Interpreting accident reports by integrating a heterogeneous graph neural network and factor analysis

Junyu Chen, Hung-Lin Chi, Bo Xiao, Rongyan Li
The Hong Kong Polytechnic University, Hong Kong SAR
hung-lin.chi@polyu.edu.hk

Abstract. Occupational safety in the construction industry is one highly prioritized concern around the globe. Accident reports are considered valuable recourses preserving information about corresponding risk factors. Many efforts in the literature have demonstrated that deep learning models are readily applicable to processing and analyzing narrative reports. However, the heterogeneous semantic information was rarely considered. This research utilizes knowledge graph-based accident analysis to provide a machine-assisted approach for construction accident report interpretation. To validate the proposed approach, this research labels 320 crane-related accident reports from the US OSHA database and develops a Crane Safety Knowledge Graph (CSKG) as a case study. Then, a Heterogeneous Graph Attention Network (HAN) is trained to explore the accident features and the importance of various risk factors. Through mapping and clustering the accident data points, the results reveal the capability of the proposed approach to learn the accident patterns and generate safety rules for construction cranes.

1. Introduction

Improving occupational safety is a challenging issue across the globe. In Europe, a total of 3581 occupational fatal accidents were recorded in the year 2018, where 716 were from the construction industry (Eurostat, 2022). In the United States, 154 construction work-related fatalities and catastrophes were reported in the year 2022 (Occupational Safety and Health Administration, 2023). The situations are even worse in developing countries and one common feature revealed by statistical data from different countries is that the construction industry is liable for a significant proportion of occupational fatalities and injuries (Mohandes et al., 2022).

The risk factors leading to construction accidents may have latent interrelations and coupling effects that should not be investigated from an isolated aspect. Hence, to promote occupational safety in the construction industry towards "Zero Accident Vision", it is vital to identify potential risk factors for prevention purposes through comprehensive accident analysis. In this direction, it has been an interest of scholars to investigate the causal factors of construction accidents by combining expert knowledge and informative accident reports. For instance, Dhalmahapatra et al. (2020) put forward an integrated modeling approach for accident data, which was based on categorical variables extracted from 179 crane-related incidents and numerical values obtained from expert surveys. Recently, this field has gradually broadened to deploy deep learning methods to extract information from massive accident reports and provide effective and replicable analytics (Sarkar and Maiti, 2020). For example, some research utilized Natural Language Processing (NLP) and advanced neural networks to process the textual accident data and achieve automatic text classification (Fang, Luo, et al., 2020; Gupta et al., 2022). However, the heterogeneous semantic information in the accident reports, such as features of construction activities, human errors, mechanical problems, and environmental hazards, is rarely systematically considered.

Knowledge Graph (KG) is a technique originating from the development of modular instructional systems for education as early as 1972 (Schneider, 1972). Google launched its KG in 2012 to enhance Google Search and has brought this term into many areas like

recommendation systems (Wang et al., 2019). In knowledge representation and reasoning, KG has become an effective tool to retrieve, analyze and visualize heterogeneous semantic information. A generic KG is composed of a data layer and a schema layer. The data layer contains domain knowledge, utilizing entities, attributes, and relations to represent semantic information. And the schema layer integrates information from the data layer into an ontology model for logic inference. As a pioneering research work to apply KG in construction safety management, Fang, Ma, et al. (2020) utilized computer vision technology to extract knowledge for ontology modeling and developed a KG for identifying construction hazards, such as the lack of safety harnesses for workers working at height. However, there are limited unsafe behaviors identified in the small-scale KG. As noted by Liu et al. (2022), more research on KG-based accident analysis for construction safety management is needed.

