Feasibility Analysis of Optimization Models for Natural Gas Distribution Networks Using Machine Learning

Junhao Liu[1] and Xiaoyong Gao [1] and Xiaozheng Chen[1]

[1] China University of Petroleum, Beijing, Beijing 102249, China

Abstract. As natural gas pipeline networks expand, the complexity of pipeline scheduling models grows, making feasibility analysis increasingly difficult. This study focuses on the feasibility analysis of optimization models for natural gas distribution network scheduling, treating it as a classification problem. Models grounded in traditional neural networks, parallel branch neural net-works, and graph neural networks are developed and assessed. Two distinct scales of natural gas distribution networks are explored by collecting a limited dataset of sample cases to train and validate the proposed feasibility analysis models through empirical case studies. The results demonstrate that the parallel branch neural network exhibits superior predictive performance. Additionally, this study introduces an innovative methodology for traceability diagnosis of infeasible cases, offering a practical framework for engineering applications.

Keywords: Natural Gas Distribution, Feasibility Analysis, Machine Learning.

1 Introduction

Natural gas, characterized as a low-carbon, clean, and high-quality energy source, occupies a significant position within the global energy structure. With increasing environmental awareness and rising energy demands, the consumption of natural gas has steadily escalated. According to the International Energy Agency, natural gas is projected to account for 25% of the global energy demand by 2040, positioning it as the second-largest energy source following crude oil [1]. Due to its efficiency, cost-effectiveness, and reliability, pipelines have become the essential mode of transportation linking gas sources to consumer markets. As per statistics from the Statista database[2], China leads globally in the operation of natural gas pipelines. As of February 2024, China's natural gas network comprises 442 functional pipelines, including various pipeline segments, with an additional 302 pipelines either proposed or under construction. The total number of operational natural gas pipelines worldwide exceeds 1,500.

Over the past decades, scholars have proposed numerous optimization algorithms to address a wide range of problems. Dynamic Programming (DP), Generalized Reduced Gradient (GRG), and Linear Programming (LP) are three commonly used precise algorithms for optimizing natural gas distribution pipeline scheduling models. Moreover, in large-scale optimization, emerging heuristic algorithms such as Genetic Algorithms (GA), Ant Colony Optimization (ACO), Simulated Annealing (SA), and Particle Swarm Optimization (PSO) have shown significant advantages over traditional precise methods. These heuristics offer novel approaches and tools for optimizing natural gas pipeline operations.

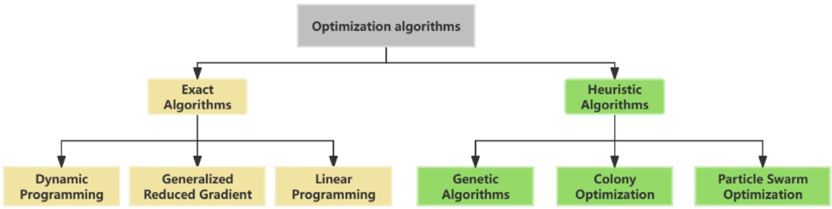

Fig. 1. Optimization algorithms applied to natural gas distribution network optimization models

However, the implementation of optimization algorithms assumes the existence of feasible solutions. Without a feasible solution, even the most advanced algorithms are ineffective. Additionally, actual production scheduling is a complex, multifaceted, and dynamic system, where feasible solutions in mathematical models are not always guaranteed. In engineering practice, determining model feasibility and identifying conflicting constraints when a model is infeasible are time-consuming and challenging tasks, which limits their practical application. Consequently, the feasibility analysis of optimization models has become increasingly crucial, especially for practical engineering optimization problems, and is now an indispensable part of engineering applications. John W. Chinneck in his book "Feasibility and Infeasibility in Optimization"[3] poses several pertinent questions: What happens when algorithms fail to find an initial feasible solution? How do we identify the source of the problem? How can we amend the model? These questions also guide the feasibility analysis of the natural gas sales pipeline network model discussed in this paper. By analyzing the feasibility of the model, identifying reasons for its infeasibility, and modifying the model to ensure it has a feasible solution domain, this process constitutes the primary research approach and content of this study.

The remainder of this paper is organized as follows. Section II introduces the background of natural gas distribution network scheduling optimization issues, research approaches to feasibility analysis, and methods for extracting characteristic data. Section III designs feasibility analysis models based on traditional neural networks, parallel branch neural networks, and graph neural networks, describing the structures of these models. Section IV presents case studies demonstrating the performance of the proposed models and introduces specific methods for traceability diagnosis of infeasible cases. The conclusions and future work are discussed in Section V. The structure of this paper is illustrated in Figure 2.

