processes

MDPI

Article

Prediction Model of Corrosion Rate for Oil and Gas Pipelines Based on Knowledge Graph and Neural Network

Ruohan Xie [1], Zheng Fan [1,*], Xinyu Hao [2], Weibin Luo [3], Yaoxiang Li [1], Yuntian Zhao [1] and Jie Han [4]

[1] College of Chemistry & Chemical Engineering, Xi'an Shiyou University, Xi'an 710065, China; 22212071134@stumail.xsyu.edu.cn (R.X.); 21212070950@stumail.xsyu.edu.cn (Y.L.); 18991329068@163.com (Y.Z.)
[2] Shaanxi Yanchang Petroleum (Group) Company Limited Refining and Chemical Company, Yan'an 727406, China; 13279521930@163.com
[3] SINOPEC Qilu Petrochemical Company, Zibo 255400, China; 22211070142@stumail.xsyu.edu.cn
[4] Shaanxi Mingze Yisheng Energy Technology Company Limited, Xianyang 712000, China; hanjiesxmz@163.com
* Correspondence: fanzheng@xsyu.edu.cn

Abstract: Aiming to solve the problems of high cost and large error, a prediction model of oil and gas pipeline corrosion rate was established by using a knowledge graph and neural network. The knowledge map of oil and gas pipeline corrosion was constructed by the seven-step method and the fusion of Bidirectional Encoder Representations from Transformers, Bidirectional Long Short-Term Memory, and Conditional Random Fields models (BERT-Bi LSTM-CRF) combined extraction, and the triplet information was stored by the Neo4j graph database. With eight factors, such as carbon dioxide content, as input signals, three kinds of neural networks were used to screen the basic prediction model, and the genetic algorithm (GA) and particle swarm optimization algorithm (PSO) were used to optimize and compare. The results show that the extraction accuracy of BERT-Bi LSTM-CRF is 94.46%, the recall rate is 90.39%, and the F1 value is 92.38%. The optimized PSO-BP model has the best effect, and the coefficient of determination in the training stage and the test stage is as high as 0.9954 and 0.9935, and the total mean square error of the sample is reduced by 92.8% compared with that before optimization. The corrosion rate model proposed in this paper ensures the safe operation of oil and gas pipelines and offers a theoretical foundation for corrosion prediction.

Keywords: oil and gas pipeline; uniform erosion rate; knowledge graph; neural network; optimization

check for updates

Citation: Xie, R.; Fan, Z.; Hao, X.; Luo, W.; Li, Y.; Zhao, Y.; Han, J. Prediction Model of Corrosion Rate for Oil and Gas Pipelines Based on Knowledge Graph and Neural Network. *Processes* **2024**, *12*, 2367. https://doi.org/10.3390/pr12112367

Academic Editor: Massimo Caruso

Received: 27 September 2024
Revised: 20 October 2024
Accepted: 26 October 2024
Published: 28 October 2024

Copyright: © 2024 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Oil and gas form the main fossil energy pillar, accounting for 27% of primary energy consumption [1]. At present, the transport modes using oil and gas mainly include air transport, railway transport, road transport, and pipeline transport, etc. Compared with other modes of transport, pipeline transport has a lot of advantages, such as high efficiency, safety, economic, and environmental protection. With the increasing scale and quantity of oil and gas pipeline networks, higher requirements have been put forward for their safe and reliable operation. According to statistics, in the past ten years, China's corrosion-induced oil and gas pipeline leakage accidents accounted for more than 40% of accidents. This has not only brought great harm to people's lives and property, but has also caused great pollution of the ecological environment. Although the rapid development of corrosion science and protection technology provides a powerful means to control corrosion, not all corrosion can be completely eliminated, so discovering how to quickly and accurately establish the corrosion rate prediction model for oil and gas pipelines is of great significance to guarantee the safe use of oil and gas pipelines during their service life [2,3].

Due to the increase in pipeline projects, corrosion data has also exploded, and artificial neural network models are commonly used in corrosion rate prediction methods. Lei Xu et al. [4] proposed a novel hybrid machine learning model, along with an

improved particle swarm optimization algorithm to optimize the parameters of support vector regression, to investigate corrosion rate prediction of natural gas pipelines. Pawan Bohane et al. [5] used a variety of machine learning models including linear regression, decision tree, k-nearest neighbor, random forest, and support vector regression to predict uniform and pitting corrosion in geothermal energy systems; Ali Hussein Khalaf et al. [6] proposed an adaptive network fuzzy inference system (ANFIS) model combined with simulated annealing (SA) optimization for improving the prediction accuracy of the corrosion rate of pipeline steel; and Bingyan Cui et al. [7] studied and developed an integrated Bayesian neural network (BNN) model which showed superior accuracy in predicting corrosion depth. In addition to this there are many artificial intelligence studies on corrosion prediction [8–10].

Meanwhile the knowledge graph has done well in this regard, and its applications with neural networks have come into the public eye. For example, Jintao Liu et al. [11] proposed a knowledge graph-based hazard prediction method for railroad operation accidents, which aims to capture the hazards that may lead to accidents in advance and block the potential path of accident triggers. Yuntao Shi et al. [12] proposed a risk prediction method for gas systems based on the temporal knowledge graph, which is capable of predicting the level of risk of the gas system in a certain period of time in the future. Pengbo Duan et al. [13] constructed a large-scale knowledge graph covering the clinical therapeutic knowledge of TCM attributes and herbs, and designed a deep knowledge embedding component to learn the deep knowledge embedding of herbs and targets, along with supervised training and target prediction. Chakraborty Nilesh et al. [14] introduced a neural network-based knowledge graph Q&A system and outlined new trends in the field, simplifying the process of making informed decisions when creating a Q&A system. Wei Mingzhu et al. [15] used complex semantic information from the knowledge graph of technology patents, relying on deep neural network algorithms to train a prediction model of industrial-emerging technologies and achieve accurate industrial-emerging technology prediction. Yu Yan et al. [16] proposed a set of neural symbolic integrated reasoning frameworks based on the knowledge graph, which can organically integrate symbolic representations and neural networks for more effective knowledge reasoning and decision-making. Fan Baochao [17] used the knowledge graph technology, neural network technology, and implicit feedback collaborative filtering recommendation ideas to construct a tongue diagnosis of breast cancer using a traditional Chinese medicine recommendation model, and achieving the simulation of grouping and dispensing of traditional Chinese medicine recommendations.

Existing research is related to each field, but there is less research for the joint application of knowledge graphs and neural networks in the corrosion field, which need more expansion. In this paper, we first constructed a knowledge graph through a seven-step method, the BERT-Bi LSTM-CRF joint extraction method, and graph database Neo4j to visualize the corrosion data of oil and gas pipelines, and we reorganized and preserved the structured data required to establish neural networks, which effectively solved the result error caused by the lack of corrosion data. The neural network was then used to establish the oil and gas pipeline corrosion rate prediction model and optimize it. This is very meaningful for the artificial intelligence prediction research in the field of corrosion protection, providing theoretical support for corrosion prediction, protection, and national safe production.