As exemplified in Figure 1(a) and (b), an accident database can be considered a heterogeneous KG. Each accident case is represented as an ontology entity (e.g., the accident entity a_1); the common accident causations shared by two accidents are represented as heterogeneous ontology relations connecting the paired entities (e.g., accident entities a_1 and a_3 are connected via two different accident causations). Heterogeneous graph neural networks (HGNNs) were proposed in recent years to capture various semantics in real-world graphs (Xiao Wang et al., 2019; Fu et al., 2020). As shown in Figure 1(c) and (d), a typical HGNN uses the meta-paths, which convey different semantic information, to enlarge the receptive field of each entity in passing information to its neighbors. This research proposes a KG-based risk factor modeling framework for accident report interpretation. The key contributions are two-fold: 1) In the case study, an accident-enabled Crane Safety Knowledge Graph (CSKG) is developed through the pre-processing of textual data and the extraction of causal factors from multiple aspects of the accident reports. The ontology modeling process can lay a foundation for crane safety management and support KG development in other research fields. 2) This research combines an HGNN and factor analysis to infer safety knowledge in the CSKG, which contains comprehensive information and rich semantics. The risk analysis framework is tested and validated as an effective tool that can assist in identifying essential risk factors and recognizing accident patterns.

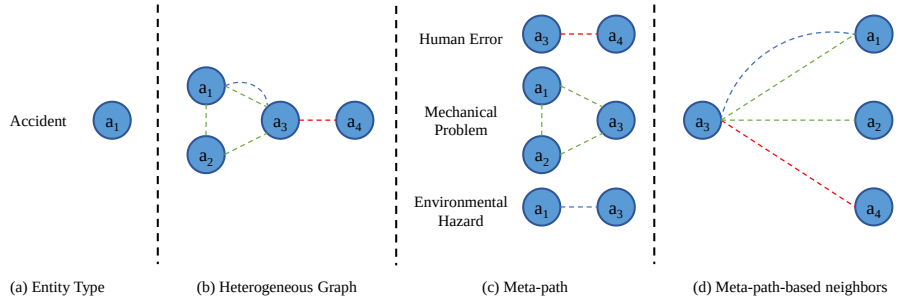

Figure 1: Exemplification of the heterogeneous crane safety knowledge graph (CSKG)

2. Problem Formulation

The problem tackled in this paper is to identify essential risk factors and accident patterns by extracting information from massive accident reports. The utilization of the textual accident data and the problem formulation are introduced below.

Risk factors revealed in accident reports may include human errors, mechanical problems, and environmental hazards, often involving the interrelationships and causations among these factors. The current post-accident investigation often emphasizes the risk factors on a case-by-

case basis, seeking attention to corresponding precautionary measures. However, for different accident cases, there is a variety of risk scenarios that present a wide spectrum of on-site conditions such as misoperations, malfunctions, and constraints. Hence, the first challenge associated with the utilization of textual accident data can be identified as how to properly extract and manage the accident information for inference, such as eliciting general safety rules. As defined by Gruber, ontology denotes the specification of conceptualization in a formal and explicit manner; ontology modeling is the process to model a set of concepts and their relationships into ontologies within a knowledge domain (Gruber, 1993). In this research, ontology modeling is potentially a promising approach to accommodate the multifariousness of different accidents while preserving the generalization ability of the developed KG, transforming the unstructured textual accident data into a structured accident database.

The latent features of accidents (e.g., the importance of various risk factors and the accident patterns) preserved in the accident-enabled KG can be learned by an HGNN, which is generally comprised of layers for input features, knowledge inference, and output predictions. Referring to previous research: 1) one-sentence accident summaries can be used as inputs through word embedding. This process utilizes a corpus of text and an embedding method to reconstruct the word sequence of accident summaries into a vector space. 2) the prediction of accident consequences, represented as accident types, are the expected outputs from model training. This process utilizes a Multi-Layered Perceptron (MLP) to train the proposed HGNN and evaluates the model through its performance in the learning task of semi-supervised node classification. Taking the one-sentence accident summaries as inputs, and the accident consequences as outputs, the second challenge in the exploration of textual accident data refers to how to devise the hidden layers of an HGNN to automatically extract the heterogeneous semantic information without requiring complicated reasoning or decision-making. The Heterogeneous Graph Attention Network (HAN), deploying a hierarchical attention framework to encapsulate the heterogeneous semantic information, can be considered a well-suited approach to extract valuable accident entities and semantics. Combined with clustering analysis, which aims to group accident data, hence the accident entities are closely associated with each other in the same cluster while separated from those in other clusters. Consequently, accident patterns can be recognized from the accident-enabled KG.

