Petroleum Science 20 (2023) 3194–3209

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

A reliability-oriented genetic algorithm-levenberg marquardt model for leak risk assessment based on time-frequency features

Ying-Ying Wang [a, *], Hai-Bo Sun [a], Jin Yang [a], Shi-De Wu [b], Wen-Ming Wang [b], Yu-Qi Li [c], Ze-Qing Lin [a]

[a] College of Safety and Ocean Engineering, China University of Petroleum, Beijing, 102249, China
[b] College of Mechanical and Transportation Engineering, China University of Petroleum, Beijing, 102249, China
[c] College of Instrument Science and Engineering, Harbin Institute of Technology, Harbin, 150006, Heilongjiang, China

ARTICLE INFO

Article history:
Received 25 July 2022
Received in revised form
18 April 2023
Accepted 18 April 2023
Available online 18 April 2023

Edited by Jia-Jia Fei

Keywords:
Leak risk assessment
Oil pipeline
GA-LM model
Data derivation
Time-frequency features

ABSTRACT

Since leaks in high-pressure pipelines transporting crude oil can cause severe economic losses, a reliable leak risk assessment can assist in developing an effective pipeline maintenance plan and avoiding unexpected incidents. The fast and accurate leak detection methods are essential for maintaining pipeline safety in pipeline reliability engineering. Current oil pipeline leakage signals are insufficient for feature extraction, while the training time for traditional leakage prediction models is too long. A new leak detection method is proposed based on time-frequency features and the Genetic Algorithm-Levenberg Marquardt (GA-LM) classification model for predicting the leakage status of oil pipelines. The signal that has been processed is transformed to the time and frequency domain, allowing full expression of the original signal. The traditional Back Propagation (BP) neural network is optimized by the Genetic Algorithm (GA) and Levenberg Marquardt (LM) algorithms. The results show that the recognition effect of a combined feature parameter is superior to that of a single feature parameter. The *Accuracy*, *Precision*, *Recall*, and *F1score* of the GA-LM model is 95%, 93.5%, 96.7%, and 95.1%, respectively, which proves that the GA-LM model has a good predictive effect and excellent stability for positive and negative samples. The proposed GA-LM model can obviously reduce training time and improve recognition efficiency. In addition, considering that a large number of samples are required for model training, a wavelet threshold method is proposed to generate sample data with higher reliability. The research results can provide an effective theoretical and technical reference for the leakage risk assessment of the actual oil pipelines.

1. Introduction

The pipeline is the most economical and effective way for transporting oil to the terminal. Mobilization and long-distance distribution of crude oil resources are primarily realized via the use of pipelines (Oyedeko and Balogun, 2015; Xu et al., 2010). In recent years, to ensure the safety of pipeline transportation, the pipeline risk assessment has attracted more and more attention (Hu et al., 2014; Zhang et al., 2020b). Many researchers are very concerned about the remaining life of the pipeline (Chen et al., 2020), the buckling of the pipeline (Zhao et al., 2010; Li et al., 2017), corrosion of the pipeline (Mazumder et al., 2021b; Cui et al., 2016; Zeng et al., 2014), and the maintenance strategies. Yet, the risk assessment of pipeline leakage during maintenance should also be of concern, as leakage can cause not only economic losses but also the risk of explosion. Oil and gas pipeline leakage is one of the main causes of resource loss and is also one of the common types of pipeline accidents (Lu et al., 2020a). The pipeline leakage can cause serious problems such as explosions, economic loss, and environmental pollution. For example, in November 2013, the Donghuang Petroleum Pipeline in Qingdao leaked crude oil into a municipal drainage culvert, causing an explosion. The accident caused 62 deaths, 136 injuries, and a direct economic loss of 750 million yuan (Lu et al., 2020b). Therefore, risk assessment of the maintenance process in the oil pipeline network is crucial to ensure the safety of the system. In the detection process, not only the accuracy and efficiency of leak detection should be considered, but also the appropriate method of leak detection should be selected

* Corresponding author.
E-mail address: wyy@cup.edu.cn (Y.-Y. Wang).

https://doi.org/10.1016/j.petsci.2023.04.016

from an economic, environmental protective, and portability perspective.

The current state-of-the-art leak detection techniques used in oil and gas pipelines are divided into hardware-based methods and software-based methods based on the detection technology characteristics (Lu et al., 2020b). The leak detection sensors mainly include acoustic and vibration. The hydrophone could acquire excellent measurement results in an environment with a low signal-to-noise ratio while the vibration sensor obtained a more significant peak value of correlation coefficient. Pipeline leakage caused local energy loss and stress waves propagating in the pipe wall. A leak detection system is proposed based on a negative pressure wave that used harmonic wavelet analysis to identify the extracted signal (Hu et al., 2011). According to the results, harmonic wavelet analysis presented advantages over similar methods in terms of extracting the weak non-stationary negative pressure wave signal. A new leak location method is proposed based on the propagation characteristics of leakage acoustic waves for oil and gas pipelines. Then, the dominant energy frequency bands of leakage acoustic waves are obtained based on the wavelet transform analysis (Liu et al., 2015). Some monitoring systems have been successfully applied in actual pipe networks (Harmouche and Narasimhan, 2020). In addition, computational fluid dynamics (CFD) was used to study the dynamic characteristics of leakage. The results showed that the pressure change at the leak location was not obvious but more distinct after the gradient transform (Fu et al., 2020). The frequency spectrum of pressure fluctuations was analyzed for indicating that the leakage signals were concentrated in a 220–500 Hz frequency band (Ben-Mansour et al., 2012). A fast Fourier transform on the collected signals was performed to understand the vibration characteristics of leakage signals (Mostafapour and Davoudi, 2013). The results revealed that the leakage signal energy was concentrated in a range of 150–300 kHz, while the theoretical and experimental errors were below 6%.

Several studies have applied machine learning to pipeline risk assessment (Kang et al., 2018; Wang et al., 2022). Eight data-driven machine learning algorithms are evaluated based on the generated dataset to identify the best failure prediction model (Mazumder et al., 2021a). The artificial neural network can learn the fault online and can also adapt to the dynamic background noise environment. A neural network was used to predict leakage online (Waleed et al., 2019). A microphone was installed on a 60-m pipe to capture the leakage noise signal. The frequency decomposition of the noise signal was used as the input of the neural network model, with the output represented by leakage flow rates of 1 mm, 2 mm, and 3 mm. The model recognition rate reached 100%, except for the 1 mm leakage calibre, indicating that the network model fully predicted the degree of leakage. A neural network based on a leak identification framework that could be used to verify the leak identification efficacy is proposed (Santos et al., 2014). Pressure data is converted into Markov chains and performed feature extraction based on statistical indicators, such as variance (Liu et al., 2019). Furthermore, two decision models were built for long-term and short-term detection. The short-term detection model used pressure data which were transformed over a short period to rapidly determine the pipeline state and detect pipeline abnormalities, while the long-term detection model could more accurately identify leak signals. A long-distance pipeline leakage model was constructed for simulating leakage data, using two support vector machine (SVM) models to predict the leakage occurrence and location, respectively (Xie et al., 2019). Various operating modes in pipeline transportation were considered to establish various leakage models, effectively reducing the false alarm rate (Zhou et al., 2019). A new pipeline leak detection technique based on data field theory was proposed (Liu, 2019). This method not only identified the leak but also predicted its location. It could reliably detect and locate singular oil pipeline leak signals. Moreover, a transient leak detection method was applied to accurately identify the leak locations. This method accurately determined the location of single leaks even in pipe flows (Aamo, 2016). MATLAB software was used to model and simulate unidirectional flow pipelines and combined this method with artificial neural networks to identify leaks (Omojugba et al., 2020). Variational modal decomposition of the signal was conducted, reconstructing the signal after removing the noise to extract the leakage characteristics. Finally, leakage pattern recognition was carried out using SVM (Diao et al., 2020).

Obtaining large amounts of real leak data is difficult. Existing data could not provide enough effective leakage information to train high-precision models of leakage prediction. Many data derivation methods using Generative Adversarial Networks have been proposed to obtain additional data (Wang et al., 2019; Zhang et al., 2020a, 2022; Hu et al., 2021). The credibility of the derived data was high, exhibiting an abundance of data types. Data derivation models based on small samples have been proposed (Gao et al., 2022). Feature extraction of the signal is crucial for leak prediction. Time-domain statistical features have been perfectly combined with neural networks (Lang and Yuan, 2020). The method of multi-scale analysis was used to extract the leakage characteristics while using the Gaussian mixed model to detect pipeline leakage (Rai and Kim, 2021). Spectrum enhancement (SE) and convolutional neural network (CNN) are combined for predicting pipeline leakage, and SE was used to enhance the signal (Ning et al., 2021). The results demonstrated that the recognition rate of CNN reached 94.3%. The contourlet neural network is optimized by an improved grey optimization algorithm, and the second curvelet neural network is optimized based on an improved firefly algorithm to improve prediction precision (Zhao et al., 2019a, 2019b, 2020; Zhao and Song, 2021).

Process monitoring plays an important role in pipeline safety management and risk assessment. A reliable leakage risk assessment can assist in developing an effective pipeline maintenance plan and avoiding unexpected incidents. Actual pipeline operating conditions are complicated. Although several studies have employed artificial neural networks for pipeline leak identification, not many leak detection models can guarantee both the accuracy and efficiency of leak detection. The detection difficulties in the pipeline operation process can be summarized as three aspects: (1) The complexity of pipeline operating conditions; (2) The extracted feature parameters cannot fully represent the original leakage signal. Leak identification is a classification problem, the leakage prediction accuracy based on single or few feature parameters is often low; (3) The actual leaked data is limited, so the number of samples is not enough to train a reliable prediction model. Therefore, data derivation is necessary for pipeline leak identification.

Inspired by the above three points, a data-driven detection method is proposed based on time-frequency feature extraction, BP neural network optimized by improved algorithm, and data derivation. The contributions of this paper mainly include three aspects. Firstly, the feature extraction method based on the time-frequency feature is applied to detect pipeline leaks. By extracting the time-frequency feature, the raw pressure data can be effectively represented. Secondly, the traditional BP neural network optimized by the GA and the LM algorithm can improve the accuracy and reduce the training time of the model respectively. The proposed GA-LM model can solve the contradiction between fast detection and accurate detection. Finally, a wavelet threshold method is proposed for data derivation. The proposed model shows high reliability for the raw and generated data, which can be applied in pipeline process safety and leakage risk engineering.

2. Modeling

2.1. GA-LM model

The artificial neural network has been widely used in oil fields of scientific research and engineering applications (Gao et al., 2021; Heravi and Hodtani, 2018; Nitta and Kuroe, 2018; Yan et al., 2020). The optimization methods of the BP network are gradually increasing, such as optimizing the BP training process through a dynamic learning rate (Zhang et al., 2012). The traditional BP neural network is used for local search optimization, causing the algorithm to fall into a local extremum while exhibiting low network convergence speed. To address these challenges, the GA and LM are combined to optimize the traditional BP neural network model. The GA is used to adjust the parameters and structure of the neural network (Yu et al., 2020). The basic elements of the GA algorithm include chromosome coding, fitness functionality, genetic operation, and operational parameters (Tao et al., 2021). The genetic operations include selection, crossover, and mutation operations.

