Research on Risk Early Warning Model of Oil and Gas Pipeline Graph Neural Network Based on Knowledge Graph

Yixin Wei[1], *Feng Yan*[1], *Chunming Wang*[1], *Yunteng Wen*[1], *Tiantian Liu*[2*]

[1] *PipeChina North Pipeline Company, China*
[2] *Beijing Information Science and Technology University, China*
weiyx@pipechina.com.cn, yanfeng@pipechina.com.cn, wangcm@pipechina.com.cn, wenyt@pipechina.com.cn, 2023020504@bistu.edu.cn

Abstract

With the rapid expansion of oil and gas pipeline networks, their operational safety faces increasingly complex risk threats. Conventional accident risk assessment methods mainly rely on manually defined rules and static indicators, making it difficult to uncover the deeper causal logic and structural patterns of accidents. To address this challenge, a risk early-warning model for oil and gas pipelines is proposed, integrating knowledge graph techniques with graph neural networks. Specifically, pipeline accident reports from Pipeline and Hazardous Materials Safety Administration (PHMSA) are consolidated to construct a comprehensive knowledge graph of pipeline risks. Based on this graph, a relation-aware graph neural network node classification approach is designed, which incorporates both structural features and numerical attributes to enable risk prediction. Within this framework, the Composition-based Multi-Relational Graph Convolutional Networks (CompMRGCN) model is further developed, extending the relation-aware graph convolutional network by embedding a Markov random field-based dependency mechanism to capture correlations among node labels during prediction. Experimental results demonstrate that the proposed early-warning model and CompMRGCN method achieve 96.2% accuracy, 95.8% F1-score, and 97.1% mAP, representing improvements of 6.7%, 5.6%, and 5.4% over the best existing baselines, respectively. Comparative analysis indicates that this approach substantially outperforms competing models in terms of accuracy, generalization, and interpretability, offering an effective and practical technical support for intelligent accident early warning and safety management of oil and gas pipelines.

Keywords: Accident risk prediction, Knowledge graph, Oil and gas pipelines, Relation-aware graph neural network

1 Introduction

As the lifeline of national energy infrastructure, the safe and stable operation of oil and gas pipelines is directly linked to economic efficiency, public safety, and environmental protection [1]. Traditional risk assessment techniques for pipelines primarily rely on statistical monitoring of physical sensor parameters such as pressure, flow rate, and temperature, combined with expert experience or fixed thresholds for alarm generation. However, such approaches are insufficient for capturing the coupled failures arising from multi-factor and multi-entity interactions within complex pipeline networks. They also exhibit limited capabilities in multi-source data integration and semantic modeling, thereby lacking systematic support for risk cognition [2]. In particular, when confronted with atypical risk inducers such as intentional damage or geological hazards, the timeliness and accuracy of traditional monitoring models decrease significantly [3].

In recent years, the rapid development of emerging technologies such as artificial intelligence, big data, and graph learning has opened up new directions for pipeline safety monitoring and risk early warning. Among them, Graph Neural Networks (GNNs), which can effectively model both the topological structure and attribute dependencies among entities, have become a powerful tool for structural modeling in complex networked systems [4]. Previous studies have demonstrated their adaptability in diverse application scenarios, including fault identification in power systems [5], water resource monitoring [6], and safety assessment of urban rail transit [7].

2 Related Work

2.1 Research Status of Graph Neural Network in Industrial Scene

GNNs as a deep learning paradigm designed for graph-structured data, have demonstrated outstanding performance in various domains in recent years [8]. The core idea of GNNs is to treat each entity (node) as a component of the graph and capture structural dependencies among nodes through iterative message-passing mechanisms, thereby generating high-quality node representations. In industrial applications, GNNs are particularly suitable for scenarios where devices exhibit topological structures or interaction relationships, such as power grid systems [9], water resource scheduling [10], and industrial process control [11]. For example, Wang et al. developed a Graph Convolutional Network (GCN) monitoring graph for wind farms, enabling turbine

*Corresponding Author: Tiantian Liu; Email: 2023020504@bistu.edu.cn

DOI: https://doi.org/10.70003/160792642025112606009

state classification and fault prediction [12]. Similarly, Veličković et al. employed a Graph Attention Network (GAT) to model the graph structure among sensor nodes, achieving effective representation and prediction of key indicators within workshops and improving both the flexibility and interpretability of the model [13].

2.2 Trends and Challenges of the Integration of GNN and Knowledge Mapping

The integration of GNNs and Knowledge Graphs (KGs) has been widely recognized as a promising direction for enhancing systematic knowledge modeling and intelligent reasoning, and recent studies have explored applications in finance, healthcare, and industrial manufacturing [14-15]. For instance, Zhao et al. proposed a Multi-Scale Dynamic Graph Neural Network (MSDG) model that integrates graph semantics with embedding features to address anomaly detection tasks in industrial sensors, demonstrating improved stability and generalization [14]. Similarly, Soler et al. designed a contextual safety modeling framework that combines ontology-based reasoning with GNNs, enabling more sensitive state recognition and responsive mechanisms [16]. To address the challenge of dynamic evolution modeling, Ma et al. developed a temporal graph-based predictive model that significantly enhanced response speed and prediction accuracy in industrial processes [17]. In terms of interpretability, Yuan et al. introduced an explainable GNN structure (xGNN) capable of explicitly modeling risk propagation paths, thereby providing decision-makers with causal chain–level analytical support [18].

Therefore, constructing a risk prediction model for oil and gas pipelines that leverages the semantic representation capabilities of KGs together with the structural modeling power of GNNs not only holds substantial theoretical research value but also offers broad engineering application prospects for advancing industrial intelligence and improving risk control capacity.

2.3 Research Objectives and Technical Route

Against the backdrop of frequent oil and gas pipeline accidents and the limited effectiveness of traditional early-warning approaches, advancing risk monitoring methods from a purely “data-driven” paradigm to an integrated framework of “structural modeling + knowledge reasoning” has become one of the core directions for intelligent development in the industry [19]. GNNs have demonstrated notable performance in tasks such as fault detection and risk propagation modeling [20]. In contrast, KGs excel at semantic-level entity modeling and causal logic representation [21]. The integration of these two approaches enables a joint representation of “structure + semantics” in complex systems, thereby providing a theoretical foundation for the development of predictive and interpretable intelligent risk identification systems.