3. Methodology

Following the research purpose of exploring the importance of risk factors and patterns of construction accidents, this research proposes a KG-based risk factor modeling framework for accident report interpretation while overcoming the two challenges mentioned above: 1) the unstructured textual accident data and 2) the latent accident features preserved in the heterogeneous semantic information. The proposed framework is comprised of three steps: ontology modeling, construction and implementation of the HAN model, and clustering analysis.

3.1 Ontology Modeling

Within the context of construction safety, the entities of the accident-enabled ontology model are construction accident cases, which contain heterogeneous semantic information that specifies the basic information of the involved construction sites, the identified causal factors, and the accident consequences. Accordingly, an ontology relation in the model is defined as the accident causation that leads to the occurrence of its connected accident entities. Different ontology relations convey different semantic information and comprise the schema layer of the

heterogeneous KG. A careful interpretation process is then conducted to extract the safety knowledge from the narrative accident reports and develop it into the data layer of the heterogeneous KG.

3.2 Construction and Implementation of The HAN Model

The construction and implementation of the HAN model are described as follows.

The Node-Level Attention Mechanism. Previous to utilizing the accident information learned from the neighbor entities for a specific accident entity, we should notice that different neighbor entities may play different roles and bear diverse importance in forming the representation of the studied entity. Figure 2(a) shows the structure of the node-level attention mechanism, which takes three steps to learn the weights of node pairs under each specific meta-path:

Step 1: For each attention head *K*, use the transformation matrix to embed the accident features into a common vector space:

$$W_{node} h_i = h_i', \tag{1}$$

where h_i is the original representation of entity *i*; h_i' is the embedded representation of entity *i*.

Step 2: Calculate the normalized weight coefficients for entities $j \in \mathcal{N}_i^{\Phi}$:

$$\alpha_{ij}^{\Phi} = \frac{\exp\left(\sigma\left(\mathrm{a}_{\Phi}^{\mathrm{T}} \cdot \left[h_i' \parallel h_j'\right]\right)\right)}{\sum_{k \in \mathcal{N}_i^{\Phi}} \exp\left(\sigma\left(\mathrm{a}_{\Phi}^{\mathrm{T}} \cdot \left[h_i' \parallel h_k'\right]\right)\right)}, \tag{2}$$

where $\mathcal{N}_i^{\Phi}$ refers to neighbor entities of entity *i* under the meta-path Φ; σ refers to the LeakyReLU function; a_{Φ} refers to the vector containing node attention values of node pairs under the meta-path Φ; $\parallel$ refers to the concatenation operation.

Step 3: Under the specified meta-path, generate the embedding z_i^{Φ} for entity *i* using the neighbors' projected features with corresponding weight coefficients:

$$z_i^{\Phi} = \sigma'\left(\sum_{j \in \mathrm{N}_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot h_j'\right), \tag{3}$$

where σ' denotes the ELU function.

The Semantic-Level Attention Mechanism. In general, each accident entity contains various semantic information, reflecting the accident features from a variety of aspects. The semantic-specific embedding of an accident entity cannot capture the rich semantics. As shown in Figure 2(b), with the obtained embeddings from each semantic aspect as input, at the semantic level, the attention mechanism aims to learn different meta-path weights, fuse information from various semantic aspects, and finalize each accident entity embedding through the three steps as follows:

Step 1: Transform the semantic-specific embedding $z_i^{\Phi_p}$ through a one-layer MLP:

$$tanh\left(W_{sem} \cdot z_i^{\Phi_p} + b\right) = z_i^{\Phi_P''}, \tag{4}$$

where W_{sem} refers to the weight matrix and b refers to the bias vector.

