

Article

Construction Safety Risk Identification and Coupling Analysis Based on Data Mining

Guozong Zhang ¹, Dexin Yang ^{2,*} and Yuan Sun ³

¹ School of Urban Economics and Management, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; zhangguozong@bucea.edu.cn

² Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Beijing 100005, China

³ Value Engineering Society of China, Beijing 100191, China; bhsttsunyan@outlook.com

* Correspondence: dexiny611@gmail.com; Tel.: +86-186-0101-1704

Abstract

Frequent accidents in the construction sector arise from the dynamic coupling of multiple risk factors, while conventional single-factor approaches fail to capture the underlying complexity. Drawing on 702 accident investigation reports, this study develops an intelligent, data-driven framework that integrates large language model-based risk identification with association rule mining to systematically uncover risk factors and their coupling patterns. A DeepSeek-based model is employed to extract risk factors from unstructured text, followed by cosine similarity-based optimization to refine factor representations. The FP-Growth algorithm is then applied to identify strong association rules among risk factors. The results reveal that deficiencies in the management dimension account for 68.30% of all identified risks, with inadequate safety education and training emerging as the central hub in the risk coupling network, which is further corroborated by complex network analysis. Moreover, a cascading transmission pathway is identified, whereby environmental deficiencies induce weakened safety awareness, which in turn leads to unsafe behaviors. These findings further demonstrate the nonlinear amplification effects arising from concurrent management failures. By establishing a transformation pathway from unstructured textual data to structured risk knowledge, this study provides a robust, data-driven foundation for precise risk identification and systematic prevention in construction safety management.

Keywords: data mining; safety risk; coupling analysis; accident investigation reports

1. Introduction

The construction industry, as a cornerstone of the global economy, generates substantial economic output and employment, serving as a critical driver of infrastructure development, economic growth, and job creation [1]. Despite its significance, construction is widely recognized as one of the most hazardous industries, characterized by disproportionately high rates of accidents and fatalities. According to the International Labour Organization (ILO), the sector employs approximately 7% of the global workforce yet accounts for 25–40% of all occupational fatalities [2]. Each year, an estimated 100,000 workers lose their lives on construction sites, representing roughly 35% of global work-related deaths [2]. The consequences of such incidents are severe, encompassing extensive human casualties, economic losses, and societal disruption. For instance, the 2013 Rana Plaza collapse in Bangladesh—one of the deadliest industrial disasters in history—resulted in 1134 fatalities and compensation claims totaling \$12.4 million [3]. Similarly, the 2019 construction hoist collapse in Hebei, China, caused 11 deaths, two injuries, and direct economic



Academic Editor: Jaewook Jeong

Received: 10 April 2026

Revised: 2 May 2026

Accepted: 7 May 2026

Published: 12 May 2026

Copyright: © 2026 by the authors.

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.

losses of approximately RMB 18 million, exposing critical deficiencies in emergency response capacity [4]. Post-accident analyses consistently demonstrate that such catastrophic events rarely stem from a single cause; rather, they emerge from the dynamic interaction and coupling of multiple risk factors [5]. This intrinsic complexity renders traditional single-factor or static analytical approaches increasingly inadequate for capturing the underlying mechanisms of construction safety incidents.

Considerable progress has been made in construction safety research, with multi-dimensional investigations conducted from diverse perspectives. Nevertheless, critical bottlenecks persist in two core areas: risk identification and risk coupling analysis. Risk identification focuses on efficiently and accurately extracting potential safety hazards from large-scale, heterogeneous datasets. The paradigm has gradually shifted from reliance on expert judgment toward data-driven approaches, including machine learning and deep learning techniques such as text mining and named entity recognition for extracting risk factors from unstructured text [6]. However, these methods remain constrained by subjectivity, limited semantic understanding, domain adaptation challenges, and the high cost of annotated data, thereby hindering the achievement of efficient, accurate, and generalizable automated identification [7]. More importantly, they are largely incapable of capturing the spatial relationships and interactive effects among risk factors [8].

At a deeper level, risk coupling analysis remains underdeveloped. Mainstream risk assessment models typically rely on the simplifying assumption of factor independence, thereby neglecting dynamic interdependencies and coupling effects. Although association rule mining has been employed to explore inter-factor relationships, conventional algorithms suffer from substantial computational inefficiencies and often remain confined to rule discovery, without adequately revealing underlying interaction mechanisms or nonlinear amplification effects among risk factors [9–11].

To address these limitations, this study develops an integrated analytical framework that combines intelligent risk identification with association rule mining to systematically investigate the complex mechanisms of multi-factor coupling in construction safety. Specifically, a large language model is innovatively introduced to enable automated extraction of risk factors from unstructured textual data, while the FP-Growth algorithm is employed for efficient mining of strong association rules. Through this approach, the study identifies key coupling patterns and transmission pathways among risk factors, aiming to advance the theoretical and methodological foundations of construction safety risk management and to provide actionable, data-driven solutions for mitigating safety incidents.

2. Literature Review

2.1. Risk Identification

Risk identification constitutes the foundation of risk management and represents its most critical initial stage, enabling the detection of unsafe conditions and behaviors that may lead to accidents [12,13]. From a safety science perspective, it extends beyond the mere extraction of risk elements to an initial delineation of the causal architecture underlying accidents. Methodologically, the field has evolved from expert-driven approaches toward data-driven paradigms. Early studies relied heavily on subjective expert judgment, offering operational simplicity and intuitive interpretability. For example, Zhang et al. [14] employed questionnaires and semi-structured interviews with construction professionals to assess safety risk management performance in metro construction projects in China. However, despite their interpretability, such approaches are fundamentally constrained by their dependence on individual expertise and engineering intuition, resulting in pronounced subjectivity and incomplete coverage. More critically, they are inherently incapable of systematically uncovering latent knowledge embedded in large-scale historical

data [15]. To overcome these limitations, subsequent research has increasingly incorporated machine learning and text mining techniques. Goh et al. [16] evaluated the performance of six machine learning algorithms in automatically extracting risk features, marking a pivotal shift toward automated risk identification. Similarly, Kim [17] developed a construction accident knowledge management system with a case-based retrieval model to extract implicit knowledge from accident reports. While these approaches substantially enhance the ability to process large volumes of textual data and detect hidden patterns and associations, they remain confined to shallow analytical paradigms—such as frequency-based methods—lacking deep semantic understanding [16]. Moreover, their reliance on domain-specific lexicons severely limits cross-context adaptability.

With the rapid advancement of deep learning, pretrained model-based named entity recognition (NER) has emerged as a dominant paradigm for risk entity extraction, particularly those built on Transformer architectures, which demonstrate superior capabilities in semantic representation. Recent studies have developed automated Transformer-based frameworks to monitor construction defects, enabling efficient identification and classification of fire door deficiencies through deep encoding of inspection records and defect descriptions [18]. In parallel, graph neural network (GNN)-based text classification methods have been applied to construction scenarios by modeling structural relationships among textual data, thereby improving the accuracy of defect identification in pre-completion stages [19]. These approaches, by incorporating contextual semantics and structural dependencies, significantly improve the precision of text classification and information extraction, offering new technical pathways for construction safety risk identification. Nevertheless, such methods still rely heavily on annotated datasets and task-specific fine-tuning, resulting in limited transferability and generalization. More importantly, they remain inadequate in capturing complex semantic reasoning and implicit causal relationships embedded in unstructured text. In contrast, large language models (LLMs) enable generative semantic modeling and exhibit superior flexibility and generality in complex text understanding [20]. Conceptually, LLM-based risk identification can be regarded as an advanced evolution of NER in the era of generative artificial intelligence [20]. Mustafa Oral [21] demonstrated the effectiveness of models such as ChatGPT 4.5 in identifying and assessing occupational safety risks in construction. By alleviating key bottlenecks associated with conventional pretrained models—such as high training costs, limited inference efficiency, insufficient depth of semantic understanding, and poor domain adaptability [20,22]—LLMs enable substantial improvements in comprehension, automation, and learning capabilities. Crucially, they allow the extraction of implicit causal knowledge without large-scale labeled data, facilitating a paradigm shift from entity recognition to deep semantic understanding.

