Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-3, March 2025
DOI: 10.53469/jrse.2025.07(03).09

A Novel PCA_ERF Method for Leakage Detection of District Heating System

Youen Zhao[1],*, Shoujun Zhou[2]

[1]School of Computing and Artificial Intelligence, Shandong University of Finance and Economics, Jinan 250014, Shandong, China
[2]School of Thermal Engineering, Shandong Jianzhu University, Jinan 250000, Shandong, China
**Correspondence Author, zhaoyouen@sdufe.edu.cn*

Abstract: ***The leakage of district heating system can lead to serious consequences. Therefore, the leakage detection of the district heating system has always been the focus of research in various industries. Relying on the intelligent heating experimental pipe-network system in Shandong Jianzhu University, this paper takes 4 topological heating experimental pipe-networks as the research objects, constructs the real-time operation datasets, simulation datasets and the cross datasets of the above two, creatively proposes a PCA-ERF (Principal Component Analysis-Extremely Random Forest) based method for the leakage detection task. The method adopts PCA to map the original pressure and flow data of the heating network into the vector space, which has a stronger feature expression ability firstly; then the decision trees for classification are trained by ERF with stronger randomness; finally, the final classification results are obtained by integrating the judgment of all the decision trees. The experimental results show that the PCA_ERF method shows excellent performance under different cross-data ratios, especially when the cross-data ratio is 2:1, the accuracy of the proposed PCA-ERF method in the leakage prediction for 4 different topologies is 98.08%, 97.1%, 98.92% and 97.64% respectively, which can complete the leakage detection task of complex heating network with multiple topologies.***

Keywords: Primary Component Analysis, Extremely Random Forest, District heating system, Leakage detection, Multiple topologies**.**

1. Introduction

With the rapid development of Chinese heating industry, the scale of district heating system continues to expand, and the topology of it is becoming complex. At the end of 2022, the heating area was 11.125 billion square meters [1]. As a consequence, the number of operational failures is constantly increasing. Among many operation failures, pipeline leakage is the most serious one. Once a leak occurs in the pipeline network, it will not only result in insufficient heating supply meeting user demand, but also increase the energy consumption of the heat source and the workload of the heat exchange station equipment. For thermal power plants, a leakage in the heating pipeline network will affect the normal operation of the power plant system, and, in severe cases, result in downtime. Therefore, district heating system leakage has become an urgent problem to be solved in the heating industry. Timely and accurate diagnosis of the leakage are necessary to fully guarantee the long-term safe and stable operation of the district heating system.

In recent years, the issue of leakage detection has gradually received attention from relevant scholars. At present, many scholars have conducted extensive research on the issue of pipeline leakage detection and achieved many results. Various leak detection methods can be roughly divided into three categories: hardware-based detection methods, physical model-based detection methods, and data-driven detection methods [2].

The first category mainly constructs leakage detection methods based on the changes in acoustic, optical, thermal and other operational characteristics of the pipeline network when a leakage occurs, including manual inspection methods, infrared methods, tracer detection methods, fiber optic sensing detection methods [3], etc. These methods rely too much on the arrangement of sensor arrays, and their detection effect may be affected by climate conditions. Among them, ultrasonic methods and manual inspection methods are not suitable for leakage detection tasks in urban pipeline networks with large heating areas.

The second category is to construct a physical model of the hydraulic condition of the pipeline network based on the signal response (pressure and flow changes, etc.) when a leakage occurs in the pipeline network [4] - [6]. During the physical model building process, various hydraulic conditions need to be fully considered, so it is difficult to establish a physical model for ideal conditions, which may lead to deviations between the simulated results and actual values of the model, and may have a certain impact on the accuracy of the leak detection method.

The third category is to convert the signal changes of pipeline leakage into certain characteristic information, and then use relevant mathematical theories to detect the leakage, mainly including material balance detection method, cross-correlation detection method [7], fuzzy clustering detection method [8], wavelet denoising detection method [9], etc. However, the detection effect is sometimes limited by environmental conditions and method factors. For example, for long-distance pipelines, material balance method and negative pressure wave method have poor detection effect on slight leakage, and cross-correlation detection method and wavelet denoising method have high data requirements and require complex data preprocessing work. Scholars have also applied optimization algorithms to the research of pipeline leakage detection, such as genetic algorithm [10-11], M-SPRT algorithm [12], particle swarm optimization algorithm [13-14], firefly swarm optimization algorithm [15], and differential evolution algorithm [16] etc. Although all of which have achieved certain detection results, this type of method requires specific denoising and data augmentation for specific data, and does not effectively solve the problem of missing actual leakage operation data in the pipeline network.

Therefore, the detection effect of this method on the latest leakage conditions in the pipeline network is poor, and the data used in the method needs to be continuously updated.

With the rapid development of computer technology, especially machine learning algorithms, machine learning algorithms such as neural networks, random forests, and support vector machines have been widely applied due to their excellent learning abilities. The so-called "learning" refers to the ability to continuously learn from a large amount of seemingly unrelated and chaotic data to improve the learning ability, in order to cope with future prediction tasks [17] - [19]. At present, machine learning has been applied in the research of traditional energy fields, mainly in load forecasting [20], system simulation [21], fault warning [22], and regulation and evaluation of energy systems [23].

Among the numerous prediction algorithms in the field of machine learning, random forests are widely used in pattern recognition tasks due to their advantages such as simple structure, parallel training, fewer hyperparameters, and resistance to overfitting. This algorithm is based on the idea of voting classification and consists of multiple decision trees, each of which is trained on a different random training subset. The random forest makes the final prediction based on the number of occurrences of the classification prediction results of numerous decision trees. This method can effectively avoid classification errors caused by using a single decision tree [24-25].

This article regards leakage detection in heating pipelines as a pattern recognition problem, and creatively applies principal component analysis method and extreme random forest algorithm in machine learning field to leakage detection in the district heating pipelines, achieving good detection results.