Step 2: Calculate and normalize the weights of different meta-paths:

$$\beta_{\Phi_i} = \frac{\exp\left(\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} q^T \cdot z_i^{\Phi_P''}\right)}{\sum_{i=1}^{P}\exp(\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} q^T \cdot z_i^{\Phi_P''}))}, \tag{5}$$

where q is the vector containing semantic attention values of node pairs from various meta-paths between the entity set $\mathcal{V}$; β_{Φ_i} evaluates the information contribution of the meta-path Φ_i.

Step 3: Fuse all the semantic-specific embeddings with the corresponding meta-path weights to generate the final embedding Z_i of node i:

$$Z_i = \sum_{l=1}^{P} \beta_{\Phi_l} \cdot z_i^{\Phi_l}. \tag{6}$$

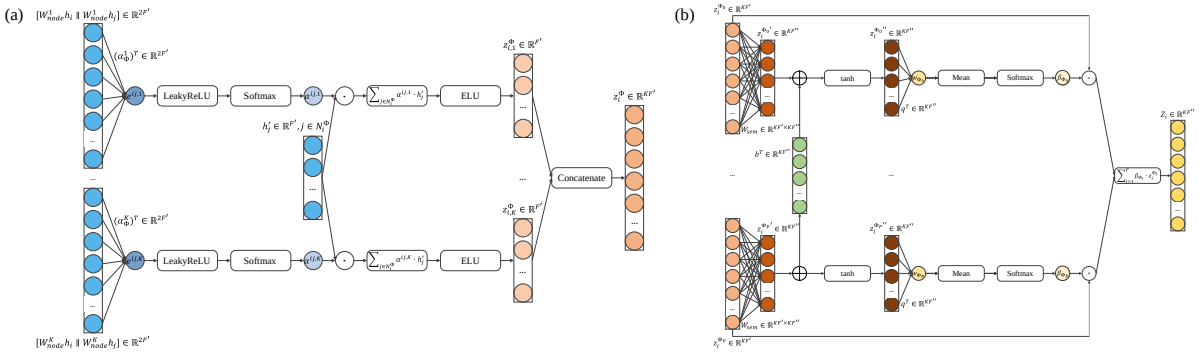

Figure 2: The hierarchical attention structure in the HAN model

Analysis of The Overall HAN Model. The final embedding of each accident entity is obtained by aggregating the embeddings carrying different semantics. For the learning task of semi-supervised node classification in this research, the labeled ground truths ($\boldsymbol{Y_l}$) and the prediction outputs from the HAN model of the accident types are utilized to minimize the loss (L):

$$\mathrm{L} = -\sum_{l\in y_L} Y_l \cdot \ln(\mathrm{C} \cdot Z_l)\,, \tag{7}$$

where C is the classifier parameter; y_L and Z_l are the labeled accident entities and embeddings.

Finally, the final embeddings of the accident nodes (Z), the node weight coefficients (α), and the semantic weight coefficients (β) can be obtained. To assess the model performance in terms of the accuracy of predicting accident consequences, the micro F1 is computed by calculating the numbers of True Positives, False Negatives, and False Positives.

3.3 Clustering Analysis

To further explore the accident information preserved in the heterogeneous KG and learned by the HAN model, clustering analysis is performed to explore the accident patterns and contributing risk factors. This research utilizes the t-distributed stochastic neighbor embedding (t-SNE) method to map the accident entities, retaining the information preserved in the high-dimensional embeddings (Maaten and Hinton, 2008). Afterward, this research compares different clustering algorithms based on their performance in partitioning the accident data points into clusters. The validity indices (e.g., DBI, DI, and SW) can identify the cohesion of data points in the same cluster and their separation from other clusters, and help retain optimal clustering results for identifying the accident patterns and eliciting safety rules (Dunn, 1974; Davies and Bouldin, 1979; Rousseeuw, 1987).