Against this backdrop, this study employs the DeepSeek model, leveraging its strengths in long-text comprehension in Chinese, complex instruction execution, and semantic consistency to effectively process multi-causal chains and non-standardized expressions in accident reports. This enables the automated normalization and structured representation of risk factors, thereby achieving robust identification of construction safety risks [23]. Given the authenticity and completeness of construction accident investigation reports, LLM-based methods are particularly well-suited to this data source. Theoretically, this process represents a data-driven realization of “multi-factor identification” in classical accident causation models, laying a rigorous foundation for subsequent risk coupling analysis.

2.2. Risk Coupling Analysis

Theories of construction accident causation consistently emphasize that modern engineering accidents rarely result from a single factor; rather, they arise from the interaction

and interdependence of multiple risk factors under specific spatiotemporal conditions—a mechanism broadly conceptualized as risk coupling [24]. The inherent dynamism, complexity, and multi-stakeholder nature of the construction industry further amplify the likelihood of cascading effects, effectively creating a fertile ground for such coupling phenomena [25]. To uncover these complex interdependencies and enable proactive risk prevention and control, systematic risk coupling analysis is essential.

Existing studies have employed models such as the NK model [26], system dynamics [27], and coupling coordination models [28] to characterize risk interactions. However, these approaches typically rely on predefined system structures, which significantly constrain their applicability in the context of complex, high-dimensional construction accident data. With the rapid advancement of information technologies, research has increasingly shifted toward intelligent and data-driven approaches [29]. In particular, the growing scale of accident datasets necessitates the adoption of association rule mining techniques to uncover intrinsic relationships among risk factors. Han et al. (2024) [30] integrated association rule mining into complex network modeling to reveal interaction mechanisms among project risks.

Current association rule mining methods primarily include the Apriori and FP-Growth algorithms. The Apriori algorithm identifies potential association rules through iterative database scanning, thereby uncovering hidden relationships in accident data. For instance, Liu et al. [31] combined digital twin technology with Apriori, integrating IoT and BIM to analyze safety risk coupling in prefabricated building hoisting operations, significantly improving rule mining efficiency and accuracy. Similarly, Fu et al. (2022) [32] integrated Apriori with weighted network theory to provide a scientific basis for more targeted and systematic risk management strategies. However, Apriori suffers from critical limitations in large-scale, high-dimensional datasets, including candidate itemset explosion and low computational efficiency [33,34], which severely restrict its practical applicability. These limitations strongly motivate the adoption of the FP-Growth algorithm. By abandoning the “candidate generation-and-test” paradigm of Apriori and introducing the FP-tree data structure, FP-Growth effectively eliminates candidate explosion and repeated database scanning, achieving superior performance in execution time, memory consumption, and CPU utilization. Through efficient data compression and pattern extraction, FP-Growth has been successfully applied across domains such as metro and tunnel engineering, and has been increasingly integrated with big data, artificial intelligence, and knowledge graph technologies to enable comprehensive risk management [22,35–37]. Notably, it is particularly well-suited for high-dimensional risk datasets extracted by LLMs.

Building upon this foundation, association rules can be further transformed into complex risk networks, enabling the structural characterization of risk interactions through network metrics. Complex network analysis complements FP-Growth by moving beyond local association patterns to reveal the global organizational structure and propagation pathways of risk factors. This integrated framework allows for multi-scale characterization of accident causation structures and cascading risk mechanisms without imposing predefined system assumptions. Accordingly, this study leverages construction accident investigation reports, employs LLMs for risk factor identification, and integrates FP-Growth with complex network analysis to systematically deconstruct the coupling effects of construction safety risks.

3. Methods

This study employed an LLM to extract risk factors from construction accident investigation reports and integrated the FP-Growth algorithm to mine association rules among these factors. The overall analytical workflow is illustrated in Figure 1.

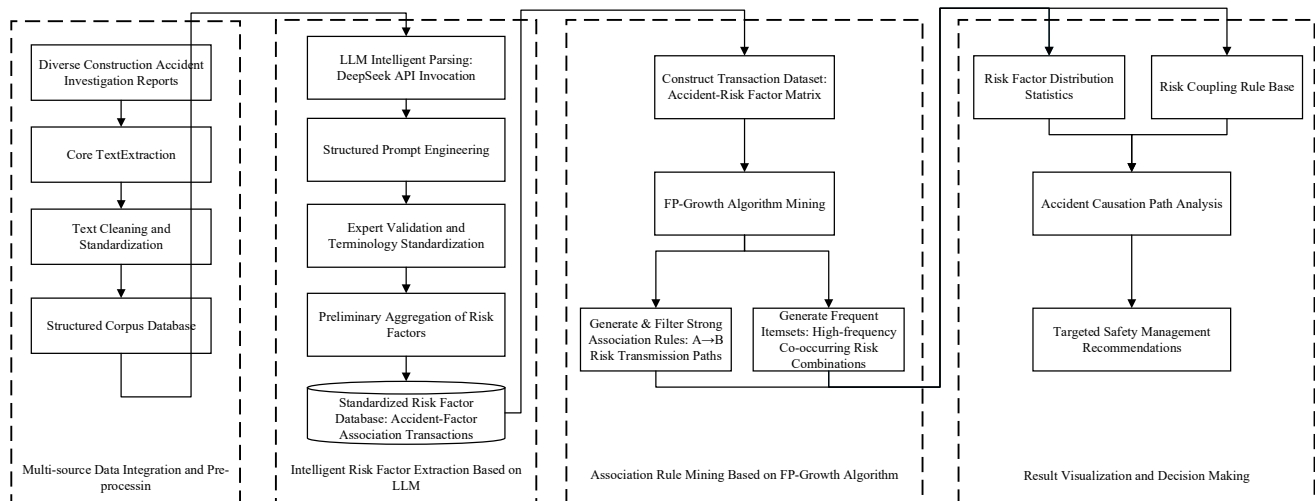


Figure 1. Overall Analytical Framework.

3.1. Data Source and Preprocessing

Construction accident investigation reports provide detailed information on accident timing, location, direct causes, and indirect causes, covering multiple major accident types. Owing to their authenticity and completeness, such reports constitute a high-quality data source for both risk identification and risk coupling analysis. In this study, a total of 810 construction safety accident investigation reports were collected from the official websites of local emergency management authorities in China.

To enhance the efficiency and accuracy of subsequent analytical procedures, the unstructured textual data were systematically preprocessed to address issues such as inconsistent formatting, redundant information, and non-standardized expressions. Reports were deemed eligible only if they contained complete information, including basic accident details, a description of the accident process, and a structured analysis of causes. Reports were excluded if they exhibited severe information deficiencies, duplication or high similarity, irrelevance to the construction sector, unclear accident types, or severely disordered or unparseable text formats, or if they consisted solely of brief notifications rather than full investigation records. Following this screening procedure, 108 reports were excluded, resulting in a final dataset of 702 construction safety accident investigation reports.

3.2. LLM-Based Identification of Construction Safety Risk Factors

3.2.1. Theoretical Framework for Risk Factor Classification

To enable systematic and structured analysis, a unified classification framework for risk factors was established. This study adopted the widely validated “4M1E” model from engineering safety management as the theoretical basis [38,39]. The model categorizes risk sources into five dimensions based on the fundamental components of production systems: Man, Machine, Material, Method, and Environment, as shown in Table 1.

Table 1. The Five-Dimension Structure of the 4M1E Theoretical Model.

Dimensions	Meanings	Examples
Man	Factors related to workers and management personnel	e.g., safety awareness, operational skills, violations, physical and psychological conditions
Machine	Factors related to equipment, tools, and facilities	e.g., equipment aging, safety device failure, tool defects, inadequate maintenance

Table 1. *Cont.*

Dimensions	Meanings	Examples
Material	Factors related to construction materials and components	e.g., material defects, lack of protective equipment, insufficient structural strength, non-compliant specifications
Method	Factors related to construction processes, technical plans, and management procedures	e.g., flawed plans, unclear technical briefings, inadequate supervision, incomplete safety systems
Environment	Factors related to the working environment and external conditions	e.g., hazardous site conditions, insufficient lighting or ventilation, adverse weather, site disorder, complex geological conditions

3.2.2. LLM Invocation and Prompt Engineering

Use the DeepSeek-V3.2 model as the core information extraction engine for risk factor extraction, and perform batch extraction of structured information through its API interface. The text chunk length is set to 1000 tokens, with an overlap of 200 tokens between adjacent chunks, and the temperature is set to 0.5. The quality of the prompt directly determines the accuracy and consistency of information extraction. After multiple rounds of iterative optimization, a structured prompt was finalized, as shown in Figure 2.