2. Introduction to the Experimental Heating-network System

The experimental heating-network system and its topology structure in the Hydraulic Balance Laboratory of Shandong Jianzhu University are shown in Figure 1 and Figure 2.

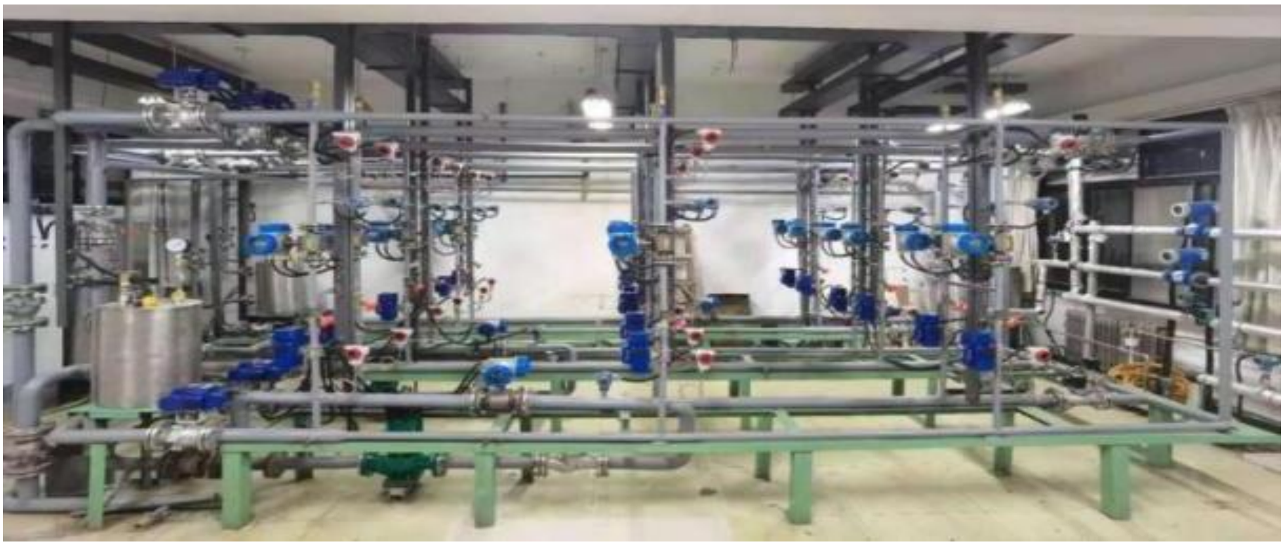


Figure 1: The physical image of the experimental heating-network system

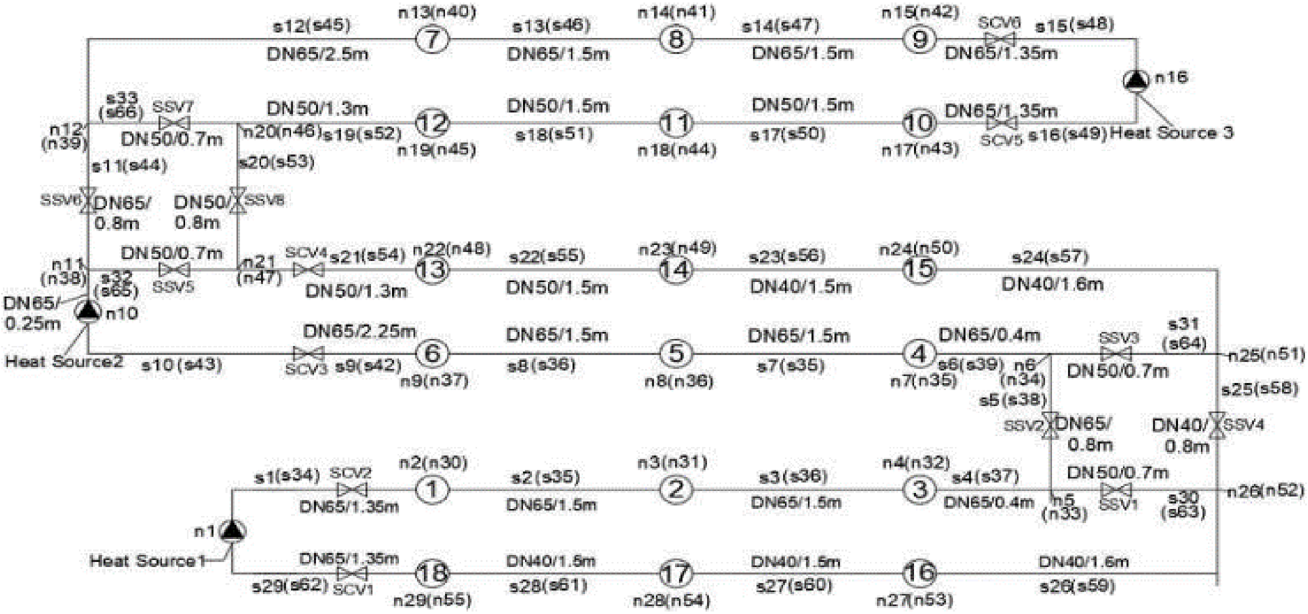


Figure 2: The topology diagram of the experimental heating-network system

A total of 18 heat users (up supply and down return), 3 heat sources (water pump), 3 basic loops are involved in the experimental heating-network system, in which different loops are connected with each other though the electromagnetic conversion valves. The constant pressure point is set up at the entrance of each heat source by regulating the filling water pump. The meters installed in user pipes are as follows: supply pressure meter (Range:0–200kPa, Accuracy:0.5%), thermal meter (Range:0–100℃, Accuracy:0.5%), electric control valve and return pressure meter (Range:0–200kPa, Accuracy:0.5%). The data of all meters are recorded by the database system in real time.