This study focuses on the high-risk scenario of oil and gas pipeline accident risk prediction and proposes a modeling framework that integrates Relational Graph Convolutional Network (R-GCN) with multi-source knowledge graphs, thereby establishing an intelligent reasoning pathway from data to knowledge and from knowledge graphs to risk. The main contributions are as follows:

(1) A relation-aware graph neural network node classification method is designed which termed the Composition-based Multi-Relational Graph Convolutional Networks (CompMRGCN). Building upon an accident knowledge graph that incorporates both structural information and numerical attributes, a multi-relational graph convolutional framework with relation-aware mechanisms is developed to achieve unified modeling of node features and complex relationships.

(2) Within the CompMRGCN architecture, relation-aware mechanisms and label dependency modeling are introduced to enhance node semantic representations and capture potential associations among risk nodes, thereby substantially improving the accuracy and robustness of pipeline accident risk prediction.

(3) Systematic experiments conducted on a real pipeline accident dataset demonstrate that CompMRGCN significantly outperforms multiple baseline methods in terms of accuracy, F1 score, and mAP. Furthermore, ablation studies confirm the effectiveness of each component design and the overall superiority of the proposed.

3 Data Set Construction and Knowledge Mapping Preprocessing

3.1 Data Source and Structure Introduction

The dataset used in this study is primarily derived from the pipeline accident reports publicly released by the Pipeline and Hazardous Materials Safety Administration (PHMSA) under the U.S. Department of Transportation. As one of the most comprehensive pipeline failure databases worldwide, the PHMSA dataset documents accident information related to natural gas and liquid pipelines in North America since 2010. It contains detailed records of more than 10,000 pipeline accidents, covering multiple attributes such as accident occurrence time, geographic location, equipment type, operating company, economic losses, and casualties. This dataset is widely regarded as one of the most authoritative and complete oil and gas pipeline accident databases available internationally.

The raw data is provided in CSV format, and consists of both structured fields (e.g., state name, pipeline type, incident ID) and semi-structured fields (e.g., incident description, cause statement). Certain fields exhibit missing values or redundancies. Preliminary statistics indicate that the dataset contains more than 1,800 complete accident samples, with each record on average associated with multiple devices, companies, and impact indicators, thereby exhibiting a clear multi-source heterogeneous nature.

At the structural level, the raw data presents two major challenges. First, semantic redundancy exists across certain fields, such as “Location_Description” and “Incident_Location”, which convey similar meanings and thus require filtering and normalization during data processing. Second, some fields have relatively high missing rates; for instance, economic loss attributes such as "Total_Investigation_Cost" and "Property_Damage" have a missing rate of nearly 30%, necessitating careful handling during subsequent feature construction.

In addition to the PHMSA dataset, multi-source data from other channels were integrated, including:

(1) Pipeline basic attribute data, such as diameter, wall thickness, material, and years of service;

(2) Environmental data, including soil type, humidity, temperature, and geological hazards.

Table 1. Data sources and description

Type	Source	Key attributes
Accident Data	PHMSA	Time, Location, Cause, Consequence, Economic Loss
Pipeline Attribute Data	Enterprise Database	Diameter, Wall Thickness, Material, Years of Service
Environmental Data	Geographic Information System	Soil Type, Humidity, Temperature, Geological Hazards

To address the characteristics described above, this study implements a unified data processing workflow, encompassing the extraction of core entities and attribute fields, field reduction, and label normalization preprocessing. The sources and key attributes of the data used in this process are summarized in Table 1. Based on this, a standardized knowledge graph structure is then designed, enabling the transformation of raw accident data into inputs suitable for graph neural networks. This process establishes a solid foundation for subsequent model training and inference analysis.

3.2 Entity and Relationship Extraction

Knowledge extraction forms the foundation of knowledge graph construction and aims to extract entities, relations, and attributes from multi-source data. To fully leverage the structured knowledge embedded in oil and gas pipeline accident data and provide graph neural network models with structurally clear input graphs, this study designs and implements a systematic data pre-processing and knowledge graph construction work-flow based on domain priors and data analysis results. This workflow extracts entities and relations for risk modeling, facilitating the transformation of raw accident records into graph-structured inputs suitable for GNN modeling [22].

The extracted entity types include: pipeline entities (pipeline segments, valves, pumping stations, etc.), accident entities (leakages, ruptures, explosions, etc.), environmental entities (soil, climate, terrain, etc.), and factor entities (corrosion, external damage, material defects, etc.) [23].

The relation types include: part-of relations (e.g., "pipeline–located in–region"), causal relations (e.g., "corrosion–causes–leakage"), temporal relations (e.g., "accident–occurred at–time"), and spatial relations (e.g., "pipeline–crosses–river") [24].

3.3 Data Preprocessing and Atlas Generation Process

Due to the presence of redundant fields, missing values, and semantic inconsistencies in the raw data, an initial data cleaning and reduction process was performed, which mainly includes the following steps:

(1) Field selection and reduction: Key information relevant to graph construction (e.g., equipment type, accident time, operator, geographic location, and loss details) was retained, while fields unrelated to statistical analysis were removed.

(2) Missing value handling: For numerical fields with missing values (e.g., Property_Damage_Cost), the mean or median was used for imputation; missing text entries were set as empty and excluded during graph construction.

(3) Enumeration and discretization: Continuous fields such as pipeline service life (Pipe_Age) and economic loss (Cost) were binned to facilitate the creation of discrete entities.

(4) Text Standardization: Free-text fields, such as Cause_Description, were processed through lower-casing, stop-word removal, and keyword extraction to normalize cause-related entity descriptions.

The cleaned entities (e.g., accident events, equipment facilities, geographic locations, operators) were assigned unique entity identifiers (IDs) using a unified mapping table, following these strategies:

(1) Independent dictionaries were created for each entity type, with consistent naming prefixes to avoid conflicts.

(2) Entity IDs were encoded consecutively to ensure consistent tensor dimensions, facilitating efficient model processing.

(3) All nodes and edges in the graph were represented by numeric indices rather than original strings, reducing storage overhead.

After entity ID assignment, standard triples were constructed in the form of (head_id, relation_id, tail_id) according to the predefined relation types (e.g., operated_by, caused_by). The characteristics of the graph structure are as follows:

(1) All edges are represented as unidirectional, with the option to extend to bidirectional relations if necessary.

(2) Relation types are stored with unified numbering, enabling efficient graph convolution propagation and parameter sharing.

(3) The final triples, along with the entity mapping table, are output together for use in deep learning frameworks such as PyTorch Geometric.