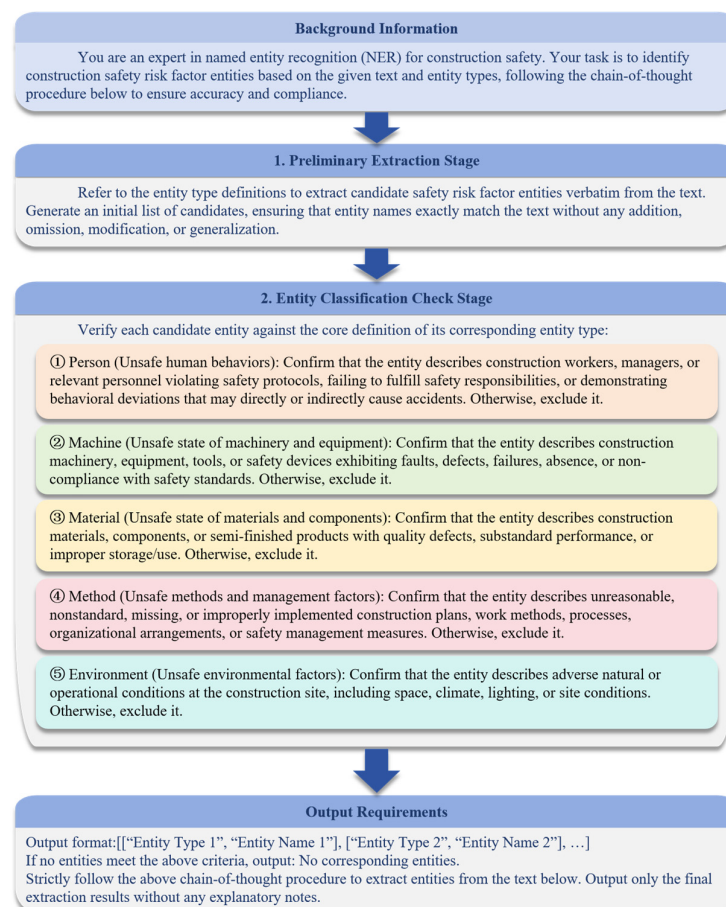


Figure 2. Prompt Design for LLM-Based Identification of Construction Safety Risk Entities.

3.3. Association Rule Mining of Risk Factors Based on the FP-Growth Algorithm

To uncover latent relationships among risk factors, the FP-Growth algorithm was applied to mine association rules from the extracted factors. This approach enabled the au-

tomatic identification of frequently co-occurring risk factor combinations and the detection of potential risk transmission pathways in construction accidents.

3.3.1. Construction of the Transaction Dataset

Association rule mining requires transforming data into a “transaction-item” format. In this study, each independently reported and investigated accident was defined as a transaction, and all unique, standardized risk factors identified within the corresponding report constituted the itemset of that transaction. A total of 702 accidents ($n = 702$) were included, with M representing the total number of standardized risk factors. The transaction dataset can be formally represented as

$$D = \{T_1, T_2, \dots, T_n\}, T_i = \{I_1, I_2, \dots, I_m\} \quad (1)$$

where I_j denotes the j -th standardized risk factor. The unstructured textual information was thus transformed into a Boolean matrix suitable for quantitative pattern mining, in which rows represent accidents, columns represent risk factors, and each matrix element indicates the presence or absence of a given factor in a specific accident.

3.3.2. Principles of the FP-Growth Algorithm

FP-Growth is an efficient algorithm for frequent itemset mining. It improves mining efficiency by constructing an FP-tree. The algorithm is intended to identify risk factors that tend to occur together, as well as potentially strong associations among them, with the objective of discovering rules in the form of $X \Rightarrow Y$. Here, X and Y are disjoint itemsets of risk factors, where X is referred to as the antecedent and Y as the consequent. The implementation of the FP-Growth algorithm consists of two stages: frequent itemset mining and strong association rule generation. By setting a minimum support threshold, itemsets with excessively low occurrence frequencies are filtered out. The support of an itemset X is defined as the proportion of transactions containing X in the total number of transactions, that is, $\text{Support}(X) = P(X)$. The algorithm scans the transaction dataset D twice in total, computes the support of all individual items, sorts them in descending order of support, and inserts them into the FP-tree. By recursively constructing conditional pattern bases and conditional FP-trees, all frequent itemsets satisfying the minimum support threshold are mined. Subsequently, all possible association rules can be generated from each mined frequent itemset L . For every non-empty proper subset S of L , the rule $S \Rightarrow (L - S)$ is considered a candidate rule. Two core metrics, confidence and lift, are introduced to measure the reliability of the rule and evaluate the strength of correlation between the antecedent and the consequent, respectively, thereby assessing the credibility and practical usefulness of the rule. The formulas are as follows:

$$\text{Confidence}(S \Rightarrow (L - S)) = \frac{\text{Support}(L)}{\text{Support}(S)} \quad (2)$$

$$\text{Lift}(S \Rightarrow (L - S)) = \frac{\text{Support}(L)}{\text{Support}(S) \times \text{Support}(L - S)} \quad (3)$$

Higher confidence indicates a greater conditional probability of observing $L-S$ given S . A lift value equal to 1 implies independence between S and $L-S$; values greater than 1 indicate positive correlation (with larger values representing stronger associations), while values less than 1 indicate negative correlation.

3.4. Structural Characterization of Risk Factors Based on Complex Network Analysis

To further characterize the coupling relationships among risk factors from a global structural perspective, complex network analysis was introduced as a complementary ap-

proach to the association rules derived from the FP-Growth algorithm. By identifying critical nodes within the network, this method facilitates the differentiation of functional roles among risk factors and reveals their structural properties in the process of risk propagation.

3.4.1. Construction of the Risk Factor Network

Standardized risk factors were represented as nodes. Based on association rules that satisfied predefined minimum support and confidence thresholds, edges were established between co-occurring risk factors, yielding a risk factor network $G = (V, E)$, where V denotes the set of nodes (risk factors) and E represents the set of edges (associations). Edge weights were quantified using support or lift to capture the strength of associations. To reduce noise and enhance interpretability, only statistically significant associations were retained, resulting in a network that reflects the dominant structure of risk coupling.

3.4.2. Centrality Measures

Based on the constructed network, node importance was quantified using degree centrality and betweenness centrality. Degree centrality captures the extent of direct connectivity of a node and is defined as

$$k_i = \deg(i) = \sum_{j \in N(i)} A_{ij} \quad (4)$$

where k_i denotes the degree of node i , $N(i)$ represents the set of neighboring nodes of i , and A_{ij} is the element of the adjacency matrix (with $A_{ij} = 1$ if nodes i and j are connected, and 0 otherwise). A higher degree centrality indicates that the corresponding risk factor co-occurs more frequently with other factors, reflecting stronger systemic connectivity.

The clustering coefficient was employed to characterize the local connectivity among neighboring nodes. A higher clustering coefficient indicates stronger interconnections among adjacent risk factors, implying a higher potential for localized risk propagation. It is defined as

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \quad (5)$$

where C_i denotes the clustering coefficient of node i , e_i is the number of edges that actually exist among the neighbors of node i , and k_i is the degree of node i . If node i has fewer than two neighbors (i.e., $k_i < 2$), the clustering coefficient is defined as zero. Through these measures, key risk factors can be identified from a network-structural perspective, providing complementary evidence to the association rule mining results.

3.5. Experimental Environment

The implementation of the LLM and FP-Growth algorithm required a dedicated computational environment. The LLM component was executed via remote access to the DeepSeek cloud-based API. Data preprocessing, algorithm execution, and output processing were conducted on a system equipped with an Intel® Core™ i7-8550U CPU @ 1.80 GHz (up to 2.00 GHz). The software environment consisted of Windows 11 Home Edition (Chinese version), with all code developed and executed in Python 3.13.5 using PyCharm Community Edition 2025.1.3.

4. Results

4.1. Risk Factor Identification Results

4.1.1. Initial Identification of Risk Factors

A total of 702 valid accident reports were analyzed using the large language model, which enabled the automated extraction and standardization of risk factors. In total,

9803 valid standardized risk factors were identified, with an average of 13.96 key risk factors per accident. This finding quantitatively supports the premise that construction safety accidents are inherently characterized by multi-factor coupling. To systematically present the composition and distribution of risk factors, all identified factors were classified and statistically analyzed according to the “4M1E” framework (Man, Machine, Material, Method, Environment). The distribution is shown in Figure 3.