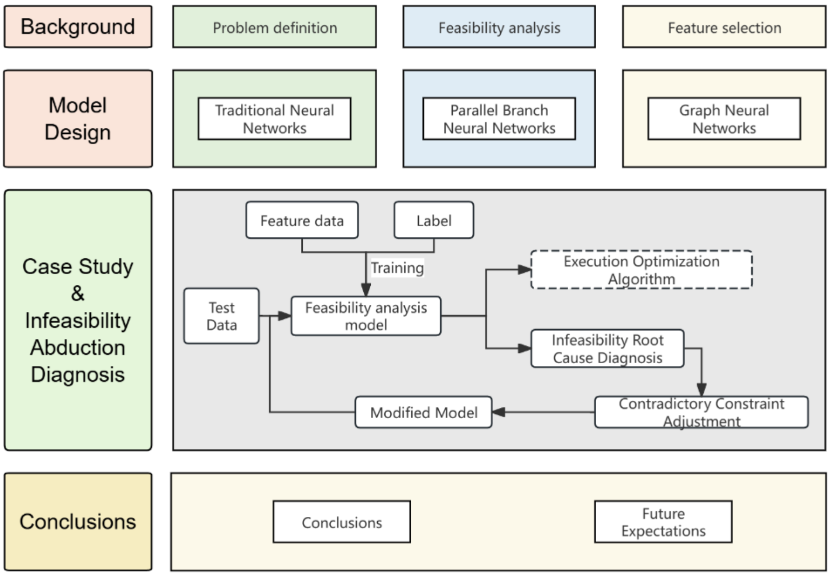

Fig. 2. The structure of this paper

2 Background

2.1 Problem definition

Natural gas pipeline networks typically comprise gathering pipelines, transmission pipelines, distribution pipelines, compressors, control valves, and resistance components. Gathering pipelines collect raw natural gas from production wells and transport it to processing plants. After impurities are removed, the transmission pipelines deliver the processed gas to city gate stations, often located thousands of kilometers away. Subsequently, the distribution pipeline distributes the natural gas to customers. Due to frictional losses during transportation, compressors are installed in series or parallel to compensate for the pressure loss of the gas. Control valves are used to regulate the demand for natural gas by customers. The structural topology of the natural gas distribution network is illustrated in Figure 3.

When optimizing scheduling strategies for natural gas distribution networks, constructing an accurate mathematical model is essential. The mathematical model of a natural gas sales network typically includes constraints such as upper and lower limits on gas flow, material balance constraints within the pipeline network, physical capacity constraints of storage components, pipeline pressure constraints, and an objective function. Since the focus of this project is on the feasibility analysis of the model rather than optimizing economic benefits, this paper adopts a fictitious objective function to simplify the solution process. In this scenario, the objective function can be set as minimize obj=0. The primary purpose of this setting is to enable the solver to quickly find a feasible solution that meets all constraints, rather than seeking an optimal solution. This approach allows for the rapid collection of sample case data on the feasibility/infeasibility of labels.

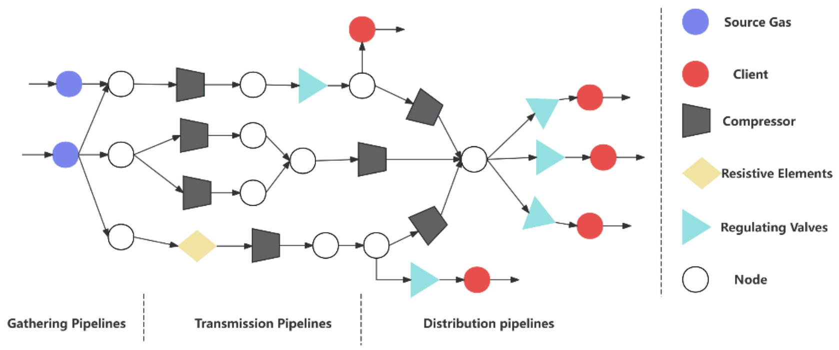

Fig. 3. The structural topology of the natural gas distribution network

2.2 Feasibility analysis

The feasibility analysis of the model involves two key components. The first is the feasibility assessment, which determines whether a feasible initial solution exists, confirming the model's viability. The second component is the infeasibility analysis, which identifies the root causes of infeasibility and suggests modifications to the model to generate feasible solutions.

In the context of design variables, feasibility analysis seeks to determine whether a process can satisfy all constraints within an uncertain space by adjusting control variables. This mathematical formulation was initially proposed by Grossmann and colleagues in 1983[4], who introduced a systematic approach to handle the optimization issues related to design and control variables. More recent research by Wang and Ierapetritou in 2017[5] presented a feasibility analysis based on surrogate models, which not only enhances prediction accuracy but also effectively identifies and confirms the feasibility regions of the process within a limited sampling budget. In 2019, Dias and Ierapetritou [6] developed a framework interpreting feasibility analysis as a classification problem, employing Support Vector Machines, Neural Networks, and Decision Trees to define a predictor that classifies any point in the uncertain space as feasible or infeasible. Chinese scholar Rong Han [7] proposed a branch-and-bound approach for feasibility analysis based on a 0-1MILP model with the context of refinery optimization scheduling, providing a relatively systematic analysis of the model's feasibility.