In Figure 2, ***s*** and ***n*** are the number of pipes and nodes respectively and the number in brackets are their corresponding number in return pipes; ***DN*** is the diameter of pipes with the unit mm; SCV and SSV are the electric control valves and the electromagnetic conversion valves in supply (return) pipes.

The system is equipped with leakage valves that simulate pipe network leakage, installed at user nodes: n2, n17, n55 and pipe sections: s2, s35, s17, s50, s22, s28, s61. Four different heating-network topologies can be achieved by regulating state of the electric control valves and the electromagnetic conversion valves, which are shown in Table 1.

Table 1: State of the electric control valves and the electromagnetic conversion valves under four different topology

Label of valve	B-SHS (Branch network with single heat source)	B-DHS (Branch network with double heat sources)	SR-SHS (Single-ring network with single heat source)	DR-DHS (Double-ring network with double heat sources)
SCV1	Close	Open	Open	Open
SCV2	Open	Close	Open	Open
SCV3	Open	Close	Open	Open
SCV4	Open	Open	Open	Open
SCV5	Open	Open	Open	Close
SCV6	Open	Open	Open	Close
SSV1	Close	Close	Close	Open
SSV2	Open	Close	Open	Open
SSV3	Close	Close	Close	Close
SSV4	Open	Open	Open	Open
SSV5	Close	Close	Close	Open
SSV6	Open	Open	Open	Close
SSV7	Close	Close	Close	Close
SSV8	Open	Open	Open	Close

3. Construction of Datasets

In this article, we generated three types of datasets: experimental datasets, simulation datasets, and their cross datasets (consisting of experimental data and simulation data).

Each data sample is a line vector with the order of user supply pressure, user return pressure, user flow rate and label. The dimensions of datasets (the features number of data sample+1) for four different topology heating-networks are different due to the number of users, specifically, a B-SHS with 18 users, a B-DHS with 12 users, a SR-SHS with 18 users and a DR-DHS with 12 users. So, the dimensions of datasets respectively are B-SHS: 55(18 +18 +18 +1), B-DHS: 37 (12 +12 +12 +1), SR-SHS: 55(18 +18 +18 +1), DR-DHS: 37(12 +12 +12 +1).

3.1 Experimental Datasets

The experimental datasets are obtained through the real-time data acquisition module of the experimental heating-network system, including the pressure and flow data. We conducted a total of 148 experiments, each running for 16 minutes, with a data collection frequency of 2 times per second. These 148 experiments cover different operating conditions of pipeline networks under different topological structures, as shown in Table 2. Considering the situation of data loss, the experimental data corresponding to different topological structures we obtained were 287756 (B-SHS), 255093 (B-DHS), 272013 (SR-SHS), and 254300 (DR-DHS).

The data for each operating condition includes both leak free data and leak data (including user node leakage and pipe segment leakage). To ensure maximum coverage of the entire pipeline network in terms of leakage range, both user node leakage and pipe segment leakage include three leakage locations in the pipeline network: near, medium, and far from the heat source. Each leakage location is divided into four leakage conditions based on different leakage rates (percentage of leakage to total flow), with leakage rates of 1.1%, 2.5%, 4%, and 5.5%, respectively. Therefore, the total number of operating conditions for the four different topology structures studied in this article are: 148=(4 × (9 × 4+1)), 132=(4 × (8 × 4+1)), 148=(4 × (9 × 4+1)), 132=(4 × (8 × 4+1)).

Table 2: Method to regulate operation parameters under different Hydraulic Working-Conditions (HWC)

Name of HWC	B-SHS	B-DHS	SR-SHS	DR-DHS
C0 (Normal HWC)	Water pump head1 is 120 kpa	Water pump head1, head2 are both 30 kpa	Water pump head1 is 45kPa	Water pump head1, head2 are both 20 kpa
C1 (Intermediate-user changing HWC)	Flow rate of user 10 is 0.5m^3/h	Flow rate of user 13 is 0.5m^3/h	Flow rate of user 10 is 0.5m^3/h	Flow rate of user 5 is 0.5m^3/h
C2 (Pipe control-valve changing HWC)	Opening of SCV4 is 60%	Opening of SCV1 is 60%	Opening of SCV2 is 60%	Opening of SCV2 is 60%
C3(Centralized-adjustment changing HWC)	Water pump head1 is 110 kpa	Water pump head1, head2 are both 35 kpa	Water pump head1 is 50kPa	Water pump head 1, head 2 are 22.5 and 24.5 kPa

3.2 Simulation Datasets

3.2.1 Basic modeling theory of HWC

The heating network is similar to the power supply network, with flow rate, pressure drop, and resistance characteristic coefficients similar to current, voltage, and resistance in the power grid, respectively. Based on graph theory and Kirchhoff's law that characterizes pipeline characteristics, the basic calculation model for hydraulic conditions of any heating pipeline network with m branches and n+1 nodes can be derived as [10-12]:

$$AG = Q \tag{1}$$

$$B_f \Delta H = 0 \tag{2}$$

$$\Delta H = S|G|G + Z - DH \tag{3}$$

A represents the unique correlation matrix of the pipeline network topology structure ($n \times m$); G represents flow vector of pipeline section, m^3/h; Q represents net outflow vector of each node in the pipeline, with positive inflow and negative outflow, m^3/h; B_f represents the basic circuit matrix of the pipeline network, with an order of $(m-n)\times m$; ΔH represents pipeline pressure drop vector, kPa; S represents diagonal matrix of resistance characteristic coefficient of pipe section, kPa/(m^3/h)^2; $|\boldsymbol{G}|$ represents the absolute value vector of the flow rate of the pipeline section, m^3/h; Z represents vector of potential difference between two nodes in the branch, kPa; DH represents head vector of the water pump in the pipeline section, kPa.

