Research Article

Research on an Oil Pipeline Anomaly Identification Method for Distinguishing True and False Anomalies

Lirong Chen,1,2 Dingxiong Wang,3 Hui Wang,1,4 Xiaodong Wu,1 Landing Gu,1 Fen Ma,1 Fei Xie,1 Muqun Feng,1 Zhaofeng Yin,1 Chao Liu,5 and Tiancheng Jin6

1Development and Research Center of China Geological Survey (National Geological Archives of China), Beijing 100037, China
2School of Earth Science and Resource, China University of Geosciences, Beijing 100083, China
3Meituan, Beijing 100083, China
4College of Geoscience and Surveying Engineering, China University of Mining & Technology, Beijing 100083, China
5No. 1 Oil Production Plant, Changqing Oilfield Company 716000, Yan’an, China
6College of Humanities and Social Science, Kangwon National University, Samcheok 25913, Republic of Korea

Correspondence should be addressed to Lirong Chen; chenlirong@mail.cgs.gov.cn

Received 21 April 2022; Accepted 27 June 2022; Published 12 August 2022

Academic Editor: Sebastian Podda

Copyright © 2022 Lirong Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Anomaly identification is important to ensure the safe and stable operation of oil pipelines and prevent leaks. Leak identification is performed to divide abnormal samples from normal oil transfer samples in monitoring data, and it is a dichotomous problem. However, the traditional machine learning binary classification method is no longer suitable for identifying leak anomalies in complex production environments. The main problem is that leaks in production environments are very rare, and traditional methods cannot directly identify the leaking pattern with their generalizability. The recognized normal pattern lacks the ability to adapt to dynamic environmental changes and an artificial adjustment of the pump frequency, instrument calibration, and other monitoring data mutations. These are known as false anomalies, and they are difficult to distinguish from true anomalies. This results in a lower recall rate for leak anomaly identification and a higher rate of false positives. To solve this problem, this study proposes a leak anomaly recognition method based on the distinction between true and false anomalies. A one-class SVM is used to learn the normal working mode of oil pipelines, and the model is used to screen out suspected pipeline anomalies, namely, true and false anomalies. It increases the morphological difference between true and false anomaly curves by superimposing multisource data and uses similarity clustering to discover anomaly patterns that indicate leak events. The results show that the leakage anomaly recall rate is 100%, and the false anomaly exclusion rate is 83.49%. This method achieves real-time and efficient monitoring of pipeline leaking events in complex production environments, and it is practical for the application of machine learning methods in production environments.

1. Introduction

Oil pipeline leak monitoring is a key step to ensure sustainable and safe production in oil fields. Accurate and timely monitoring of leaking events allows time for pipeline repair and can prevent environmental pollution, national property loss, and fire and explosion accidents caused by an oil leak [1–4].

With the rapid development of big data analysis, artificial intelligence, machine learning, and other technologies, abnormal event monitoring methods based on real-time data (such as pressure, flow, temperature, and other monitoring data) have gradually become used for pipeline leak event identification [5–8]. At present, there are several oil pipeline anomaly detection methods based on real-time monitoring data, such as (1) the volume and mass balance method, which diagnoses abnormal events by observing the degree of balance between volume and flow at both ends of the pipeline. When a leak accident occurs, the transport medium is lost, resulting in a large difference in monitoring
indicators at both ends of the pipeline. When the difference exceeds the threshold value, the accident is identified as the occurrence of the leak. However, this method requires highly accurate data, and a slight error leads to a false leaking event [9–12]. (2) In the pressure gradient method, leak monitoring is carried out by monitoring the gradient change of the pipeline pressure curve. In the actual pipeline transportation environment, the pressure gradient is affected by many factors, and its variation is nonlinear and uncertain, which makes it difficult to identify anomalies. At the same time, this method requires highly accurate data, and it is usually used as an auxiliary means to conduct joint monitoring with other methods [13, 14]. (3) The wavelet transform method involves digital signal filtering and denoising through the threshold to complete the task of fault identification. The advantage of this method is that it does not require highly precise input data from the signal, and its disadvantage is that it can easily identify leaks with a single threshold value, leading to a large number of false faults [15, 16]. (4) The negative pressure wave analysis method identifies abnormal events by monitoring the negative pressure wave generated by a pipeline leak. The advantage of this method is that it does not need a mathematical model and the amount of calculation required is small. However, as small slow leak events produce almost no negative pressure waves, the monitoring effect for small slow leak events is not ideal [17–19]. (5) The accident state simulation method records various pipeline monitoring data indicators at the time of an accident by simulating the leak accident, enabling the severity of the leak accident to be determined. However, this method has limitations and cannot identify the unsimulated leak events [20]. (6) The method based on the neural network makes use of the advantages of nonlinear fitting of the neural network method to model the common working conditions in the pipeline, classify different working conditions, and use the classification results from real-time data of the model to determine the running state of the pipeline [21]. This method has a strong ability to withstand harsh environments and noise interference [22–25], but it requires substantial manual label costs for leak identification, as well as skills and experience in labeling time series data. (7) The statistical method is based on the energy balance and mass balance equations. It also involves statistical modeling of the oil pipeline monitoring data to determine the characteristic model and the state estimation model of the oil pipeline. Then, these models are used to diagnose the real-time state of the oil pipeline [19]. However, due to the diverse causes of pipeline leak accidents, it is easy to ignore some potential variable characteristics when general statistical modeling methods are used, and false positives are often generated in complex production environments [21]. (8) The anomaly detection method is based on machine learning. This method uses the machine learning method to model oil pipeline monitoring data and then uses the model to detect anomalies from real-time oil pipeline data. The deep learning method models oil pipeline monitoring data and then uses the model to classify and identify the state of the pipeline to determine the pipeline’s running state. This method has high timeliness and accuracy for anomaly detection, but it requires highly complete data samples. As abnormal samples are generally rare in reality, the constructed model is prone to a precision offset [22–25].