Considering the given N samples (x_k, y_k) $(k = 1, 2, \ldots, N)$, the output of the network is y_k for a certain x_k output. The output of node i is $O_{ik,}$ and the j-th unit of the t-th layer is examined. At a k-th sample input, the output of node j is expressed by:

$$net_{jk}^{t} = \sum_{j} w_{ij}^{t} O_{jk}^{t-1} + q_j^{t} \tag{1}$$

where $O^{t-1}{}_{jk}$ represents the output of the j-th unit node when the k-th sample is entered into the t-1 layer. q_j is the threshold of the j-th neuron. The loss function is expressed by:

$$E = \frac{1}{N}\sum_{k=1}^{N}\left(y_{jk} - y_{jk}^{*}\right)^{2} \tag{2}$$

The weight correction reduces E:

$$W_{ij}^{*} = W_{ij} - \eta\frac{\partial E}{\partial W_{ij}} \tag{3}$$

where, (η>0), $W^{*}{}_{ij}$ is the correction value, and η is the learning rate. The neural network is trained to find the weight and threshold when E is the smallest.

Assuming that there are n_L nodes in the t-th layer, the loss function E is expressed by:

$$E(W) = \frac{1}{2}\sum_{i=1}^{n_L} e_i^2(W) = e^T(W)e(W) \tag{4}$$

The individual contains all the weights and thresholds of the neural network. The BP neural network is cyclically trained according to individual weights and thresholds. The code length S is expressed by:

$$S = n_1{*}m + m{*}n_2 + m + n_2 \tag{5}$$

where n_1 is the number of input layer nodes, m is the number of hidden layer nodes, and n_2 is the number of output layer nodes. Part of the neural network algorithm is selected as the objective function of the GA. After the training data is passed through the neural network to predict the output, the calculation formula of the individual fitness value f is expressed by:

$$f = k\left(\sum_{i=1}^{n} abs(y_i - o_i)\right) \tag{6}$$

where n is the number of network output nodes. y_i is the expected output of the ith node of the BP neural network. o_i is the actual output of the i-th node, and k is the coefficient.

The selection operator is improved based on the optimal preservation strategy, which can effectively select the better individuals in the population. The process of selecting individuals is shown in Fig. 1. The steps of selecting individuals are as follows: (a) Determine an initial population and calculate the fitness value of each individual in the population; (b) Sort the individuals in the population according to their fitness from small to large; (c) Divide the population equally into 3 segments; (d) Each segment is randomly selected according to the ratio of 0.6, 0.8, and 1; (e) Randomly select a lost individual from the individuals in the tail segment; (f) Insert the lost individuals at the end of the proportionally selected population to obtain a new population. After the above operation process, the better individuals in the population can be selected, while maintaining the diversity of the population. The average fitness value of the final population is improved compared to the initial population.

Assume that the h-th chromosome and the l-th chromosome intersect at the j-th position.

$$d_{hj} = d_{hj}(1-b) + d_{lj}b \tag{7}$$

$$d_{lj} = d_{lj}(1-b) + d_{hj}b \tag{8}$$

where b is a random number between 0 and 1, d_{hj} is the j-th gene of the h-th chromosome and d_{lj} is the j-th gene of the l-th chromosome after hybridization.

The crossover operation can ensure that the excellent genes of each evolution are retained, but it is only a selection of the original result set, and the calculation result is closer to the local optimal solution, and cannot reach the global optimal solution. To solve this problem, a mutation operation is introduced as follows.

$$C = \begin{cases} \dfrac{k_1(f_{\max} - f)}{f_{\max} - f_a} & f \geq f_a \\ k_2 & f < f_a \end{cases} \tag{9}$$

$f_{\max}$ is the maximum value of the fitness of the population, f_a is the average of the fitness of the population, f is the fitness of the individual, k_1 and k_2 are random numbers between 0 and 1, and C is the mutation operator.

The LM algorithm combined the advantages of the Gauss-Newton and gradient descent methods (Wilamowski and Yu, 2010). The ω_k factor is added to the Gauss-Newton method. ω_k is equivalent to the gradient descent method at a high value and equates to the Gauss-Newton method at a low value.

$$W(k+1) = W(k) - [G]^{-1}J^T(W_k)e(W_k) \tag{10}$$

$$G = J^T(W_k)J(W_k) + \omega_k I = H + \omega_k I$$

where I is the identity matrix, while the proportional coefficient ω_k represents a tiny parameter greater than zero. The $\omega_k I$ term guarantees the reversibility of G, otherwise, J^TJ may be irreversible and cannot be calculated. The LM algorithm can continuously adjust the model according to the ω_k parameter changes.

The gradient of $E(W)$ is expressed by:

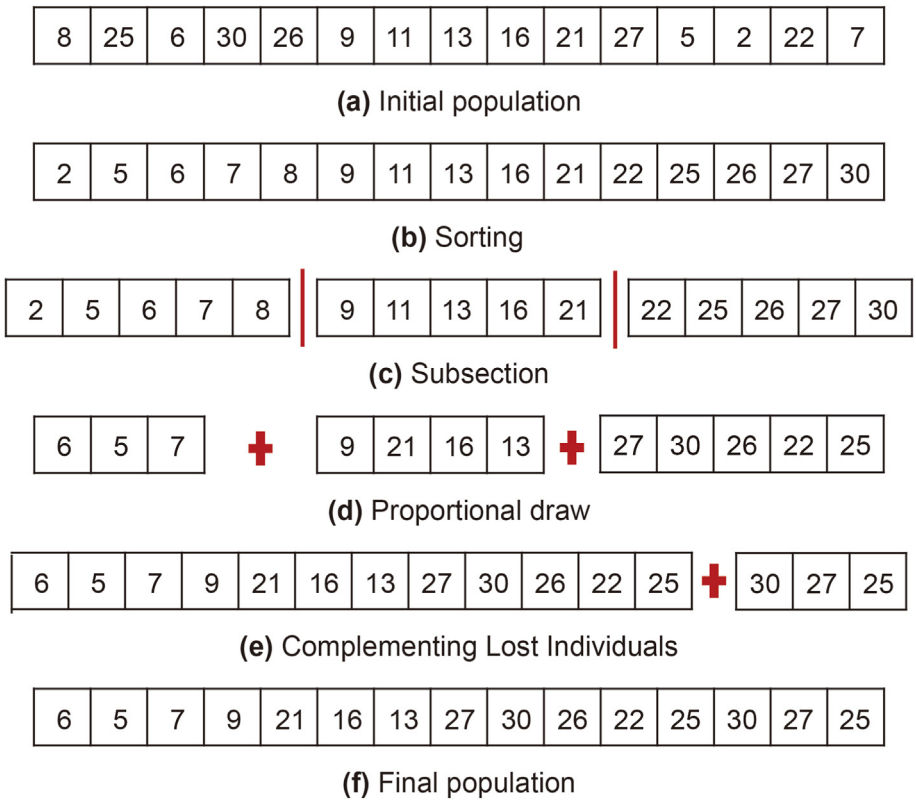


Fig. 1. Demonstration of selection.

$$E(W)=\sum_{i=1}^{n_L} e_i(W)\frac{\partial e_i(W)}{\partial W}=J^T(W)e(W) \tag{11}$$

where $J^T(W)$ is known as the Jacobi matrix.

$$E^2(W)_{kj}=J^T(W)\sum_{i=1}^{n_L}\left(\frac{\partial e_i^2(W)}{\partial W_j^2}\partial e_i(W)+\frac{\partial e_i(W)}{\partial W_k}\frac{\partial e_i(W)}{\partial W_k}\right)=J^T(W)J(W)+S(W) \tag{12}$$

where $S(W)=e(W)e^2(W)$ is difficult to calculate, while the LM algorithm disregards it. The LM algorithm displays a second-order convergence rate and requires a small number of iterations. Therefore, the convergence speed and stability of the algorithm are significantly improved, meanwhile, the local minimum value can be avoided.

The flowchart of the GA-LM model is shown in Fig. 2. After initializing the network, the samples are initially trained with the GA algorithm to obtain the optimized initial weight and threshold. If the requirements are satisfied at the end of the cycle, the training is terminated. Otherwise, the LM algorithm is used to re-correct the network weights and thresholds while the training is repeated until the results meet the requirements.

The algorithmic structure of the GA-LM model can be expressed in Table 1.

2.2. Identification process

Leakage causes local energy loss and produces negative pressure waves propagating to both sides of the pipeline. Therefore, sensors positioned on both sides of the pipe can capture the negative pressure wave signal. The location of the leak can be determined by the time taken to detect it, the velocity of the negative pressure wave, and the length of the pipe. The structure of leak detection is shown in Fig. 3.

The pipeline leakage signal identification process is shown in Fig. 4. Firstly, the raw training and testing data are derived by wavelet threshold method. Secondly, feature extraction of the data is performed, and the extracted feature parameters are normalized to the feature vector matrix of the model. Thirdly, a feature vector matrix of training data is used to train the model and a data-driven classifier is obtained to identify pipeline leakage. Finally, the feature vector matrix of testing sample is input to the classifier for leakage signal identification. The classifier outputs the final classification result.

3. Leak identification based on the GA-LM classification model

3.1. Experiment platform and data acquisition

Because the leakage of water and oil pipelines belongs to pressure pipeline leakage, to reduce the experimental cost, some scholars have simulated the leakage of oil pipelines through the leakage of water pipelines. To complete the testing and verification of the leakage algorithm, Liu et al. have simulated the leakage of oil pipelines through leakage experiments of water pipelines (Liu et al., 2019). The experiment platform is shown in Fig. 5. The platform can also be used to test three flow experiments (Ruiz-Cárcel et al., 2016). Equipment used in the platform includes dynamic pressure transducers (DPT), flow valves, flowmeters, pressure gauges, and pumps. The test loop is 200 m and the internal diameter of the test

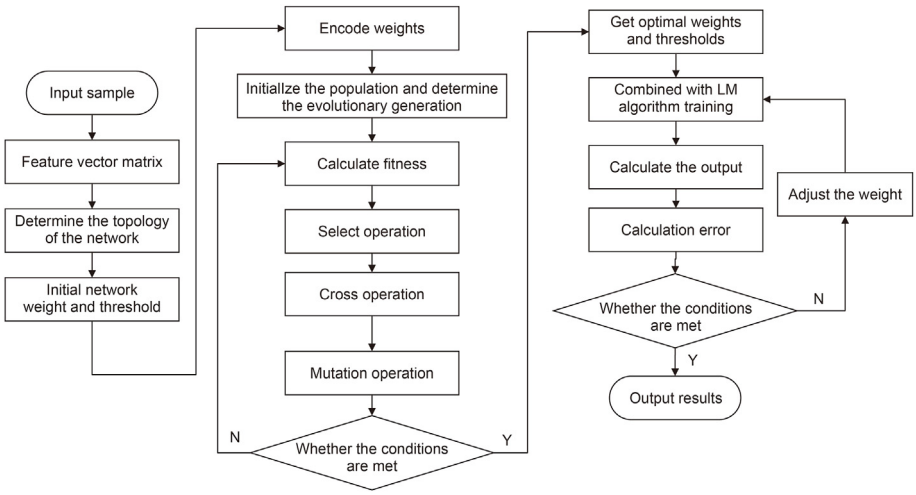


Fig. 2. Flow chart of the GA-LM model.

Table 1
GA-LM model for leak risk assessment.

Input: Training set D (input, output).
Output: leak accuracy.