4. Case Study

Data Collection. Fatality and Catastrophe Investigation Summaries reported by the US OSHA were considered in this research. The case study focused on construction crane safety in the past two decades. The scope of data collection was hereby specified by using the keyword "construction crane" for retrieval on the US OSHA website. Through a careful interpretation process, the accident reports not involving crane usage or construction activities were ruled out and a total of 320 cases were compiled in the final accident database in this research.

Ontology-Based Knowledge Extraction. As shown in Table 1, the labeling for the accident consequences was following the Top Four construction hazards identified by the US OSHA. The basic accident information considered the features of involved construction sites and activities. For accident causations, the primary causes were categorized as human errors, mechanical problems, and environmental hazards. The original accident narratives were labeled from each information aspect and category, determining the corresponding attributes of the accident entities.

Implementation of The HAN Model. Referring to (Xiao Wang et al., 2019), the configuration of parameters to implement the HAN model in this research was set and adjusted as follows: the learning rate: 0.005; the regularization parameter: 0.001; the dimension of final embeddings: 64; the number of attention heads: 8; the dropout ratio of attention: 0.6; the ratios of training data, validation data, testing data: 80%, 10%, and 10%. An early stopping mechanism was adopted to terminate the model training when the decrease in validation loss is not observed in 100 consecutive epochs.

5. Results and Discussion

CSKG in The Construction Industry. For an illustration of a sub-graph of the heterogeneous CSKG, the accident nodes and their associations under the meta-path "Operator misoperation" (A_HEa_A) are visualized using Gephi 0.10 in Figure 3. There were 16 categories of causal factors under this meta-path, connecting their corresponding accident nodes.

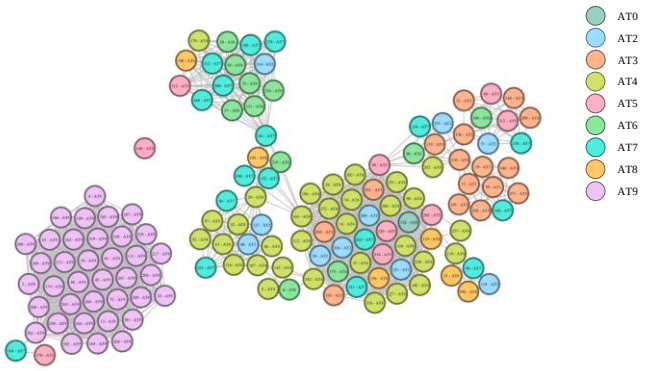

Figure 3: The accident nodes connected via the meta-path"Operator misoperation"

Training Performance of The HAN Model. To investigate the importance of different safety concerns, the HAN model was implemented using the 320 datasets. For the task of predicting crane accident consequences, the higher accuracy indicates the higher reliability of the model; the smaller loss indicates the prediction is closer to the ground truth. As shown in Figure 4, the training accuracy was close to 1; the validation accuracy was larger than 0.8. Whereas the

training loss was close to 0; the validation loss was smaller than 0.3. The training results demonstrated the reliability of the trained HAN model for the sampling accident data.

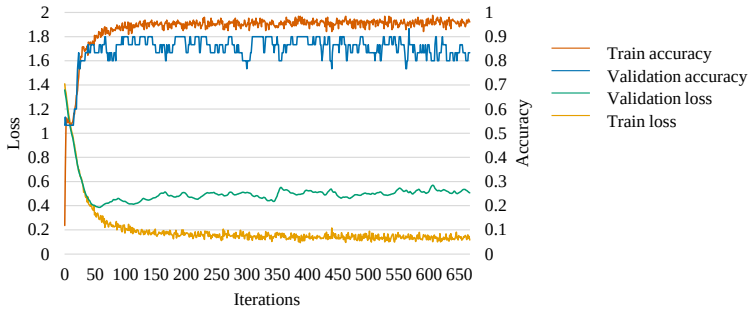

Figure 4: The iterative process of model training

Table 1: The aspects of accident information and the corresponding ontology relations.