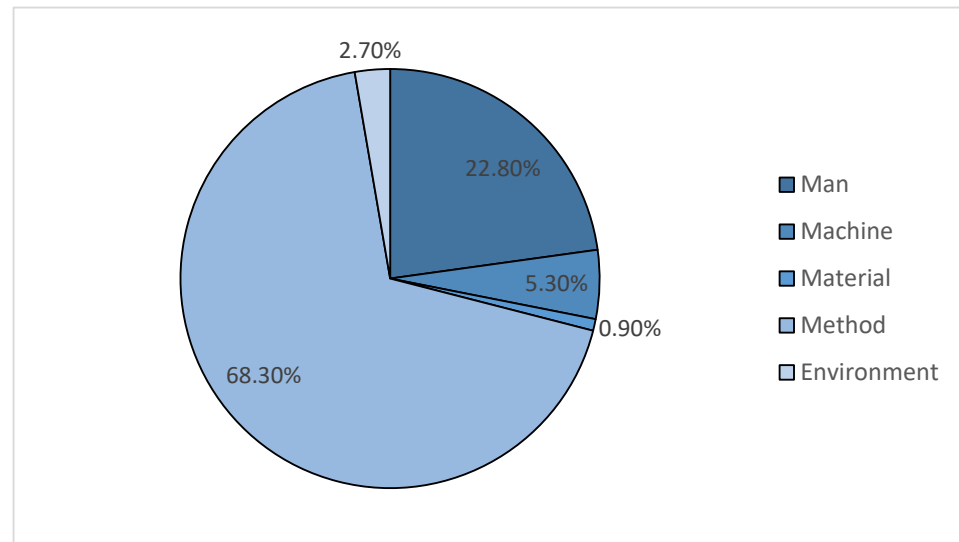


Figure 3. Distribution of Construction Safety Risk Factors.

The findings unequivocally demonstrate that construction safety risks are overwhelmingly concentrated in the Method dimension, which accounts for 68.30% of all identified factors, far exceeding the contributions of Man (22.80%), Machine (5.30%), Environment (2.70%), and Material (0.90%). This pronounced dominance indicates that construction processes, technical schemes, and managerial procedures constitute the primary sources of safety risk.

More fundamentally, this distribution reveals that a substantial proportion of risks do not originate from isolated entities, but are inherently embedded within organizational arrangements, operational workflows, and technological implementation processes. Such risks are typically articulated in accident reports as process-oriented narratives, which systematically channel them into the Method category and generate a strong aggregation effect. At the same time, this structural prominence is reinforced by both reporting conventions—where causal attribution tends to emphasize managerial and procedural deficiencies—and the heightened sensitivity of large language models to processual and institutional semantics during extraction. Taken together, these results not only redefine the locus of construction safety risk as process-centric rather than entity-centric, but also substantiate the methodological advantage of LLM-driven analysis in uncovering deeply embedded, system-level vulnerabilities.

4.1.2. Optimization of Risk Factors

Due to the diversity of Chinese expressions, many of the extracted risk factors appear different in wording but share identical meanings. Therefore, the Qwen3-Embedding-8B model was employed to vectorize these terms and perform cosine similarity detection. A statistical analysis was conducted on the average number of elements within each similarity interval, and the results are shown in Figure 4. Based on the Pareto principle (the 80/20 rule) [40], terms with a similarity greater than 0.7 were merged. Partial results of

the similarity detection are presented in Table 2. The formula for similarity calculation is as follows:

$$S_C(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}} \quad (6)$$

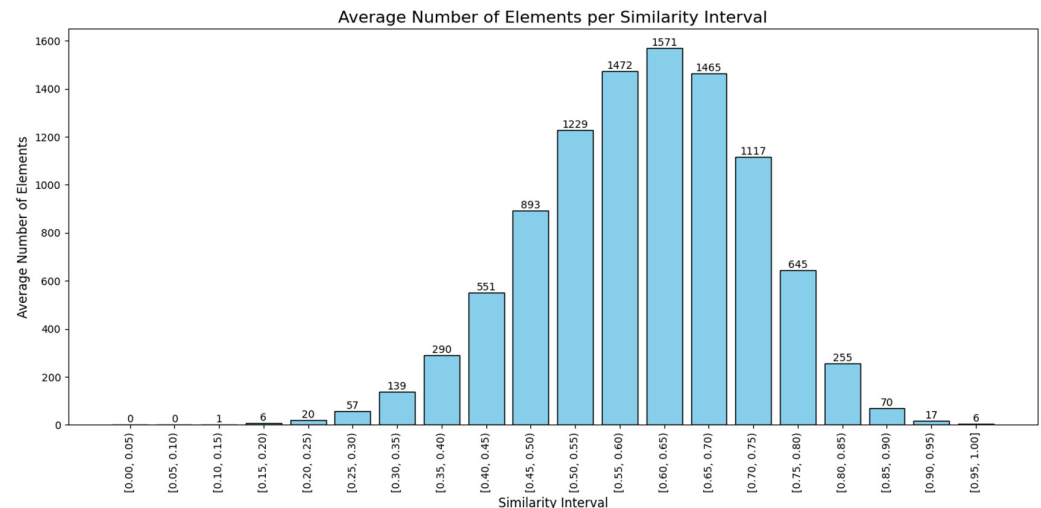


Figure 4. Average Number of Elements per Similarity Interval.

Table 2. Similarity Detection of Risk Factors (Excerpt).

Text	Similarity
Unauthorized bundling of loads by a construction hoisting signal and sling operator without the relevant special operations certificate.	1.000000000
Performing slinging operations without a construction hoisting signal and sling operator certificate	0.858886719
Illegally performing tower crane signaling without the required certificate.	0.838867188
Non-compliance with personnel-to-certificate alignment regulations for tower crane signal and sling operators.	0.779296875
Operator Sun used incorrect sling rigging, violating technical standards, directed unstable lifts, and failed to warn personnel below.	0.772460938
Failure to timely detect and stop unauthorized operations by signal and sling operators.	0.771484375
Unauthorized personnel without a construction hoisting signal and sling operator certificate directing Tower Crane 7C.	0.76953125
Tower crane operator Li followed instructions from non-certified personnel and lifted loads not properly rigged.	0.763183594
No working-at-height or hoisting signal and sling operation certificates.	0.759765625
Liu, lacking operator certification, directed the crane via hand signals.	0.755859375
Assigning unqualified personnel to direct lifts.	0.749023438
Liu had no certificate for special sling operations.	0.74609375
Absence of the required slinging qualification.	0.744628906
No certified personnel assigned for supervision during lifting operations.	0.740234375
Failure to recognize subcontractors who did not assign certified signal and sling supervisors.	0.739257813
Non-compliance with prescribed allocation of signal and sling operators.	0.736816406
Blindly following unqualified instructions of carpenter Wang without verifying crane signal qualification.	0.731933594
Crane operators and signalers lacked essential occupational safety knowledge for lifting operations	0.73046875
On-site supervisors and signalers not certified.	0.728027344
Failure to detect and correct lifting operations conducted with one signaler missing.	0.7265625

4.1.3. Construction of the Risk Factor Indicator System

After consolidating synonymous risk entities, a validation panel comprising five domain experts (Table 3) was convened. A random sample representing 10% of the total

dataset was selected to audit the model-extracted results. Misclassifications and inaccurately identified factors were systematically corrected. Based on expert feedback, the mapping schema and prompting strategy were iteratively refined, and the entire dataset was subsequently reprocessed. The identified risk factors were then systematically categorized under the 4M1E framework and further standardized in accordance with the Classification and Code for Hazardous and Harmful Factors in Process (GB/T 13861–2022) [41]. This procedure yielded the final construction safety risk factor index system (Table 4).

Table 3. Expert Panel Information.

Expert	Domain	Affiliation	Title	Years of Experience
A	Construction Management	University	Professor	20
B	Construction Management	University	Associate Professor	18
C	Intelligent Construction	University	Professor	21
D	Intelligent Construction	Construction Firm	Senior Engineer	15
E	Construction Management	Construction Firm	Senior Engineer	15

Table 4. Construction Safety Risk Factor Indicator System.

