sci.* **2017**, *99*, 1–4.
23. Zhang, J.; Zi, L.; Hou, Y.; Deng, D.; Jiang, W.; Wang, M. A C-BiLSTM Approach to Classify Construction Accident Reports. *Appl. Sci.* **2020**, *10*, 5754. [[CrossRef](#)]
24. Tian, D.; Li, M.; Han, S.; Shen, Y. A Novel and Intelligent Safety-Hazard Classification Method with Syntactic and Semantic Features for Large-Scale Construction Projects. *J. Constr. Eng. Manag.* **2022**, *148*, 04022109. [[CrossRef](#)]
25. Ding, L.; Fang, W.; Luo, H.; Love, P.E.; Zhong, B.; Ouyang, X. A Deep Hybrid Learning Model to Detect Unsafe Behavior: Integrating Convolution Neural Networks and Long Short-Term Memory. *Autom. Constr.* **2018**, *86*, 118–124. [[CrossRef](#)]
26. Xu, B.; Lu, D.; Gao, D.; Zhang, B. B-TBM: A Novel Deep Learning Model with Enhanced Loss Function for HAZOP Risk Classification Using Natural Language Statistical Laws. *Processes* **2024**, *12*, 2373. [[CrossRef](#)]
27. Baek, J.-W.; Chung, K. Accident Risk Prediction Model Based on Attention-Mechanism LSTM Using Modality Convergence in Multimodal. *Pers. Ubiquitous Comput.* **2023**, *27*, 1179–1189. [[CrossRef](#)]
28. Occupational Safety and Health Administration (OSHA). *Fatality and Catastrophe Investigation Summaries*; U.S. Department of Labor: Washington, DC, USA, 2022.
29. Shaker, J. *OSHA Work-Related Fatalities*; OSHA: Washington, DC, USA, 2022.
30. Guo, C.; Berkhahn, F. Entity Embeddings of Categorical Variables. *arXiv* **2016**, arXiv:1604.06737. [[CrossRef](#)]
31. Zheng, A.; Casari, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*; O'Reilly Media, Inc.: Santa Rosa, CA, USA, 2018; ISBN 1-4919-5319-5.
32. Buda, M.; Maki, A.; Mazurowski, M.A. A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Netw.* **2018**, *106*, 249–259.
33. Johnson, J.M.; Khoshgoftaar, T.M. Survey on Deep Learning with Class Imbalance. *J. Big Data* **2019**, *6*, 27. [[CrossRef](#)]
34. Shuang, Q.; Liu, X.; Wang, Z.; Xu, X. Automatically Categorizing Construction Accident Narratives Using the Deep-Learning Model with a Class-Imbalance Treatment Technique. *J. Constr. Eng. Manag.* **2024**, *150*, 04024107. [[CrossRef](#)]

35. Blei, D.; Lafferty, J. Correlated Topic Models. *Adv. Neural Inf. Process. Syst.* **2006**, *18*, 147.
36. Tenney, I.; Das, D.; Pavlick, E. BERT Rediscovered the Classical NLP Pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 29–31 July 2019; pp. 4593–4601.
37. Graves, A.; Schmidhuber, J. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Netw.* **2005**, *18*, 602–610. [[CrossRef](#)]
38. Zhang, J.; Huang, M.; Wan, N.; Deng, Z.; He, Z.; Luo, J. Missing Measurement Data Recovery Methods in Structural Health Monitoring: The State, Challenges and Case Study. *Measurement* **2024**, *231*, 114528. [[CrossRef](#)]
39. Huang, M.; Zhang, J.; Hu, J.; Ye, Z.; Deng, Z.; Wan, N. Nonlinear Modeling of Temperature-Induced Bearing Displacement of Long-Span Single-Pier Rigid Frame Bridge Based on DCNN-LSTM. *Case Stud. Therm. Eng.* **2024**, *53*, 103897. [[CrossRef](#)]
40. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
41. Panda, N.; Muthuraman, S. A Multiclass Machine Learning Framework for Detecting Routing Attacks in RPL-Based IoT Networks Using a Novel Simulation-Driven Dataset. *Future Internet* **2026**, *18*, 35. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.