This paper treats the feasibility analysis of the natural gas distribution network scheduling optimization model as a classification problem, utilizing machine learning methods to determine feasible regions for scheduling issues. To comprehensively assess the performance of the feasibility analysis models developed in this paper, we adopt feasibility metrics proposed by Wang and Ierapetritou in 2017[8]. The entire range of uncertainty parameters is divided into four regions as shown in Figure 4: CF

(Correctly Feasible) represents the region correctly predicted as feasible by the model; CIF (Correctly Infeasible) denotes the region correctly predicted as infeasible; ICF (Incorrectly Feasible) indicates the region predicted as feasible but is actually infeasible; ICIF (Incorrectly Infeasible) refers to the region predicted as infeasible but is actually feasible.

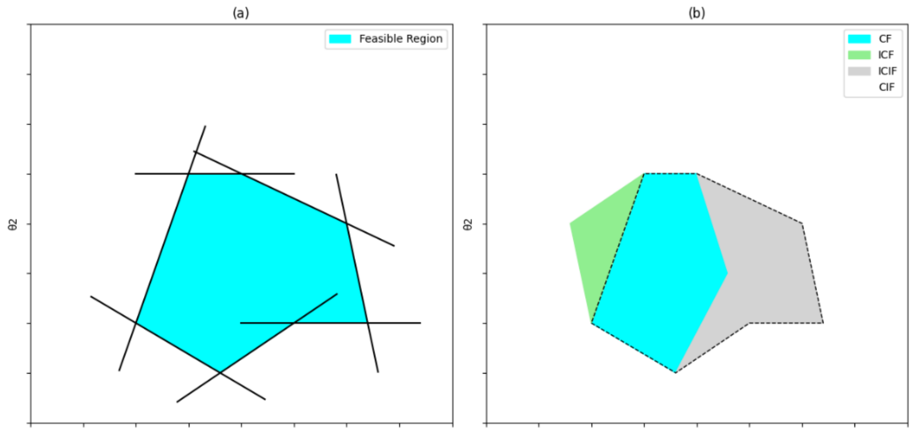

Fig. 4. Four regions in the feasibility analysis

In the diagram, the left image (a) represents the feasible region within the uncertain space as a shaded blue area. The right image (b) shows the four evaluation metrics for the model in the given feasible region. Based on the aforementioned definitions, four metrics are derived to evaluate the accuracy of the model:

$$CF\% = \frac{CF}{CF+ICIF} \times 100 \tag{1}$$

$$CIF\% = \frac{CIF}{CIF+ICF} \times 100 \tag{2}$$

$$NC\% = \frac{ICF}{ICF+CF} \times 100 \tag{3}$$

$$Total\ Error = \frac{ICF+ICIF}{CF+CIF+ICF+ICIF} \times 100 \tag{4}$$

The first two metrics, CF% and CIF%, describe the extent to which the uncertain parameter space has been correctly explored and classified in terms of feasibility. CF% indicates the percentage of the feasible region in the original space that is correctly identified by the classifier; CIF% indicates the percentage of the infeasible region in the original space that is correctly identified by the classifier. The third metric, NC%, represents the percentage of the feasible region overestimated by the classifier and is used to evaluate the conservativeness of the classifier's predictions. The fourth metric, Total Error, measures the total number of misclassifications relative to the size of the test set. When CF% and CIF% approach 100%, and NC% and Total Error approach 0, the feasibility analysis model is demonstrated to accurately approximate the feasible region.

2.3 Feature selection

In the feasibility analysis of natural gas sales network scheduling optimization models, selecting characteristic data is a crucial step, especially when applying machine learning techniques for prediction and classification. The primary goal of feature selection is to identify the most influential subset of features from the original dataset that impacts the label data y (i.e., whether the network has a feasible solution). This process not only enhances the model's predictive accuracy but also significantly reduces the computational complexity of model training and improves interpretability.

This paper conducts a chi-square test on variable data within the network (such as the upper and lower limits of pipeline flow) and model label data (feasible or infeasible). The chi-square test is a method used to examine the association between two categorical variables. In this study, we employ the chi-square test to assess the correlation between various features and the feasibility (label) of the network scheduling model. Specifically, our aim is to identify which features exhibit a significant statistical association with the model's ability to find feasible solutions. Here are the steps involved in the chi-square testing process:

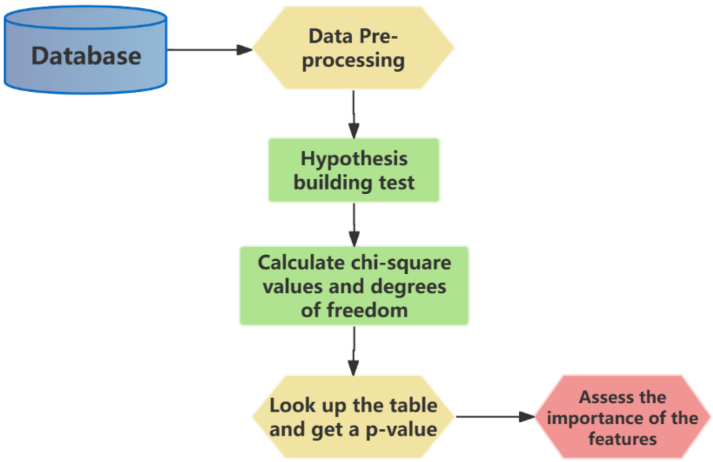

Fig. 5. The steps of the chi-square test

In this section, we first extract the variable data and label data from the mathematical model and set up hypothesis testing to assume that the feature data and label data are independent of each other. Next, we insert each feature data and corresponding label data into the chi-square statistical formula to calculate the chi-square value x^2 and degrees of freedom σ, which typically equals (number of categories -1)×(number of features -1).The chi-square statistic formula is shown below:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{5}$$

where O_i is the observed frequency, and E_i is the expected frequency under the assumption that the feature and label are independent. A larger chi-square statistic for a feature indicates a stronger correlation between that feature and the label. Based on this, we extract the characteristic data of the model.

3 Feasibility Analysis Model Design

In this chapter,we approach the feasibility analysis of natural gas distribution network scheduling optimization as a classification problem. Three models have been developed for this purpose: a traditional neural network, a parallel branch neural network, and a graph neural network. These models assess the feasibility of optimization solutions across different scales in practical applications.

3.1 Design of Feasibility Analysis Model Based on Traditional Neural Networks

This paper initially employs a traditional neural network to identify feasible regions for scheduling. The structure of the model consists of an input layer, hidden layers, a Dropout layer, and an output layer, as illustrated in Figure 6. In the input layer, neurons multiply each input by a weight $w_{input} \in R^{1\times N}$ and add a bias term $b_{input} \in R^{1\times N}$, which is then passed through a "ReLU" activation function to serve as the input for the hidden layers. The ReLU function, short for Rectified Linear Unit, effectively prevents the problem of vanishing gradients and enhances the convergence efficiency of the gradient descent method. It is one of the most widely used activation functions today. The formula for the ReLU function is as follows:

$$f(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{6}$$

The hidden layers of the model consist of three fully connected layers, where each neuron's input is successively multiplied by hidden weights w_{hidden} and hidden biases b_{hidden}, and processed through a ReLU activation function to serve as the input for the subsequent layer. Additionally, during the training of neural networks, models are highly susceptible to overfitting. To mitigate this issue, a Dropout layer is incorporated between the hidden layers and the output layer, effectively reducing overfitting and providing a degree of regularization. In this project, a dropout rate of 0.3 is set, deactivating approximately 30% of the neurons randomly to prevent model overfitting. The structure diagram of traditional neural network model is illustrated in Figure 6.

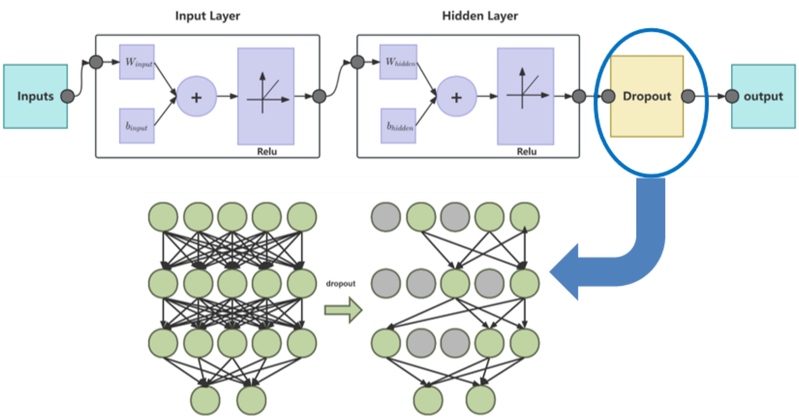

Fig. 6. Structure diagram of traditional neural network model

Finally, the output layer consists of two neurons, utilizing the Softmax activation function. The Softmax function is a normalized exponential function commonly employed in multi-class classification problems, allowing model outputs to be interpreted as probability distributions. In this model, the design of the output layer enables the differentiation between two states: feasible or infeasible for the scheduling problem. The formula for the Softmax function is as follows:

$$p(x) = \frac{1}{1+e^{-x}} \tag{7}$$

3.2 Design of Feasibility Analysis Model Based on Parallel Branch Neural Networks

The second method of feasibility analysis employed in this study utilizes a Parallel Branch Neural Network (PBNN) architecture to delineate the feasible regions for scheduling tasks. A PBNN is characterized by a neural network structure with multiple independent branches that process input data simultaneously. This design allows the network to learn different types of features concurrently across branches, enabling more effective handling of complex or heterogeneous datasets.