As shown in Figure 1, the complete data preprocessing workflow clearly illustrates the entire pipeline from raw accident data to GNN graph data, encompassing five core steps: cleaning, standardization, entity construction, relation definition, and data output.

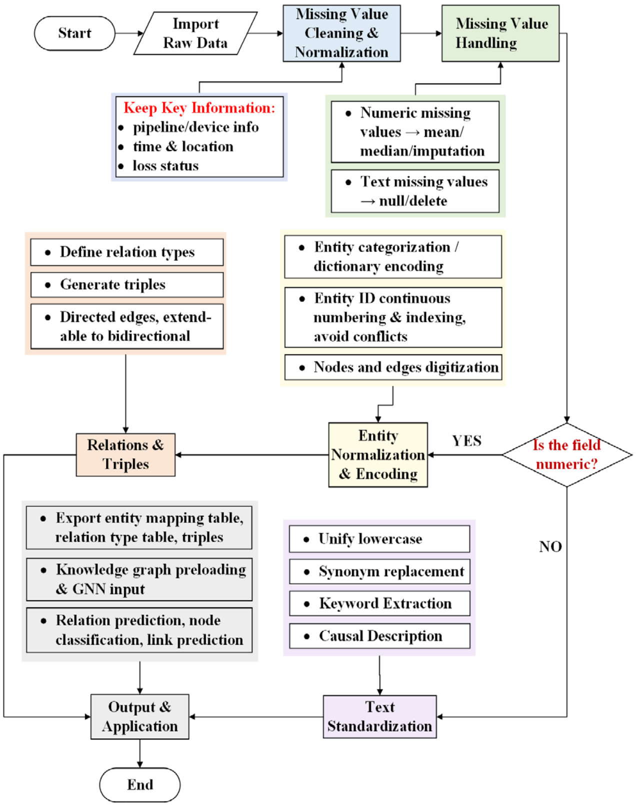

Start
Import Raw Data
Missing Value Cleaning & Normalization
Missing Value Handling
Keep Key Information:
pipeline/device info
time & location
loss status
Numeric missing values → mean/ median/imputation
Text missing values → null/delete
Define relation types
Generate triples
Directed edges, extend-able to bidirectional
Entity categorization / dictionary encoding
Entity ID continuous numbering & indexing, avoid conflicts
Nodes and edges digitization
Relations & Triples
Entity Normalization & Encoding
YES
Is the field numeric?
NO
Export entity mapping table, relation type table, triples
Knowledge graph preloading & GNN input
Relation prediction, node classification, link prediction
Unify lowercase
Synonym replacement
Keyword Extraction
Causal Description
Output & Application
Text Standardization
End

Figure 1. Flow chart of data cleaning and graph structure generation

4 Model Design and Implementation

4.1 Overall Model Architecture

To enhance the accuracy and structural interpretability of oil and gas pipeline accident risk prediction, this study proposes a relation-aware graph neural network node classification method based on accident knowledge graph construction, termed CompMRGCN. The overall design constructs a heterogeneous graph input of entities and their semantic relationships using historical multi-field accident data. Through a relation-aware convolution mechanism, structural information and node features are jointly modeled, ultimately enabling risk prediction for accident nodes.

Unlike traditional classification models that rely solely on numerical feature vectors, the proposed method first extracts key information from accident data, such as equipment type, operator, geographic location, and accident cause to construct entity nodes and define semantic relationships (edges) between them, including operated_by, located_in, and caused_by [14]. This results in a heterogeneous graph that integrates structural connections and attribute features. The graph not only captures the inter-entity associations but also preserves the attributes of each node (e.g., equipment parameters, accident loss metrics), providing rich input information for the graph neural network [15].

Figure 2 shows the overall model architecture proposed in this study, which mainly includes the following three modules.

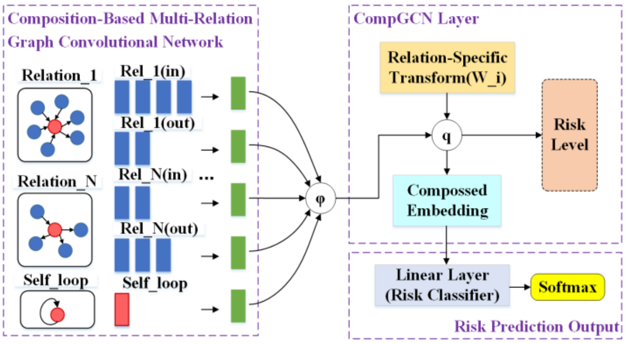


Figure 2. Overall framework of accident risk prediction

(1) Input Layer

The input layer constructs an oil and gas pipeline accident knowledge graph based on pipeline accident report data, unifying the representation of accident entities, attribute information, and multiple types of relationships. Node structural features and numerical attributes are integrated to provide foundational support for subsequent modeling.

(2) Relation-Aware Graph Convolution Layer

Building on the concept of multi-relational combination modeling, the relation-aware graph convolution layer incorporates a relation-aware mechanism, which fully leverages information from different types of relational edges. By integrating label-dependency constraints, the layer learns node representations with enhanced semantic expressiveness, enabling better capture of complex latent connections among accident nodes [16].

(3) Output Layer

The output layer feeds the node representations into a classifier to predict the probability distribution across three risk levels: low, medium, and high. Through threshold optimization, the model's recognition capability under class-imbalanced conditions is improved, achieving precise risk early warning for pipeline accidents [17].

The overall workflow enables intelligent prediction of oil and gas pipeline accident risks. By constructing an accident knowledge graph that encompasses entities, attributes, and multiple relationship types at the input stage, the model effectively integrates structural information with numerical features. In the intermediate stage, the relation-aware graph convolution layer leverages multi-relational combination modeling and the relation-aware mechanism to semantically enhance complex relationships, while incorporating label-dependency modeling to obtain more expressive node representations. At the output stage, the model classifies node risk levels, providing low, medium, and high predictions, and employs threshold optimization to improve performance under class-imbalanced conditions [18].

The modules illustrated in the figure reflect the progressive process from data to knowledge, from knowledge to representation, and from representation to prediction: the input module emphasizes data integration and graph construction, the convolution module highlights deep modeling of relations and semantics, and the output module corresponds to explicit determination of risk levels. This workflow not only ensures accuracy and robustness in risk prediction but also demonstrates the scalability and application potential of graph neural networks in complex industrial safety scenarios [25].