Aspects	Ontology relations	Aspects	Ontology relations
Accident Type	Fall from the personnel basket (AT0)	Human Error	Operator misoperation (A_HEa_A)
	Fall from the extension ladder (AT1)		Rigger misoperation (A_HEb_A)
	Fall from the constructed structure (AT2)		Signaller misoperation (A_HEc_A)
	Fall from the crane (AT3)		Assembly/disassembly misoperation (A_HEd_A)
	Struck by loads (AT4)		Maintenance or inspection misoperation (A_HEe_A)
	Struck by falling crane parts (AT5)	Mechanical Problem	Crane collapse (A_MPa_A)
	Struck by the moving machine (AT6)		Crane tip-over (A_MPb_A)
	Body caught in/between (AT7)		Fall of crane jib/boom (A_MPc_A)
	Finger/hand/foot caught in/between (AT8)		Fall/Shift of crane loads (A_MPd_A)
	Electrocutions (AT9)		Malfunction/failure of crane (A_MPe_A)
Basic Information	Involved Staff (A_IS_A)	Environmental Hazard	Poor weather/operation conditions (A_EHa_A)
	Work Process (A_WP_A)		A lack of standard procedure (A_EHb_A)
	Project Feature (A_PF_A)		A lack of clear division of work area (A_EHc_A)
	Work Shift (A_WS_A)		A lack of sufficient PPE devices (A_EHd_A)
	Machine Type (A_MT_A)		A lack of sufficient inspections (A_EHe_A)

Representation of Accident Entities and Weights of Meta-Paths. Through the training process, each accident node was represented as a 64-dimensional vector. To visualize the 320 accident entities as data points as well as retain the information from the final embeddings, the t-SNE method was used to map the 320 data points as shown in Figure 5(a). The normalized weights of meta-paths were also obtained from the trained HAN model, indicating the different importance of various causal factors leading to crane accidents. As indicated in Table 2, considering the information revealed by meta-paths with a weight higher than 0.01, research findings were summarized from three aspects: 1) from the human error aspect, the misoperation of maintenance workers or inspectors was identified as an important risk factor for crane operation; 2) from the mechanical problem aspect, the fall of the crane boom or jib was revealed as an essential risk factor for crane operation, followed by malfunction or failure of the crane, and crane collapse; 3) from the environmental hazard aspect, poor weather or operation conditions were essential for crane operation, followed by a lack of clear division of work area, and a lack of PPE or communication devices.

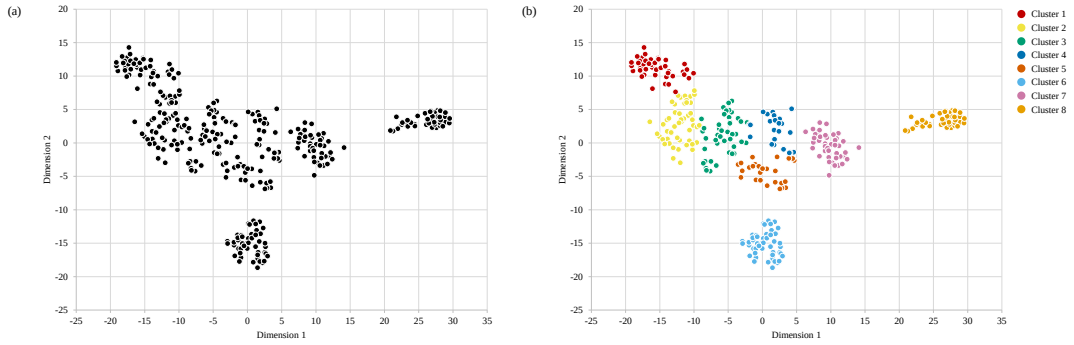

Figure 5: Visualization of the 320 accident data points (a) using t-SNE; (b) using fuzzy c-means

Table 2: Obtained weights of meta-paths for three time periods.