3.2.2 Generate simulation datasets

Calculate the resistance characteristic coefficient of the pipeline section and construct the correlation matrix and loop matrix that characterize the unique topology structure of the pipeline network. Then, based on formula (1)-(3), establish a simulation model of the normal operating conditions of the pipeline network to simulate the pressure and flow distribution of users under the conditions of no leakage and leakage in the pipeline network, that is, the simulation data.
Due to the inevitable fluctuations in data reading during the actual operation of the pipeline network, in order to make the simulation data more realistic in simulating the actual operation of the pipeline network, this paper sets the iteration times for each working condition with different topology structures between 1500-1900. Therefore, the total simulation data for all working conditions of B-SHS, B-DHS, SR-SHS, and DR-DHS are 222244, 237218, 266400, and 237600, respectively.

3.3 Cross Datasets

Sort the experimental data and simulation data out of order, and construct their cross data with dimensions consistent with the experimental data and simulation data. Cross data ratio refers to the ratio of simulated data to experimental data in a cross dataset. To investigate the impact of cross data comparison on the prediction accuracy of classification algorithms, different cross data ratios were used in this paper. The distribution of cross data at each ratio is shown in Table 3.

3.4 Setting Labels

The end column of each dataset is the label. The labels are encoded by triple digits: The 1–4 hundred digits indicate four kinds of HWCs (C0, C1, C2, C3); the 0–4 single digits indicate five leakage degrees. 0 represents no leakage, and 1-4 represents four degrees of leakage (leakage rates are 1.1%, 2.5%, 4%, and 5.5%, respectively). Ten digits indicate leakage locations, represented by 0-9, where 0 indicates that there is no leakage at that location. The leakage location labels for the four different topology pipe networks are different (1-9), as shown in Table 4.

Table 3: Data distribution under different cross data ratios.

	B-SHS		B-DHS		SR-SHS		DR-DHS	
Cross-data ratio	Simulation data amount	Experimental data amount	Simulation data amount	Experimental data amount	Simulation data amount	Experimental data amount	Simulation data amount	Experimental data amount
2:1	222244	111122	237218	118609	266400	133200	237600	118800
4:1	222244	55561	237218	59305	266400	66600	237600	59400
6:1	222244	37041	237218	39536	266400	44400	237600	39600
8:1	222244	27781	237218	29652	266400	33300	237600	29700
10:1	222244	22224	237218	23722	266400	266640	237600	23760
20:1	222244	11112	237218	11861	266400	13320	237600	11880
30:1	222244	7408	237218	7907	266400	8880	237600	7920
40:1	222244	5556	237218	5930	266400	6660	237600	5940
50:1	222244	4445	237218	4744	266400	5328	237600	4752
60:1	222244	3704	237218	3954	266400	4440	237600	3960
70:1	222244	3715	237218	3389	266400	3806	237600	3394
80:1	222244	2778	237218	2965	266400	3330	237600	2970
90:1	222244	2469	237218	2636	266400	2960	237600	2640
100:1	222244	2222	237218	2372	266400	2664	237600	2376

Table 4: Setting of leakage location labels for four different topological structures

B-SHS		B-DHS		SR-SHS		DR-DHS	
Leakage location	Label	Leakage location	Label	Leakage location	Label	Leakage location	Label
n2	1	n17	1	n2	1	n2	1
n17	2	n29	2	n17	2	n55	2
n55	3	s17	3	n55	3	s2	3
s2	4	s50	4	s2	4	s35	4
s35	5	s22	5	s35	5	s22	5
s17	6	s55	6	s17	6	s55	6
s50	7	s28	7	s50	7	s2	7
s2	8	s61	8	s2	8	s35	8
s35	9			s35	9		

4. PCA_ERF Algorithm

The pressure and flow data of a district heating system are functions of time and constantly change over time, which are a set of strongly correlated data. Principal component analysis (PCA) method is a multi-variate statistical analysis method. The purpose of PCA is to map the original data to a new feature space, which is called the feature mapping. Different data feature distributions are unified during this process. Therefore, the data transformed to the new feature space will have stronger expressive power. In addition to reducing the dimensionality of the original data, this method can also rearrange the original data from large to small according to the eigenvalues.

Random Forest is a supervised ensemble learning algorithm that has better accuracy than most individual machine learning algorithms due to the use of ensemble algorithms.

Additionally, the introduction of two randomness factors makes it less prone to overfitting, making it widely used in classification and regression tasks. Extreme Random Forest (ERT) is similar to the random forest method in that it consists of many decision trees. The difference between the two is that ERT is more random. It randomly selects a subset of features at each node and randomly splits to obtain the optimal branching attributes and thresholds. This increased randomness helps to create more independent decision trees and train a better performing learning model.

4.1 Principal Component Analysis

The principal component analysis [26] - [27] method is a multivariate statistical analysis method that performs linear transformation on input data to select a small number of important feature vectors. This article uses this method to map data features to unify the feature distribution of simulation data and experimental data. The calculation process is as follows:

1) Standardization transformation of raw data. The calculation method for standardizing the original data is as follows:

$$X = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{k1} \\ X_{21} & X_{22} & \dots & X_{k2} \\ \dots & \dots & & \dots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \qquad X_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{4}$$

In the above equation, x_{ij} is the original data, $\bar{x}_j$ is the mean of the *jth* column of the original data, and s_j is the standard deviation of the *jth* column of the data.

2) Calculation of correlation coefficient matrix

The purpose of calculating the correlation coefficient matrix is to obtain the degree of correlation between each column of data. The calculation method is as follows:

$$R = C\,ov(X) = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{k1} \\ r_{21} & r_{22} & \cdots & r_{k2} \\ \cdots & \cdots & & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nk} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{k1} \\ r_{21} & 1 & \cdots & r_{k2} \\ \cdots & \cdots & & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{bmatrix} \tag{5}$$

$$r_{ij} = \frac{\sum_{k=1}^{n}(X_{kj} - \bar{X}_j)(X_{ij} - \bar{X}_j)}{\sqrt{\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)^2}\sqrt{\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)^2}}$$

r_{ij} presents the correlation coefficient between the *i-th* column data and the *j-th* column data, with the numerator being the covariance of the corresponding data and the denominator being the product of the standard deviations of the two columns. According to the above formula, the higher the correlation coefficient, the greater the degree of correlation between data. If the correlation coefficient is 1 or -1, it indicates a completely linear correlation between the data.