At present, with the upgrading of monitoring equipment, leak monitoring focuses on acoustic and infrared (IR) signal anomaly processing in the technical research of liquid pipeline transportation leak monitoring [26–28]. For old industrial oil production well sites, if new technology is to be applied to improve the leak monitoring effect, the monitoring equipment needs to be upgraded first, and the traditional pressure monitoring equipment needs to be replaced by acoustic and infrared monitoring equipment, which requires huge costs and may lead to a serious imbalance between output and income. At present, there are basic monitoring equipment and pipeline pressure monitoring equipment in the old well pads of oil fields, and upgrading monitoring technology based on existing equipment is an important direction of research. At present, the abnormal leak monitoring of pressure data focuses on the application innovation of machine learning methods. For example, Waleed et al. [29] use a neural network algorithm to extract the pressure characteristics of a 0.203 m transportation pipeline under normal oil transportation conditions. Xu et al. [30] use a support vector machine algorithm to deal with a leak signal noise. Yeo et al. [28] use a depth self-coding algorithm to extract the leakage-free state features of transportation pipelines. These methods make use of the advantages of machine learning, which is independent of data distribution and good for nonlinear fitting [31–35]. However, in the production environment, the pressure monitoring data may appear abnormal due to manual calibration of instruments and manual pressurization of oil pumps, which may lead to a large number of false alarm leaks in the leak monitoring model, thus reducing the effectiveness of the leak alarms.

At present, leakage event identification is oriented at two methods, one of which is for leakage identification and leakage type identification (such as pitting corrosion, man-made damage, and pipeline rupture caused by earthquake), and the other is for false anomaly identification of suspected leakage (such as sudden change of pressure curve caused by instrument calibration, instrument replacement, and pressure frequency adjustment). In the production environment, “misstatement rather than omission” is required as per regulations, which leads to a serious shortage of collected leakage data and a high false alarm rate of leakage events.

In production environments, monitoring methods based on statistics and machine learning have become the main development direction for pipeline leak monitoring, but they have several problems: (1) there is a large amount of oil pipeline monitoring data, but abnormal samples are rare. Normal patterns identified by traditional machine learning methods lack adaptability to environmental changes, and anomaly patterns lack generalization. (2) Oil pipeline anomalies have a variety of modes, mainly those caused by manual operations and accidents. In industrial production environments, anomaly patterns are complex and difficult to distinguish. Moreover, when the application of machine learning methods is transferred from the experimental
expressed as follows:

\[
\text{the red box represented a real leak was confirmed that these fluctuations were caused by the number of violent fluctuations. After field investigation, it pipeline pressure curve were basically stable with a small collection time. As shown in Figure 1, the fluctuations in the pressure in megapascals, and the abscissa presented the data collected over half a month. The ordinate presented the million pressure test records. Pressure monitoring data had collected from the 19 pipelines was approximately 22.8 environment generates approximately 1.2 million pressure collection frequency of 2 seconds. Each pipeline in this environment generates approximately 1.2 million pressure monitoring records per month, and thus, the data volume collected from the 19 pipelines was approximately 22.8 million pressure test records. Pressure monitoring data had collected over half a month. The ordinate presented the pressure in megapascals, and the abscissa presented the data collection time. As shown in Figure 1, the fluctuations in the pipeline pressure curve were basically stable with a small number of violent fluctuations. After field investigation, it was confirmed that these fluctuations were caused by manual operations. The red box represented a real leak exception. The process of collecting leak event involved manual inspection. The pipelines were inspected once every half day. Once a leak event was discovered, the inspector checked the abnormal fluctuation of the pressure curve of the pipeline and marked the abnormality in the pressure curve by visual interpretation.}