Meta-path code	Normalized Weight	Meta-path code	Normalized Weight	Meta-path code	Normalized Weight	Meta-path code	Normalized Weight
A_IS_A	0.0005	A_HEa_A	0.0015	A_MPa_A	0.0238	A_EHa_A	0.2237
A_WP_A	0.0004	A_HEb_A	0.0008	A_MPb_A	0.0008	A_EHb_A	0.0008
A_PF_A	0.0005	A_HEc_A	0.0006	A_MPc_A	0.2393	A_EHc_A	0.1030
A_WS_A	0.0006	A_HEd_A	0.0026	A_MPd_A	0.0008	A_EHd_A	0.1769
A_MT_A	0.0005	A_HEe_A	0.0752	A_MPe_A	0.1458	A_EHe_A	0.0020

Accident Pattern Detection. As shown in Table 3, considering that a lower DBI value, a higher DI value, and a higher SW value lead to better clustering, the optimal number of clusters was eight, which was achieved by using the fuzzy c-means clustering algorithm. As illustrated in Figure 5(b), the accident pattern detection can be summarized as follows:

1) From Cluster 1, Cluster 2, and Cluster 3, it was found that some cases of being struck by loads (AT4) and being struck by falling crane parts (AT5) were highly associated, indicating the most significant contributing factor as the fall of the crane jib or boom due to a variety of reasons such as failed connections, the improper disassembly process, and system failures.

2) From Cluster 4 and Cluster 5, it was found that some accident cases from various accident types could also be highly associated, but emphasizing a significant common contributing factor as the lack of fall protection devices.

3) From Cluster 6, it was found that some cases of the body caught in or between (AT7) and the finger/hand/foot caught in or between (AT8) were highly associated, indicating the most significant contributing factor as not properly separating the operating crane from surrounding workers.

4) From Cluster 7, it was found that some cases of fall from the extension ladder (AT1), fall from constructed structure (AT2), and fall from the crane (AT3), were highly associated. In this cluster, the lack of fall protection devices was also emphasized as the most significant contributing factor.

5) From Cluster 8, it was found that the cases of electrocutions (AT9) were typical accident type that was not associated with other accident types, indicating the environmental hazard of overhead power lines across the site's air space as the most significant contributing factor.

Table 3: Comparison of validity indices for different clustering methods.

Number of clusters	**Clustering algorithms**					
	Fuzzy c-means			**K-means**		
	DBI	DI	SW	DBI	DI	SW
6	1.6812	0.0393	0.3575	1.6708	0.0284	0.5418
7	1.9370	0.0409	0.5598	2.0386	0.0245	0.5032
8	*1.8577*	*0.0411*	*0.5665*	2.1742	0.0337	0.4524
9	1.7858	0.0735	0.5545	1.9562	0.0406	0.4792
10	2.0383	0.0615	0.5096	1.9385	0.0445	0.4910

6. Conclusion

This research proposes an accident-enabled risk analysis modeling framework that utilizes ontology modeling, knowledge extraction, and knowledge inference for identifying risk factors and accident patterns preserved in massive accident reports. To achieve the proposed semi-automated construction accident report interpretation, the authors first developed an ontology model to specify the entities, attributes, and relations that should be considered for construction accident analysis. In addition, careful manual extraction and classification of the information in construction crane accident reports in the case study were conducted. The developed heterogeneous CSKG can support crane safety management and KG development in other realms. Furthermore, the authors constructed and implemented the HAN model for capturing heterogeneous accident information. The training and validation results indicated that the HAN model combined with factor clustering analysis can be used as an effective tool to elicit implications on construction safety management.