3) Eigenvalue calculation of correlation coefficient matrix

The purpose of calculating the eigenvalues of the correlation coefficient matrix is to rank the eigenvalues to determine the principal components. The calculation method is as follows:

$$Cov(X)L = L\begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \cdots & \\ 0 & & & \lambda_k \end{bmatrix} \tag{6}$$
$$L = [l_1 \quad l_2 \quad \dots \quad l_k]$$
$$l_i = [l_{i1} \quad l_{i2} \qquad l_{ik}]'(i = 1,2,\dots,k)$$

l_i represents the eigenvector of the correlation coefficient matrix, and λ_i represents the corresponding eigenvalue of the eigenvector. After sorting all feature values in descending order, the distribution of feature vectors corresponding to the order can be obtained.

4.2 Extreme Random Forest

The basic idea of extreme random forest is to combine multiple classifiers with weaker classification ability to form a classifier group with stronger classification ability. The core training process of this algorithm lies in the decision tree algorithm. The decision tree is based on a top-down hierarchical structure, which sequentially judges one or more features of the sample until the leaf node, and derives the final prediction label. The calculation process is as follows [28] - [31]:

1) Construct root nodes

Place all training samples on the root node, which determines the decisive feature after evaluating multiple data features and divides the training dataset according to this feature, so that each segmented subset has the best classification and distributes these subsets on all branches of the root node;

2) Construct leaf nodes

The principle of constructing leaf nodes for decision-making lies in whether the segmented subset is correctly classified. If the segmented subset can be correctly classified, leaf nodes are constructed and the segmented subset is assigned to the corresponding leaf nodes; If the segmented subsets cannot be classified correctly, then the optimal features are re selected for these subsets, and the segmentation process is repeated to construct the corresponding leaf nodes; The classification operation of the decision tree follows the above two steps recursively until all training subsets are classified correctly. At this point, each subset is assigned to a leaf node, that is, all subsets have a clear category, and the decision tree training is complete.

The purpose of decision tree partitioning of a dataset is to make unordered raw data more ordered. The partitioning criteria are mainly based on information gain, that is, obtaining the feature with the highest information gain is the best choice [32] - [33].

3) Flow chart of PCA_ERF algorithm

The flow chart of PCA-ERF algorithm includes two parts: the training part and the test part, as shown in Figure 3 and Figure 4.

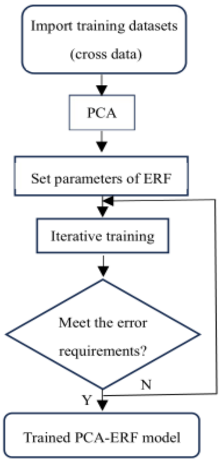


Figure 3: Training section of PCA_ERF algorithm

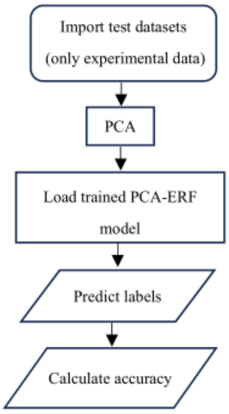


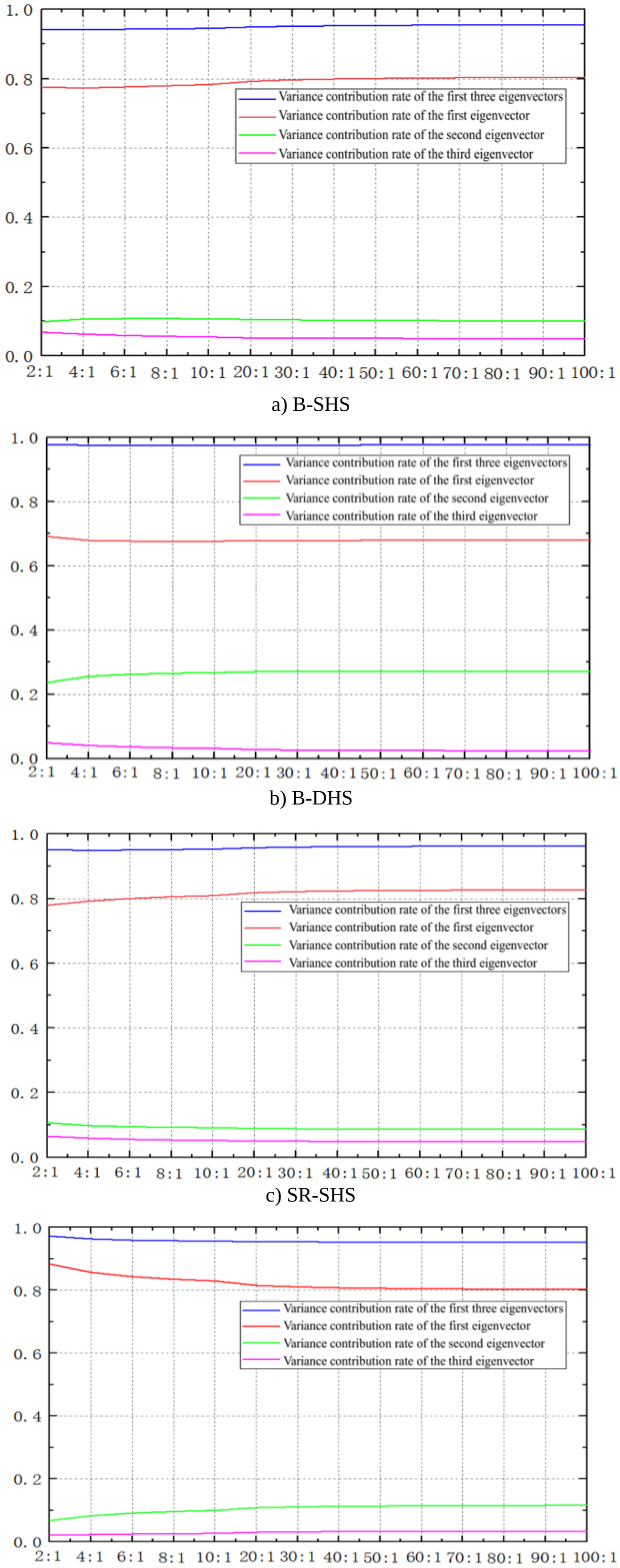
Figure 4: Test section of PCA_ERF algorithm