4.2 Graph Neural Network Modeling

This section presents the implementation details of the graph neural network model, including the network architecture, feature fusion mechanism, and loss function definition. The modeling framework is not based on traditional static graphs or rule-based networks; rather, it leverages the structured knowledge graph constructed previously, in which entities represent elements such as accidents, equipment, and locations, while relations capture various types of semantic associations. Consequently, the graph neural network is employed to perform node representation learning and risk inference directly on the knowledge graph structure.

4.2.1 Heterogeneous Graph Structure Encoding Based on Knowledge Graph

The input graph structure is derived from knowledge graph triples (head, relation, tail), encompassing multiple types of entity nodes (e.g., accidents, equipment, geographic locations, operators) and semantic relation types (e.g., operated_by, caused_by, has_location). Unlike general graph models, this graph exhibits typical knowledge graph characteristics:

Heterogeneous entities: Different node types are widely distributed, with varying attribute dimensions.

Explicit semantics: Each edge type carries a clearly defined semantic meaning.

Sparse structure with high information density: Accident nodes serve as anchors, connecting to contextual entities through multi-hop structures.

R-GCN designed for heterogeneous graphs typical of knowledge graphs, assigns independent message-passing transformations for each relation type, enabling semantically differentiated information aggregation within the graph structure [19]. The node update mechanism of R-GCN, allows the model to automatically learn higher-order contextual embeddings from graph relations such as “an accident caused by a specific reason” or “a company operates a specific pipeline segment,” thereby enhancing its ability to perceive semantic contexts of accidents.

4.2.2 Graph Neural Network Encoding Structures

After the construction of the knowledge graph structure, R-GCN is adopted as the core graph neural network architecture. Building upon the standard GCN, R-GCN introduces a relation-type–aware mechanism, enabling the model to learn a distinct transformation matrix for each type of edge relation, thereby enhancing its capacity to model multi-relational graph structures [21].

(1) Basic Notations and Definitions

The graph structure is denoted as $G = (V, E, R)$, where V denotes the set of nodes, E is the set of edges, and R denotes the set of relation types.

The node feature matrix is denoted as $X \in R^{|V| \times d}$, where d is the feature dimension.

The adjacency matrix is denoted as $A_r \in R^{|V| \times |V|}$, for each relation type $r \in R$. The entry $A_r[i, j]=1$ if and only if there exists an edge of type r from node i to node j.

(2) Relation-Specific Transformation Matrices

Traditional GCN employs a single weight matrix W, whereas R-GCN introduces relation-specific transformation matrices for each relation type:

$$h_i^{(l+1)} = \sigma\left(\Sigma_{r \in R} \Sigma_{j \in N_i^r} \frac{1}{c_{i,j}} W_r^{(l)} h_i^{(l)} + W_0^{(l)} h_i^{(l)} \right), \qquad \textbf{(1)}$$

where $h_i^{(l)}$ denotes the representation of node i at the l-th layer, N_i^r represents the set of neighboring nodes connected to node i via relation r, $c_{i,j}$ is a normalization constant, typically defined as $c_{i,j} = \left| N_i^r \right|$. $W_r^{(l)}$ is the relation-specific weight matrix corresponding to relation r, $W_0^{(l)}$ is the weight matrix for self-connections and σ denotes a non-linear activation function.

(3) Relation-Aware Graph Convolutional Network

GCNs are primarily designed for simple undirected graphs and are incapable of handling directed multi-relational graphs in knowledge graphs. To address this issue, this study adopts Composition-based Graph Convolutional Network (CompGCN) as the foundational architecture.

The model comprises two R-GCN layers, enabling the transition from semantic aggregation of local neighborhood information to higher-order modeling of global structural information [26]. CompGCN simultaneously learns representations of both nodes and relations. For an edge (u, v, r), which denotes the existence of an edge of type r directed from node u to node v, the convolution operation is defined as follows:

$$h_v^{(l+1)} = f\left(\Sigma_{(u,r) \in N(v)} W_{\lambda(r)}^{(l)} \phi\left(h_u^{(l)}, h_r^{(l)} \right) \right), \qquad \textbf{(2)}$$

where $h_v^{(l)}$ denotes the representation of node v at the l-th layer, $N(v)$ is the neighborhood set of node v, $W_{\lambda(r)}^{(l)}$ is the relation-specific weight matrix corresponding to relation r, ϕ is the composition operator, and $h_r^{(l)}$ denotes the representation of relation r.

The composition operator ϕ can take one of the following three forms:

Subtraction: $\phi\ (h_u, h_r) = h_u - h_r$

Multiplication: $\phi\ (h_u, h_r) = h_u \cdot h_r$

Circular Convolution: $\phi\ (h_u, h_r) = h_u \times h_r$

To reduce the parameter complexity introduced by a large number of relations, CompGCN employs a set of basis vectors $V_b = v_1, v_2, \ldots, v_B$ to represent all relations:

$$h_r = \sum_{b=1}^{B} \alpha_{rb} v_b \quad \textbf{(3)}$$

where α_{rb} denotes the coefficient of relation r with respect to the basis vector v_b.

$V_b = \{ v_1, v_2, \ldots, v_B \} \in \mathbb{R}^{d\times B}$ (where d is the dimension of relation representation), ensuring that the dimension of the basis vectors matches the dimension of the relation representation.

Logic of Coefficient Learning, the coefficients $\alpha_{rb} \in \mathbb{R}$ are usually learned through the following two methods:

- Fixed basis vectors: The basis vectors are predefined (e.g., fixed after random initialization), and only α_{rb} is learned (obtained by mapping relation IDs through a fully connected layer).
- Learnable basis vectors: Both the basis vectors v_b and the coefficients α_{rb} are updated through backpropagation. In this case, α_{rb} is often constrained to be non-negative (e.g., via softmax activation) to ensure the physical meaning that "the relation representation is a weighted combination of basis vectors".

In this model, the structural embeddings learned by the R-GCN are concatenated or integrated with the numerical attribute features of the incident nodes and subsequently fed into a Multi-Layer Perceptron (MLP) architecture to perform the final risk level prediction.

4.2.3 Numerical Feature Fusion Strategy and Loss Function

Although knowledge graphs provide abundant structural and semantic information, nodes themselves also contain important numerical attributes, such as the time of accident occurrence, economic losses, and equipment parameters. These attributes are normalized to form the attribute vector x_i^{attr} of node *i*.