2. Materials and Methods

2.1. Data. The data in this study were collected from a real oil production environment in northwest China. The data used were pressure monitoring data from 19 oil pipelines, which were collected over a time span of 1 month and a data collection frequency of 2 seconds. Each pipeline in this environment generates approximately 1.2 million pressure monitoring records per month, and thus, the data volume collected from the 19 pipelines was approximately 22.8 million pressure test records. Pressure monitoring data had collected over half a month. The ordinate presented the pressure in megapascals, and the abscissa presented the data collection time. As shown in Figure 1, the fluctuations in the pipeline pressure curve were basically stable with a small number of violent fluctuations. After field investigation, it was confirmed that these fluctuations were caused by manual operations. The red box represented a real leak exception.

The process of collecting leak event involved manual inspection. The pipelines were inspected once every half day. Once a leak event was discovered, the inspector checked the abnormal fluctuation of the pressure curve of the pipeline and marked the abnormality in the pressure curve by visual interpretation.

2.2. Time Series Data Preprocessing Method. Pipeline pressure data are time series data. Time series data are a series of numbers [29] that are arranged according to the sequence of time, and the interval of the numbers is consistent with the frequency of data collection. Assuming that the length of the timing sequence \( s \) is \( m \), the timing sequence data can be expressed as follows:

\[
S = \{<s_1, t_1>, <s_2, t_2>, \ldots, <s_m, t_m>\},
\]

\(<s_i, t_i> <s_m, t_m>\) represents the monitoring value \( s_m \) at moment \( t_m \). If there is only one observed object, then \( s_m = x_m \). If there are \( n \) observed objects, then \( s_m = x_1, x_2, \ldots, x_n \), which represents the vector that records the values of \( n \) observed objects at time \( t_m \).

2.2.1. Sliding Time Window Method. The main purpose of time series data analysis is to find the rule that describes how the observation value changes with time. Generally, the sliding time window method is adopted [36], which intercepts and analyzes time series data fragments and helps to find the regularities in the observed values through feature extraction of fragments. The most important parameter in the sliding time window method is setting the length of the time window. The process of intercepting data fragments with this length is called time series slicing, and this directly affects the feature extraction effect of time series data [37–39].

The time sliding window method adopts the idea of time series piecewise dimension reduction. It divides the time series into an equal-length subseries, extracts the features from each subseries, and replaces the original subseries with the features to achieve data dimension reduction [40]. The temporal sliding window method has the characteristics of dimensionality reduction, morphological retention, and noise removal. It is easy to implement due to its low time complexity. It is also widely used for the feature extraction of temporal data [41].

2.2.2. Data Loss Processing. Time series data in the production environment are incomplete and discontinuous. Thus, the data need to be cleaned to meet the modeling requirements. For small amounts of discrete data or short-time sequence, fragmental data, and discontinuous data, interpolation is adopted [42]. If the data were not continuous over a long period, the data would be deleted.

2.3. Leakage Anomaly Identification Method Based on True and False Anomaly Differentiation. In this study, leak anomaly identification was performed by taking pressure monitoring data as the research object. This was mainly done to distinguish between true and false anomalies in complex production environments. A true anomaly was an abrupt change in the pressure curve caused by tubing leakage, and a false anomaly was an abrupt change in the pressure curve caused by an artificial operation (such as an artificial adjustment of the pump frequency, instrument calibration, and pipeline replacement). The traditional binary classification method could not handle the complexity of production environments because of the similar curve mutation patterns between true and false anomalies. The idea adopted in this method was to distinguish anomalies layer by layer. First, a class of the support vector machine was used to classify
normal events and abnormal events, and then, the hierarchical clustering method was used to distinguish between true and false anomalies. Finally, a multisource data joint analysis method was adopted to further eliminate the false anomalies that were not identified in the previous step. A flow chart of the method is presented in Figure 2.