5. Experiment

This article uses a cross dataset as the training set for this simulation experiment, as shown in Table 3. At the same time, in order to more accurately obtain the actual leakage detection performance of the PCA-ERF method in pipeline networks, the test set is only composed of experimental datasets.

5.1 PCA Variance Contribution Analysis

In the PCA method, the variance contribution rate can explain the amount of data information that feature vectors can reflect. The larger the variance contribution rate, the more data information the feature vector covers, and the stronger its feature representation. In most cases, the first three eigenvectors can represent most of the data information. Therefore, this article selects the first three eigenvectors for analysis, and the trend of their variance contribution rate with the change of cross data ratio is shown in Figure 5.



a) B-SHS

b) B-DHS

c) SR-SHS

d) DR-DHS

Figure 5: The variation of the first three eigenvectors with cross data ratio under four different topological structures

From Figure 5, it can be seen that for the four different topological structures of the pipeline network, as the cross-data ratio continues to increase (i.e., the amount of simulated data in the cross data continues to increase), the sum of the variance contribution rates of the first three feature vectors slowly increases and closes to 1, indicating that it retains more feature information from the original data.

5.2 PCA-RF Method

After preliminary experiments and analysis, this article sets the number and depth of decision trees in the random forest for leak detection of four different topological structures of pipeline networks, with values of 216= (4 × 54), 144= (4 × 36), 216= (4 × 54), and 144= (4 × 36), respectively. To ensure that each leaf node can quickly find the best features, this article sets the maximum number of features traversed by each node to be equal to the number of features in the input dataset. The prediction accuracy of the algorithm varies with the cross-data ratio as shown in Figure 6.

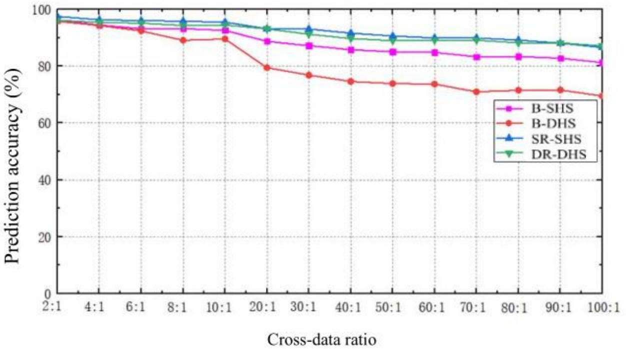


Figure 6: Prediction accuracy of PCA-RF method under different cross data ratios

From Figure 6, it can be seen that this method has a high prediction accuracy and changes slowly with the cross-data ratio. However, for B-DHS pipeline network, the prediction accuracy of its leakage conditions is significantly lower than the other three pipeline networks, indicating that the prediction performance of PCA-RF is not suitable for leak detection tasks in various topological structures. Therefore, we consider using the Extremely Random Forest (ERT) method to solve our problem.

5.3 PCA-ERF Method

In this article, the number of extreme random trees is set to be twice the number of features in the dataset, where the depth of the trees and the number of features traversed by nodes are the same as described in Section 4.2.

For different cross data ratios, the prediction accuracy of the algorithm also varies, and its values and distribution trends are shown in Table 5 and Figure 6.

Table 5: Prediction accuracy of PCA-ERF under different cross data ratios

Cross data ratios	B-SHS	B-DHS	SR-SHS	DR-DHS
2:1	98.08%	97.17%	98.92%	97.64%
4:1	97.57%	96.58%	98.61%	97.68%
6:1	97.40%	96.07%	98.38%	97.23%
8:1	97.11%	95.05%	98.19%	96.98%
10:1	96.79%	95.48%	98.04%	96.80%
20:1	95.81%	95.08%	97.26%	96.27%
30:1	95.23%	92.94%	96.54%	95.95%
40:1	94.25%	91.51%	95.91%	95.53%
50:1	93.23%	90.09%	95.46%	95.15%
60:1	92.05%	88.67%	94.11%	94.54%
70:1	90.97%	86.96%	93.05%	94.27%
80:1	89.60%	85.15%	91.89%	93.68%
90:1	88.85%	84.03%	90.91%	93.15%
100:1	86.65%	81.68%	89.31%	92.62%

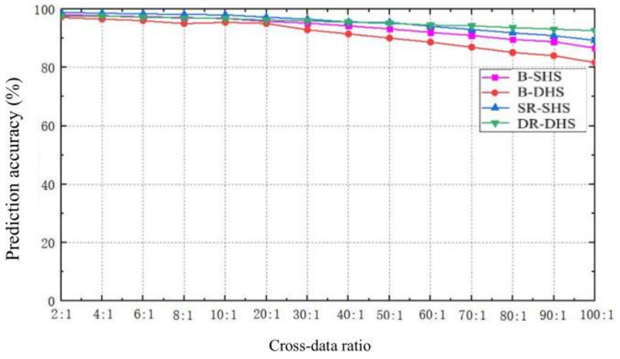


Figure 7: Prediction accuracy of PCA-ERF under different cross data ratios

According to Table 6, even in the case of the highest cross data ratio (100:1), the prediction accuracy of this method is still high. From Fig 7, it can be seen that the distribution trend of prediction accuracy is stable and relatively small with changes in pipeline network topology. This result indicates that the randomness in PCA-ERF further enhances the algorithm's performance in finding the best data features, and its feature transfer learning ability is further improved. The pipeline leakage detection method constructed using PCA-ERF is suitable for leakage detection tasks in multi heat source circular complex pipelines.