Dimension		Risk Factor
Construction Safety Risk Factor Indicator System	Man	Improper use of personal protective equipment
		Unsafe operations
		Unauthorized command or supervision
		Weak safety awareness
		Uncertified operation of special tasks
		Abnormal health conditions
	Machine	Deficiencies in safety protection
		Design flaws
		Equipment operating with defects
		Improper selection or overloading of equipment
		Use of obsolete or non-compliant equipment
		Incorrect installation or fixing of equipment
		Insufficient maintenance and upkeep
	Material	Inadequate material strength
		Material degradation or aging
		Non-compliant material specifications
		Poor structural stability
	Method	Non-implementation of safety responsibility system
		Lack of safety education and training
		Deficiencies in emergency management
		Inadequate safety inspection and hazard mitigation
		Flaws in operating procedures
	Environment	Incomplete setup of safety management structures and staffing
		Insufficient lighting and visibility
		Constrained or disorderly worksite
		Adverse weather conditions
		Slippery work surfaces
		Inadequate safety distance in overlapping operations
		Deficient safety signage and markings

4.2. Association Rule Mining Results Based on FP-Growth

4.2.1. Threshold Setting

In association rule mining, support and confidence are two key thresholds whose settings are typically not fixed a priori, but instead require dynamic adjustment according to the research objectives and data characteristics. A review of existing studies indicates that, in order to balance rule quantity and quality, the minimum support is commonly set within the range of 0.1–0.3, while the minimum confidence is typically set between 0.7 and 0.8.

In this study, drawing on relevant literature [11,31–33] and considering both the distributional characteristics of the dataset and analytical requirements, the minimum support was set to 0.2 and the minimum confidence to 0.75. This configuration ensures that the extracted rules are both sufficiently representative and statistically reliable. Furthermore, to control the scale of frequent itemsets, an additional constraint was imposed on the output size. Specifically, to prevent the generation of an excessive number of rules that may exhibit low support or semantic redundancy—thereby reducing marginal information gain and increasing interpretative complexity and decision-making noise—the number of frequent itemsets was limited to five. This setting not only preserves the generality and reliability of the extracted rules, but also effectively eliminates redundant and less informative associations. The generation process of strong association rules is as follows:

$$\text{Asso}(X, Y) = \{\text{Support} \geq 20\%, \text{Confidence} \geq 75\%, \text{Number of FI} = 5\} \quad (7)$$

4.2.2. Strong Association Rules

Under the specified thresholds, several thousand association rules were initially generated. However, many of them lacked practical relevance. Therefore, rules were further filtered based on lift values ($\text{lift} > 1$), logical consistency, and practical interpretability. Ultimately, ten strong association rules were selected for detailed analysis (Table 5).

All selected rules exhibit lift values greater than 1, indicating positive correlations between antecedents and consequents. Specifically, Rules 1–4 reveal strong associations between the lack of safety education and training and multiple fundamental management deficiencies. For example, Rule 1 shows that inadequate safety training undermines the effectiveness of safety inspections and hazard identification. Rule 2 indicates that failure to implement the safety responsibility system fundamentally prevents the execution of safety training, leading to a breakdown in the management cycle. Rules 3 and 4 further demonstrate that both emergency management and operational procedures depend critically on adequate safety training. Collectively, these rules ($\text{lift} > 1.015$) form a management deficiency cluster centered on safety education and training, consistent with previous findings [42–45].

Rule 5 exhibits a notably high lift value (2.571), indicating that weak safety awareness is a direct behavioral driver of improper use of personal protective equipment. Rule 6 further identifies the root cause of diminished safety awareness: inadequate or unclear safety signage reduces continuous risk communication, thereby reinforcing unsafe cognitive states. Together, these rules establish a causal chain in which environmental deficiencies lead to weakened safety awareness, ultimately resulting in unsafe behaviors. This finding aligns with the theoretical perspective in safety psychology that environmental conditions shape individual awareness [46].

Table 5. Strong Association Rules among Construction Safety Risk Factors (Excerpt).

No.	Association Rule (Antecedent → Consequent)	Support	Confidence	Lift
1	{Lack of safety education and training} → {Inadequate safety inspections and hazard mitigation}	0.967	0.987	1.016
2	{Failure to implement safety responsibility system} → {Lack of safety education and training}	0.920	0.998	1.019
3	{Deficiencies in emergency management} → {Lack of safety education and training}	0.846	0.995	1.015
4	{Flaws in operating procedures} → {Lack of safety education and training}	0.838	0.995	1.015
5	{Weak safety awareness} → {Improper use of personal protective equipment}	0.313	0.806	2.571
6	{Deficient safety signage and markings} → {Weak safety awareness}	0.229	0.752	1.935
7	{Unsafe operations} → {Insufficient safety distance in overlapping operations}	0.349	0.878	1.511
8	{Failure to implement safety responsibility system, inadequate safety inspections and hazard mitigation} → {Lack of safety education and training}	0.911	0.989	1.009
9	{Deficiencies in emergency management, lack of safety education and training} → {Failure to implement safety responsibility system}	0.795	0.939	1.019
10	{Improper selection or overloading of equipment} → {Insufficient safety distance in overlapping operations}	0.249	0.911	1.568

Rules 7 and 10 highlight the coupling between unsafe behaviors and hazardous environmental conditions as direct triggers of accidents. Both rules indicate that reduced safety distances serve as a critical mechanism linking unsafe human actions and unsafe physical conditions, thereby creating high-risk environments. This finding is consistent with prior research emphasizing the coupling effect between compressed safety distances, unsafe behaviors, and hazardous conditions [47].

Rules 8 and 9 demonstrate the superposition effects of multiple management deficiencies, revealing complex chain-like transmission relationships among management factors. Rule 8 indicates that under conditions of ineffective safety responsibility systems and inadequate inspections, the absence of safety education and training becomes almost inevitable. Similarly, Rule 9 shows that deficiencies in both emergency management and training are highly likely to co-occur with ineffective responsibility systems. These findings confirm that accidents are not caused by isolated factors but rather result from the simultaneous failure of multiple components within the safety management system [48,49].

Overall, the results indicate that construction safety risk factors exhibit systematic and multi-level evolutionary patterns. Deficiencies in safety education and training emerge as the central root cause, triggering cascading failures across the safety management system, including ineffective responsibility implementation, superficial inspections, and inadequate emergency management. Furthermore, environmental deficiencies indirectly induce unsafe behaviors by weakening risk perception and safety awareness, forming a clear transmission pathway. The interaction between human and material risks, particularly through reduced safety distances, acts as a direct precursor to accidents. Importantly, risk factors do not operate through simple linear accumulation; instead, concurrent management failures generate nonlinear amplification effects, potentially leading to systemic breakdowns. These findings demonstrate that construction accidents are the result of the co-evolution of root-level management failures, intermediate behavioral deviations, and direct risk coupling,

providing robust theoretical and empirical support for the development of systematic and targeted safety prevention strategies.

4.3. Results of Complex Network Analysis

Based on the construction of a construction safety risk network derived from association rules, complex network analysis was further introduced to characterize the interdependencies among risk factors from a structural perspective, as illustrated in Figure 5. Degree centrality and clustering coefficient were selected as key metrics to respectively measure the number of direct connections of each node and the connectivity within its local neighborhood, thereby enabling the identification of critical risk nodes and their localized coupling characteristics.

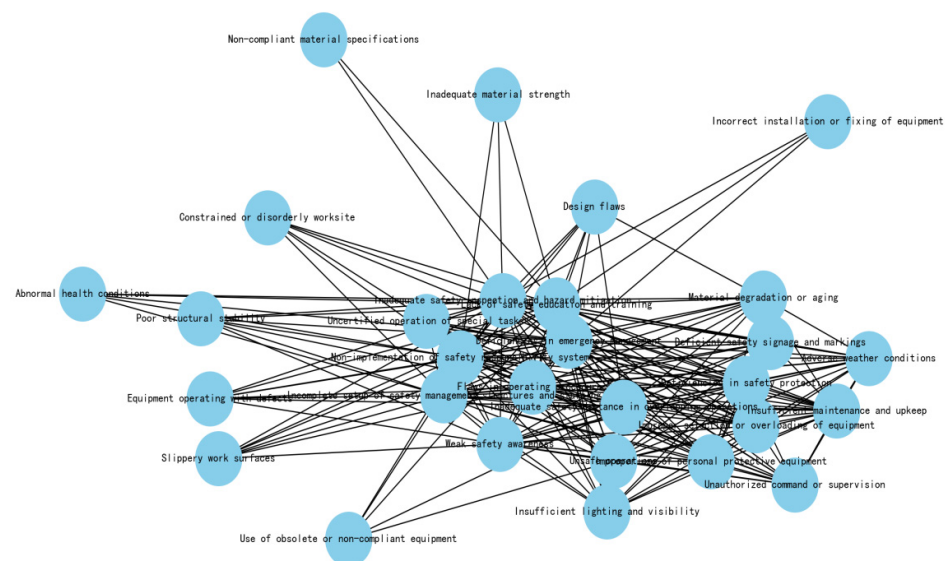


Figure 5. Risk Factor Network Model.