2. Materials and Methods

2.1. Construction of a Knowledge Graph for Oil and Gas Pipeline Corrosion

2.1.1. Oil and Gas Pipeline Ontology Construction

The knowledge map of the oil and gas pipeline corrosion field constructed in this paper belongs to the industry knowledge map, which refers to the Chinese National Corrosion Evaluation Criteria [18–20], which is mainly stored in the form of (entity-relationship-entity) and (entity-attribute-value) triples, for example, 'a thick oil area-contains-pipeline

T732CH-20-4 station', 'pipeline Station T732CH-20-4—belongs to—20# steel", "Pipeline Station T732CH-20-4—CO_2 content of—2.9%", "Pipeline Station T732CH-20-4—water content of—63.2%", etc. Its development has a great potential and, moreover, its development is highly targeted, the entities and attributes are the professional terms of this industry, the scale of knowledge and the attributes possessed by the entities are more complete, and the knowledge map in the field of oil and gas pipeline corrosion is formed through the interconnection between the entities. Manual extraction is used to obtain all the data of the ontology and draw on the relatively mature seven-step method of ontology construction [21]. The constructed oil and gas pipeline uniform corrosion domain comprehensively considers realistic application scenarios, factors affecting the corrosion of oil and gas pipelines, the quality and scale of the original data, and so on. By asking expert opinions, finding relevant data, learning and analyzing the characteristics of the oil and gas pipeline corrosion domain, the composition of the ontology is determined, and the attributes and relationships of hierarchical relationships and conceptual classes are constructed. Oil and gas pipelines are classified into oil field pipelines and gas field pipelines, and more subdivided according to the characteristics of the oil and gas area in which they are located, in order to accurately carry out knowledge extraction. Meanwhile, relationships and attributes are defined for each ontology class to describe the relationships and inherent characteristics between concepts, limit the use of attributes, avoid attribute errors, and ensure data consistency.

2.1.2. Oil and Gas Pipeline Corrosion Knowledge Extraction

For unstructured and semi-structured raw corpuses, such as the monitoring data of oil and gas pipeline corrosion and pipeline perforation measured data, the Maximum Entropy Classification Model (MEM), Hidden Markov Model (HMM), Conditional Random Field Model (CRF), Hybrid Model, and BILSTM model are commonly used for entity information extraction [22]. The LSTM-CRF model combines the long and short-term memory capability of LSTM and the labeled sequence optimization capability of CRF, which is suitable for dealing with sequence annotation tasks; LSTM is able to learn long-term dependencies, while the CRF layer considers the dependencies between labels to optimize the prediction results; the Bi LSTM model is able to consider the forward and backward information of the sequence at the same time, which makes it more effective in dealing with the text's contextual information. Compared to unidirectional LSTM, Bi LSTM is able to capture the information in the sequence more comprehensively, which improves the model's understanding and prediction of the sequence data; the BERT-CRF model uses the pre-training ability of BERT to extract the text features, and then optimizes the labeled sequences through the CRF layer. In this paper, the advantages of each model are fully combined, and the BERT-Bi LSTM-CRF model is used as the entity extraction model for oil and gas pipeline corrosion data information to achieve the entity and attribute extraction model for oil and gas pipeline corrosion. Firstly, the text is pre-trained using the BERT model to generate dynamic word vectors. Secondly, these word vectors are input into the bidirectional LSTM (Bi LSTM) model for training to further extract text features, and then the outputs of the Bi LSTM module are passed into the CRF module for decoding to obtain the predicted labelled sequences. Finally, the whole Chinese entity recognition process is completed by extracting and classifying each entity in the sequence, as shown in Figure 1.

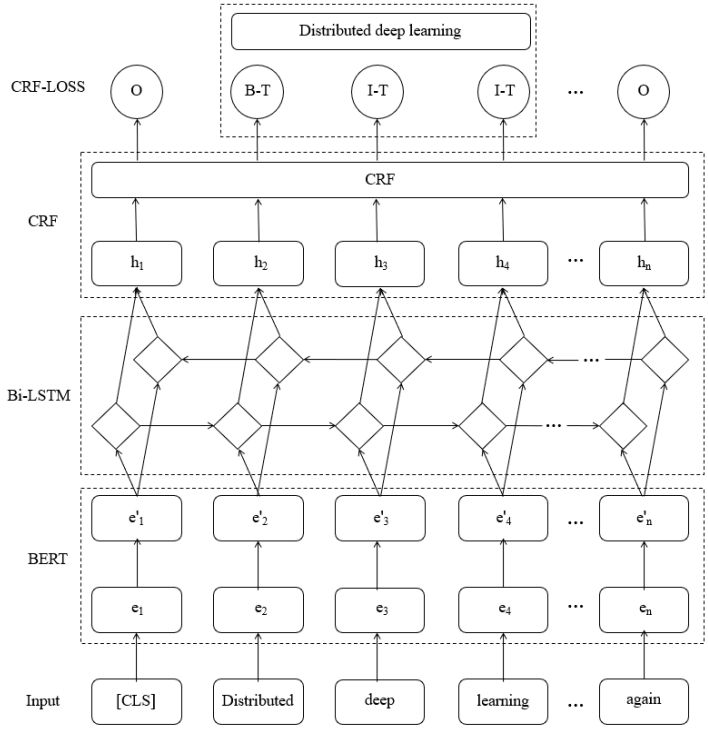


Figure 1. BERT-Bi LSTM-CRF model framework.

It is worth mentioning that, when using the BERT model, the original text is firstly processed to obtain the segmentation sequence, and then the whole-word Mask is performed for the specific words in the segmentation sequence. At the same time, tokens need to be added at the beginning of the sequence [CLS], and also at the end of the sequence or at the intervals of the sentences [SEP]. At this point, the output Embedding of all words then consists of three parts: Token Embedding, Segment Embedding and Position Embedding. Finally, the sequence vectors are input to the encoder to extract features to obtain a more semantically rich sequence vector representation. Bi LSTM is able to process contextual information at the same time, which improves the recognition capability and performance. CRF can comprehensively take into account the relationship between neighboring tags to obtain the optimal predicted sequence, which makes up for the shortcomings of Bi LSTM. For any sequence, the Bi LSTM model generates an output score matrix, P, P of size, where n is the number of words and k is the number of tags, denoting the score of the j-th tag of the i-th word. For the predicted sequence, its score function is as follows:

$$s(X, Y) = \sum_{i=0}^{n} Ay_i, y_{i+1} + \sum_{i=0}^{n} Pi, y_i \tag{1}$$

A denotes the transfer score matrix, and A_{ij} represents the score transferred from label i to label j. The size of A is $k + 2$. The probability of generating the predicted sequence Y is shown below.

$$p(Y|X) = \frac{e^{s(S,Y)}}{\sum_{\overline{Y}} S(X, \overline{Y})} \tag{2}$$

$\overline{Y}$ denotes the real labeled sequence, Y_X denotes all possible labeled sequences, and the likelihood function of the predicted sequence is obtained by taking the logarithm on both sides as follows:

$$\ln(P(Y|X)) = s(X|Y) - \ln\left(\sum_{\overline{Y} \in Y_X} S(X, \overline{Y})\right) \tag{3}$$

The output sequence that was decoded to yield the highest score is shown below.

$$Y^* = \text{argmax}\, s(X, \overline{Y}) \tag{4}$$

2.1.3. Datasets and Dataset Labelling

In this paper, we use the BIO annotation system to carry out the recognition task for entities, attributes and attribute values, and define a set of seven tokens, i.e., L = (B-GD, I-GD, B-SX, I-SX, B-SXZ, I-SXZ, O), where the meanings of the individual tokens are Pipeline Name Header, Pipeline Name Tail, Pipeline Attribute Header, Pipeline Attribute Tail, Other Attribute Header, and Other Attribute Tail, for example: 'Pipeline AT2 station average erosion rate 0.6591 mm/a'. The labelling example is shown in Figure 2.