In this study, we adopt a concatenation-based fusion strategy between structural embeddings and attribute vectors to construct the final node representation:

$$z_i = Concat\left(h_i^{(2)}, x_i^{attr}\right), \quad \textbf{(4)}$$

where $h_i^{(2)}$ denotes the structural embedding of node i obtained through two R-GCN layers, and x_i^{attr} represents the original attribute vector of the accident node. Attention-based Fusion: $z_i = w_h \cdot h_i^{(2)} + w_x \cdot x_i^{attr}$), where w_h and w_x are attention weights (obtained through learning), which are suitable for scenarios where the importance of different features varies with samples (e.g., numerical features are more important in high-risk incidents, while structural features are more important in low-risk incidents). The fused representation simultaneously incorporates both "structural semantic context" and "static risk features."

The loss function of the CompMRGCN model consists of two components:

$$L = L_{task} + \beta L_{KL}, \quad \textbf{(5)}$$

where L_{task} is the task-related loss (e.g., cross-entropy loss), L_{KL} denotes the KL divergence between the posterior distribution and the prior distribution, and β is a balancing hyperparameter.

The explanations of prior distribution and posterior distribution are as follows:

Posterior Distribution: $q(h|G,X)$, which represents the distribution of node embeddings h given the graph G and node features X (usually modeled as a multivariate normal distribution $\varkappa(\mu,\sigma^2 I)$, where μ and σ are output by the GCN layer).

Prior Distribution: $p(h)$, which is usually set as a standard normal distribution $\varkappa(0,I)$ to constrain the smoothness of the embedding distribution.

Calculation of KL Divergence:

$$L_{KL} = \frac{1}{2}\sum_{i=1}^{|V|}\left(u_i^2 + \sigma_i^2 - \log\sigma_i^2 - 1\right), \quad \textbf{(6)}$$

Its function is to make the posterior distribution approximate the prior distribution and prevent the embeddings from overfitting to the local structure of the training data.

The model parameters are updated using stochastic gradient descent with backpropagation to compute gradient. To mitigate overfitting, regularization techniques such as dropout and weight decay are applied.

4.3 Risk Prediction and Output Module

The output of the model is the predicted risk level for each accident node. The labels are divided into three categories (low, medium, and high risk), and are generated by integrating historical casualties, losses, and other fields. During the prediction phase, the model outputs the probability distribution over the three risk levels for each accident node, and the class with the highest probability is selected as the predicted category.

Considering the scarcity of high-risk accident samples and the highly imbalanced class distribution, the training phase adopts a weighted binary cross-entropy loss (BCEWithLogitsLoss) function, assigning different loss weights to different risk levels in order to enhance the recognition capability for the high-risk category [27].

In the output stage, a single-layer MLP is employed to map the fused representation into the risk-level space, and the model is trained using BCEWithLogitsLoss. To address the severe class imbalance problem (with high-risk samples accounting for only a small proportion), a weighting mechanism is incorporated into the loss function to enhance the model's ability to identify high-risk accidents [28].

For the node classification task, a softmax classifier is added to the last layer of R-GCN:

$$\hat{y}_{i,c} = soft\max\left(W_{class} h_i^{(L)} + b_{class}\right). \quad \textbf{(7)}$$

Supplementary Explanations on Dimension and Logic:

- $h_i^{(L)}$: Represents the final structural embedding of node i after passing through L layers of R-GCN. Here, L = 2, so $h_i^{(L)} = h_i^{(2)} \in \mathbb{R}^d$.
- Parameter dimensions: $W_{class} \in \mathbb{R}^{C \times d}$ (where C = 3, corresponding to three risk levels: low, medium, and high), and $b_{class} \in \mathbb{R}^C$.
- Softmax calculation: $\hat{y}_{i,c} = \frac{\exp\left(W_{class,c} \cdot h_i^{(L)} + b_{class,c}\right)}{\Sigma_{c=1}^{3} \exp\left(W_{class,c} \cdot h_i^{(L)} + b_{class,c}\right)}$

  , where $\hat{y}_{i,c}$ represents the probability that node i belongs to the c-th risk level, which is consistent with the logic of selecting the category with the highest probability as the prediction result.

The training and inference of the entire model are implemented using PyTorch Geometric, together with the graph-structured data generated in the preceding sections, achieving a complete modeling process from structure construction and embedding propagation to label prediction [29]. The entire model can be regarded as a structure-aware predictor for risk learning on knowledge graph structures, which not only models explicit causal paths among entities but also incorporates numerical factors that affect the severity of accidents [30]. The entire algorithm flow is shown in Algorithm 1.

Algorithm 1: The Main Procedure of Method Implementation.

```
Data: Node Representation Learning on Relational Graph
Input: Graph $G = (V, E, R)$, node features $X$, number of
       layers $L$
Output: Node representations $H^{(L)}$
Set $H^{(0)} = X$;
for $l = 0$ to $L - 1$ do
    for each relationship $r \in R$ do
        Calculate the normalized normalized adjadjacency
        matrix:$\widehat{A}_r = \widehat{D}_r^{-1}\widehat{A}_r$
    end
    for each relationship $i \in V$ do
        Aggregate neighbor information:
                $h_{N(i)}^{(l)} = \Sigma_{r\in R} \Sigma_{j\in N_i^j} \frac{1}{c_{i,j}} W_r^{(l)} h_j^{(l)}$
            Add self-connections:
                $h_i^{(l+1)} = \sigma(h_{N(i)}^{(l)} + W_0^{(l)} h_i^{(l)})$
                if $l > 0$ then
                    $h_i^{(l+1)} = h_i^{(l+1)} + h_i^{(l)}$
                end
    end
end
return $H^{(L)}$
```

5 Experiments and Analysis of Experimental Results

5.1 Dataset Partitioning and Evaluation Metrics

This study employs the oil and gas pipeline risk knowledge graph constructed in Section 3 as the experimental dataset, which is randomly divided into training, validation, and test sets in a ratio of 7:2:1, as shown in Table 2.

Three evaluation metrics are employed to assess the performance of the model: the Average Precision (AP), F1-score, Accuracy.

Table 2. Dataset statistics

Statistic	Quantity
Total number of nodes	32,000
Total number of edges	48,000
Node types	12
Edge types	25
Training set nodes	22,400
Validation set nodes	6,400
Test set nodes	3,200

5.2 Experimental Results and Analysis

In this study, node classification prediction is conducted on the constructed oil and gas pipeline accident risk knowledge graph using the CompMRGCN model based on the relation-aware graph convolutional network. To verify the effectiveness of the proposed method, multiple comparison models are selected for performance evaluation, including the standard GCN, GAT, R-GCN, CompGCN, and Graph Markov Neural Network (GMNN).