Based on the degree centrality results (Table 6), significant heterogeneity is observed in the structural importance of different risk nodes within the network. Specifically, insufficient safety education and training and inadequate safety inspection and hazard rectification exhibit the highest degree centrality ($k = 28$), occupying the core positions of the network. This indicates that they are directly connected to the majority of other risk factors and function as dominant hub nodes. In addition, weak implementation of safety responsibility systems ($k = 26$) and insufficient emergency management capability ($k = 25$) also demonstrate relatively high connectivity.

Overall, nodes with higher degree centrality are primarily concentrated in the domains of safety management and organizational governance, indicating that managerial risks possess stronger connectivity and greater potential for systemic propagation. These high-centrality nodes play a critical mediating and amplifying role in risk transmission; their failure can readily trigger multi-path cascading diffusion and systemic chain reactions.

The clustering coefficient results (Table 6) further reveal that several risk nodes reach a value of 1, indicating that their neighboring nodes are fully interconnected. Such nodes are highly concentrated within densely coupled substructures centered on management and on-site execution factors, including deficiencies in safety management systems, inadequate site supervision, and unsafe human behaviors. This suggests that these risks do not occur in isolation; instead, they exhibit strong coupling and co-evolution under specific factor combinations. Once triggered, they can rapidly propagate within local clusters, generating pronounced amplification effects.

Table 6. Results of Complex Network Analysis of Risk Factors.

Node	Degree (k)	Clustering Coefficient (C)
Lack of safety education and training	28	0.494708995
Inadequate safety inspection and hazard mitigation	28	0.494708995
Non-implementation of safety responsibility system	26	0.566153846
Deficiencies in emergency management	25	0.6
Incomplete setup of safety management structures and staffing	24	0.644927536
Flaws in operating procedures	23	0.687747036
Inadequate safety distance in overlapping operations	21	0.766666667
Weak safety awareness	21	0.771428571
Unsafe operations	19	0.859649123
Material degradation or aging	16	0.933333333
Deficiencies in safety protection	18	0.934640523
Improper use of personal protective equipment	18	0.934640523
Insufficient maintenance and upkeep	18	0.934640523
Improper selection or overloading of equipment	18	0.934640523
Deficient safety signage and markings	16	0.983333333
Adverse weather conditions	16	0.983333333
Insufficient lighting and visibility	14	1
Unauthorized command or supervision	14	1
Uncertified operation of special tasks	13	1
Equipment operating with defects	8	1
Design flaws	8	1
Slippery work surfaces	7	1
Poor structural stability	7	1
Constrained or disorderly worksite	6	1
Use of obsolete or non-compliant equipment	5	1
Abnormal health conditions	5	1
Incorrect installation or fixing of equipment	3	1
Inadequate material strength	3	1
Non-compliant material specifications	2	1

Integrating the results of degree centrality and clustering coefficient, it is evident that high-degree nodes ($k \geq 25$) serve as critical hubs governing risk transmission pathways across the system, while nodes with high clustering coefficients—particularly those with $C = 1$ —form tightly coupled local subnetworks that intensify risk aggregation and diffusion. This confirms that construction safety risks tend to accumulate and propagate within localized structures, especially those dominated by management and operational execution factors.

Moreover, when combined with the strong association rules derived from the FP-Growth algorithm, it is observed that risk combinations with high support and confidence correspond closely to either high-degree nodes or highly clustered substructures. This consistency further validates the robustness and reliability of the association rule mining results. Overall, complex network analysis not only identifies key hub factors in construction safety risks but also uncovers the intrinsic mechanisms of local risk clustering and propagation, thereby providing a solid theoretical foundation for targeted risk intervention and coordinated governance.

5. Discussion

5.1. Theoretical Implications

From a theoretical perspective, this study advances the understanding of construction safety accidents as complex, evolving systems by uncovering latent interaction patterns through systematic association rule mining. The findings clearly establish the dominant

and root-cause role of management deficiencies in accident causation, while simultaneously revealing chain-like transmission mechanisms among risk factors. In particular, the results provide empirical evidence for the nonlinear amplification effect arising from concurrent management failures, whereby multiple deficiencies interact to magnify overall system risk.

These insights extend existing causation theories, which have traditionally emphasized either linear factor relationships or isolated interactions, by demonstrating that construction safety risks evolve through multi-level, dynamic coupling processes. More importantly, this study offers a data-driven explanation of how localized failures within safety management subsystems can propagate and escalate into systemic breakdowns. In doing so, it bridges the gap between micro-level risk factors and macro-level system failures, thereby enriching the theoretical framework of construction safety science.

Methodologically, the proposed integrated framework, combining LLM-based risk identification with FP-Growth-based association rule mining, introduces a novel paradigm for extracting structured knowledge from unstructured textual data in the construction safety domain. By tightly coupling natural language processing techniques with data mining algorithms, this approach provides a scalable and reusable analytical pipeline. The use of large language models reduces the subjectivity and inefficiency associated with manual coding in traditional risk identification, while substantially enhancing semantic understanding and automation. Meanwhile, the application of the FP-Growth algorithm markedly improves computational efficiency in large-scale data mining, laying a solid foundation for subsequent network-based and system-level analyses. Collectively, these methodological contributions position the study at the intersection of artificial intelligence and safety science, highlighting its broader applicability to complex risk systems.

5.2. Practical Implications

The findings of this study provide a clear data-driven basis for decision-making in construction safety management. Compared with traditional experience-based approaches, this study translates risk co-occurrence patterns into differentiated management strategies by leveraging three key metrics derived from association rules: support (capturing the prevalence of risk combinations), confidence (reflecting the triggering probability along risk chains), and lift (measuring the amplification effect of risk coupling). Combined with the structural characteristics of the risk co-occurrence network, key risk nodes and highly clustered substructures are further identified, leading to the following targeted intervention pathways:

1. Strengthen source-oriented control along high-confidence pathways. High-confidence rules reveal pronounced chain-like failure mechanisms within the management system. Together with the identification of hub nodes in the network, safety training should be elevated from a basic measure to a source-control lever, with targeted programs for critical positions and high-risk operations, reinforced by process-based assessments to interrupt cascading managerial failures.
2. Establish routine system-level governance for high-support combinations. Multiple rules exhibit support values exceeding 0.8, indicating that responsibility systems, training, and inspection form a high-frequency co-occurrence structure, reflecting systemic institutional deficiencies. Accordingly, safety management should shift from single-factor control to coordinated system governance, strengthening the linkage among responsibility implementation, training, and hazard inspection to reduce recurring risks.
3. Implement coordinated interventions for high-lift coupling relationships. High-lift rules indicate strong amplification effects between human and environmental factors. Combined with nodes exhibiting high clustering coefficients, these risks tend to form

dense local substructures. Management should therefore move from single-factor interventions to integrated strategies, optimizing site signage, working environments, and behavioral norms to weaken coupling conditions.

4. Control critical coupling nodes in cross-operation scenarios. Rule mining and network analysis identify cross-operation contexts as key nodes where human and material risks intersect. Their high degree and clustering coefficients highlight their bridging role in risk transmission. Targeted control measures should be implemented at high-risk interfaces, including defined safety distances and dynamic monitoring mechanisms, to prevent physical risk convergence.
5. Develop multi-indicator monitoring for concurrent management failures. When multiple managerial deficiencies co-occur, systemic risk increases significantly. A multi-indicator monitoring system based on medium- to high-confidence rules can trigger early warnings when key management factors simultaneously deviate, enabling timely identification of concurrent risk conditions and preventing local failures from escalating into systemic accidents.
6. Adopt stratified and network-informed precision governance. Homogeneous governance approaches should be avoided. Based on network metrics, high-degree nodes should be prioritized for intervention to weaken their connectivity with other risks, while highly clustered substructures should be addressed through coordinated governance to disrupt tight interdependencies. This enables dynamic risk identification and precise control on construction sites.