In the proposed PBNN model for feasibility analysis, the network receives n distinct types of input data from various segments of the natural gas distribution network. For example, inputs pertaining to gas sources include parameters such as "maximum gas source flow" and "minimum gas source flow." Inputs from the client side encompass "maximum client flow," whereas pipeline-related inputs cover "maximum forward flow," "minimum forward flow," "maximum reverse flow," and "minimum reverse flow," among others. Each branch of the network processes the same feature data specific to each component type. The outputs from all branches are subsequently amalgamated into a unified feature vector through a concatenation operation. The final output layer employs a Softmax activation function to determine the class with the highest probability. To augment the generalization capability of the model and mitigate the risk of overfitting, a Dropout layer with a dropout rate of 30% is integrated. The architecture of the model is depicted in Figure 7.

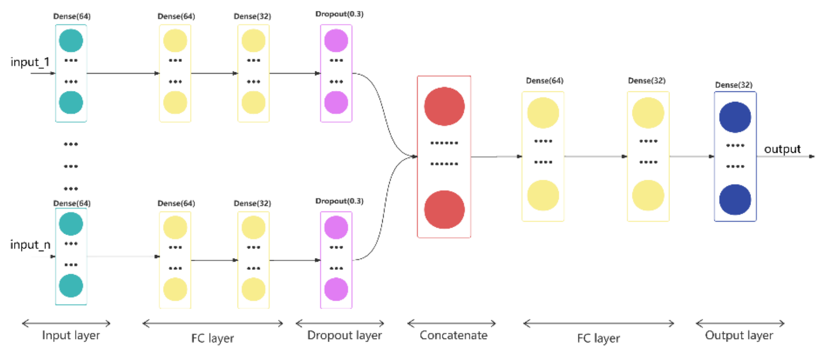

Fig. 7. Structure diagram of parallel branch neural network model

This model design effectively addresses the complexity and diversity of the natural gas network by employing a PBNN architecture, where each branch is specialized to capture the essential features of different network components. This targeted approach enhances the model's efficacy in extracting and utilizing information from each component type, surpassing traditional single-network models. By integrating specialized branches, the model provides a robust framework for precise feasibility classification, significantly improving prediction accuracy over conventional methods. This refinement in neural network design represents a major advancement in feasibility analysis for complex systems.

3.3 Design of Feasibility Analysis Model Based on Graph Neural Networks

The third method for feasibility analysis incorporates a Graph Neural Network (GNN) to delineate the feasible region for scheduling within the complex pipeline system. This approach conceptualizes the pipeline system as a graph, framing the feasibility analysis as a graph classification issue. Within this model, various components of the natural gas network, including clients, gas sources, pipelines, and regulating valves, are represented as nodes, while the physical interconnections between these components are depicted as edges.

Each component's feature data is employed as node features, collectively comprising the feature matrix X and the adjacency matrix A to construct the dataset. These matrices are integral to the graph representation, enabling the GNN to process the intricate relationships and interactions within the network.

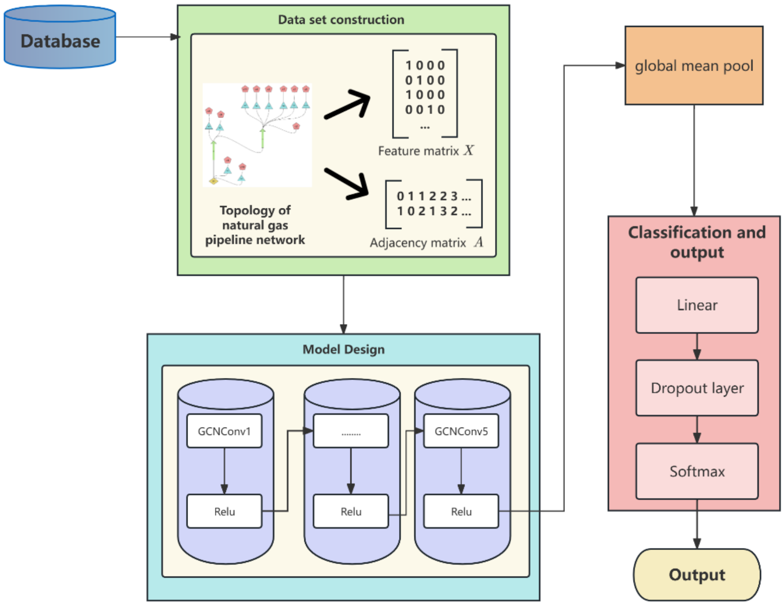

Fig. 8. Structure diagram of graph neural network model

The GNN model is designed with five graph neural network layers, each followed by a ReLU (Rectified Linear Unit) activation function. This configuration introduces non-linearity, enhancing the model's ability to capture complex relationships and feature propagation among nodes. Such an arrangement allows the network to assimilate the collective behavior of the nodes effectively. Following the node-level processing, a global mean pooling operation aggregates the features of individual nodes into a unified representation of the graph's global features, categorized as feasible or infeasible. These global features are then processed through a linear layer, which outputs the graph classification prediction results. The structural schematic of this model is shown in Figure 8.