All experiments are performed under a unified data processing pipeline. Negative sampling is employed to control the ratio of positive to negative samples at approximately 1:1.5, and the dataset is randomly partitioned into training, validation, and test sets in a ratio of 7:2:1. During training, weighted BCEWithLogitsLoss is used, while the AP on the validation set serves as the criterion for early stopping. In the testing phase, the optimal classification boundary is determined by traversing different decision thresholds, in order to achieve the best F1-score and Accuracy.

The results of the comparison experiments are shown in Table 3.

Table 3. Comparison experiment results

Model type	mAP (%)	F1-score (%)	Optimal threshold	Acc (%)
CompMRGCN	97.1	95.8	0.39	96.2
GCN	80.3	81.6	0.45	82.4
GAT	83.6	84.2	0.48	85.7
R-GCN	87.3	86.5	0.41	85.9
CompGCN	89.2	88.4	0.43	87.8
GMNN	91.7	90.2	0.44	89.5

On the test set, the CompMRGCN model achieved excellent performance with an mAP of 97.1%, an F1-score of 95.8%, and an Accuracy of 96.2%, significantly outperforming the other comparison models. Among them, GCN obtained relatively good results with an mAP of 80.3% and the F1-score of 81.6%, highlighting the importance of graph structural information in risk prediction. The GAT model, however, performed poorly in this task, with an mAP of 83.6%, the results are shown in Figure 3, possibly due to the overall sparsity of the oil and gas pipeline accident graph and the weak local neighborhood information of nodes, which limit the effectiveness of the attention mechanism.

R-GCN exhibited relatively stable performance in this task, achieving an mAP of 87.3% and an F1-score of 86.5%, indicating that the incorporation of relational modeling has a positive effect on enhancing the expressive capacity of the model. However, since it adopts only simple linear relational transformations, it fails to sufficiently capture the compositional characteristics among multiple relations, and therefore still shows certain limitations in exploiting complex semantic relationships.

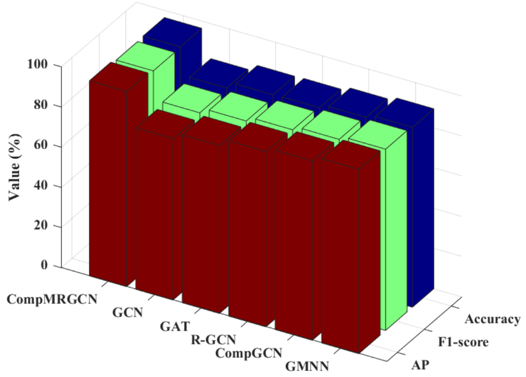


Figure 3. Comparison of experimental results

The CompGCN model performed better than R-GCN in terms of results, achieving an mAP of 89.2% and an F1-score of 88.4%, showing a clear improvement compared with R-GCN. This demonstrates that relation composition operations can effectively enhance feature interaction capability and better exploit neighboring relational information. However, since this method does not fully integrate numerical attributes in accident scenarios, its overall performance remains slightly lower than that of the CompMRGCN proposed in this article.

GMNN also achieved relatively good results in the experiments, with an mAP of 91.7% and an F1-score of 90.2%. As a method that incorporates graph structural priors and label propagation mechanisms, GMNN can alleviate, to some extent, the negative impact caused by label scarcity. However, compared with the method proposed in this research, its capability of modeling multi-relational knowledge is insufficient, and therefore its overall performance is still inferior to that of CompMRGCN.

It is worth noting that, in order to improve the discrimination between positive and negative samples, decision threshold optimization was applied during the testing phase. The optimal classification threshold for the CompMRGCN model was 0.39, which is lower than the conventional default threshold of 0.5. In the case of class imbalance, the output probabilities of the model often exhibit certain biases. By appropriately adjusting the threshold, a better balance between recall and accuracy can be achieved, thereby significantly improving the overall F1-score. This treatment does not alter the label definitions or optimization objectives during training, ensuring the fairness and credibility of the evaluation metrics.

In summary, the CompMRGCN model proposed in this article significantly outperforms the comparison methods across all three metrics. Compared with the best comparison method, GMNN, accuracy is improved by 6.7%, the F1-score by 5.6%, and mAP by 5.4%. This demonstrates that, by integrating relation-aware graph convolutions, the model can better exploit graph structural information and label dependencies. The experimental results fully verify that the proposed CompMRGCN method, by incorporating knowledge graph structures and numerical attributes in the domain of oil and gas pipeline accidents, can effectively enhance node-level risk prediction performance, showing strong potential for application and generalization.

5.3 Ablation Study

To further verify the importance of the design of each module in the model, ablation experiments were conducted on graph structure, relational information, node attributes, and the negative sampling mechanism. The results are shown in Table 4.

From the overall results, the complete CompMRGCN model achieved the highest F1-score (95.8%) and AUC (97.14%) on the test set, demonstrating the effectiveness of each module design. The detailed analysis is as follows:

First, after removing relational type information (replacing R-GCN with GCN), the model performance declined, with the F1-score dropping from 0.9583 to 0.8970 and the AUC from 0.9623 to 0.8794, indicating that properly modeling semantic relations among entities plays an important role in high-risk prediction.

The results of the ablation experiments indicate that each component of the model plays a critical role in overall performance. Among them, the contribution of the relation-aware mechanism is the most significant; after its removal, the model's accuracy dropped by 8.3%, the F1-score by 6.1%, and the mAP by 6.9%, demonstrating that introducing relation awareness is indispensable when dealing with multi-relational knowledge graphs.

When the graph structure was completely removed and only node attributes were retained MLP, the model performance declined drastically, with the F1-score reduced to 0.5797, the AUC to 0.5848, and the mAP to 58.4%, which is nearly equivalent to random guessing. This result highlights the central role of graph structural information in oil and gas pipeline risk modeling.

When node numerical features were removed (w/o Feature) and only structural information was preserved, the F1-score decreased to 0.9082 and the mAP decreased by 5.5%. Although the impact was smaller than that of removing graph structure, this still demonstrates that node attribute features provide important complementary information to the model.

In addition, when negative sampling was eliminated (w/o Negative Sampling) and the model was trained directly on the complete dataset, its ability to handle class imbalance declined significantly, with the F1-score reduced to only 0.7523 and the mAP decreased by 4.7%. Although the AUC remained relatively high (0.9277), the actual classification performance deteriorated considerably, indicating that the negative sampling strategy is particularly important for training stability and for identifying high-risk categories.