5.3. Limitations and Future Research

Although this study introduces methodological innovations enabled by advances in information technology, several limitations remain in terms of data sources, parameter settings, and model design, which warrant further refinement in future research. While official construction accident investigation reports offer strong advantages in terms of authenticity and authority, their practical application is still subject to certain systematic biases. For instance, the exclusion of near-miss incidents may lead to an underestimation of the true frequency of potential risks. In addition, variations across regions in reporting standards, information disclosure completeness, and investigation depth may compromise the consistency and comparability of the dataset. Future studies should therefore carefully account for these latent biases and integrate multi-source heterogeneous data to enhance the comprehensiveness and robustness of the constructed risk landscape. Moreover, although large language models and embedding-based methods were introduced in this study to address ambiguity resolution and feature optimization, contextual dependency in Chinese textual expressions may still introduce identification biases. This highlights the need for further methodological refinement to achieve more fine-grained and semantically robust classification. In addition, the threshold settings for association rule mining inevitably involve a degree of subjectivity, which may affect the robustness of the results. Future research should incorporate sensitivity analysis and explicitly consider temporal dynamics to better capture the evolutionary patterns of risk coupling relationships.

Finally, the current complex network analysis primarily focuses on node-level importance identification, while quantitative modeling of risk propagation pathways remains insufficiently explored. Future work should further develop dynamic simulation approaches to more rigorously characterize risk diffusion mechanisms across the network structure.

6. Conclusions

This study developed an integrated, data-driven framework that combines large language model-based extraction with FP-Growth association rule mining to systematically

identify construction safety risk factors, construct a structured risk indicator system, and uncover strong inter-factor associations from accident investigation reports. Three principal conclusions emerge:

1. The distribution of construction safety risks shows strong systemic and highly uneven characteristics, with the Method dimension playing a dominant role. At a deeper level, organizational and institutional factors consistently occupy core positions in the risk network. On average, each accident involves 11.81 risk factors, of which management-related factors account for 68.3%, highlighting the dominant role of managerial deficiencies in construction safety accidents. Association rule analysis further shows that the lack of safety education and training acts as a key hub in the risk system. It is strongly associated with inadequate safety inspection and hazard identification and rectification (support = 0.967, confidence = 0.987), weak implementation of safety responsibility systems (support = 0.920, confidence = 0.998), and deficiencies in emergency management (support = 0.846, confidence = 0.995). These results indicate that the underlying mechanisms of risk factors are largely universal. However, differences in safety culture, regulatory intensity, and organizational modes across countries may influence their specific manifestations.
2. Risk evolution follows two characteristic transmission pathways. One is a cascading cognitive-behavioral pathway, whereby environmental deficiencies erode workers' safety awareness and subsequently induce unsafe behaviors, providing empirical support for the safety psychology principle that "environment shapes cognition." The other is a direct interaction pathway between human and material factors, in which reduced safety distances act as a critical interface, allowing unsafe behaviors and hazardous conditions to converge and trigger accidents.
3. The co-occurrence of management deficiencies exhibits a pronounced "amplification effect." When multiple managerial factors fail simultaneously, the resulting systemic risk is substantially greater than that caused by any single-factor failure. In extreme cases, such co-failures may escalate localized breakdowns of the safety management system into systemic collapse. This finding further substantiates that construction safety accidents are the outcome of multi-factor interaction and dynamic coupling. It also confirms that construction safety risks do not exist in isolation; instead, they form a structured network system through complex interdependencies. Complex network analysis further validates the structural robustness of the association rule mining results. At the network level, it reveals the key coupling mechanisms and critical coupling factors underlying construction safety risks.

Collectively, these results establish a robust empirical and methodological foundation for advancing precision risk identification and systemic prevention in construction safety, offering a scalable paradigm for analyzing complex risk systems in high-hazard industries.

Author Contributions: Methodology, writing—original draft preparation, and visualization, G.Z. Conceptualization, validation, writing—review and editing, supervision, project administration, and resources, D.Y. Formal analysis, investigation, and data curation, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Giang, D.T.H.; Pheng, L.S. Role of construction in economic development: Review of key concepts in the past 40 years. *Habitat Int.* **2011**, *35*, 118–125. [CrossRef]
2. Collie, A. Disparities in death at work: Reflections on occupational injury fatality data. *Occup. Environ. Med.* **2024**, *81*, 167–168. [CrossRef]
3. Barua, U.; Wiersma, J.W.F.; Ansary, M.A. Can rana plaza happen again in Bangladesh? *Saf. Sci.* **2021**, *135*, 105103. [CrossRef]
4. Hebei Department of Emergency Management. Investigation Report on the “4·25” Major Construction Hoist Car Fall Accident at Feicui Huating in Hengshui City [OL] Hebei Department of Emergency Management, 2019. Available online: <https://yjgl.hebei.gov.cn/portal/index/getPortalNewsDetails?id=93a0c0cc-4ffd-4688-afeb-ced41ae43c86&categoryid=3a9d0375-6937-4730-bf52-febb997d8b48> (accessed on 12 March 2026).
5. Wang, B.; Wang, Y.; Xu, F.; Shi, Z. Intelligence-led accident prevention and its application in petrochemical Enterprises. *Process Saf. Environ. Prot.* **2024**, *184*, 690–702. [CrossRef]
6. Olimat, H.; Alwashah, Z.; Abudayyeh, O.; Liu, H. Data-Driven Analysis of Construction Safety Dynamics: Regulatory Frameworks, Evolutionary Patterns, and Technological Innovations. *Buildings* **2025**, *15*, 1680. [CrossRef]
7. Wang, D.; Yin, K.; Wang, H. Risk identification in prefabricated building construction safety systems based on STPA-TM. *Reliab. Eng. Syst. Saf.* **2026**, *268*, 112004. [CrossRef]
8. Liu, C.; Yang, S. Using text mining to establish knowledge graph from accident/incident reports in risk assessment. *Expert Syst. Appl.* **2022**, *207*, 117991. [CrossRef]
9. Hai, N.; Gong, D.; Liu, S.; Dai, Z. Dynamic coupling risk assessment model of utility tunnels based on multimethod fusion. *Reliab. Eng. Syst. Saf.* **2022**, *228*, 108773. [CrossRef]
10. Xu, X.; Zou, P.X.W. Discovery of new safety knowledge from mining large injury dataset in construction. *Saf. Sci.* **2021**, *144*, 105481. [CrossRef]
11. Chowdhury, A.M.; Park, S.I.; Choi, J.-H. Safety Scheduling Through Integrated Accident Analysis Using Multiple Correspondence Analysis and Association Rule Mining: A Construction Engineering Perspective. *Buildings* **2025**, *15*, 4020. [CrossRef]
12. Nwafor, M. Mitigating uncertainty: Analyzing the role of risk management in the construction industry. *Harrisbg. Univ. Sci. Technol.* **2024**. Available online: <https://digitalcommons.harrisburgu.edu/dandt/24/> (accessed on 12 March 2026).
13. Liu, J.; Yan, X.; Gao, W. A Dualistic Perspective of Opportunity and Risk: The Impact of Head-Mounted Augmented Reality on Construction Onsite Hazard Identification of Workers. *J. Constr. Eng. Manag.* **2024**, *150*, 04024160. [CrossRef]
14. Zhang, S.; Loosemore, M.; Sunindijo, R.Y.; Galvin, S.; Wu, J.; Zhang, S. Assessing Safety Risk Management Performance in Chinese Subway Construction Projects: A Multistakeholder Perspective. *J. Manag. Eng.* **2022**, *38*, 05022009. [CrossRef]
15. Liang, Y.; Xu, N.; Chang, H.; Qian, S.; Liu, Y. Automatic construction of risk transmission network about subway construction based on deep learning Models. *Sci. Rep.* **2025**, *15*, 16383. [CrossRef]
16. Goh, Y.M.; Ubeynarayana, C.U. Construction accident narrative classification: An evaluation of text mining Techniques. *Accid. Anal. Prev.* **2017**, *108*, 122–130. [CrossRef]
17. Kim, T.; Chi, S. Accident Case Retrieval and Analyses: Using Natural Language Processing in the Construction Industry. *J. Constr. Eng. Manag.* **2019**, *145*, 04019004. [CrossRef]
18. Wang, S. Development of an automated transformer-based text analysis framework for monitoring fire door defects in buildings. *Sci. Rep.* **2025**, *15*, 43910. [CrossRef]
19. Wang, S. Graph neural network-driven text classification for fire-door defect inspection in pre-completion construction. *Sci. Rep.* **2025**, *15*, 44382. [CrossRef]
20. Zhu, Y.; Yuan, H.; Wang, S.; Liu, J.; Liu, W.; Deng, C.; Chen, H.; Liu, Z.; Dou, Z.; Wen, J.-R. Large language models for information retrieval: A survey. *ACM Trans. Inf. Syst.* **2025**, *44*, 1–54. [CrossRef]
21. Oral, M.; Alboga, Ö.; Aydinli, S.; Erdis, E. Usability of large language models for building construction safety risk assessment. *Eng. Constr. Archit. Manag.* **2025**.
22. Wu, W.; Wen, C.; Yuan, Q.; Chen, Q.; Cao, Y. Construction and application of knowledge graph for construction accidents based on deep learning. *Eng. Constr. Archit. Manag.* **2023**, *32*, 1097–1121. [CrossRef]
23. Deng, Z.; Ma, W.; Han, Q.L.; Zhou, W.; Zhu, X.; Wen, S.; Xiang, Y. Exploring DeepSeek: A Survey on Advances, Applications, Challenges and Future Directions. *IEEE/CAA J. Autom. Sin.* **2025**, *12*, 872–893. [CrossRef]
24. Ma, G.; Wu, Z.; Jia, J.; Shang, S. Safety risk factors comprehensive analysis for construction project: Combined cascading effect and machine learning Approach. *Saf. Sci.* **2021**, *143*, 105410. [CrossRef]
25. Fu, L.; Li, X.; Wang, X.; Li, M. Safety risk propagation in complex construction projects: Insights from metro deep foundation pit projects. *Reliab. Eng. Syst. Saf.* **2025**, *257*, 110858. [CrossRef]
26. Jiang, J.; Liu, G.; Ou, X. Risk Coupling Analysis of Deep Foundation Pits Adjacent to Existing Underpass Tunnels Based on Dynamic Bayesian Network and N-K Model. *Appl. Sci.* **2022**, *12*, 10467. [CrossRef]