The limitations of this research should also be noted. The first limitation concerns the accident database, which may not contain all crane accident mechanisms. And the description of some crane accident cases lacks the information necessary for analysis. The second limitation comes from the manual labeling process for classifying the accident contributing factors and consequences into different categories, which is laborious and time-consuming, as well as existing a certain amount of subjectivity. The authors will consider taking advantage of the power of Large Language Models (LLMs) to develop and evaluate a ChatGPT-assisted accident analysis framework to extend this research study in the future.

References

Davies, D. L. and Bouldin, D. W. (1979) 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2), pp. 224–227.

Dhalmahapatra, K., Shingade, R. and Maiti, J. (2020) 'An innovative integrated modeling of safety data using multiple correspondence analysis and fuzzy discretization techniques', *Safety Science*, 130(January), p. 104828. doi: 10.1016/j.ssci.2020.104828.

Dunn, J. C. (1974) 'Well-separated clusters and optimal fuzzy partitions', *Journal of Cybernetics*, 4(1), pp. 95–104. doi: 10.1080/01969727408546059.

Eurostat (2022) *Fatal Accidents at Work by NACE Rev. 2 activity*. Available at: https://ec.europa.eu/eurostat/databrowser/view/HSW_N2_02/default/table?lang=en&category=hlth.hsw.hsw_acc_work.hsw_n2 (Accessed: 15 September 2022).

Fang, W., Luo, H., *et al.* (2020) 'Automated text classification of near-misses from safety reports: An improved deep learning approach', *Advanced Engineering Informatics*, 44(March), p. 101060. doi: 10.1016/j.aei.2020.101060.

Fang, W., Ma, L., *et al.* (2020) 'Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology', *Automation in Construction*, 119(May), p. 103310. doi: 10.1016/j.autcon.2020.103310.

Fu, X. *et al.* (2020) 'MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding', *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, pp. 2331–2341. doi: 10.1145/3366423.3380297.

Gruber, T. R. (1993) 'A translation approach to portable ontology specifications', *Knowledge Acquisition*, 5(2), pp. 199–220. doi: 10.1006/KNAC.1993.1008.

Gupta, A. K. *et al.* (2022) 'A novel classification approach based on context connotative network (CCNet): A case of construction site accidents', *Expert Systems with Applications*, 202(July 2021), p. 117281. doi: 10.1016/j.eswa.2022.117281.

Liu, J., Luo, H. and Liu, H. (2022) 'Deep learning-based data analytics for safety in construction', *Automation in Construction*, 140(February), p. 104302. doi: 10.1016/j.autcon.2022.104302.

Maaten, L. van der and Hinton, G. (2008) 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*, 9(1), pp. 2579–2605. doi: 10.1007/s10479-011-0841-3.

Mohandes, S. R. *et al.* (2022) 'Causal analysis of accidents on construction sites: A hybrid fuzzy Delphi and DEMATEL approach', *Safety Science*, 151(June 2021), p. 105730. doi: 10.1016/j.ssci.2022.105730.

Occupational Safety and Health Administration (2023) *Reports of Fatalities and Catastrophes*. Available at: https://www.osha.gov/ords/imis/accidentsearch.html (Accessed: 10 April 2023).

Rousseeuw, P. J. (1987) 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, 20(C), pp. 53–65. doi: 10.1016/0377-0427(87)90125-7.

Sarkar, S. and Maiti, J. (2020) 'Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis', *Safety Science*, 131(June), p. 104900. doi: 10.1016/j.ssci.2020.104900.

Schneider, E. W. (1972) *Course Modularization Applied: The Interface System and Its Implications For Sequence Control and Data Analysis*. Available at: https://eric.ed.gov/?id=ED088424.

Wang, Xiao *et al.* (2019) 'Heterogeneous graph attention network', *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, pp. 2022–2032. doi: 10.1145/3308558.3313562.

Wang, Xiang *et al.* (2019) 'KGAT: Knowledge graph attention network for recommendation', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 950–958. doi: 10.1145/3292500.3330989.