4 Case Analysis and Infeasibility Root Cause Diagnosis

In this section, we demonstrate and validate the effectiveness of the three feasibility analysis models through specific case studies. Specifically, this chapter selects two natural gas sales network models of different scales (a small-scale model with 13 nodes and a large-scale model with 131 nodes) as the subjects of study. Through these two cases,we explore the performance differences of the models on datasets of varying sizes. For cases predicted as infeasible, we also perform detailed infeasibility root cause diagnoses to identify and analyze the key factors contributing to the infeasibility.

4.1 Case Analysis of the 13-Node Natural Gas Distribution Network Model

In this section, we first select a natural gas sales network model comprising 13 nodes. This model includes 11 customer nodes, 1 gas source node, 2 pipelines, and 10 regulating valve components. Utilizing the 'networkx' package from the Graph Neural Network library, we render a visualization of the network's topology. The depicted topology is then methodically compared with the actual structural diagram to ensure accuracy and consistency in representation, as illustrated in Figure 9.

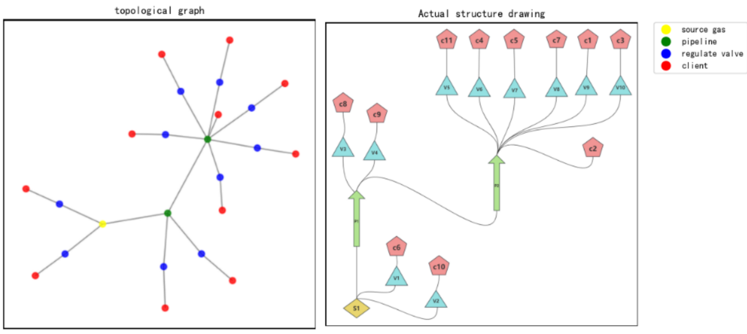

Fig. 9. Topology of the 13-node model

This paper divides the collected 864 sample cases into training, validation, and testing sets in a 7:1:2 ratio, respectively. To identify the variables most relevant to predicting feasibility, chi-square testing is employed to determine specific feature input data. The output data consist of labels indicating whether the cases are feasible or infeasible, and each model is trained using the defined training dataset. During the training process, the three feasibility analysis models utilize the Adam optimization algorithm and employ sparse categorical cross-entropy as the loss function. Each model is configured with a batch size of 32 and is run for 300 epochs. The training results are depicted in Figures 11 and 12.

Fig. 10. Comparison of model training loss values

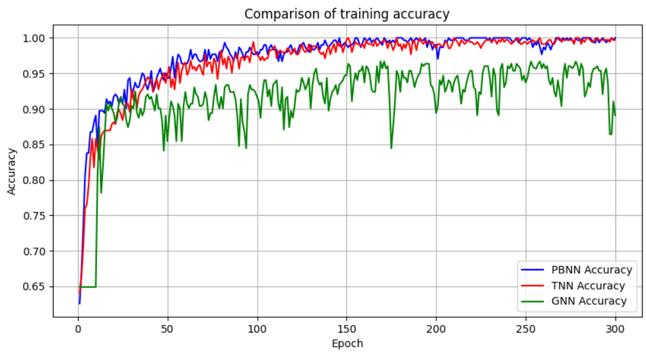

Fig. 11. Comparison of model training accuracy

We then input the test set data into the three trained feasibility analysis models and further evaluate their performance using feasibility metrics CF%, CIF%, NC%, and Total Error. The results are presented in Table 1.

Table 1. Performance of different models in the test set.

	CF%	CIF%	NC%	Total Error
PBNN Model	**90.91**	**98.21**	**3.5**	**5.75**
TNN Model	89.09	96.61	7.54	5.78
GNN Model	89.28	93.24	9.09	8.46

According to the data in the table, the Parallel Branch Neural Network model demonstrates the best predictive performance among the trained models when applied to the test set. Both CF% and CIF% are close to 100%, while NC% and Total Error approach 0, indicating that this feasibility analysis model can accurately approximate feasible regions. The traditional neural network performs better than the Graph Neural Network, but both are overall less effective compared to the Parallel Branch Neural Network model.

4.2 Case Analysis of the 131-Node Natural Gas Distribution Network Model

This section features a case study analysis of a natural gas sales network model composed of 131 nodes. The model includes 121 customer nodes, 10 gas source nodes, 53 pipelines, 3 compressors, 73 regulating valves, and 2 resistors among other components. We have visualized the network topology of this model, and the resulting topology diagram is compared with the actual structural diagram as shown in Figure 12.