Initializing parameters of weights and thresholds.
for i = 1: population size(x)
individuals (i) = **Function** Code (x, bound)
x = individuals (i)
individuals fitness(i) = **Function** error (x, net, input, output)
end for
[best fitness and best index] = min(individuals fitness)
best = individuals (best index)
average fitness = sum(individuals fitness)/x
trace = [average fitness, best fitness]
for num = Maximum iteration(k)
individuals = **Function** select (individuals, x)
average fitness = sum (individuals fitness)/x
individuals = **Function** Cross (position, length, individuals, x, bound)
individuals = **Function** Mutation (position, length, individuals, x, num, k, bound)
for j = 1:x
x = individuals (j)
individuals fitness(j) = **Function** error (x, net, input, output)
end for
[new best fitness, new best index] = min(individuals fitness)
[max fitness, max index] = max(individuals fitness)
if best fitness > new best fitness
best fitness = new best fitness
best = individuals (new best index)
end for
individuals (max index) = best
individuals fitness (max index) = best fitness
average fitness = sum (individuals fitness)/x
trace = [trace; average fitness; best fitness]
end for
reshape parameters of weights and thresholds
accuracy = **Function** sim (net, input);

section is 108.2 mm. A 12.7 mm (1/2 in.) needle valve is used to simulate leakage. The leakage is simulated by switching on the needle valve. Two sensors are arranged upstream and downstream of the pipeline respectively. The leak location is 50 m away from the upstream sensor and 30 m away from the downstream sensor. The product manufactured by HBM has a pressure range of 0.01–10 MPa, an accuracy class of 0.2%, and an output voltage range of 0–5 V. A data acquisition (DAQ) NI USB6009 is selected to collect the output of DPT, and the sampling rate is set to 1000 Hz. Data is recorded and saved by LabVIEW software.

The data comes from experiments and fields. Among them, part of the data under normal operation comes from the scene. Since the actual leakage data is very limited and the cost of simulating the leakage of oil pipelines is very high, the leakage data is obtained by simulating the leakage of water pipelines, and the operating conditions of field pipelines are used for experimental simulation. A total of 320 sets of original data are obtained, the number of positive samples is equal to the number of negative samples, and 80 sets of conditional operating data are also considered positive sample data. In addition, all data are denoised by the moving

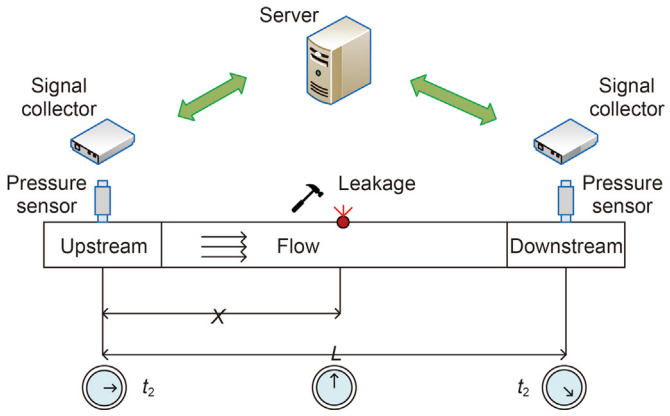


Fig. 3. Structure of leak detection.

average method.

3.2. Feature extraction

Since the leaked signal is a stationary random signal, the power spectrum is used in this paper to describe the frequency-domain characteristics of the signal. The signal after time-domain and frequency-domain processing is shown in Fig. 6. From Fig. 6, we can see that the leak signal and the normal signal are different in the time-domain and frequency-domain. Therefore, the characteristics of the leakage signal can be well described by extracting the parameters in the time-domain and frequency-domain.

The process of feature extraction is shown in Fig. 7. Firstly, wavelet thresholding is performed on the original signal. Secondly, appropriate samples are selected to extract time-domain features and frequency-domain features of the data. Then the extracted features are combined as the input vectors of the model.

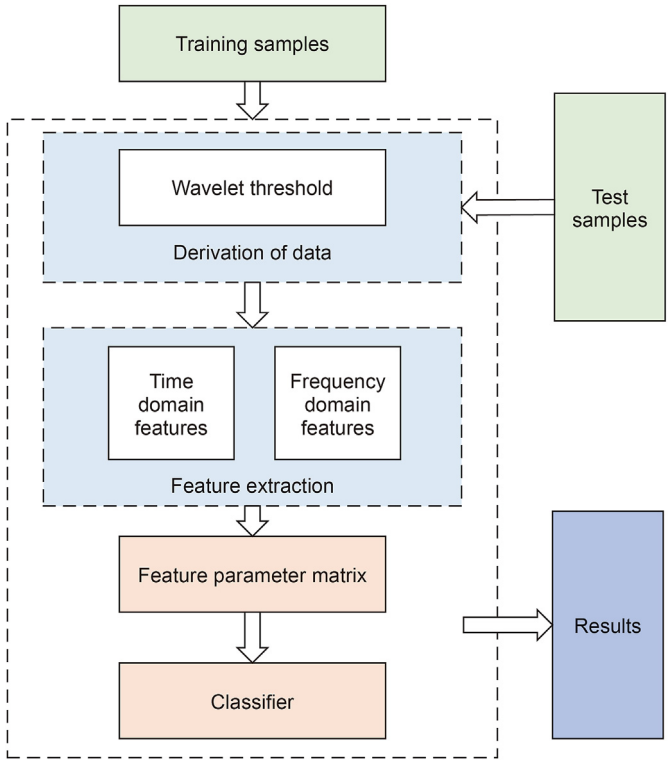


Fig. 4. Identification flow chart.

The signal captured upstream and downstream is shown in Fig. 8. Due to process loss, the upstream pressure is significantly greater than the downstream pressure. When the leakage valve is opened, there will be a certain pressure fluctuation. Therefore, leak detection can be converted into vibration signal identification. By collecting vibration signals from upstream and downstream, a data-driven leak identification model is proposed in this paper.

Signals under different operating conditions are shown in Fig. 9. Different operating conditions can also cause fluctuation of pressure, which is also one of the reasons for the high false alarm rate of leakage. Therefore, an efficient and accurate leakage prediction model is crucial. The length of selected samples will directly affect the identification effect of leakage. Short samples will decrease the accuracy of identification, while long samples will increase the time cost. The identification results of different sample lengths have been discussed in Section 4.1.1.

3.2.1. Derivation of data

An accurate prediction model requires sufficient data, therefore, a wavelet threshold method is used to increase the sample size. The evaluation of derived data is discussed in Section 4.2.2. The wavelet transform method is widely used in various fields due to its simplicity, rapid calculations, and excellent denoising ability (Xu et al., 2021). Wavelet transform uses the Mallat algorithm for rapid signal decomposition and reconstruction. Here, the Daubechies-3(db3) wavelet is selected as the wavelet function. The subsequently introduced smoothing error can be easily disregarded, rendering the signal smoother during the final reconstruction.

Firstly, after the wavelet basis function is shifted t, it does an inner product with the signal $x(t)$ at different scales a.

$$WT_x(a,t)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty}x(t)\psi(\tau-t)dt \tag{13}$$

The selection of the threshold requires the estimation of noise variance σ, as shown in Eq. (14).

$$\sigma=\frac{median(w_j(k))}{0.6745} \tag{14}$$

Where $w_j(k)$ is the coefficient of *the* j-th layer wavelet. If the signal length is L, the threshold λ is expressed by:

$$\lambda=\sigma^2\log_{10}L \tag{15}$$

To combine the advantages of hard and soft threshold functions in denoising, an improved threshold function is selected to process the wavelet coefficients. The improved threshold function is expressed by:

$$w_y=\begin{cases}sign(w_x)\left(|w_x|-\left(1-\dfrac{1-e^{-w_x}}{1+e^{-w_x}}\right)\lambda\right) & |w_x|\geq\lambda\\ 0 & |w_x|<\lambda\end{cases} \tag{16}$$

where w_y and w_x are the wavelet coefficients before and after signal processing respectively, and λ is the threshold value. To evaluate the similarity between the original sample and the generated sample, the Pearson correlation coefficient (PCC) is introduced as the index of similarity evaluation. It is shown in Eq. (17).

$$\rho(X,Y)=\frac{E\left[(X-\bar{x})(Y-\bar{y})\right]}{\sqrt{\sum_{i=1}^{L}(x_i-\bar{x})^2\sum_{i=1}^{L}(y_i-\bar{y})^2}} \tag{17}$$

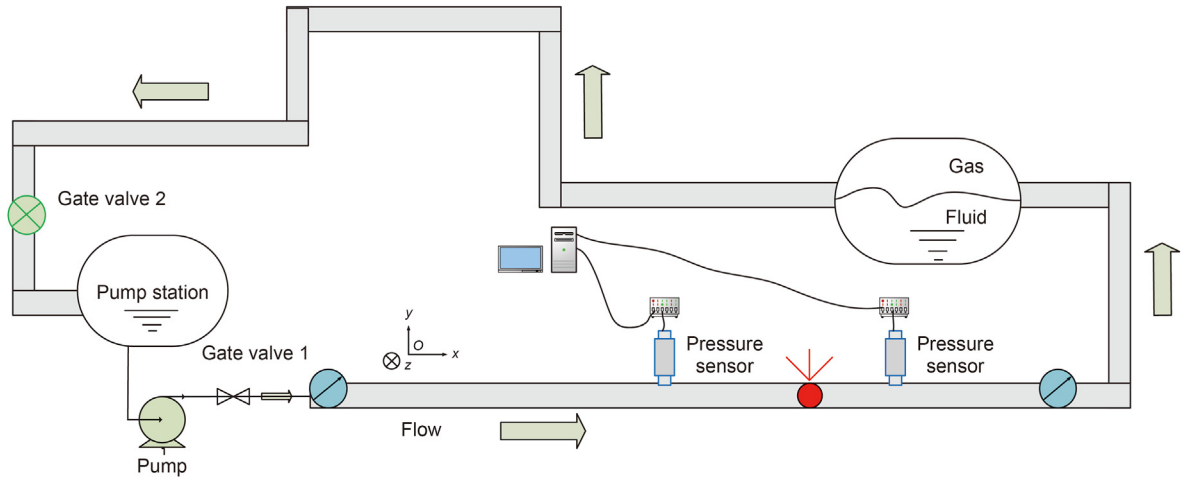


Fig. 5. Experiment platform.

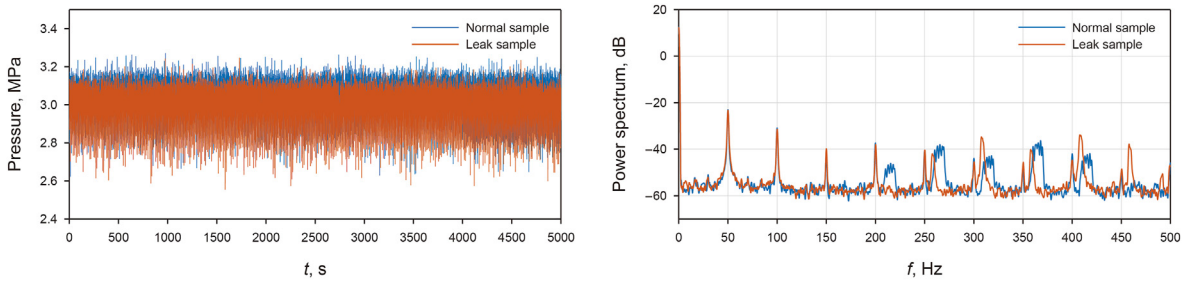


Fig. 6. Signal processing.