In summary, relation awareness, graph structure modeling, node numerical features, and the negative sampling strategy are all indispensable components of the overall framework. Together, they contribute to accurate prediction of high-risk accidents in oil and gas pipelines and demonstrate significant advantages across multiple evaluation metrics.

Table 4. Ablation experiment results

Model variant	Structural features	Numerical features	Relation modeling	Negative sampling	AUC (%)	Acc (%)	F1 (%)	mAP (%)
CompMRGCN	Yes	Yes	Yes	Yes	0.9714	0.9623	95.8	97.1
w/o Relation-aware	Yes	Yes	No	Yes	0.8950	0.8794	89.7 (–6.1)	90.2 (–6.9)
w/o Graph	No	Yes	No	Yes	0.4081	0.5848	57.9 (–37.9)	58.4 (–38.7)
w/o Feature	No	No	Yes	Yes	0.9570	0.9530	90.8 (–5.0)	91.6 (–5.5)
w/o Negative sampling	Yes	Yes	Yes	No	0.9370	0.9277	75.2 (–20.6)	92.4 (–4.7)

6 Conclusion

This study investigates the problem of intelligent risk identification for oil and gas pipeline accidents and proposes a graph neural network modeling approach that combines structural information with semantic relations. Based on the pipeline accident report data released by the U.S. PHMSA, a heterogeneous knowledge graph for risk prediction was constructed.

In addition, this study proposes an oil and gas pipeline risk early warning model based on knowledge graphs and graph neural networks, with the main contributions summarized as follows: First, a comprehensive oil and gas pipeline risk knowledge graph was constructed using the publicly released PHMSA pipeline accident report data. An experimental dataset with a total of 32,000 nodes was established. Second, a relation-aware graph neural network node classification method (CompMRGCN) was designed. By introducing a relational convolution mechanism and a node attribute fusion strategy, this method simultaneously leverages node features and label dependencies, thereby effectively improving the accuracy of risk prediction and enhancing structural interpretability. Finally, extensive experiments on real datasets demonstrate that the CompMRGCN model achieves an accuracy of 96.2%, an F1-score of 95.8%, and a mean average precision (mAP) of 97.1%, significantly outperforming the comparison methods. The ablation study further validates the effectiveness of each component of the model.

Although the proposed method achieved good performance in experimental validation, certain limitations remain. For example, the data dimension is limited: the current experiments construct the knowledge graph solely from structured accident data, without fully utilizing multi-source information such as raw text descriptions and sensor sequence data, leading to insufficient exploitation of available information. Moreover, model interpretability is still limited. Although structural information and relational modeling have been incorporated, a systematic causal reasoning pathway for accident mechanisms has not yet been established.

In summary, this research provides a new method and perspective for knowledge graph–based risk identification of oil and gas pipeline accidents. The study lays a solid foundation and demonstrates broad application prospects, while still leaving room for improvement and valuable directions for future research.

Acknowledgement

This work is supported by National Key Research and Development Program Project (Grant No. 2023YFB3107700); Provincial Science and Technology Plan Project of Hebei Province (Grant No. 253A7634D).

References

[1] H. Jing, L. Huang, H. Liu, W. Jiang, Q. Deng, and R. Niu, A proposal for rapid assessment of long-distance oil and gas pipelines after earthquakes, *Applied Sciences*, Vol. 15, No. 7, Article No. 3595, April, 2025. https://doi.org/10.3390/app15073595

[2] M. Zhan, Y.-L. Li, Risk evaluation of submarine pipelines using improved FMEA model based on social network analysis and extended GLDS method under a linguistic Z-number preference relation environment, *Journal of Loss Prevention in the Process Industries*, Vol. 96, Article No. 105611, August, 2025. https://doi.org/10.1016/j.jlp.2025.105611

[3] P. Chen, Advancements and future outlook of safety monitoring, inspection and assessment technologies for oil and gas pipeline networks, *Journal of Pipeline Science and Engineering*, Article No. 100267, March, 2025. https://doi.org/10.1016/j.jpse.2025.100267

[4] M. J. Hasan, M. Arifeen, M. Sohaib, A. Rohan, S. Kannan, Enhancing gas pipeline monitoring with graph neural networks: A new approach for acoustic emission analysis

under variable pressure conditions, *Proceedings of the 20th International Conference on Condition Monitoring and Asset Management*, Oxford, U.K., June, 2024, pp. 1–10. https://doi.org/10.1784/cm2024.4b3

[5] J. Hao, Z. Zhang, Y. Ping, Power system fault diagnosis and prediction system based on graph neural network, *International Journal of Information Technologies and Systems Approach*, Vol. 17, No. 1, pp. 1–14, 2024. https://doi.org/10.4018/IJITSA.336475

[6] R. Anaadumba, Y. Bozkurt, C. Sullivan, M. Pagare, P. Kurup, B. Liu, M. A. U. Alam, Graph neural network-based water contamination detection from community housing information, *Frontiers in Environmental Engineering*, Vol. 4, Article No. 1488965, March, 2025. https://doi.org/10.3389/fenve.2025.1488965

[7] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32, No. 1, pp. 4–24, January, 2021. https://doi.org/10.1109/TNNLS.2020.2978386

[8] H. Lu, L. Wang, X. Ma, J. Cheng, M. Zhou, A survey of graph neural networks and their industrial applications, *Neurocomputing*, Vol. 614, Article No. 128761, January, 2025. https://doi.org/10.1016/j.neucom.2024.128761

[9] Y. Zhang, P. M. Karve, S. Mahadevan, Graph neural networks for power grid operational risk assessment under evolving unit commitment, *Applied Energy*, Vol. 380, Article No. 124793, February, 2025. https://doi.org/10.1016/j.apenergy.2024.124793

[10] H. Truong, A. Tello, A. Lazovik, V. Degeler, Graph neural networks for pressure estimation in water distribution systems, *Water Resources Research*, Vol. 60, No. 7, Article No. e2023WR036741, July, 2024. https://doi.org/10.1029/2023WR036741

[11] L. Zhang, Optimization of oil and gas pipeline leakage data and defect identification based on graph neural processing, *Annals of Data Science*, Vol. 12, No. 4, pp. 1413–1430, August, 2025. https://doi.org/10.1007/s40745-025-00619-7

[12] R. Xie, Z. Fan, X. Hao, W. Luo, Y. Li, Y. Zhao, J. Han, Prediction model of corrosion rate for oil and gas pipelines based on knowledge graph and neural network, *Processes*, Vol. 12, No. 11, Article No. 2367, November, 2024. https://doi.org/10.3390/pr12112367

[13] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018, pp. 1–12. https://openreview.net/forum?id=rJXMpikCZ

[14] Z. Zhao, Z. Xiao, J. Tao, MSDG: Multi-scale dynamic graph neural network for industrial time series anomaly detection, *Sensors*, Vol. 24, No. 22, Article No. 7218, November, 2024. https://doi.org/10.3390/s24227218

[15] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, RotatE: Knowledge graph embedding by relational rotation in complex space, *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019, pp. 1–18.