27. Guo, Q.; Amin, S.; Wang, H.; Yan, H. Coupling Simulation of Human-Environmental Safety Risk Factors in Metro Construction—a Case Study of Rongjiazhai Station at Xi'an Metro Line 5 in China. *Int. J. Constr. Educ. Res.* **2023**, *20*, 26–42. [\[CrossRef\]](#)
28. Yan, K.; Jin, L.; Yu, X. Ordered weighted evaluation method of lifting operation safety risks considering coupling effect. *Sci. Rep.* **2024**, *14*, 5776. [\[CrossRef\]](#) [\[PubMed\]](#)
29. He, Y.; Li, J. Analysis of coal mining accident risk factors based on text mining. *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.* **2025**, *239*, 630–644. [\[CrossRef\]](#)
30. Han, Y.; Shen, J.; Zhu, X.; An, B.; Bao, X. Interaction mechanisms of interface management risks in complex systems of high-speed rail construction projects: An association rule mining-based modeling framework. *Eng. Constr. Archit. Manag.* **2024**, *31*, 2101–2127. [\[CrossRef\]](#)
31. Liu, Z.; Meng, X.; Xing, Z.; Jiang, A. Digital Twin-Based Safety Risk Coupling of Prefabricated Building Hoisting. *Sensors* **2021**, *21*, 3583. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Fu, L.; Wang, X.; Zhao, H.; Li, M. Interactions among safety risks in metro deep foundation pit projects: An association rule mining-based modeling framework. *Reliab. Eng. Syst. Saf.* **2022**, *221*, 108381. [\[CrossRef\]](#)
33. Li, H.; Xiao, J.; Gan, L.; Liu, K. Prediction of navigation aid malfunction based on hash chain-optimized FP-growth and gradient boosting random forest. *Reliab. Eng. Syst. Saf.* **2026**, *269*, 112046. [\[CrossRef\]](#)
34. Lawal, M.M.; Matthew, O.T. FP-Growth Algorithm: Mining Association Rules without Candidate Sets Generation. *Kasu J. Comput. Sci.* **2024**, *1*, 392–411. [\[CrossRef\]](#)
35. Yang, Y.; Wang, Y.; Easa, S.M.; Yan, X. Risk factors influencing tunnel construction safety: Structural equation model Approach. *Heliyon* **2023**, *9*, e12924. [\[CrossRef\]](#)
36. Hunyadi, I.D.; Constantinescu, N.; Ticleanu, O.A. Efficient Discovery of Association Rules in E-Commerce: Comparing Candidate Generation and Pattern Growth Techniques. *Appl. Sci.* **2025**, *15*, 5498. [\[CrossRef\]](#)
37. Lin, K.C.; Liao, I.E.; Chen, Z.S. An improved frequent pattern growth method for mining association rules. *Expert Syst. Appl.* **2011**, *38*, 5154–5161. [\[CrossRef\]](#)
38. Xiang, P.; Yang, Y.; Yan, K.; Jin, L. Identification of Key Safety Risk Factors and Coupling Paths in Mega Construction Projects. *J. Manag. Eng.* **2024**, *40*, 04024023. [\[CrossRef\]](#)
39. Zhu, Y.; Li, C.; Li, L.; Yang, K.; Yang, Y.; Zhang, G. Dynamic assessment and system dynamics simulation of safety risk in whole life cycle of coal mine. *Environ. Sci. Pollut. Res.* **2023**, *30*, 64154–64167. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Tanabe, K. Pareto's 80/20 Rule and the Gaussian Distribution. *Phys. A* **2018**, *510*, 635–640. [\[CrossRef\]](#)
41. GB/T 13861-2022; Classification and Code for the Hazardous and Harmful Factors in Process. State Administration for Market Regulation of China, Standardization Administration of China: Beijing, China, 2022.
42. Yoon, Y.G.; Ahn, C.R.; Yum, S.G.; Oh, T.K. Establishment of Safety Management Measures for Major Construction Workers through the Association Rule Mining Analysis of the Data on Construction Accidents in Korea. *Buildings* **2024**, *14*, 998. [\[CrossRef\]](#)
43. Liu, J.; Wang, Y.; Deng, C.; Jin, Z.; Wang, G.; Yang, C.; Li, X. Research on safety supervision and management system of China railway based on association rule and DEMATEL. *PLoS ONE* **2023**, *18*, e0295755. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Liu, Q.; Ding, Y.; Luo, X. Automated knowledge graph-based risk assessment for fall-from-height accidents in construction. *Autom. Constr.* **2025**, *179*, 106482. [\[CrossRef\]](#)
45. Liu, W.; Kang, X.; Ye, Q.; Xie, J. Unraveling hierarchical penetration mechanisms and coupling relationships of safety risks in major transportation infrastructure construction using text mining and complex networks. *Sci. Rep.* **2026**, *16*, 7313. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Edmondson, A.C.; Bransby, D.P. Psychological Safety Comes of Age: Observed Themes in an Established Literature. *Annu. Rev. Organ. Psychol. Organ. Behav.* **2023**, *10*, 55–78. [\[CrossRef\]](#)
47. Li, X.; Liu, W.; Chen, B.; Zhou, N.; Huang, W.; Liang, Y.; Yuan, X.; Li, Z. Domino Effect Risk Modeling and Analysis of Tank Area Accidents Based on Accident Chain and Multifactor Coupling. *ACS Chem. Health Saf.* **2025**, *32*, 413–425. [\[CrossRef\]](#)
48. Zhou, F.; Zhang, J.; Fu, C. Generation paths of major production safety accidents—A fuzzy-set qualitative comparative analysis based on Chinese data. *Front. Public Health* **2023**, *11*, 1136640. [\[CrossRef\]](#)
49. Niu, H.; Yang, X.; Zhang, J.; Guo, S. Risk coupling analysis of causal factors in construction Fall-from-height Accidents. *Eng. Constr. Archit. Manag.* **2024**, *32*, 6045–6067. [\[CrossRef\]](#)

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.