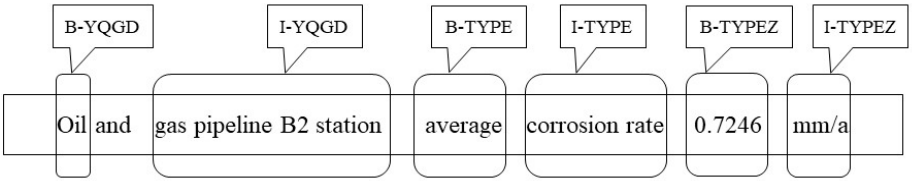


Figure 2. Annotated instance.

Oil and gas pipeline corrosion mapping belongs to the specific corrosion of oil and gas pipelines in the field of knowledge map, and access to and processing of raw data are the basic conditions for the construction of the map. The data in this paper mainly comprise a pipeline under the Northwest Oilfield Branch of the Tahe Oilfield historical corrosion monitoring data and pipeline perforation measured data. The Tahe Oilfield corrosion environment is harsh, corrosive media types are many and significant, and pipe types are complete, so this paper chose Tahe Oilfield oil and gas pipeline corrosion data as the knowledge map to build the data more persuasively. In order to make the experimental statements contain better semantic information as much as possible, this paper adopts a manual way to remove part of the invalid data records, and the labelled data records total 3905 texts. Among them, 3555 entries were used for the training process and 350 entries were used for the testing process. The statistical information of sentences, words, and various entities in the corpus is listed as shown in Table 1.

Table 1. Oil and gas pipeline corrosion data statistics.

	Quantity	B-YQGD	B-TYPE	B-TYPEZ	I-YQGD	I-TYPE	I-TYPEZ
Training set	3555	3500	39,850	40,126	30,542	149,755	75,774
Test set	350	350	4100	4100	2633	13,657	7428

In this paper, we use the TensorFlow deep learning framework for the model design. In terms of model evaluation, this study uses the probability of matching between the labelled results and the actual results as the evaluation result, and usually uses Precision (P), Recall (R), and F1 value to evaluate the model performance (see Equations (5)–(7)). The higher these three evaluation metrics are, the better the performance of the model is represented.

$$P = \frac{a}{B} \times 100\% \tag{5}$$

$$R = \frac{a}{A} \times 100\% \tag{6}$$

$$F1 = \frac{2PR}{P+R} \times 100\% \tag{7}$$

where a is the number of correct knowledge entities identified, B is the number of knowledge entities identified, and A is the number of all labeled terms.

2.1.4. Oil and Gas Pipeline Corrosion Knowledge Storage

The logical idea of knowledge graph construction is to extract unstructured and semi-structured knowledge data and represent them in the form of ternary groups, and after completing the construction of the ternary groups, it is also necessary to store the entities, attributes, and relationships that make up the ternary groups. Based on this, it is necessary to construct a graph database that can accept and store ternary data, as well as reason with and calculate the stored data. The construction of a graph database not only helps to deeply understand the association between entities, but also can be applied to subsequent knowledge retrieval, helping us to obtain the required information more quickly and improve work efficiency. Through the connections between various types of entities and the corresponding relationships between them shown in the graph database, the data information in the field of oil and gas pipeline corrosion can be understood more clearly. Constructing a knowledge map in the field of oil and gas pipeline corrosion not only clarifies the complex data information, but also helps us to better understand the relevant knowledge in the field of corrosion, further improving the level of corrosion protection, and ensuring the safe operation of pipelines.

The corrosion knowledge map of oil and gas pipeline in one area of thick oil includes nine kinds of entity node information of mean corrosion rate, H_2S content, CO_2 content, water content, pressure, temperature, flow rate, pH value, and Cl^- content, and nine kinds of entity node information of 'Mean corrosion rate is', 'H_2S content is', 'CO_2 content is', 'pressure is', 'temperature is', 'pressure is', and 'temperature is', "CO_2 content", "water content", "pressure", "temperature", "flow rate", "pressure", "temperature", "flow rate", and "pressure", "flow rate is", "pH value is" and "Cl^- content is" comprises the nine relationship edge information, as shown in Figure 3, for some oil and gas pipeline corrosion. As shown in Figure 3, it is a part of the oil and gas pipeline corrosion knowledge graph. In this paper, the Neo4j graph database [23] is used to construct the knowledge graph in the field of oil and gas pipeline corrosion. Compared with other relational databases, the Neo4j graph database provides more professional graph database functions. A total of 488 entities and 681 pairs of relationships were obtained after importing the data files of a zone of thick oil into the Neo4j graph database. Through the graph database, the relationship information between entities in the field of oil and gas pipeline corrosion can be displayed more intuitively.

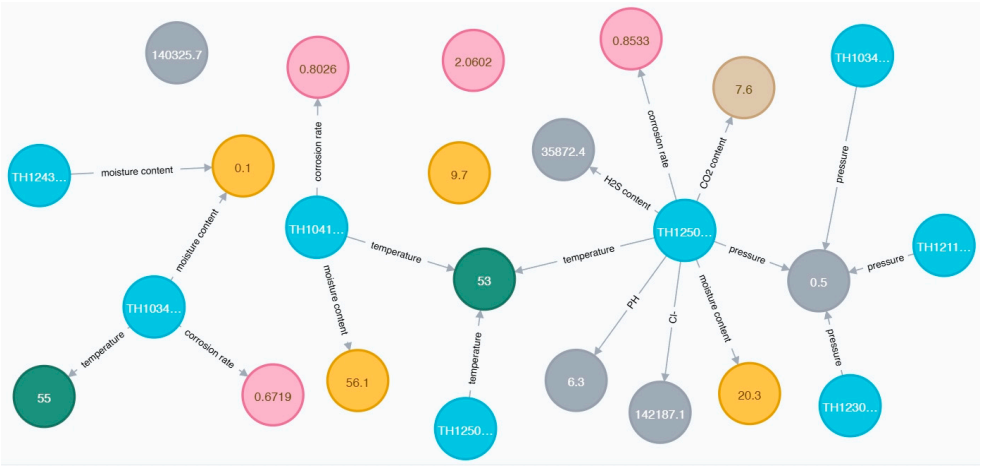


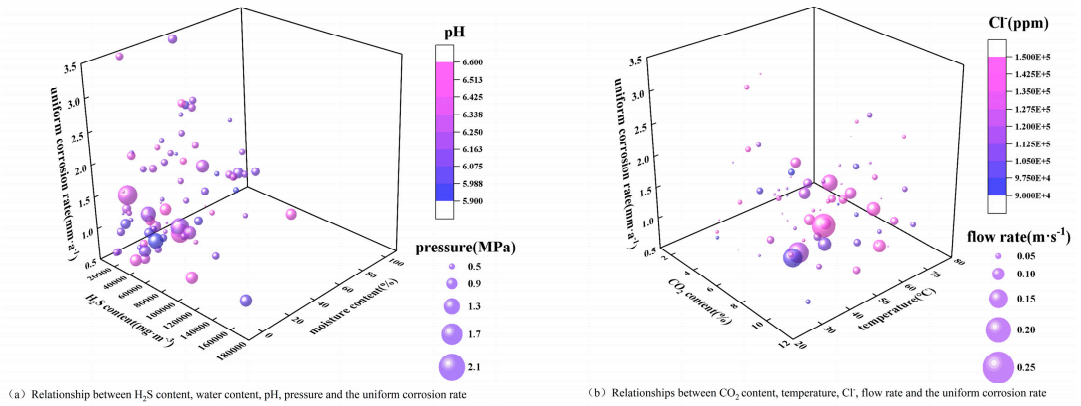
Figure 3. Schematic diagram of knowledge map of corrosion of some oil and gas pipelines.