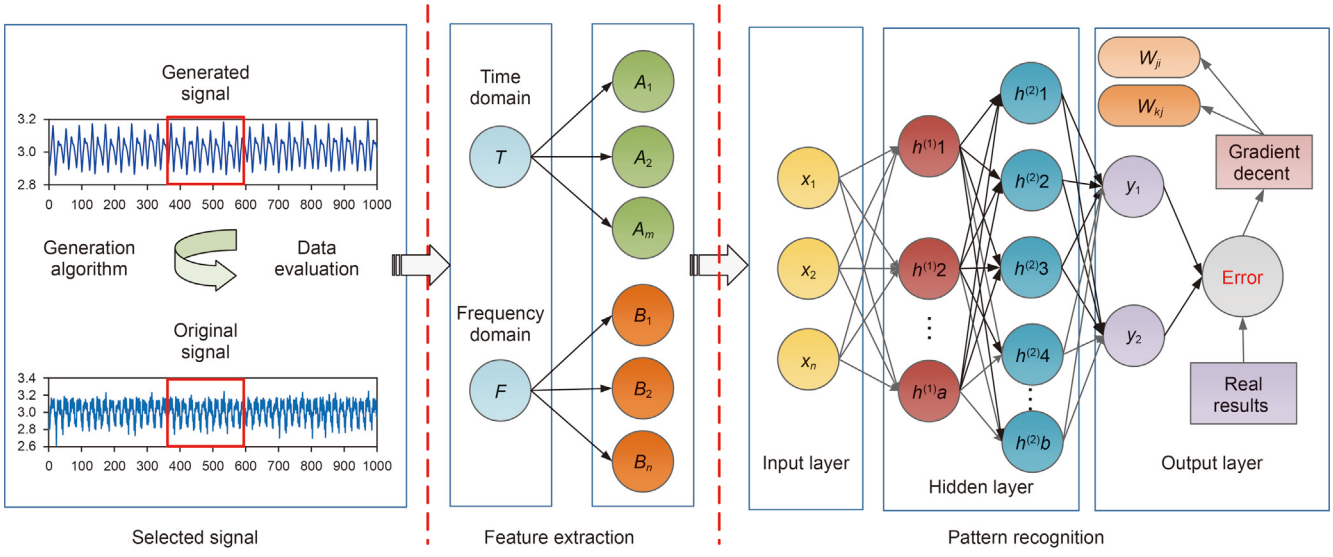


Fig. 7. Process of feature extraction.

Usually, the correlation intensity between variables is judged by the range of PCC. The larger the PCC value, the greater the correlation between samples. As shown in Fig. 10, the values of PCC are mainly distributed between 0.6–0.8, which proves that the generated data is strongly correlated with the original data. Set a lower limit of 0.6 so that generated data with high correlation is retained.

As shown in Fig. 11, the original data collected under different working conditions is compared with the generated data, and the original data can be well simulated by the wavelet threshold method. In addition, in Section 4.2.2, the GA-LM model will be used to evaluate the reliability of the generated data.

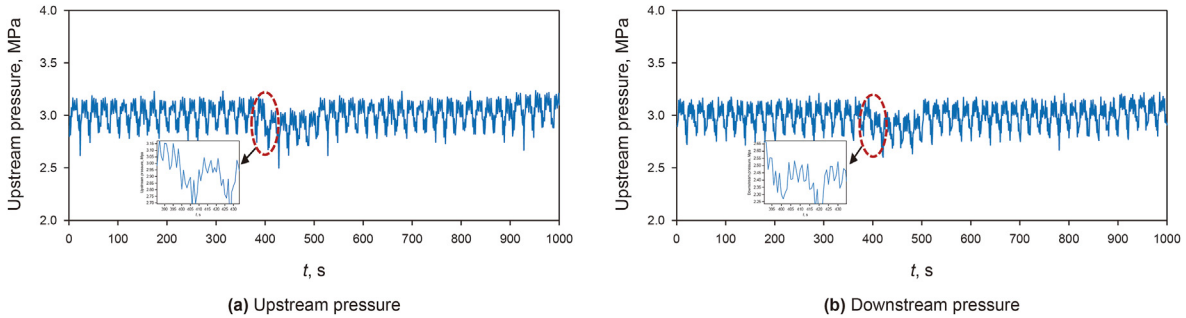


Fig. 8. The signal captured upstream and downstream.

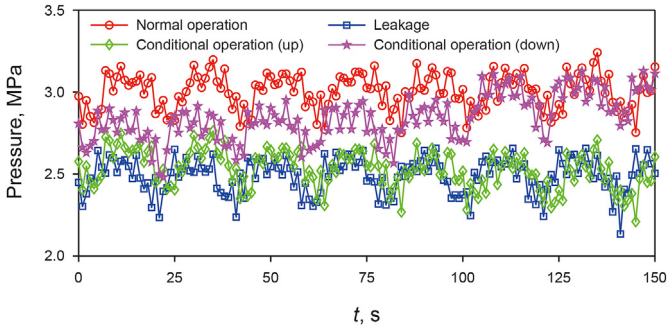


Fig. 9. Signals under different operating conditions.

3.2.2. Time-domain

The waveform of the signal in the time-domain can intuitively analyze some of the leakage signal characteristics. The central tendency of the signal is described by the Mean and Mean Square which can detect the energy of the vibration signal when leakage occurs. Variance represents the dynamic component of signal energy and reflects the degree of dispersion between the leaked data, displaying better model prediction and experimental data description accuracy. Furthermore, the Effective Value described the energy of the vibration signal, while the Shape Factor is used to delineate the shape characteristics. The Crest Factor represented the extreme degree of the peak in the waveform, while the Kurtosis Factor indicates the smoothness of the leakage waveform and is used to describe the distribution of variables.

Seven characteristic time-domain parameters are selected to describe the leakage signal changes. The extracted time-domain feature parameters and expressions are shown in Eqs. (18)–(24). Of these, $x(n)$ represents the signal time-domain sequence, $n = 1, 2, \ldots, N$, while N denotes the number of sample points.

Mean $A1$ is expressed by:

$$A1 = \frac{1}{N}\sum_{n=1}^{N} x(n) \tag{18}$$

Mean Square $A2$ is expressed by:

$$A2 = \frac{1}{N}\sum_{n=1}^{N} |x(n)|^2 \tag{19}$$

Variance $A3$ is expressed by:

$$A3 = \frac{1}{N-1}\sum_{n=1}^{N}\left[x(n) - \frac{1}{N}\sum_{n=1}^{N} x(n)\right]^2 \tag{20}$$

Effective Value $A4$ is expressed by:

$$A4 = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x(n)^2} \tag{21}$$

Shape Factor $A5$ is expressed by:

$$A5 = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N} x(n)^2}}{\frac{1}{N}\sum_{n=1}^{N} x(n)} \tag{22}$$

Crest Factor $A6$ is expressed by:

$$A6 = \frac{\max|x(n)|}{\sqrt{\frac{1}{N}\sum_{n=1}^{N} x(n)^2}} \tag{23}$$

Kurtosis Factor $A7$ is expressed by:

$$A7 = \frac{\sum_{n=1}^{N}\left[x(n) - \frac{1}{N}\sum_{n=1}^{N} x(n)\right]^4}{\frac{1}{N-1}\sum_{n=1}^{N}\left[x(n) - \frac{1}{N}\sum_{n=1}^{N} x(n)\right]^4} \tag{24}$$

3.2.3. Frequency-domain

The power spectrum is commonly used to analyze signal characteristics. The signal energy distribution is distinguishable via frequency-domain analysis. The characteristic statistical parameters of the frequency-domain are introduced here to describe its leakage signal properties. The Spectral Mean and Spectrum Root Mean Square describe its fluctuation, while the Center of Gravity Frequency represents the entire frequency band. The leakage signal frequency distributions are described by the Mean Square Frequency and Root Mean Square Frequency.

The specific characteristic frequency-domain parameters and their formulas are shown in Eqs. (25)–(29). $s(k)$ is the power spectrum of the time signal $x(n)$, $k = 1,2, \ldots, K$, K is the number of

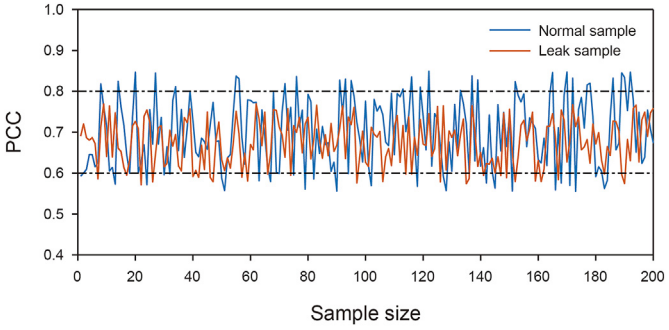


Fig. 10. PCC of different sample sizes.

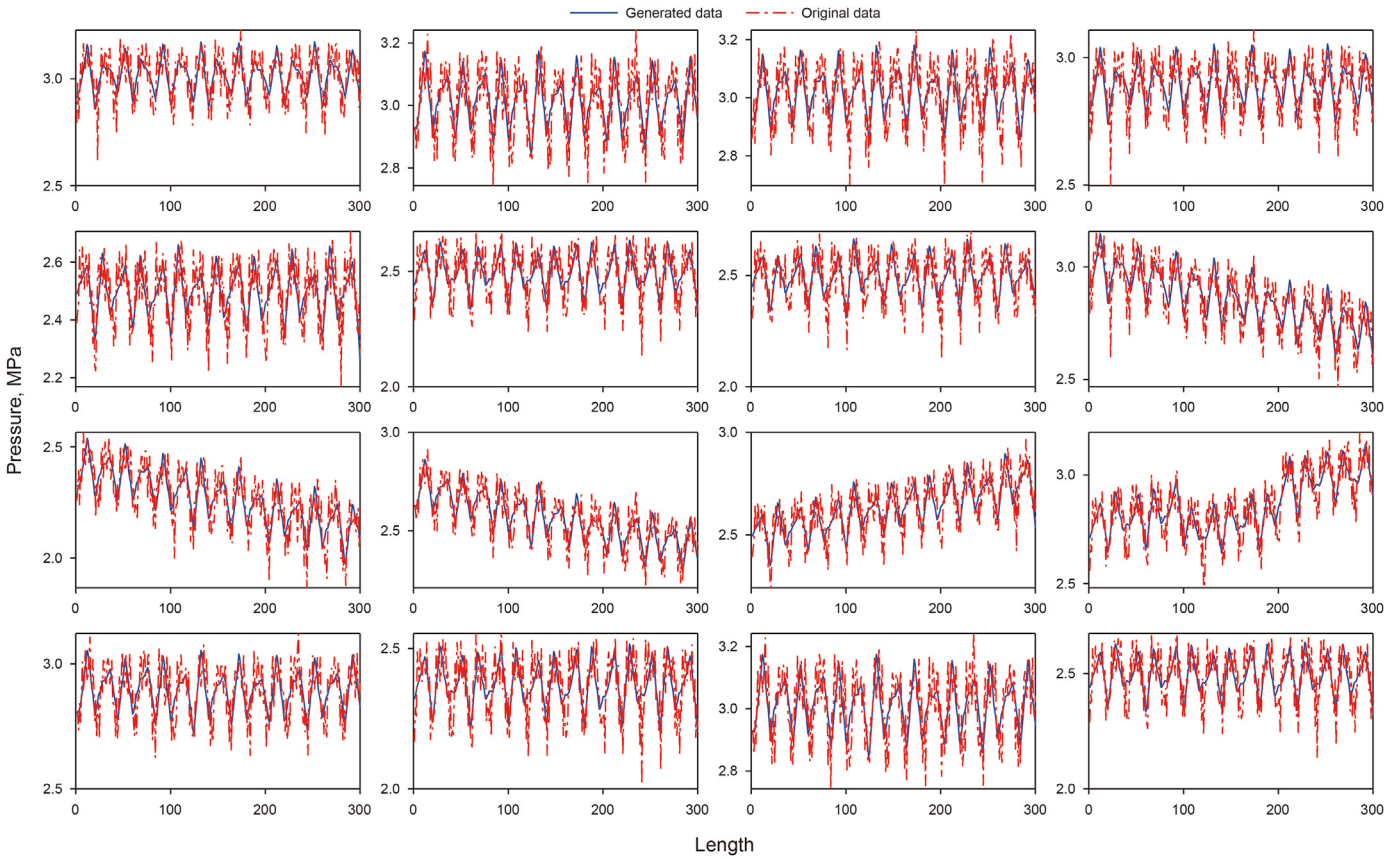


Fig. 11. Original data and generated data.

spectral lines, and f_k is the frequency value of the kth spectral line.