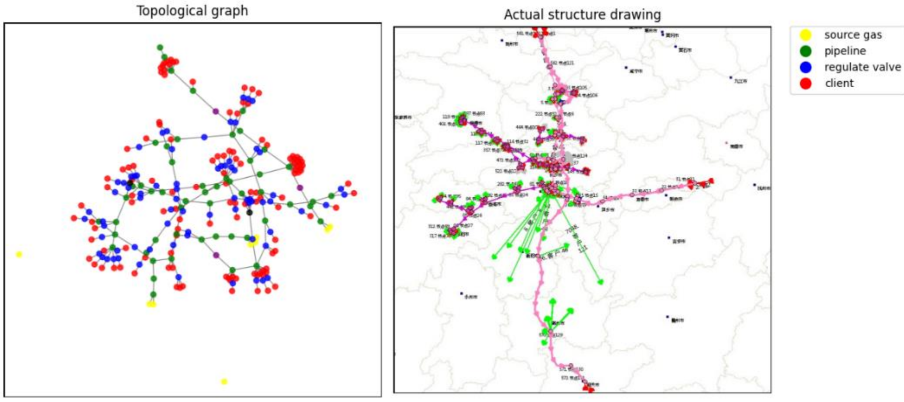

Fig. 12. Topology of the 131-node model

In this study, 432 sample cases were collected and divided into training, validation, and testing sets in a 7:1:2 ratio, following the same training procedures. The training outcomes are illustrated in Figures 13 and 14.

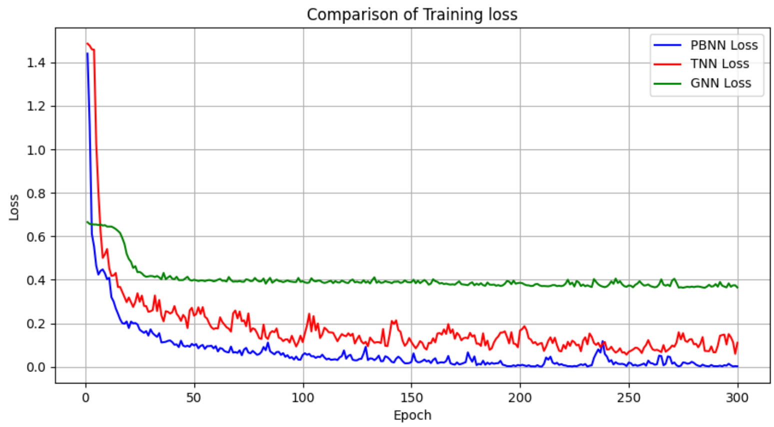

Fig. 13. Comparison of model training loss values

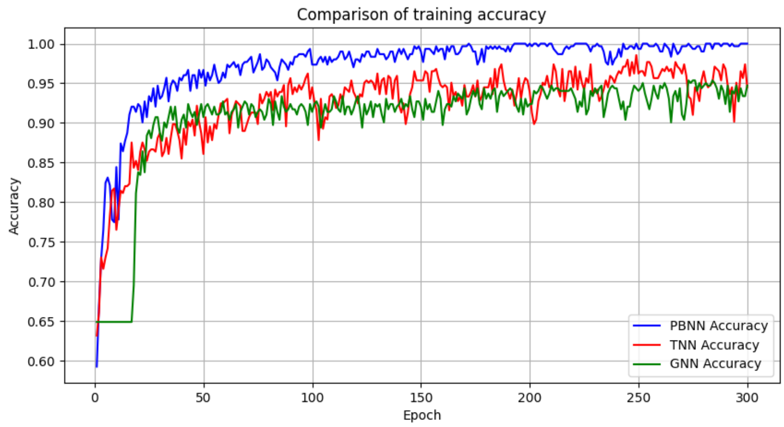

Fig. 14. Comparison of model training accuracy

Subsequently, we input the test set data into the three trained models mentioned above, compared the predicted results with the actual outcomes, and evaluated the performance of the feasibility analysis models using the feasibility metrics CF% (Correctly Feasible), CIF% (Correctly Infeasible), NC% (Necessarily Correct), and Total Error. The results are displayed in Table 2.

Table 2. Performance of different models in the test set.

	CF%	CIF%	NC%	Total Error
PBNN Model	**92.85**	**94.91**	10.34	**5.74**
TNN Model	81.25	94.54	13.79	10.34
GNN Model	91.07	93.24	**8.92**	7.69

According to these results, although the accuracy of the three feasibility analysis models decreased as the scale of the natural gas network increased, the Parallel Branch Neural Network model still outperformed the other two neural network models in predictive performance. With a CIF% nearing 95% and a minimal Total Error, the model demonstrated commendable predictive accuracy.

4.3 Infeasibility Root Cause Diagnosis

This section employs the Irreducible Infeasible Subsystem (IIS) method for infeasibility root cause diagnosis. The goal is to identify an IIS from the set of infeasible constraints, localize and centralize the problem within the conflicting constraints of the IIS, and then seek model corrections within this relatively small set of constraints.