Figure 3 shows the relationship of certain sites, such as the TH1205 temperature is 53, pH is 6.3, pressure is 0.5, etc., which are the results of saving the ternary of the entity relationship in an Excel file, and then importing the data into the Neo4j graph database using Python 3.8.5 language, and then later storing the ternary information using the Neo4j graph database.

2.2. Corrosion Rate Prediction Model for Oil and Gas Pipelines

2.2.1. Sample Data Processing

A total of 80 sets of valid samples in a zone of thick oil were obtained by searching the oil and gas pipeline corrosion knowledge map, as shown in Figure 4. For its working conditions, this paper chose H_2S content, CO_2 content, water content, pressure, temperature, flow rate, pH, and Cl^- content as the independent variables of the corrosion rate prediction model. They are all key environmental factors affecting pipeline corrosion, and H_2S and CO_2 are the main acidic gases causing corrosion in oil and gas pipelines. Water content affects the conductivity of the corrosive medium, pressure and temperature affects the rate and equilibrium of the chemical reaction, flow rate affects the contact between the corrosive medium and the pipeline surface, pH reflects the acidity and alkalinity of the medium, and Cl^- content is related to localized corrosion. Because the average corrosion rate can comprehensively reflect the corrosion condition of pipeline materials in a specific environment, which is an important indicator for assessing the life and safety of pipelines, it is set as the dependent variable of the model [24–26]. Furthermore, this study took the data from the 1st to the 64th group as the training set, and the data from the 65th to the 80th group as the test set to verify the accuracy and reliability of the prediction model. Three neural networks were used to establish the corrosion rate prediction model for oil and gas pipelines, and their prediction results were analyzed and evaluated.



（a）Relationship between H_2S content, water content, pH, pressure and the uniform corrosion rate

（b）Relationships between CO_2 content, temperature, Cl^-, flow rate and the uniform corrosion rate

Figure 4. 3D diagram of sample data.

2.2.2. Artificial Neural Network Modelling

Three kinds of neural networks were used to find out the basic network structure.

The unique advantage of BP neural networks is that they can deal with complex nonlinear characteristic functions. Compared with the traditional corrosion prediction methods, the construction of BP neural networks can study the relationship between the various factors of oil and gas pipeline corrosion in a more in-depth manner, to achieve fast and reliable prediction of the corrosion rate. The construction of the BP neural network model was divided into the creation of the network structure, determining the number of nodes of the implicit layer, selecting the appropriate activation function and the training method, data normalization processing, determining the target error, and determining the learning rate. In this paper, the network structure was set up to include 8 input layer nodes, 17 hidden layer nodes and 1 output layer node, and the tansig function and purelin function were chosen as the transfer functions between the input layer to the hidden layer and the hidden layer to the output layer. Firstly, the Levenberg-Marquardt (LM) algorithm was used for training, and the original data were normalized to improve the training speed of the model; secondly, the target error was set to 0.015 to ensure the high efficiency and stability of the training; then, through several debugging comparisons, the learning rate was set to 0.05 to obtain better prediction results; and finally, the number of iterations was set to 800 to start the training samples and test the prediction results.

RBF neural network model construction is divided into five steps: defining the RBF neural network, determining the expansion rate spread, training the network, testing, and tuning the network. In this paper, the newrb function of MATLAB 2020b was used to define the RBF neural network. The function can automatically set the architecture of the RBF network and the number of neurons in the hidden layer, and set the upper iteration limit to 800, with an error of 0.015. Since the spread parameter controls the width of the radial basis function, which affects the network's prediction accuracy and generalization ability, the optimal spread value of 100 was finally selected by using the cross-validation technique. The least squares method was used for training in order to minimize the mean square error between the output of the network and the real output, which was repeatedly adjusted until a predetermined number of training times was reached or the error reached a certain accuracy requirement. After the training was completed, it was evaluated and adjusted to improve the performance of the network.

The Random Forest algorithm constructs the decision tree using the random selection of split attribute set method. The Bootstrap method is used to resample and using each training set, the corresponding decision tree is generated. Each tree is grown intact without pruning. For test set samples, each decision tree is tested using each decision tree to obtain the corresponding category. In this paper, the TreeBagger function in MATLAB 2020b was used to construct the random forest model. The function requires the specification of the number of trees and other relevant parameters such as the maximum depth and minimum leaf size of each decision tree. By setting different numbers of decision trees and training under the error limit of 0.015, the final result was best when the number of decision trees was 100 and the minimum leaf tree was 5. Then the predict function was used for the simulation test.

2.2.3. Results and Comparative Analysis of Three Neural Network Corrosion Prediction Models

Figure 5 shows the changes in the mean square error of BP, RBF, and RF artificial neural networks with the increase in the number of iterations. As can be seen from Figure 6, with the increasing number of iterations of the artificial neural network, the mean square error of the three prediction models gradually decreases, and when the BP, RBF, and RF artificial neural networks were iterated for 652 times, their mean square errors are 0.0633, 0.0713, and 0.0924, in turn, which are all smaller than the allowable convergence error limit of 10^{-1}. Among them, the mean square error of the BP neural network is only 0.0624 when it stabilizes, which indicates that it still has the best performance among the three while meeting the specified convergence requirements.

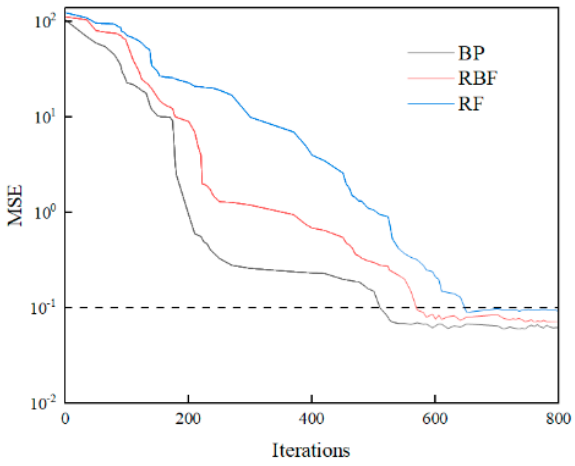


Figure 5. Variation trend of mean square error of three kinds of artificial neural networks.