Spectral Mean $B1$ is expressed by:

$$B1 = \frac{1}{K}\sum_{k=1}^{K} s(k) \tag{25}$$

Spectrum Root Mean Square $B2$ is expressed by:

$$B2 = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left[s(k) - \frac{1}{K}\sum_{k=1}^{K} s(k)\right]^2} \tag{26}$$

Center of Gravity Frequency $B3$ is expressed by:

$$B3 = \frac{\sum_{k=1}^{K} f_k \cdot s(k)}{\sum_{k=1}^{K} s(k)} \tag{27}$$

Mean Square Frequency $B4$ is expressed by:

$$B4 = \frac{\sum_{k=1}^{K} {f_k}^2 \cdot s(k)}{\sum_{k=1}^{K} s(k)} \tag{28}$$

Root Mean Square Frequency $B5$ is expressed by:

$$B5 = \sqrt{\frac{\sum_{k=1}^{K} {f_k}^2 \cdot s(k)}{\sum_{k=1}^{K} s(k)}} \tag{29}$$

3.3. Model structure

The extracted feature parameters included seven time-domain and five frequency-domain feature parameters. Therefore, the neural network input layer contained twelve nodes, and the output layer contained two, while the leakage situation is expressed by [1 0], and the no-leakage situation is expressed by [0 1].

Furthermore, to reduce the training number and improve training efficiency, it is necessary to normalize the input vector matrix, as shown in Eq. (30).

$$x_i^* = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{30}$$

where x_i* is the normalized value of the i-th feature, x_i is the value of the i-th feature before normalization, $x_{\max}$ is the maximum value of the i-th feature before normalization, and $x_{\min}$ is the minimum value of the i-th feature before normalization.

The empirical Eq. (31) indicated that the optimal number of nodes in the hidden layer is 4–13, while the number of hidden layer neurons is initially set to 8.

$$h = \sqrt{m+n} + a \tag{31}$$

where h is the number of nodes in the hidden layer, m is the number of nodes in the input layer, n is the number of nodes in the output layer, and a is a constant from 1 to 10.

During classification and recognition, the activation function generally used the non-linear “logsig” and “tansig”. The learning rate η of the BP neural network is between [0,1], affecting its learning speed. A lower learning rate is generally selected to ensure the stability of the system. The initial learning rate is set to 0.04, while the other parameters are set as follows: the maximum number of iterations is $k = 1000$, while the required training

precision is $e = 0.001$. The population size is 20, the crossover rate is 0.7, the mutation rate is 0.05, and the maximum number of evolutions is 50. The GA-LM model is constructed based on MATLAB software. After determining the network structure, 400 sets of cases are used for network training, while 240 sets are employed as network tests.

4. Results and discussion

4.1. Parameter sensitivity analysis

The parameter of the model is analyzed to improve the leak identification reliability. The parameters analyzed mainly include the length of the sample, input matrices, the number of hidden layer nodes, and activation functions. The optimal structural parameters are selected to build the model.

4.1.1. Sample length

The leakage detection of pipelines in actual operation needs high reliability and low time cost, too low accuracy and long detection time are meaningless. The leakage signal will continue after the leakage occurs. In principle, increasing the sample length will increase the accuracy and time cost of leak identification. Therefore, the selection of an appropriate sample length is very important for leak detection. Prediction accuracy (P) and response time (t) under different sample lengths are shown in Table 2. The minimum sample length is 40. Considering accuracy and time cost, a sample length of 300 is selected.

4.1.2. Activation function

Now, the influence of the activation function is explored by unchanged the other parameters of the previously constructed model. The activation function between the input and hidden layers is set to “logsig” and to “tansig” between the hidden and output layers, which is recorded as a combination of [log-tan]. In addition, the function “SoftMax” is often used for classification tasks. Ultimately, there are six combinations, namely [log-log], [log-tan], [tan-log], [tan-tan], [log-sof] and [tan-sof]. The recognition rate of the network model constructed using different combinations is shown in Fig. 12. Of these, the [tan-log] combination displayed the highest recognition rate. Therefore, the activation function between the input and hidden layers is set to “tansig” and “logsig” between the hidden and output layers.

4.1.3. Input matrix

The other parameters of the previously constructed network are kept unchanged to explore the influence of the different input matrices. A single feature parameter is used as the model input. The model recognition rate of each feature parameter input is shown separately in Fig. 13. The recognition rate of each parameter

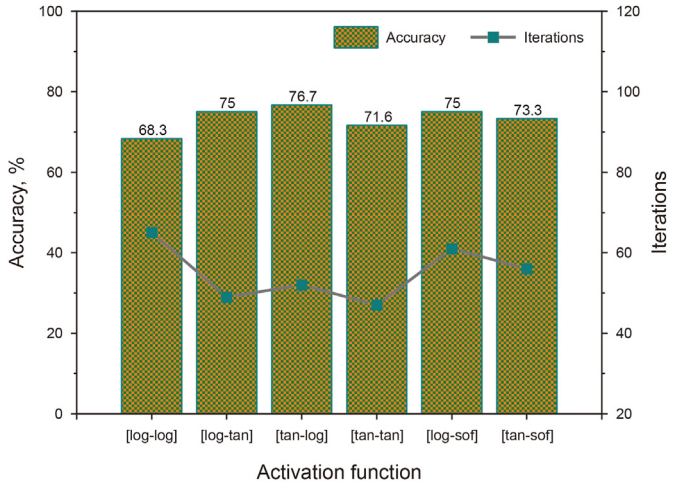


Fig. 12. The prediction results of different activation function combinations.

exceeded 50%, verifying the efficacy of the time-frequency feature extraction method.

The parameters with a recognition rate of 50% or above include $A1$, $A2$, $A3$, $A4$, $A5$, $A6$, $A7$, $B1$, $B2$, $B3$, $B4$, and $B5$, which is named $T1 = [A1, A2, A3, A4, A5, A6, A7, B1, B2, B3, B4, B5]$. The parameters with a recognition rate of above 55% include $A1$, $A2$, $A3$, $A4$, $A5$, $B1$, $B2$, $B3$, $B4$, and $B5$, which is named $T2 = [A1, A2, A3, A4, A5, B1, B2, B3, B4, B5]$. The parameters with a recognition rate of above 60% include $A1$, $A2$, $A3$, $A4$, $A5$, $B1$, and $B4$, which is named $T3 = [A1, A2, A3, A4, A5, B1, B4]$. The parameters with a recognition rate of above 65% include $A1$, $A2$, $A3$, $A4$, $B1$, and $B4$, which is named $T4 = [A1, A2, A3, A4, B1, B4]$. The parameters with a recognition rate of above 70% include $A2$ and $A3$, which is named $T5 = [A2, A3]$.

Different combination parameters ($T1$~$T5$) are used as the input matrix to build the model. The recognition rate of each parameter combination is shown in Fig. 13. The recognition rate of each feature parameter combination is higher than the recognition rate of a single feature parameter. The network model constructed by combining the feature parameters is more reliable. Too many or not enough feature parameters could not effectively represent the leakage signal. Since the overall recognition rate of $T3$ is the highest, this parameter combination is selected as the optimal input for model construction.

4.1.4. Number of hidden layer neurons

The increase in the number of hidden layers leads to a decrease in model performance for the multi-layer neural network, and the recognition ability of the one-hidden-layer model is the best (Zheng et al., 2021). Therefore, the one-hidden-layer model is considered. To explore the influence of the number of neurons in

Table 2
Prediction accuracy and response time under different sample lengths.

Sample length	40		60		80		100		200		300		400	
	P, %	t, s	P, %	t, s	P, %	t, s	P, %	t, s	P, %	t, s	P, %	t, s	P, %	t, s
1	83.3	16.7	86.7	17.9	91.7	18.4	93.3	19.7	95.0	20.1	95.8	23.7	95.8	24.2
2	84.2	15.3	88.3	16.8	91.7	17.5	93.3	18.7	94.2	18.5	95.0	21.1	95.0	25.1
3	83.3	16.4	88.3	17.2	92.5	17.9	94.2	18.1	95.0	19.3	95.8	19.3	95.8	23.2
4	85.8	15.9	87.5	18.3	90.0	18.3	92.5	19.3	94.2	20.5	95.0	19.9	95.0	23.6
5	81.0	15.1	84.2	17.6	89.2	18.8	90.0	18.9	93.3	18.7	94.2	18.9	95.0	24.5
6	82.5	16.2	86.7	17.9	91.7	17.5	94.2	18.7	95.8	19.6	95.8	19.8	95.8	23.9
7	84.2	16.8	85.8	18.5	88.3	17.8	91.7	19.6	94.2	20.7	95.0	20.8	95.0	24.5
8	81.6	15.2	87.5	18.1	90.0	18.9	93.3	17.8	95.0	20.1	95.0	21.4	95.0	25.6
Average value	83.1	16.0	86.9	17.8	90.6	18.1	92.8	18.9	94.6	19.7	95.2	20.6	95.3	24.3

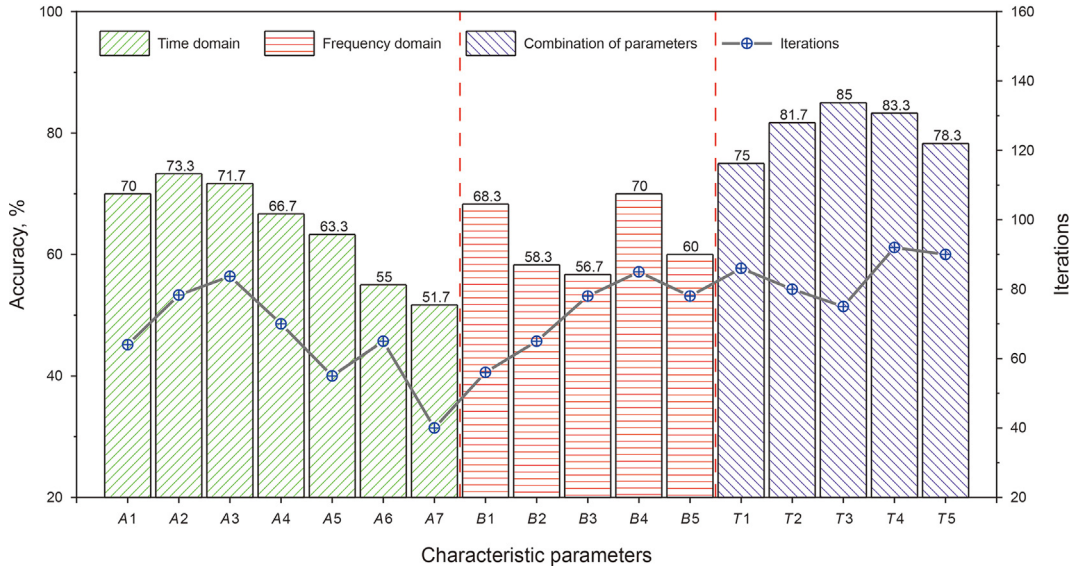


Fig. 13. The prediction results of the characteristic parameter.