Infeasibility root cause diagnosis is based on the aforementioned feasibility analysis. Each iteration requires invoking the feasibility analysis model, necessitating a very high accuracy for the feasibility analysis model. However, for any neural network prediction model, multiple iterations can cause the errors to gradually accumulate, ultimately leading to poor infeasibility diagnosis performance and inaccurate IIS identification. Therefore, this study focuses only on identifying the category of the conflicting component, determining which type of component causes infeasibility. The best-performing model, the Parallel Branch Neural Network feasibility analysis model, is used for infeasibility root cause diagnosis. The specific diagnostic process is as follows:

1. For an infeasible case Q, let q_i represent the feature data of the i-th component of Q.
2. Temporarily expand or reduce the feature data q_i in a direction that favors a feasible solution.
3. Use the trained Parallel Branch Neural Network feasibility analysis model to assess the feasibility of the modified Q'.
4. If Q' is feasible, then q_i is considered to be the cause of the infeasibility.
5. If Q' remains infeasible, then q_i is not considered the cause of the infeasibility.
6. Iterate through all feature data to identify an IIS of the infeasible model.

In this section, we perform an Infeasibility Root Cause Diagnosis on an infeasible 13-node case to identify the component category responsible for the infeasibility. First, the trained PBNN model is used to determine whether the sample case is feasible. If the model identifies the case as infeasible, each component is examined individually. We iteratively adjust the feature data of each component in a direction that favors feasibility, either increasing or decreasing the relevant parameters.The modified case is then reevaluated using the model until it is classified as feasible.Through this process, we successfully identified that the infeasibility of the case was due to the gas source's maximum output being insufficient to meet the total customer demand. The corresponding program analysis results are shown in the figure 15.

```
1/1 [==============================] - 0s 171ms/step
Predictions classes: [0]
The sample case is predicted as infeasible, proceeding with the Infeasibility Root Cause Diagnosis.
First, increase the maximum flow rate feature data for all pipeline components, and re-evaluate the feasibility.
1/1 [==============================] - 0s 20ms/step
Predictions classes: [0]
The prediction result after increasing the pipeline component feature data: Infeasible.
Reverting to the initial data, then slightly increasing the maximum flow rate feature data for all valve components, and re-evaluating the feasibility.
1/1 [==============================] - 0s 23ms/step
Predictions classes: [0]
The prediction result after increasing the valve component feature data: Infeasible.
Reverting to the initial data, then slightly increasing the maximum flow rate feature data for the gas source, and re-evaluating the feasibility.
1/1 [==============================] - 0s 21ms/step
Predictions classes: [1]
The prediction result after increasing the gas source component feature data: Feasible.
*************************************
All components have been examined. Based on the results, the cause of infeasibility in this case is likely that
gas source's maximum output being insufficient to meet the total customer demand
```

Fig. 15. Infeasibility Root Cause Diagnosis result graph

By utilizing the aforementioned method, users can modify the data within the constraint conditions to render the problem feasible. This significantly reduces the complexity associated with applying optimization scheduling to large-scale network operations.

5 Conclusions

This paper investigates the feasibility of a scheduling optimization model for natural gas distribution networks. By reviewing both domestic and international studies, we develop three models that identify feasible regions using a constrained set of sample cases and machine learning techniques. The models provide an effective way to define feasible boundaries within the parameter space. Additionally, we propose a method for diagnosing the causes of infeasibility when sample cases fall outside these boundaries, offering practical insights for engineering applications. These approaches not only improve the accuracy of feasibility analysis but also aids in optimizing the management of natural gas distribution networks.

References

1. Group C W .World Energy Outlook 2022[J].Chemical Weekly, 2022.
2. STATISTA.Countries with largest number of operational and planned natural gas pipelines worldwide as 2024
3. Chinneck J W .Feasibility and infeasibility in optimization. Algorithms and computational methods[J].Interfaces, 2009, 39(3):292-295.DOI:10.1007/978-0-387-74932-7.
4. Grossmann I E, Halemane K P, Swaney R E. Optimization strategies for flexible chemical processes[J]. Computers & Chemical Engineering, 1983, 7(4): 439-462.
5. Wang Z , Ierapetritou M .A novel feasibility analysis method for black-box processes using a radial basis function adaptive sampling approach[J].Aiche Journal, 2017.DOI:10.1002/aic.15362.
6. Dias L S , Ierapetritou M G .Data-driven feasibility analysis for the integration of planning and scheduling problems[J].Optimization and Engineering, 2019, 20(4):1029-1066.DOI:10.1007/s11081-019-09459-w.
7. Han, Rong. "Feasible Solution Domain Analysis of Mixed-Integer Linear Programming Model for Production Scheduling" [D]. Shandong University, 2010. DOI: 10.7666/d.y1790213.
8. Wang Z , Ierapetritou M .A novel feasibility analysis method for black-box processes using a radial basis function adaptive sampling approach[J].Aiche Journal, 2017.DOI:10.1002/aic.15362.