Figure 6 shows the predicted output values of BP, RBF, and RF neural networks training and testing samples compared with the predicted expected values. As can be seen

from Figure 6, the three neural networks corrode the prediction model BP with the best results. This is because the BP artificial neural network performs better in dealing with complex nonlinear problems, and it can continuously adjust the weights and biases through the back-propagation algorithm in order to gradually approximate the true distribution of the training data. The RBF neural network usually uses the Gaussian radial basis function as the hidden layer activation function, which is more suitable for dealing with discrete and target variables; the RF neural network usually performs better in the classification task, because it can combine predictions from multiple decision trees and utilize integrated learning methods to improve the performance of the model. In this paper, the data retrieved from the knowledge graph was predicted with high accuracy by the BP neural network.

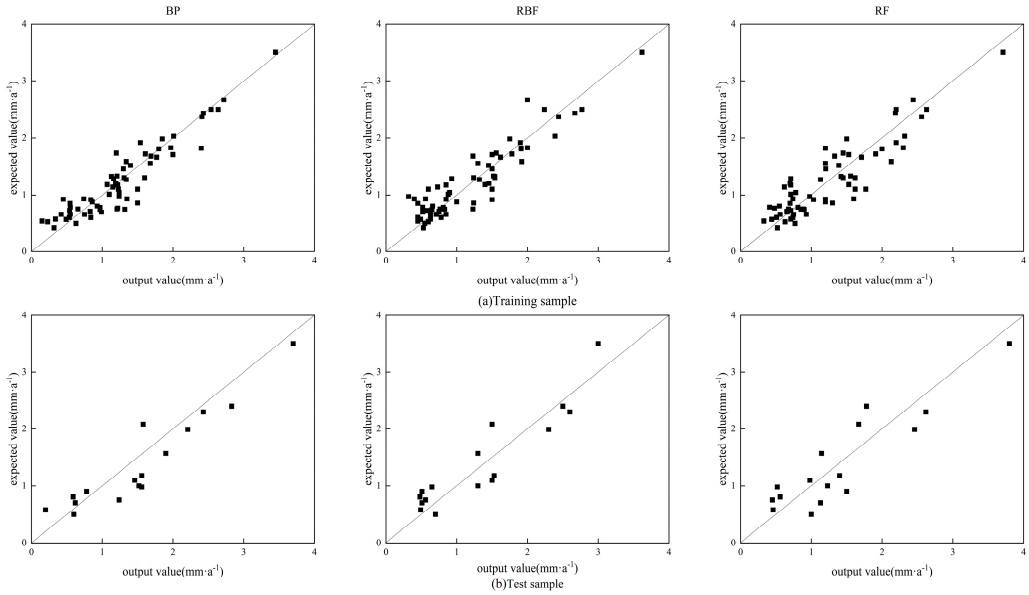


Figure 6. Output values of BP, RBF, RF neural network training and test samples compared with expected values.

Under the condition of the same mean square error limit, the coefficients of determination of the training samples and test samples of BP, RBF, and RF neural networks were 0.8447 and 0.8375; 0.8226 and 0.8110; and 0.7701 and 0.7693, respectively, and the prediction of BP was significantly better than the other two. In order to further improve the prediction accuracy and computational speed of the model, the BP neural network was optimized as follows.

2.2.4. Neural Network Optimization with Genetic Algorithms and Particle Swarm Algorithms

Two optimization algorithms were used for optimization.

As can be seen in Figure 7, GA and PSO refine, modify, and improve the prediction model in different ways. In the Genetic Algorithm section, the weights and thresholds of the BP neural network were optimized by selection, crossover, and mutation operations. In the particle swarm optimization part, individual extremes and population extremes were found by updating particle positions and velocities. These two optimization processes cooperate with each other and alternate until the end conditions are satisfied, and finally the optimal weights and thresholds are obtained, which leads to the prediction results. The whole process improves the prediction accuracy of the network by optimizing the structure and parameters of the BP neural network. The detailed construction process of the optimization model and the related parameters are explained next.

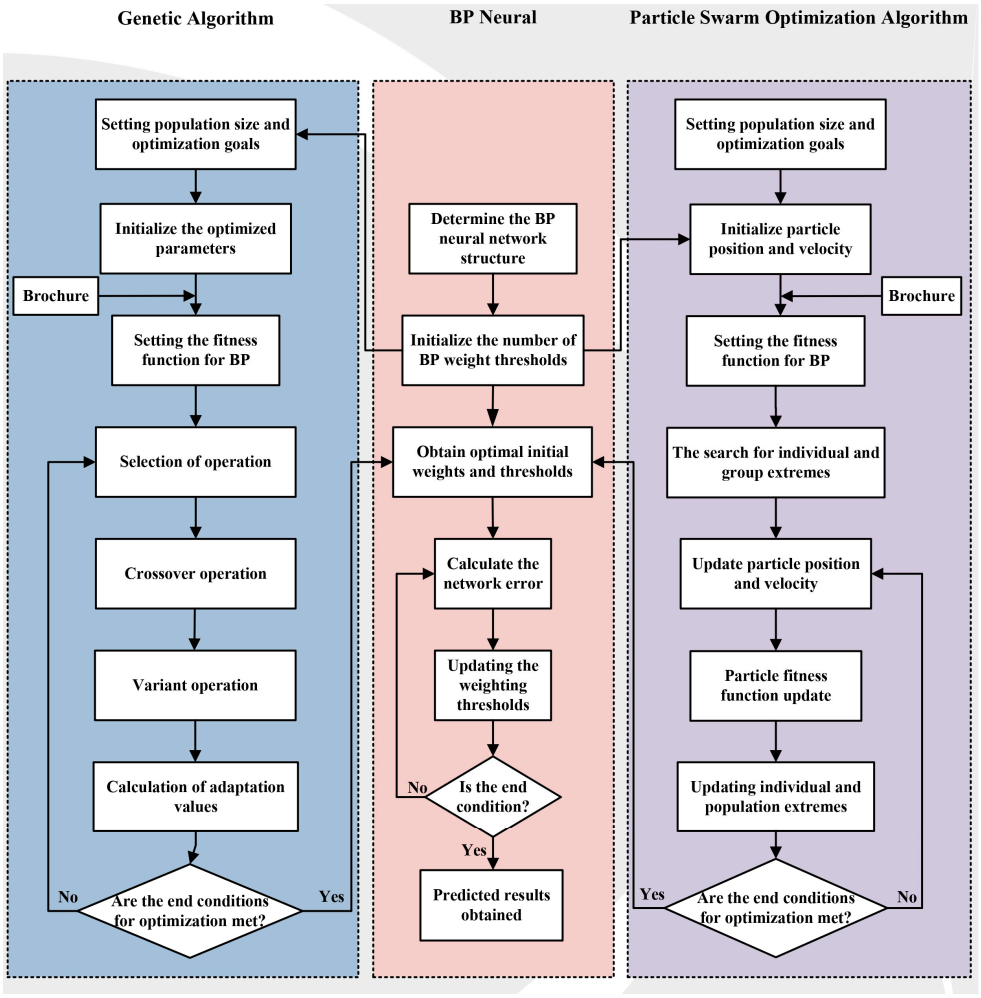


Figure 7. GA and PSO optimize the prediction modeling process.