5.4 Comparison Studies

In order to explore the connection between the PCA_ERF and the popular machine learning methods, totally five models are established for comparison. Four of them use PCA described in section 4.1 as feature sets for classification. The rest one method is a deep neural network-based model, which can carry out both feature extraction and fault classification operations.

First, the widely used classification algorithms including support vector machine (SVM), k-nearest neighbor (KNN), random forest (RF), and back propagation neural network (BPNN) are employed to classify the extracted features. After that, a convolution neural network (CNN) is designed for leakage detection. The BPNN consists of three fully connected layers with a softmax classifier while the CNN model contains two convolution layers and two max-pooling blocks, a flatten layer, and a fully connected layer. The cross-data ratio used for the comparison experiments is uniformly set to 2:1, and the results are shown in Table 6.

Table 6: Comparison results on four datasets

Method	accuracy (%)				
	B-SHS	B-DHS	SR-SHS	DR-DHS	Average
SVM	92.6	91.8	89.7	89.6	90.925
KNN	95.6	95.7	94.9	94.7	95.225
RF	96.3	95.8	96.8	96.4	96.325
BPNN	97.8	97.5	97.2	96.8	97.325
CNN	99.1	98.3	98.1	98.42	**98.48**
Proposed method	**98.08**	**97.17**	**98.92**	**97.64**	**97.95**

As can be seen from table 6, in term of the classification accuracy, the method proposed in this paper performs best among all the non-deep models, including SVM, KNN, RF and BPNN, which indicate the effectiveness for leakage detection in district heating system of the proposed method. When it comes to the deep model CNN, it performs better than the proposed method. The main reason is that the CNN which employs a complex structure with two convolution layers and two max-pooling blocks, a flatten layer, and a fully connected layer can obtain better feature representations from original pressure and flow data of heating system. Meanwhile, CNN requires more training data and longer training time. Thus, the proposed method is more suitable for industry application since it is designed to solve the practical problem with less training data and shorter training time.

6. Conclusions

Relying on the experimental heating-network system in the Hydraulic Balance Laboratory of Shandong Jianzhu University, four topological structures, including B-SHS (single heat source branch), B-DHS (double heat source branch), SR-SHS (single heat source single ring), and DR-DHS (double heat source double ring), are taken as the research object, the leakage detection of the district heating system is regarded as a pattern recognition problem, and the real-time operation dataset, simulation dataset, and their cross dataset are constructed, the principal component analysis method and extreme random forest method are creatively combined for the leakage detection task of the district heating pipeline network, and high detection results are obtained. We can draw the following conclusions:

1) The PCA-ERF method does not require a large dataset and can achieve higher leakage detection accuracy for 4 operating conditions, 5 leakage levels, and 10 leakage points in topologies of B-SHS, B-DHS, SR-SHS, DR-DHS. It has the characteristics of fast detection speed and high detection accuracy;

2) The impact of the cross- data ratios on recognition results: As the cross-data ratio increases, that is, as the amount of experimental data in the cross-data set decreases, the recognition accuracy slightly decreases, indicating that compared to simulation data, experimental data has a greater impact on recognition accuracy;

3) The impact of different topological structures on recognition results: Different pipe network topological structures have little effect on recognition results, because the randomness of the extreme random forest method in PCA-ERF further improves the performance of the algorithm in finding the best data features. Therefore, the pipe network leakage detection method constructed using PCA-ERF is suitable for leakage detection tasks in multi heat source circular complex topology pipe networks.

In the future, the research group will further study the leakage detection problem of district heating pipelines, including applying the PCA-ERF method to actual heating pipelines for leakage detection; The thermal characteristics of district heating system with different topological structures and their impact on pressure and flow data; Apply other new machine learning algorithms to the leakage detection field.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank the anonymous reviewers for valuable comments. This work was supported by the Shandong Provincial Natural Science Foundation, China (Grant no. ZR2020ME189) and Jinan Scientific Research Leader Studio, China (Grant no. 202228105).

References

[1] 2022 Statistical Yearbook of Urban and Rural Construction, Ministry of Housing and Urban-Rural Development of the People’s Republic of China. 2024.3.1

[2] Zhou, S.J., O’Neill, Z., O’Neill, C. A review of leakage detection methods for district heating networks. Appl. Therm. Eng. (2018)137, 567–574

[3] Li Jian, et al. Review of leakage monitoring and quasi real-time detection technologies for long gas & oil pipelines. Chinese Journal of Scientific Instrument, 2016.08: 1747-1758.

[4] Zhou, S.J., et al. The model of leakage fault diagnosis for the pipe network of hot-water district heating. Journal of Shandong University (Engineering Science), 2013.08: 105-110.

[5] Liu Y., Tie Y., Na S., et al. Acoustic Signal Acquisition and Analysis System Based on Digital Signal Processor. International Conference on Energy and Environmental Protection,2012.

[6] Li G.J., et al. Modeling and Analyzing Leakage of Heating Pipe Network in Park. Journal of Northeastern University (Natural Science),2020.10: 1402-1409.

[7] Xiaofang Shan, Peng Wang, Weizhen Lu. The reliability and availability evaluation of repairable district heating networks under changeable external conditions. Applied Energy,2017:686-695.