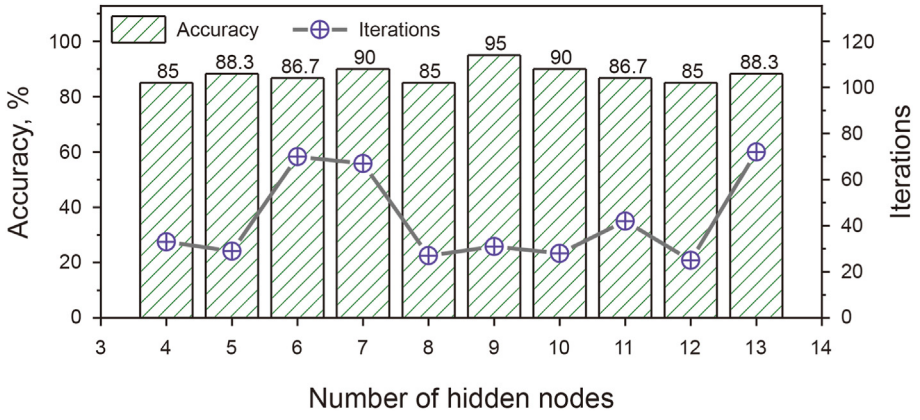
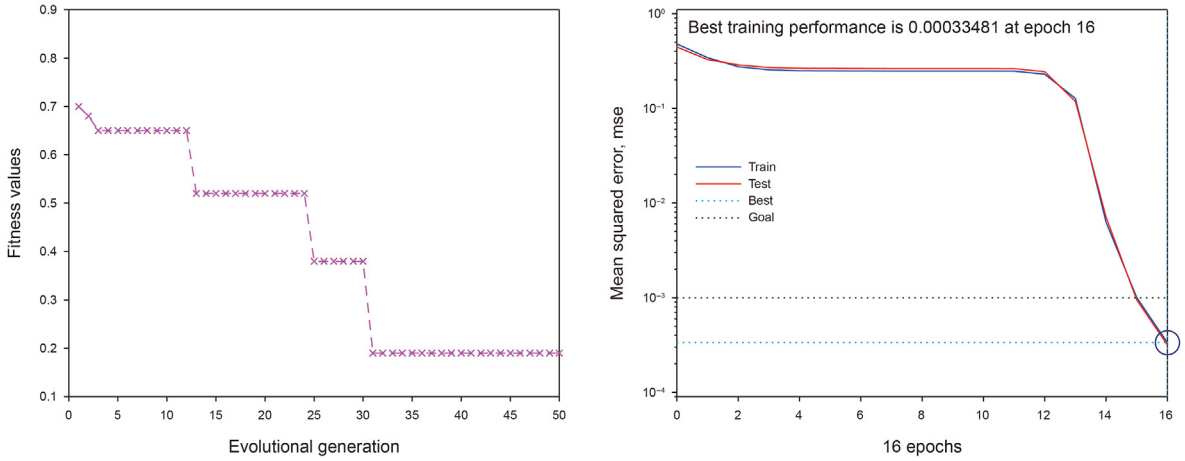


Fig. 14. The prediction results of different neurons in hidden layers.



(a) Network performance tracking curve

(b) Network training error curve

Fig. 15. Network training performance.

the hidden layer on the network, the other parameters of the previously constructed network are kept unchanged. The optimal number of nodes calculated via the empirical formula is 4–13. Therefore, ten network structures are compared to obtain the best network model. The prediction results of the number of different hidden nodes are shown in Fig. 14. The model built with a hidden layer of 9 neurons achieved an overall recognition rate of 95%, which is optimal for model building.

As shown in Fig. 15(a), after 31 evolutions, the performance tracking curve has reached stability, and the network has found better weights and thresholds. After the data in the training set have been trained 16 times, the Mean Square Error (MSE) of the model is 0.000335, as shown in Fig. 15(b). The training and prediction errors of the network are relatively small, and the network does not have overfitting.

4.2. Model evaluation

4.2.1. Evaluation indicators

To represent the performance of models, the model is evaluated based on the evaluation indicators, such as *Accuracy*, *Precision*, *Recall*, *F1score*, MSE, Network training time (NTT), and optimality. The Confusion Matrix is usually used to express the classification results of the model. The Classification Matrix is shown in Table 3. The *Accuracy*, *Precision*, *Recall*, and *F1score* are important indicators to measure the effect of model classification (Liu et al., 2019; Mazumder et al., 2021a).

(1) *Accuracy*: Representing the proportion of correct data judged by the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{32}$$

(2) *Precision*: How many of the positive examples predicted by the model are correct.

$$Precision = \frac{TP}{TP + FP} \tag{33}$$

(3) *Recall*: The ratio of positive examples judged by the model to the total positive examples in the data.

$$Recall = \frac{TP}{TP + FN} \tag{34}$$

(4) *F1score*: Representing the effect of the classification model to recognize positive class.

$$F1score = \frac{2*Precision*Recall}{Precision + Recall} \tag{35}$$

4.2.2. Evaluation of generated data

The Confusion Matrix is also used to assess the reliability of derived data. The green blocks can show the number of correct classifications and their ratios in the total data, respectively. Meanwhile, the pink blocks can present the number of misclassifications and their ratios in the total data, respectively. The correct classification accuracy and false alarm rate can be found in the dark grey blocks. Different training and testing sets are shown in Table 4. The raw data and generated data are used for network training and testing. The number of samples in the training set is 200 and the number of samples in the test set is 120.

Table 3
Confusion matrix.

	Target class 1	Target class 2
Output class 1	True Negative (*TN*)	False Negative (*FN*)
Output class 2	False Positive (*FP*)	True Positive (*TP*)

Training and testing effects in three cases are shown in Fig. 16. As shown in Fig. 16(a), (c) and (e), the training accuracy of the model is 100%, 99% and 98.5% respectively, showing excellent training effect. However, the testing effect of the model should be focused, preventing the model from overfitting. As shown in Fig. 16(b), (d) and (f), the testing accuracy of the model is 95%, 95.8% and 95% respectively, and the false alarm rate of the model is only 5%, 4.2% and 5% respectively. Therefore, there is no overfitting of the model. The ratio of the original data to the generated data has little impact on the network, and the accuracy of all test sets is above 95%, indicating that the generated data is reliable.

4.2.3. Comparison of model

Firstly, a classifier combining wavelet packet decomposition and Support Vector Machine (SVM) is studied (Qu et al., 2010). Secondly, traditional BP models based on the training function "traingd" are compared. In addition, Markov features are used for feature extraction of the data (Liu et al., 2019), which is abbreviated as the 'M-BP' model. Thirdly, the GA-BP model combined with the genetic algorithm and the traditional BP model is compared. Fourthly, a K-Nearest Neighbor (KNN) classifier is studied by selecting the *K* known samples that are closest to the unknown samples for classification (Arian et al., 2020). Fifthly, a Decision Tree (DT) classifier by partitioning the feature space is investigated (Sabah et al., 2019). All test models use the same test data, and the classification results of different models are shown in Fig. 17. The *Accuracy*, *Precision*, *Recall* and *F1score* is 95%, 93.5%, 96.7% and 95.1%, respectively. The GA-LM model has a good predictive effect for positive and negative samples. By observing the Recall, the GA-LM model has a greater advantage for the prediction of positive samples. The results also indicated that the classification effect of the GA-LM model based on time-frequency features is better than the other models.

The MSE, NTT, and Optimality are important indicators for evaluating model performance. The results of different models are shown in Table 5. Comparing GA-BP and M-BP the genetic algorithm makes the network performance more excellent. Comparing GA-LM and GA-BP the LM algorithm increases the convergence speed of the network and obtains the global optimal solution. The GA-LM model exhibited a small classification error. The MSE showed that the performance of the GA-LM model is relatively high in medium-sample prediction. The NTT is relatively short. Compared with the traditional BP and GA-BP model, the optimized model significantly reduced the NTT. It greatly improves the efficiency of leak identification. The traditional BP algorithm easily obtained the local optimal solution, while the GA-LM model employed the LM algorithm to acquire the global optimal solution.

In addition, the data collected randomly during different periods are used to verify the adaptability of the model. As shown in Table 6, 160 raw data are collected in the field and 320 data are generated by the derivative algorithm. The types of raw data include normal running samples, leak running samples, and conditional running samples.

Firstly, the performance of the leak detection method is tested using 160 raw data. To ensure the fairness of the test results, 100 test samples are randomly selected from 160 raw data. In particular,

Table 4
Different training and test sets.

	Training		Testing	
	Raw data number	Generated data number	Raw data number	Generated data number
Case 1	150	50	40	80
Case 2	50	150	80	40
Case 3	100	100	60	60

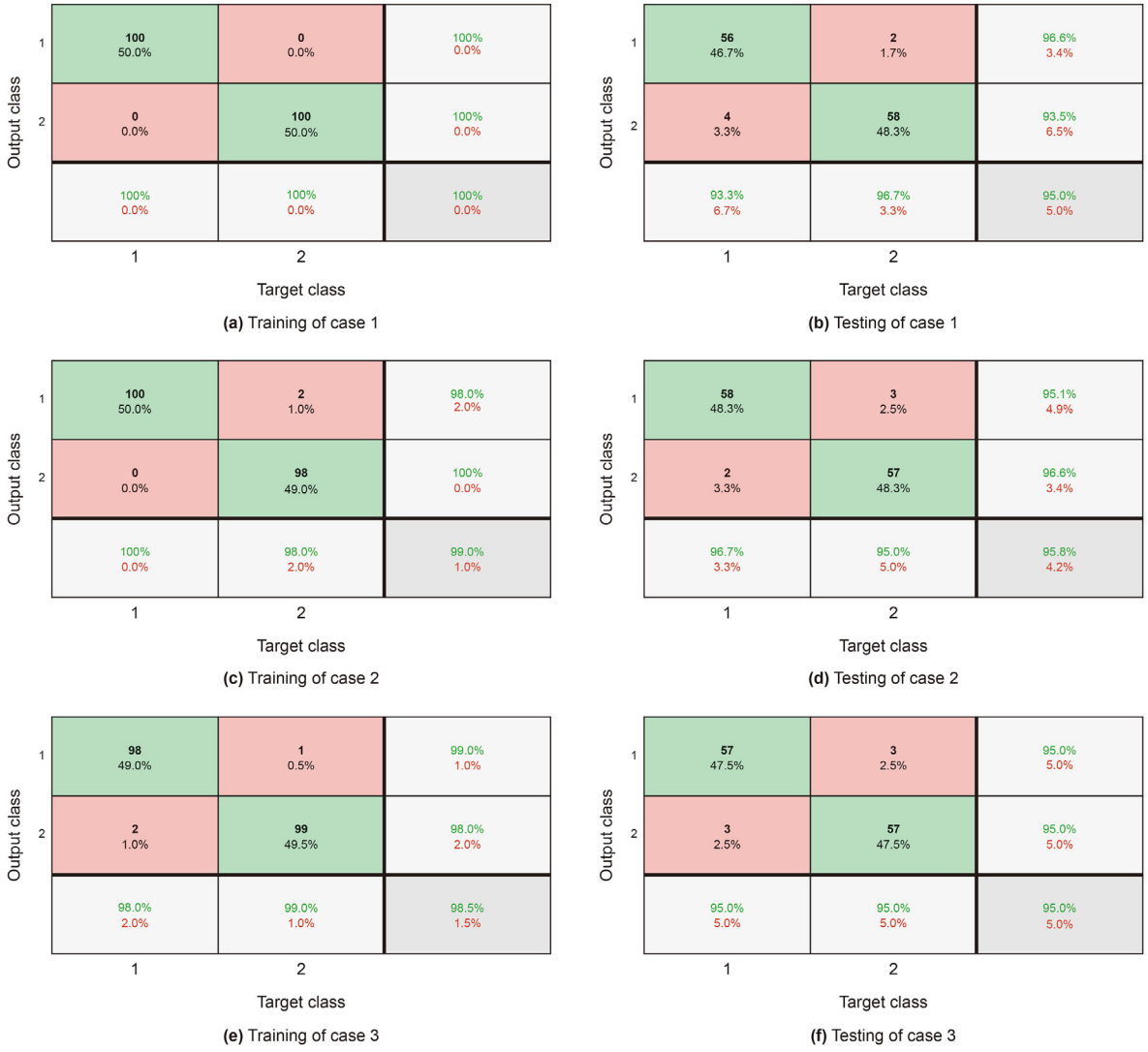


Fig. 16. Training and testing effects in different cases.

each of the eight tests is completed independently. The results are shown in Table 7. In the eight tests, the maximum detection error is 4% for all 100 samples. It shows that the model has a good prediction effect for raw data collected during different periods.