The establishment of the GA-BP neural network includes three parts: BP neural network structure determination, genetic algorithm optimization, and BP neural network prediction. The specific network structure in this paper is 8 input layer nodes, 17 hidden layer nodes, and 1 output layer node. The number of iterations was set as 1000, the learning rate was 0.05, and the target minimum error was 0.015. The parameters of the genetic algorithm were set: the number of genetic generations was 50, the population size was 5, and the selection function was initialized, along with crossover function and compilation function. Meanwhile, according to the training data set, the fitness function of the neural network was calculated for evaluating the fitting ability and performance of the network, and then the genetic algorithm was used for optimization, including crossover, mutation, selection, and other operations. Furthermore, the weights and thresholds of the neural network were used as the genes of the genetic algorithm, and the weights and thresholds were iteratively optimized through the genetic algorithm in order to improve the fitting ability and the generalization performance of the neural network. Finally, the optimized weights and threshold value were set as the initial value, and the back propagation algorithm was used to train and optimize the neural network to improve the accuracy of regression. Figure 8 shows the genetic iteration number and adaptation value change curve.

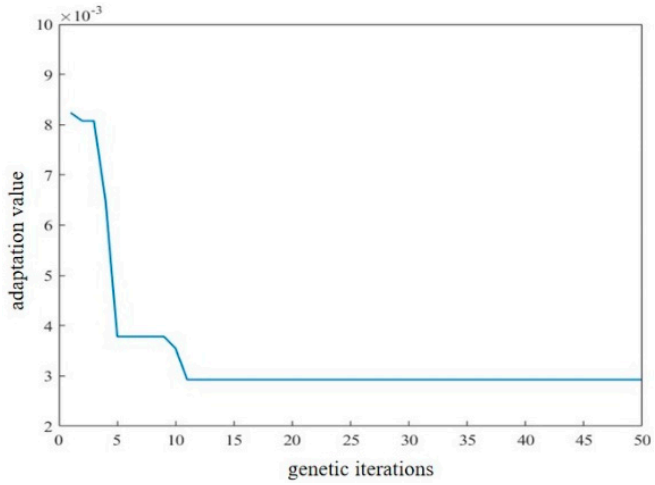


Figure 8. Fitness change curve.

The particle swarm optimization algorithm originates from the study of bird feeding behavior. Firstly, a group of particles were initialized in the solvable space, and each particle represents a potential optimal solution of the optimization problem, and the particle characteristics are represented by three indexes: the position, the speed, and the adaptation value. The specific structure comprised 17 hidden layer nodes, the number of training times was 1000, the learning rate was 0.01. The target error was 0.015, the number of population updates was 50, the population size was 5, the maximum speed was set to 1, and the minimum was −1. The neural network was initialized with the population and the speed, and there was a search for the best fitness values for the global and the individual and an iterative search for the optimization to update the optimal individual and the population and assign the values for them. The results are generated to derive the corresponding evaluation index values. Figure 9 shows the change curve of the number of iterations of the particle swarm and the fitness value.

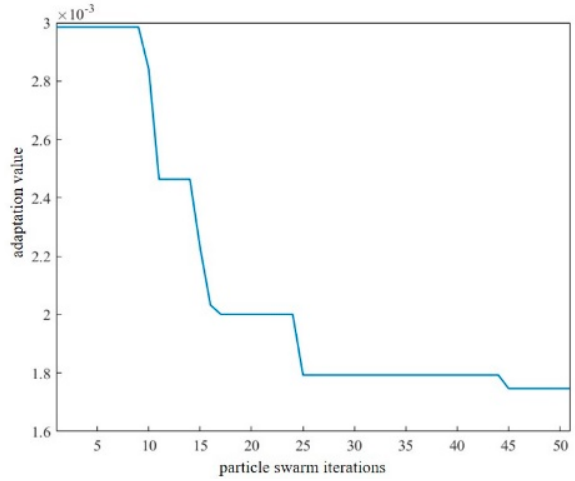


Figure 9. Model iteration change curve.

Meanwhile, in order to avoid model overfitting, cross validation and the early stopping method were used during model training. Cross validation provides more reliable model performance by dividing the data into multiple subsets, each of which is rotated as a test set and the rest as a training set; the early stopping method monitors the performance during the training process, and terminates the training early when the performance starts to degrade, which avoids the model overfitting on the training set.

3. Results

3.1. Knowledge Extraction Model Extraction Results

Under the same conditions, the LSTM-CRF model, Bi LSTM model, Bi LSTM-CRF model, and BERT-Bi LSTM-CRF were used to extract and train the corrosion information of oil and gas pipelines, respectively, and the results were compared in the three metrics of the model's precision, recall, and F1 value, and the results are shown in Figure 10.

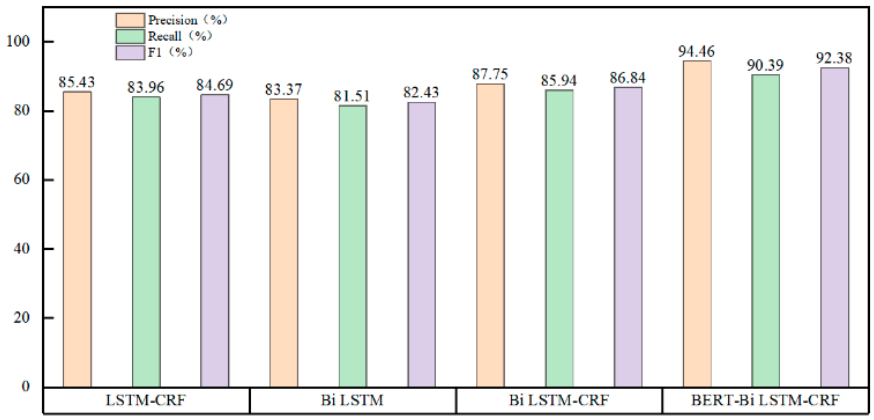


Figure 10. Effect of different models on knowledge extraction of oil and gas pipeline corrosion.

From Figure 10, it is known that the Bi LSTM-CRF model performs better in the entity task compared to the LSTM-CRF model. Specifically, the Bi LSTM-CRF model is 2.42%, 1.7%, and 2.05% higher in the three metrics of precision rate, recall rate, and F1 value, respectively. This is because the Bi LSTM-CRF model is able to acquire both forward and backward textual information, which better captures the contextual information of entities in the text and thus improves entity recognition.

The Bi LSTM-CRF model improves 4.3%, 4.4%, and 4.35% in terms of precision, recall, and F1 value, respectively, relative to the Bi LSTM model. This is due to the fact that the CRF layer globally models the tag sequences while constraining the dependencies between each tag, thus obtaining the global optimal sequence. This acquisition of global optimal sequences allows the model to be more accurate in transferring between labels, improving the model's entity and entity-relationship extraction capability and accuracy. The BERT-Bi LSTM-CRF model outperforms the Bi LSTM-CRF model in terms of precision, recall, and F1 value, reaching an improvement of 6.65%, 4.46%, and 5.52%, respectively.

3.2. Neural Network Optimization Results for Genetic Algorithms and Particle Swarm Algorithms

Figure 11 shows the comparison of the expected and output values of GA-BP and PSO-BP neural network prediction models used for training and testing. The output values of the two artificial neural networks in the training and testing phases are uniformly distributed in the vicinity, which is approximately linear, and the coefficients of determination (R^2) in the training and testing phases of GA-BP are 0.9901 and 0.9889, respectively, while those of PSO-BP are 0.9954 and 0.9935 for the training and testing phases, respectively, which shows that the expected and predicted values in the models are highly correlated and both can accurately predict the oil and gas pipelines in this thick oil area. Both algorithms combined BP with other optimization algorithms to overcome some of the shortcomings of the original algorithms. Through different methods, they were able to introduce a certain degree of randomness in the search process and improve their global search ability, which improves the robustness and search efficiency of the algorithms.