[8] Jean Duquette, Andrew Rowe, Peter Wild. Thermal performance of a steady state physical pipe model for simulating district heating grids with variable flow. Applied Energy,2016:383-393.

[9] Keng X.Y., et al. Application of Wavelet Denoising in Leakage Signal Processing of Water Supply Network. Mechanical & Electrical Technology,2018.08: 10-12.

[10] Guan Y, Lv M, Dong S. Pressure-driven Background Leakage Models and their Application for Leak Localization Using a Multi-population Genetic Algorithm. Water Resources Management, 2023, 37(1): 359-373.DOI:10.1007/s11269-022-03377.

[11] Nasirian A., Maghrebi M F., Yazdani S. Leakage Detection in Water Distribution Network Based on a

New Heuristic Genetic Algorithm Model. Journal of Water Resource & Protection, 2013, 5(3): 294-303.DOI:10.4236/jwarp.2013.53030.

[12] Fei W, Haitao Z, Hanming F, et al. The Leakage Detection Method During Storage and Transportation Process Based on Improved M-SPRT. Journal of Chongqing University of Technology (Natural Science), 2018.

[13] Chi Z, Jiang J, Diao X, et al. Novel Leakage Detection Method by Improved Adaptive Filtering and Pattern Recognition Based on Acoustic Waves. International journal of pattern recognition and artificial intelligence, 2022(2):36. DOI:10.1142/S0218001422590017.

[14] Lu Z.Q., et al. Corrosion Leakage Prediction of Submarine Oil and Gas Pipeline Based on Neural Network for Parameter Optimization. Science Technology and Engineering,2022, 22(20): 150-155.

[15] Chen Z.Q., et al. Study on inverse calculation of leakage source intensity and location based on improved glowworm swarm optimization algorithms. Journal of Safety Science and Technology,2022,10(08):150-155.

[16] Quinones-Grueiro M, Milian M A, et al. Robust leak localization in water distribution networks using computational intelligence. Neurocomputing, 2021(5): 438.DOI: 10.1016/j.neucom.2020.04.159.

[17] Kang J, Park Y J, Lee J, et al. Novel Leakage Detection by Ensemble CNN-SVM and Graph-Based Localization in Water Distribution Systems. IEEE Transactions on Industrial Electronics, 2017, PP (99): 1-1. DOI:10.1109/TIE.2017.2764861.

[18] Zhou S. J., Liu X. R., Tian Y. S., et al. Multi-fault diagnosis of district heating system based on PCA_BP neural network. Process Safety and Environmental Protection 186 (2024) 301–317

[19] Rai, K., Hojatpanah F., Ajaei F. B., & Grolinger K. Deep learning for high-impedance fault detection: convolutional autoencoders. Energies, 2021, 14. DOI:10.3390/en14123623.

[20] Banik R, Biswas A. Enhanced renewable power and load forecasting using RF-XGBoost stacked ensemble. Electrical Engineering, 2024, 106(4): 4947-4967. DOI:10.1007/s00202-024-02273-3.

[21] Weiler V, Lust D, Brennenstuhl M, et al. Automatic dimensioning of energy system components for building cluster simulation. Applied energy, 2022(5): 313.DOI: 10.1016/j.apenergy.2022.118651.

[22] Zhou Z, Xia T, Ma J, et al. Transparent Grid Visualization Surveillance and Fault Warning in High Density Distributed Power Access Areas [J]. Applied Mathematics and Nonlinear Sciences, 2024, 9(1). DOI:10.2478/amns-2024-2564.

[23] Zhao L, Yin L. Knowledge-shareable adaptive deep dynamic programming for hierarchical generation control of distributed high-percentage renewable energy systems. Renewable Energy, 2024, 228.DOI: 10.1016/ j.renene.2024.120627.

[24] Anna Hoła, Sławomir Czarnecki. Random forest algorithm and support vector machine for nondestructive assessment of mass moisture content of brick walls in historic buildings. Automation in Construction. 149(2023)104793.

[25] Wei Gao, Fan Xu, Zhi-Hua Zhou. Towards convergence rate analysis of random forests for classification, Artificial Intelligence313(2022)103788.

[26] Mohamed N.A. Meshref, Seyed Mohammad Mirsoleimani Azizi, Wafa Dastyar, Rasha Maal-Bared, Bipro Ranjan Dhar. Low-temperature thermal hydrolysis of sludge prior to anaerobic digestion: Principal component analysis (PCA) of experimental data, Data in Brief. 38 (2021) 107323.

[27] J. Zhang, D. Zhou, M. Chen, Monitoring multimode processes: A modified PCA algorithm with continual learning ability, J. Process Control. 103 (2021)76-86.

[28] Lindner C, Bromiley P A, Ionita M C, et al. Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1862-1874.DOI:10.1109/TPAMI.2014.2382106.

[29] S. Razvarz, R. Jafari, C. Vargas-JarilloA, Gegov, M. Forooshani. Leakage Detection in Pipeline Based on Second Order Extended Kalman Filter Observer, IFAC 52-29 (2019) 116-121.

[30] S. Zhou, H. Li, P. Gong, M. Tian. Hydraulic modeling of double-source and ring-shaped heating networks, Appl. Therm. Eng. 119 (2017) 215-221.

[31] Yumi Deng, Xudong Cheng, Fang Tang, Yong Zhou. The control of moldy risk during rice storage based on multi-Variate linear regression analysis ant random forest algorithm. JUSTC,2022,52(1) 1-12.

[32] Rhodes J S, Cutler A, Moon K R. Geometry- and Accuracy-Preserving Random Forest Proximities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9):10947-10959.

[33] S. Dharumarajan, Thomas F. A. Bishop. Desertification status mapping in Mutuma Watershed by using Random Forest Model, Sciences in Cold and Arid Regions. 14(1)2022: 32−42.