Secondly, the performance of the leak detection method is tested using 320 generated data. To ensure the fairness of the test results, 100 test samples are randomly selected from 320 generated data. In particular, each of the eight tests is completed independently. The results are shown in Table 8. In the eight tests, the lowest recognition rate of the model is 96%. It shows that the generated data can replace the raw data as a part of the sample database.

Finally, 100 raw data and 100 generated data are randomly selected each time to test the performance of different models. The recognition results of six models are shown in Fig. 18. The recognition effect of the unoptimized BP neural network is the worst, and the average accuracy is below 80%. The GA-LM model shows excellent stability and the average accuracy of the GA-LM model

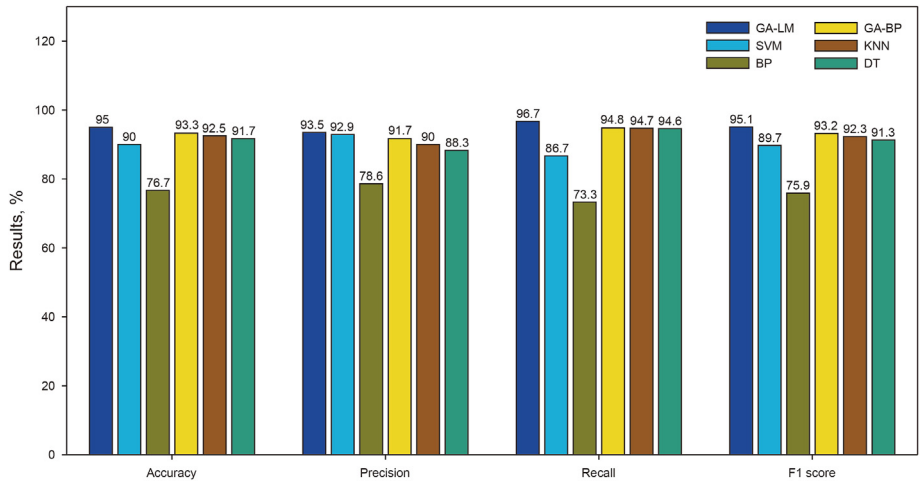


Fig. 17. Results of different models.

Table 5
Evaluation indicators of different models.

	SVM	M-BP	GA-BP	GA-LM
MSE	2.5%	47.5%	3.05%	0.033%
NTT	6.1s	65.5s	54.2s	5.4s
Optimality	Global	Local	Local	Global

Table 6
Data collected during different periods.

Sample Type	Raw data	Generated data	Length	Acquisition time
Normal running samples	40	80	300	10:05:24
Leak running samples	80	160	300	11:32:53
Conditional running samples	40	80	300	14:20:41

Table 7
Experiments test under raw data.

Test number	1	2	3	4	5	6	7	8
False alarm times	3	4	2	3	2	4	3	4
False alarm rate	3%	4%	2%	3%	2%	4%	3%	4%

Table 8
Experiments test under generated data.

Test number	1	2	3	4	5	6	7	8
True alarm times	97	96	97	99	96	96	98	97
True alarm rate	97%	96%	97%	99%	96%	97%	98%	97%

can be reached to 96%, which can greatly improve the recognition accuracy of the traditional BP network optimized by the proposed optimization algorithm. The recognition effect of the GA-LM model exceed that of the other models. It can be seen that the GA-LM model based on time-frequency features has excellent performance for actual pipeline leak detection.

5. Conclusions

This paper considers the actual oil pipeline leakage problem as the starting point to examine the characterization method for pipeline leakage signals. In addition, the BP network is optimized by the GA and LM. A classification model for pipeline leakage prediction is constructed with the main conclusions as follows:

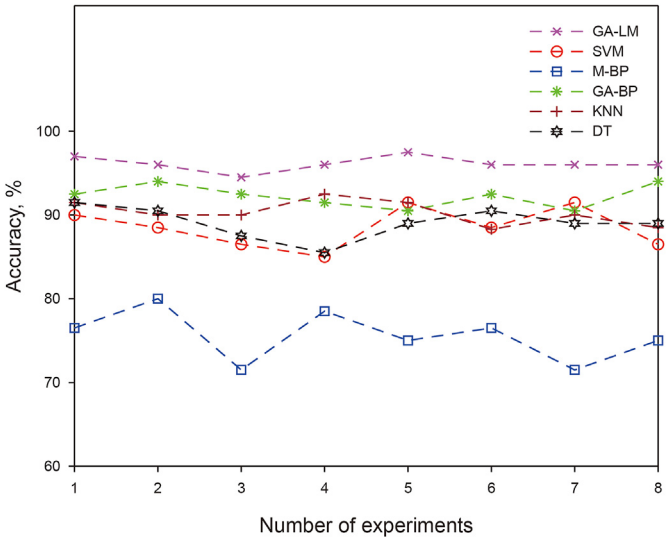


Fig. 18. Accuracy of different models.

(1) Twelve feature parameters are extracted by the time-frequency feature method to characterize the leakage signal. The recognition rate of each parameter exceeds 50%, verifying the efficacy of the time-frequency feature extraction method. The combined feature parameters are superior to single feature parameters for classification. Finally, seven parameters, namely the Mean, Mean Square, Variance, Effective Value, Shape factor, Spectral Mean, and Mean Square Frequency, are selected as the optimal input matrix of the model.

(2) The traditional BP neural network is optimized by combining the GA and LM algorithms, while a GA-LM classification model is constructed for oil pipeline leakage detection. The *Accuracy*, *Precision*, *Recall* and *F1score* is 95%, 93.5%, 96.7% and 95.1%, respectively. The average Accuracy of the GA-LM model reached 96%, showing high robustness. The recognition effect of the GA-LM model exceeded that of the other models. Compared with the traditional BP model, the GA-LM

model significantly reduced the NTT. It greatly improves the efficiency of leak identification.

(3) Considering that a large number of samples are required for model training, a wavelet threshold method is proposed to generate sample data with higher reliability. The ratio of the original data to the generated data has little impact on the network, and the accuracy of all test sets is above 95%, indicating that the generated data is reliable.

In this paper, due to the limited experimental conditions, a large amount of operating condition data cannot be obtained to establish a reliable model, and the operating condition recognition of pipeline is not conducted. In the future, the sample database should be enriched based on different operating condition and the operating condition recognition should be focused.

Declaration of competing interest

The authors declare that we have no conflict of interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the support provided by: The National Key Research and Development Program of China: Design and Key Technology Research of Non-metallic Flexible Risers for Deep Sea Mining (2022YFC2803701) and The General Program of National Natural Science Foundation of China (52071336, 52374022).

References

Aamo, O.M., 2016. Leak detection, size estimation and localization in pipe flows. IEEE Trans. Automat. Control 61 (1), 246–251. https://doi.org/10.1109/tac.2015.2434031.

Arian, R., Hariri, A., Mehridehnavi, A., Fassihi, A., Ghasemi, F., 2020. Protein kinase inhibitors' classification using K-Nearest neighbor algorithm. Comput. Biol. Chem. 86, 107269. https://doi.org/10.1016/j.compbiolchem.2020.107269.

Ben-Mansour, R., Habib, M.A., Khalifa, A., Youcef-Toumi, k, Chatzigeorgiou, D., 2012. Computational fluid dynamic simulation of small leaks in water pipelines for direct leak pressure transduction. Comput. Fluids 57, 110–123. https://doi.org/10.1016/j.compfluid.2011.12.016.

Chen, Q., Zuo, L.L., Wu, C.C., Bu, Y.R., Lu, Y.F., Hang, Y.F., Chen, F., 2020. Short-term supply reliability assessment of a gas pipeline system under demand variations. Reliab. Eng. Syst. Saf. 202, 107004. https://doi.org/10.1016/j.ress.2020.107004.

Cui, G., Li, Z.L., Yang, C., Wang, M., 2016. The influence of DC stray current on pipeline corrosion. Petrol. Sci. 13 (1), 135–145. https://doi.org/10.1007/s12182-015-0064-3.

Diao, X., Jiang, J.C., Shen, G.D., Chi, Z.Z., Wang, Z.R., Ni, L., Mebarki, A., Bian, H.T., Hao, Y.M., 2020. An improved variational mode decomposition method based on particle swarm optimization for leak detection of liquid pipelines. Mech. Syst. Signal Process. 143, 106787. https://doi.org/10.1016/j.ymssp.2020.106787.

Fu, H., Yang, L., Liang, H.R., Wang, S., Ling, K.G., 2020. Diagnosis of the single leakage in the fluid pipeline through experimental study and CFD simulation. J. Petrol. Sci. Eng. 193, 107437. https://doi.org/10.1016/j.petrol.2020.107437.

Gao, Z.K., Liu, M.X., Dang, W.D., Cai, Q., 2021. A novel complex network-based deep learning method for characterizing gas–liquid two-phase flow. Petrol. Sci. 18 (1), 259–268. https://doi.org/10.1007/s12182-020-00493-3.

Gao, Z., Guo, L.M., Ren, T.W., Liu, A.A., Cheng, Z.Y., Chen, S.Y., 2022. Pairwise two-stream ConvNets for cross-domain action recognition with small data. IEEE Transact. Neural Networks Learn. Syst. 33 (3), 1147–1161. https://doi.org/10.1109/TNNLS.2020.3041018.

Harmouche, J., Narasimhan, S., 2020. Long-term monitoring for leaks in water distribution networks using association rules mining. IEEE Trans. Ind. Inf. 16 (1), 258–266. https://doi.org/10.1109/tii.2019.2911064.

Heravi, A.R., Hodtani, G.A., 2018. A new correntropy-based conjugate gradient backpropagation algorithm for improving training in neural networks. IEEE Transact. Neural Networks Learn. Syst. 29 (12), 6252–6263. https://doi.org/10.1109/TNNLS.2018.2827778.

Hu, J.Q., Zhang, L.B., Liang, W., 2011. Detection of small leakage from long transportation pipeline with complex noise. J. Loss Prev. Process. Ind. 24 (4), 449–457. https://doi.org/10.1016/j.jlp.2011.04.003.

Hu, X.G., Zhang, H.G., Ma, D.Z., Wang, R., 2021. A tnGAN-based leak detection method for pipeline network considering incomplete sensor data. IEEE Trans. Instrum. Meas. 70, 3510610. https://doi.org/10.1109/tim.2020.3045843.

Hu, X.W., Zhou, C.F., Duan, M.L., An, C., 2014. Reliability analysis of marine risers with narrow and long corrosion defects under combined loads. Petrol. Sci. 11 (1), 139–146. https://doi.org/10.1007/s12182-014-0325-6.

Kang, J., Park, Y.J., Lee, J., Wang, S.H., Eom, D.S., 2018. Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems. IEEE Trans. Ind. Electron. 65 (5), 4279–4289. https://doi.org/10.1109/tie.2017.2764861.