[16] J. S. Garrido, D. Dold, J. Frank, Machine learning on knowledge graphs for context-aware security monitoring, *Proceedings of IEEE International Conference on Cyber Security and Resilience (CSR)*, Rhodes, Greece, 2021, pp. 1–8. https://doi.org/10.1109/CSR51186.2021.9527927

[17] S. Wu, F. Sun, W. Zhang, X. Xie, B. Cui, Graph neural networks in recommender systems: A survey, *ACM Computing Surveys*, Vol. 55, No. 5, pp. 1–37, May, 2023. https://doi.org/10.1145/3535101

[18] H. Yuan, J. Tang, X. Hu, S. Ji, XGNN: Towards model-level explanations of graph neural networks, in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Virtual Event, 2020, pp. 430–438. https://doi.org/10.1145/3394486.3403085

[19] Z. Chen, K. Huang, L. Wu, Z. Zhong, Z. Jiao, Relational graph convolutional network for text-mining-based accident causal classification, *Applied Sciences*, Vol. 12, No. 5, Article No. 2482, March, 2022. https://doi.org/10.3390/app12052482

[20] Q. Sha, T. Tang, X. Du, J. Liu, Y. Wang, Y. Sheng, Detecting credit card fraud via heterogeneous graph neural networks with graph attention, *Proceedings of the IEEE 6th International Conference on Artificial Intelligence and Information Technology (AINIT)*, Chengdu, China, 2025, pp. 1–6. https://doi.org/10.1109/AINIT65432.2025.11035158

[21] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: A. Gangemi, R. Navigli, M. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, M. Alam (Eds.), The Semantic Web. *Proceedings of the European Semantic Web Conference (ESWC)*, Vol. 10843, *Lecture Notes in Computer Science*, Cham, Switzerland: Springer, 2018, pp. 593–607. https://doi.org/10.1007/978-3-319-93417-4_38

[22] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, Z. Wang, A survey of information extraction based on deep learning, *Applied Sciences*, Vol. 12, No. 19, Article No. 9691, October, 2022. https://doi.org/10.3390/app12199691

[23] J. Du, G. Liu, J. Gao, X. Liao, J. Hu, L. Wu, Graph neural network-based entity extraction and relationship reasoning in complex knowledge graphs, *Proceedings of the IEEE International Conference on Intelligent Computing and Machine Learning (ICICML)*, Shenzhen, China, 2024, pp. 1–6. https://doi.org/10.1109/ICICML63543.2024.10958048

[24] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, A survey on knowledge graphs: Representation, acquisition, and applications, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 33, No. 2, pp. 494–514, February, 2022. https://doi.org/10.1109/TNNLS.2021.3070843

[25] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32, No. 1, pp. 4–24, January, 2021. https://doi.org/10.1109/TNNLS.2020.2978386

[26] S. Vashishth, S. Sanyal, V. Nitin, P. Talukdar, Composition-based multi-relational graph convolutional networks, *arXiv preprint*, arXiv: 1911.03082, November, 2019. https://arxiv.org/abs/1911.03082

[27] A. Alchihabi, H. Yan, Y. Guo, Overcoming class imbalance: Unified GNN learning with structural and semantic connectivity representations, *arXiv preprint*, arXiv: 2412.20656, December, 2024. https://arxiv.org/abs/2412.20656

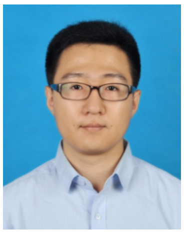
[28] A. Nippant, D. Li, H. Ju, H. N. Koutsopoulos, H. R. Zhang,

Graph neural networks for road safety modeling: Datasets and evaluations for accident analysis, *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS'23)*, New Orleans, LA, USA, 2023, pp. 52009–52032.

[29] M. Fey, J. E. Lenssen, Fast graph representation learning with PyTorch Geometric, *arXiv preprint*, arXiv: 1903.02428, March, 2019. https://arxiv.org/abs/1903.02428

[30] PyTorch Geometric Team, Heterogeneous graph learning, *PyTorch Geometric Documentation*, [Online]. Available: https://pytorch-geometric.readthedocs.io/en/latest/notes/heterogeneous.html [Accessed: Jan. 10, 2025]

Biographies



Yixin Wei works at PipeChina North Pipeline Company, mainly engaged in the production and operation of oil and gas pipelines, as well as industrial control system network security. He has experience in SCADA system regulation, pipeline process optimization, and integrity management, and is familiar with safety operation and maintenance technologies such as pressure monitoring and leak warning.

Feng Yan received a Master's degree in Automation from China University of Petroleum (Beijing), Beijing, China. He currently serves as the Director (Division Chief level) of the Production Department at China Oil & Gas Pipeline Network Corporation, Beijing, China. His current research interests include oil and gas pipeline operation and scheduling, pipeline integrity management, and industrial control network security.

Chunming Wang works at PipeChina North Pipeline Company and has long been engaged in the field of safe production and operation of oil and gas pipelines, as well as industrial control system network security. He familiar with pipeline process regulation, risk and hazard investigation, and emergency response processes; Simultaneously focusing on network security protection of industrial control systems, proficient in vulnerability assessment and research on attack and defense technologies.

Yunteng Wen works at PipeChina North Pipeline Company. He focuses on network security work at the sub-control center, with core responsibilities to ensure the stable and secure operation of the center's network system—his key work includes conducting real-time monitoring of network infrastructure (tracking traffic, logs and abnormal activities via advanced tools to timely detect threats like unauthorized access and malware intrusions).

Tiantian Liu is currently a Master's student in Information and Communication Engineering at Beijing Information Science and Technology University, China. Her research interests include knowledge graph and artificial intelligence algorithms. She has obtained one authorized patent and two software copyrights.