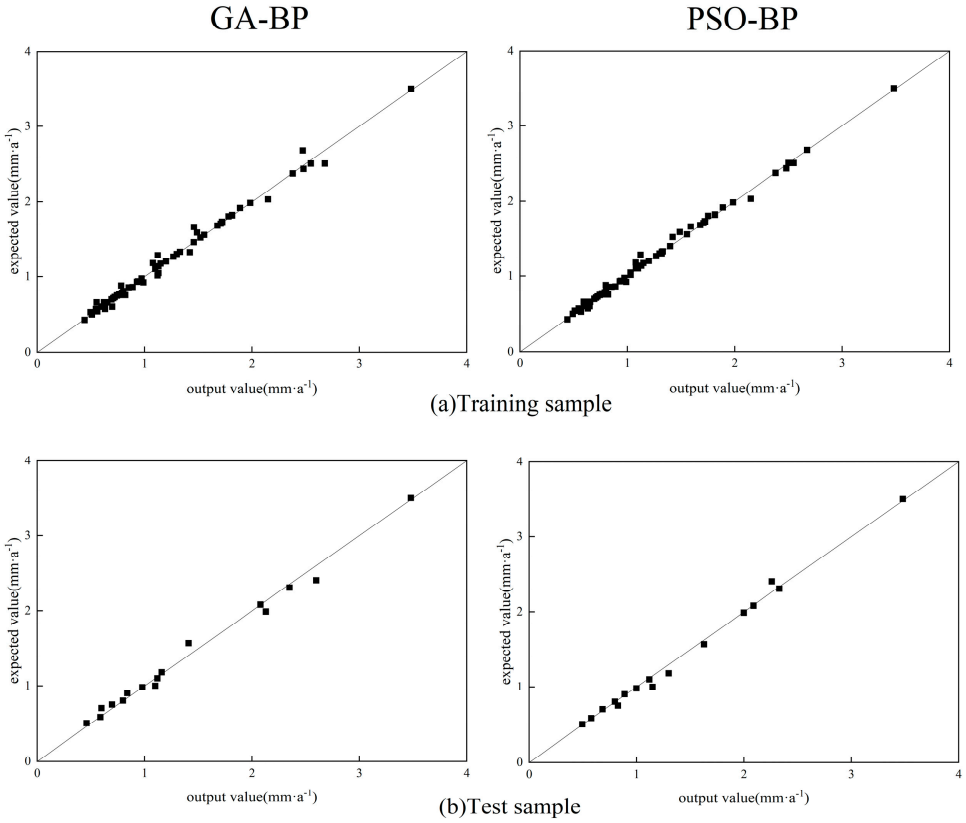


Figure 11. Comparison of output values and expected values of GA-BP and PSO-BP neural network training and test samples.

The mean square error of the corrosion rate prediction models optimized by the genetic algorithm and particle swarm algorithm are calculated to be 0.000733 and 0.000432, respectively. The particle swarm-optimized BP artificial neural network is better than the genetic algorithm, which indicates that the particle swarm optimization algorithm has higher prediction accuracy, which can better make up for the shortcomings of the prediction of the BP neural network alone.

4. Discussion

The excellent results of the BERT-Bi LSTM-CRF model in terms of precision, recall, and F1 value are due to the strong semantic understanding capability of the BERT model, as well as its ability to comprehensively characterize character-level, word-level, and inter-sentence relationship information. Pre-training the BERT model can handle syntactic and lexical information in different contexts, which improves the generalization ability of the model and further improves the recognition and extraction of entities and entity relations. The BERT model itself has a large number of parameters, plus BiLSTM and CRF layers, which makes the number of parameters of the whole BERT-BiLSTM-CRF model increase dramatically, and the model's complexity is the highest among the four models, except for the complex semantic information in the field of corrosion. To improve the semantic recognition ability is the most important; weighing the model's complexity as well as the computational efficiency, the BERT-BiLSTM-CRF model is more suitable for this study.

For model optimization, the PSO-BP algorithm is more suitable for optimization problems in continuous space, and can better deal with the balance between local optimal solutions and global optimal solutions, improve the performance and convergence speed of the model through the collaborative search ability of particle swarms, and the optimization ability of BP neural networks, and its self-cognition and social cognition can ensure its development ability. The PSO algorithm has a more flexible search strategy that can simultaneously consider more than one search direction at the same time, which allows it to

better adapt to different problems and datasets. However, for genetic algorithms, it is more sensitive to the selection of the initial population. In the same population, setting different fitness functions can result in different numbers of individuals better than the average fitness value, i.e., different problem-solving abilities. Not only that, but genetic algorithms also need more evolutionary cycles in complex engineering problems. Meanwhile, for the samples in this study, PSO-BP was better than GA-BP in terms of model complexity and computational efficiency. Although it is not as simple and fast as the BP neural network, sacrificing a small amount of resources in exchange for a more accurate model prediction is acceptable, and it is necessary for the accuracy of corrosion prediction. Based on various factors, for the sample data studied in this paper, the PSO-BP neural network corrosion rate prediction model has a higher accuracy, which can predict the desired results more accurately and provide an important theoretical basis for corrosion protection.

It also has high potential application value. Using the BERT-BiLSTM-CRF model and PSO-BP neural network model can monitor the corrosion condition of oil and gas pipelines in real time and predict potential corrosion risks, so that maintenance and repair can be carried out in advance to avoid accidents. Integration of these models into automated inspection systems can be used to automatically identify and categorize pipeline corrosion types, improving inspection efficiency and accuracy. Using the corrosion rate data predicted by the models, asset managers can better plan maintenance budgets, develop long-term maintenance strategies, and optimize resource allocation.

5. Conclusions

In general, it is feasible to use a combination of knowledge graphs and neural network models for corrosion prediction.

(1) The BERT-Bi LSTM-CRF joint extraction method and seven-step method were used to construct the knowledge graph in the field of oil and gas pipeline corrosion, and the Neo4j graph database was used to store the related ternary group information and the ontology of oil and gas pipeline ontology. Classes of corrosion influencing factors and relationships were defined by combining the historical monitoring data, pipeline perforation measured data, and pipeline information, and the pipeline is classified in partitions. The structure of the pipeline corrosion data was visualized and the effect is excellent.

(2) After comparison, it was found that the BP neural network was accurate in predicting the rate of mean corrosion of oil and gas pipelines. In order to improve the accuracy, the genetic algorithm and particle swarm optimization algorithm were used to optimize the BP neural network, which achieved a more accurate prediction of the rate of the mean corrosion of oil and gas pipelines.

(3) The model built by the method of knowledge graph for data organization and the PSO-BP neural network for corrosion prediction was highly accurate, but with the increase of oil and gas pipelines, the number of metal pipeline entities and their relationships in the knowledge graph will increase exponentially, which will lead to ambiguity problems in the relationships between entities. The problem of compatibility of a large number of entities and their relationships can be a limitation of this study, so the dataset should be expanded and a deeper study should be conducted.