Lang, X.M., Yuan, W.Q., 2020. Localization method of multiple leaks based on time-frequency analysis and improved differential evolution. IEEE Sensor. J. 20 (23), 14383–14390. https://doi.org/10.1109/jsen.2020.3009091.

Li, Y., Shuai, J., Jin, Z.L., Zhao, Y.T., Xu, K., 2017. Local buckling failure analysis of high-strength pipelines. Petrol. Sci. 14 (3), 549–559. https://doi.org/10.1007/s12182-017-0172-3.

Liu, C.W., Li, Y.X., Yan, Y.K., Fu, J.T., Zhang, Y.Q., 2015. A new leak location method based on leakage acoustic waves for oil and gas pipelines. J. Loss Prev. Process. Ind. 35, 236–246. https://doi.org/10.1016/j.jlp.2015.05.006.

Liu, J.H., Zang, D., Liu, C., Ma, Y.J., Fu, M.G., 2019. A leak detection method for oil pipeline based on markov feature and two-stage decision scheme. Measurement 138, 433–445. https://doi.org/10.1016/j.measurement.2019.01.029.

Liu, W., 2019. Oil pipeline leak signal image recognition based on improved data field theory. Cluster Comput. 22 (S5), 12949–12957. https://doi.org/10.1007/s10586-018-1816-9.

Lu, H.F., Wu, X.N., Ni, H.M., Azimi, M., Yan, X.C., Niu, Y.Q., 2020a. Stress analysis of urban gas pipeline repaired by inserted hose lining method. Compos. B Eng. 183, 107657. https://doi.org/10.1016/j.compositesb.2019.107657.

Lu, H.F., Iseley, T., Behbahani, S., Fu, L.D., 2020b. Leakage detection techniques for oil and gas pipelines: state-of-the-art. Tunn. Undergr. Space Technol. 98, 103249. https://doi.org/10.1016/j.tust.2019.103249.

Mazumder, R.K., Salman, A.M., Li, Y., 2021a. Failure risk analysis of pipelines using data-driven machine learning algorithms. Struct. Saf. 89, 102047. https://doi.org/10.1016/j.strusafe.2020.102047.

Mazumder, R.K., Salman, A.M., Li, Y., 2021b. Reliability assessment of oil and gas pipeline systems at burst limit state under active corrosion. 18th Int.Probabilistic Workshop 653–660. https://doi.org/10.1007/978-3-030-73616-3_50.

Mostafapour, A., Davoudi, S., 2013. Analysis of leakage in the high-pressure pipe using acoustic emission method. Appl. Acoust. 74 (3), 335–342. https://doi.org/10.1016/j.apacoust.2012.07.012.

Ning, F.L., Cheng, Z.H., Meng, D., Duan, S., Wei, J., 2021. Enhanced spectrum convolutional neural architecture: an intelligent leak detection method for a gas pipeline. Process Saf. Environ. Protect. 146, 726–735. https://doi.org/10.1016/j.psep.2020.12.011.

Nitta, T., Kuroe, Y., 2018. Hyperbolic gradient operator and hyperbolic back-propagation learning algorithms. IEEE Transact. Neural Networks Learn. Syst. 29 (5), 1689–1702. https://doi.org/10.1109/TNNLS.2017.2677446.

Omojugba, B., Oyetunji, S., Adetan, O., 2020. Multiproduct pipeline leak detection and localization system using artificial intelligence. SN Comp. Sci. 1, 132. https://doi.org/10.1007/s42979-020-00144-9.

Oyedeko, K.F.K., Balogun, H.A., 2015. Modeling and simulation of a leak detection for oil and gas pipelines via transient model: a case study of the Niger delta. J. Energy Technol. Pol. 5 (1), 16–27. https://doi.org/10.7176/JETP.

Qu, Z.G., Feng, H., Zeng, Z.M., Zhuge, J.C., Jin, S.J., 2010. A SVM-based pipeline leakage detection and pre-warning system. Measurement 43 (4), 513–519. https://doi.org/10.1016/j.measurement.2009.12.022.

Rai, A., Kim, J.M., 2021. A novel pipeline leak detection approach independent of prior failure information. Measurement 167, 108284. https://doi.org/10.1016/j.measurement.2020.108284.

Ruiz-Cárcel, C., Lao, L., Cao, Y., Mba, D., 2016. Canonical variate analysis for performance degradation under faulty conditions. Control Eng. Pract. 54, 70–80. https://doi.org/10.1016/j.conengprac.2016.05.018.

Sabah, M., Talebkeikhah, M., Agin, F., Talebkeikhah, F., Hasheminasab, E., 2019. Application of decision tree, artificial neural networks, and adaptive neuro-fuzzy inference system on predicting lost circulation: a case study from Marun oil field. J. Petrol. Sci. Eng. 177, 236–249. https://doi.org/10.1016/j.petrol.2019.02.045.

Santos, R.B., Sousa, E.O.D., Silva, F.V.D., Cruz, S.L.D., Fileti, A.M.F., 2014. Detection and on-line prediction of leak magnitude in a gas pipeline using an acoustic method and neural network data processing. Braz. J. Chem. Eng. 31 (1), 145–153. https://doi.org/10.1590/S0104-66322014000100014.

Tao, J.L., Yu, Z., Zhang, R.D., Gao, F.R., 2021. RBF neural network modeling approach using PCA based LM–GA optimization for coke furnace system. Appl. Soft Comput. 111, 107691. https://doi.org/10.1016/j.asoc.2021.107691.

Waleed, D., Mustafa, S.H., Mukhopadhyay, S., Abdel-Hafez, M.F., Jaradat, M.A.K., Dias, K.R., Arif, F., Ahmed, J.I., 2019. An in-pipe leak detection robot with a neural-network-based leak verification system. IEEE Sensor. J. 19 (3), 1153–1165. https://doi.org/10.1109/jsen.2018.2879248.

Wang, C.Y., Xu, C., Yao, X., Tao, D.C., 2019. Evolutionary generative adversarial networks. IEEE Trans. Evol. Comput. 23 (6), 921–934. https://doi.org/10.1109/tevc.2019.2895748.

Wang, X.W., Mazumder, R.K., Salarieh, B., Salman, A.M., Shafleezadeh, A., Li, Y., 2022. Machine learning for risk and resilience assessment in structural engineering: progress and future trends. J. Struct. Eng. 148 (8), 03122003. https://doi.org/10.1061/(Asce)St.1943-541x.0003392.

Wilamowski, B.M., Yu, H., 2010. Improved computation for Levenberg-Marquardt training. IEEE Trans. Neural Network. 21 (6), 930–937. https://doi.org/10.1109/TNN.2010.2045657.

Xie, J.Y., Xu, X.D., Dubljevic, S., 2019. Long range pipeline leak detection and localization using discrete observer and support vector machine. AIChE J. 65 (7), e16532. https://doi.org/10.1002/aic.16532.

Xu, C., Yu, B., Zhang, Z.W., Zhang, J.J., Wei, J.J., Sun, S.Y., 2010. Numerical simulation of a buried hot crude oil pipeline during shutdown. Petrol. Sci. 7 (1), 73–82. https://doi.org/10.1007/s12182-010-0008-x.

Xu, Z.D., Zhu, C., Shao, L.W., 2021. Damage identification of pipeline based on ultrasonic guided wave and wavelet denoising. J. Pipeline Syst. Eng. Pract. 12 (4), 04021051. https://doi.org/10.1061/(asce)ps.1949-1204.0000600.

Yan, S.N., Wang, T.Y., Tang, T.Q., Ren, A.X., He, Y.R., 2020. Simulation on hydrodynamics of non-spherical particulate system using a drag coefficient correlation based on artificial neural network. Petrol. Sci. 17 (2), 537–555. https://doi.org/10.1007/s12182-019-00411-2.

Yu, T., Li, C.X., Yao, B., Zhang, Z.J., Guo, Y., Liu, L.J., 2020. Standard friction prediction model of long-distance hot oil pipelines. Petrol. Sci. 17 (2), 487–498. https://doi.org/10.1007/s12182-019-00417-w.

Zeng, D.Z., Deng, K.H., Lin, Y.H., Shi, T.H., Shi, D.Y., Zhou, L.Z., 2014. Theoretical and experimental study of the thermal strength of anticorrosive lined steel pipes. Petrol. Sci. 11 (3), 417–423. https://doi.org/10.1007/s12182-014-0356-z.

Zhang, H.G., Hu, X.G., Ma, D.Z., Wang, R., Xie, X.P., 2022. Insufficient data generative model for pipeline network leak detection using generative adversarial networks. IEEE Trans. Cybern. 52 (7), 7107–7120. https://doi.org/10.1109/TCYB.2020.3035518.

Zhang, J., Peng, Y.X., Yuan, M.K., 2020a. SCH-GAN: semi-supervised cross-modal hashing by generative adversarial network. IEEE Trans. Cybern. 50 (2), 489–502. https://doi.org/10.1109/TCYB.2018.2868826.

Zhang, R., Xu, Z.B., Huang, G.B., Wang, D.H., 2012. Global convergence of online BP training with dynamic learning rate. IEEE Transact. Neural Networks Learn. Syst. 23 (2), 330–341. https://doi.org/10.1109/TNNLS.2011.2178315.

Zhang, S.Z., Wang, X., Cheng, Y.F., Shuai, J., 2020b. Modeling and analysis of a catastrophic oil spill and vapor cloud explosion in a confined space upon oil pipeline leaking. Petrol. Sci. 17 (2), 556–566. https://doi.org/10.1007/s12182-019-00403-2.

Zhao, B., Chen, H., Gao, D.K., Xu, L.Z., 2020. Risk assessment of refinery unit maintenance based on fuzzy second generation curvelet neural network. Alex. Eng. J. 59 (3), 1823–1831. https://doi.org/10.1016/j.aej.2020.04.052.

Zhao, B., Ren, Y., Gao, D.K., Xu, L.Z., 2019a. Performance ratio prediction of photovoltaic pumping system based on grey clustering and second curvelet neural network. Energy 171, 360–371. https://doi.org/10.1016/j.energy.2019.01.028.

Zhao, B., Ren, Y., Gao, D.K., Xu, L.Z., Zhang, Y.Y., 2019b. Energy utilization efficiency evaluation model of refining unit Based on Contourlet neural network optimized by improved grey optimization algorithm. Energy 185, 1032–1044. https://doi.org/10.1016/j.energy.2019.07.111.

Zhao, B., Song, H.Y., 2021. Fuzzy Shannon wavelet finite element methodology of coupled heat transfer analysis for clearance leakage flow of single screw compressor. Eng. Comput. 37 (3), 2493–2503. https://doi.org/10.1007/s00366-020-01259-6.

Zhao, T.F., Duan, M.L., Pan, X.D., Feng, X.H., 2010. Lateral buckling of non-trenched high temperature pipelines with pipelay imperfections. Petrol. Sci. 7 (1), 123–131. https://doi.org/10.1007/s12182-010-0016-x.

Zheng, J.Q., Du, J., Liang, Y.T., Liao, Q., Li, Z.B., Zhang, H.R., Wu, Y., 2021. Deeppipe: a semi-supervised learning for operating condition recognition of multi-product pipelines. Process Saf. Environ. Protect. 150, 510–521. https://doi.org/10.1016/j.psep.2021.04.031.

Zhou, M.F., Zhang, Q., Liu, Y.W., Sun, X.F., Cai, Y.J., Pan, H.T., 2019. An integration method using kernel principal component analysis and cascade support vector data description for pipeline leak detection with multiple operating modes. Processes 7 (10), 648. https://doi.org/10.3390/pr7100648.