Author Contributions: R.X.: Responsible for the main article writing and completing the whole paper. Z.F.: Responsible for the revision of the article, put forward a lot of valuable suggestions, very instructive. X.H.: Responsible for the modification of the article, and complete part of the picture drawing. W.L., Y.L., Y.Z. and J.H.: Responsible for completing part of the picture drawing and modifying part of the description. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Xi'an Shiyou University Graduate Student Innovation and Practical Ability Cultivation Programme (YCS23214233).

Data Availability Statement: Because of confidentiality and privacy restrictions, the data cannot be made publicly available.

Acknowledgments: We thank Xi'an Petroleum University, especially the School of Chemistry and Chemical Engineering, and are grateful to the faculty as well as the remaining authors for their contributions to this article, and to the editors and reviewers of Processes for their recognition of the article.

Conflicts of Interest: Author Xinyu Hao was employed by the company Shaanxi Yanchang Petroleum (Group) Company Limited Refining and Chemical Company. Author Weibin Luo was employed by the company SINOPEC Qilu Petrochemical Company. Author Jie Han was employed by the company Shaanxi Mingze Yisheng Energy Technology Company Limited. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Yu, X.Y.; Zhou, L.Z. Current situation and countermeasures for the security of China's onshore oil and gas corridors. *China Energy* **2022**, *44*, 51–55.
2. Yue, H.J. Status quo of anti-corrosion work of oil pipeline and solution countermeasures. *Chem. Manag.* **2018**, *11*, 223–224.
3. Zhang, X.L. Analysis of pipeline anti-corrosion measures in oil and gas storage and transport. *Chem. Manag.* **2019**, *8*, 156–157. [CrossRef]
4. Xu, L.; Yu, J.; Zhu, Z.; Man, J.; Yu, P.; Li, C.; Wang, X.; Zhao, Y. Research and Application for Corrosion Rate Prediction of Natural Gas Pipelines Based on a Novel Hybrid Machine Learning Approach. *Coatings* **2023**, *13*, 856. [CrossRef]
5. Pawan, B.; Trushar, B.G.; Ajeet, K.S. Machine learning-based predictive approach for pitting and uniform corrosion in geothermal energy systems. *Electrochim. Acta* **2024**, *504*, 144884.
6. Khalaf, A.H.; Lin, B.; Abdalla, A.N.; Han, Z.; Xiao, Y.; Tang, J. Enhanced prediction of corrosion rates of pipeline steels using simulated annealing-optimized ANFIS models. *Results Eng.* **2024**, *24*, 102853. [CrossRef]
7. Cui, B.Y.; Wang, H. Pipeline corrosion prediction and uncertainty analysis with an ensemble Bayesian neural network approach. *Process Saf. Environ. Prot.* **2024**, *187*, 483–494. [CrossRef]
8. Zhuang, Q.; Liu, D.; Chen, Z. Prediction on Failure Pressure of Pipeline Containing Corrosion Defects Based on ISSA-BPNN Model. *Energy Eng.* **2024**, *121*, 821–834. [CrossRef]
9. Yu, G.; Wang, W.H.; Song, H.W.; Mi, H.; Tang, J.; Zhao, Z. Prediction of external corrosion rate for buried oil and gas pipelines: A novel deep learning method with DNN and attention mechanism. *Int. J. Press. Vessel. Pip.* **2024**, *209*, 105218.
10. Li, S.; Du, H.; Cui, Q.; Liu, P.; Ma, X.; Wang, H. Pipeline Corrosion Prediction Using the Grey Model and Artificial Bee Colony Algorithm. *Axioms* **2022**, *11*, 289. [CrossRef]
11. Liu, J.T.; Chen, K.Y.; Duan, H.Y.; Li, C. A knowledge graph-based hazard prediction approach for preventing railway operational accidents. *Reliab. Eng. Syst. Saf.* **2024**, *247*, 110126. [CrossRef]
12. Shi, Y.T.; Zhao, L.; Zhou, M.; Yin, X.; Guo, W.; Li, C. A dynamic community gas risk-prediction method based on temporal knowledge graphs. *Process Saf. Environ. Prot.* **2023**, *177*, 436–445. [CrossRef]
13. Duan, P.B.; Yang, K.; Su, X.; Fan, S.; Dong, X.; Zhang, F.; Li, X.; Xing, X.; Zhu, Q.; Yu, J.; et al. HTINet2: Herb-target prediction via knowledge graph embedding and residual-like graph neural network. *Brief. Bioinform.* **2024**, *25*, bbae414. [CrossRef] [PubMed]
14. Chakraborty, N.; Lukovnikov, D.; Maheshwari, G.; Trivedi, P.; Lehmann, J.; Fischer, A. Introduction to neural network-based question answering over knowledge graphs. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2021**, *11*, e1389. [CrossRef]
15. Wei, M.Z.; Zheng, R. Research on Industry Emerging Technology Forecast Modeling Based on Knowledge Graph and Deep Neural Networks. *J. China Soc. Sci. Tech. Inf.* **2022**, *41*, 1134–1148.
16. Yu, Y.; Zhang, W. Fusing Knowledge Graph and Neural Networks to Enable Digital Intelligent Management Decision Making. *J. Manag. Sci. China* **2023**, *26*, 231–247.
17. Fan, B.C. *Study on Chinese Medicine Recommendation Technique for Breast Cancer Chinese Tongue Diagnosis Based on Knowledge Graph and Neural Network*; Guangzhou University of Chinese Medicine: Guangzhou, China, 2021.
18. *SY/T 0087.2-2012*; Corrosion Evaluation Standard for Steel Pipelines and Storage Tanks. Direct Evaluation of Corrosion in Buried Steel Pipelines. National Energy Administration: Beijing, China, 2012.
19. *SY/T 0087.3-2010*; Corrosion Evaluation Standard for Steel Pipelines and Storage Tanks, Direct Evaluation of Corrosion in Steel Storage Tanks. National Energy Administration: Beijing, China, 2010.
20. *SY/T 0087.1-2006*; Corrosion Evaluation Standard for Steel Pipelines and Storage Tanks, Direct Evaluation of External Corrosion of Buried Steel Pipelines. National Energy Administration: Beijing, China, 2006.
21. Noy, N.F.; Mc, G. Ontology Development 101: A Guide to Creating Your First Ontology. *Knowl. Syst. Lab.* **2001**, *32*, 98–103.
22. Cai, C.H. *Research of Knowledge Graph-Based Semantic Information Extraction Method and Its Application*; Harbin Institute of Technology: Harbin, China, 2018.
23. Lu, F. Construction and Application of Personnel Relationship Knowledge Graph based on Neo4j. *Softw. Eng.* **2022**, *25*, 5–8.

24. Ju, J.P. *Research on Rate Prediction and Corrosion Mechanism of Buried Long-Distance Pipeline in Oilfield*; Northwest University: Xi'an, China, 2018.
25. Obaseki, M.; Alfred, P.B.; Elijah, P.T.; Okuma, S.O. Corrosion Rate Prediction in Oil and Gas Pipelines Based on Multiphase Flow Modelling. *Int. J. Eng. Res. Afr.* **2023**, *67*, 27–48. [CrossRef]
26. Fang, J.; Cheng, X.; Gai, H.L.; Lin, S.; Lou, H. Development of machine learning algorithms for predicting internal corrosion of crude oil and natural gas pipelines. *Comput. Chem. Eng.* **2023**, *177*, 108358. [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.