2.3.1. One-Class SVM. Anomaly recognition is a dichotomous problem that aims to divide data samples into normal events and abnormal events. Among the many machine learning methods, the support vector machine (SVM) [43], neural network [44], and logistic regression [45] methods can solve binary classification problems. In production environments, the model must be easy to operate, have few parameters, require a small amount of calculation, and have a fast processing speed to ensure that it can handle massive amounts of complex real-time data. The support vector machine method is more suitable for dichotomous problems in production environments than other methods because it has fewer parameters and less dependence on manual experience during the parameter tuning process. However, the support vector machine method still has a deficiency; it cannot deal with the uneven number of positive and negative samples in production environments [46]. Generally, anomaly data are difficult to collect, the acquisition costs are high, and the occurrence of abnormal events is rare. Thus, the numbers of normal events and abnormal events are very different. This means that the support vector machine method is prone to large errors in the construction of a classification model.

To solve this problem, a support vector machine (one-class SVM, OCSVM) [47] algorithm is proposed. Its purpose is to take zero points as negative sample points and other data as positive sample points and construct a hyperplane between positive samples and zero points, which is the maximum distance from the origin. OCSVM solves the problem of having unbalanced positive and negative sample numbers, and it is widely used in the field of anomaly detection [48].

We suppose that the given training set in space is as follows:

$$D = \{(x_1, y_1), \ldots, (x_m, y_m)\}, \quad (2)$$

where instance $x_i, X_i$ belongs to the input space, $x_i \in X = \mathbb{R}_m$. There are two types of $y_i \in Y = \{-1, +1\}, \quad i = 1, 2, \ldots, m$ for corresponding tags. The total number of samples is $m$.

OCSVM finds a partition hyperplane in the sample space based on the positive sample training set $D$ and divides the categories that are different from the positive samples and zeros. In the sample space, OCSVM defines the partition hyperplane as follows:

$$w^T \varphi(x) - \rho = 0, \quad (3)$$

where $\varphi(x)$ is the nonlinear mapping function and $W = W_1; W_2; W_3; \ldots; W_d$ is the normal vector, which determines the direction of the hyperplane, $\rho$ is the intercept from the hyperplane to the origin, or the offset.

In this study, normal pressure fluctuated steadily and regularly, while abnormal pressure fluctuated sharply and violently. We divided the oil pipeline pressure samples into normal pressure samples and abnormal pressure samples. The one-class SVM first learned the characteristic model of normal pressure in the training sample, and then was able to identify the category of the test samples by calculating the level of similarity between the test samples and the characteristic model. If similar, the test samples were classified as normal; otherwise, they were deemed to be an anomaly.

2.3.2. Hierarchical Clustering Algorithm. The exceptions obtained by the one-class SVM method include true and false exceptions. In the production environment, abnormal oil pipeline events include abnormal manual operation and abnormal leak events. Abnormal artificial operation events (such as pressurization and pump stop) does not produce oil leaks and should not be identified as false anomalies. Accident anomalies (including leakage and pipe explosion) are the true anomalies that need to be identified. Here, we needed to identify the leak anomaly pattern to effectively distinguish between true and false exceptions.

Due to the difference in true and false anomalies in the pressure curve, to further distinguish differences between the two, the characteristics of tubing leak events needed to be determined. We used the hierarchical clustering algorithm to exclude false anomaly events by looking for samples with similar characteristics to real leak events. In general, the distance calculation formula was used to calculate the level of similarity between samples.

The hierarchical clustering algorithm [49] calculated the distance between the pairs of samples and merged two samples with distances of less than the threshold into clusters. This combination was repeated until the number of clusters were less than the preset number. Thus, $x_i$ and $x_j$ were any two clusters, and the level of similarity between the two clusters was calculated by the Euclidean distance formula as follows:

$$\text{dist}(x_i, x_j) = \|x_i - x_j\|^2 = \sum_{u=1}^{n} |x_{iu} - x_{ju}|^2, \quad (4)$$

where $n$ is the number of clusters.

2.3.3. Joint Analysis Anomaly Identification Method. Monitoring data from a single data source are expressed similarly on curves with true and false anomaly events. For example, adjusting the pump frequency can cause abnormal fluctuations in the pressure curve data, which is similar to what can be observed in a leak anomaly curve. This is a false exception event and can easily be classified as a true exception event. At this point, additional information is needed to enlarge the distinction between true and false exception events. In this study, we used oil pump frequency data to exclude false anomalies in the pressure data.

Since variance can reflect the degree of data deviation from the mean value, the variance of the oil pump frequency
value in the time period when abnormal events occur is used as the basis for determining false anomalies [50, 51]. When the variance is 0, the abnormality has nothing to do with the adjustment of the oil pump frequency; when the variance is not 0, the anomaly is a false anomaly caused by the adjustment of the oil pump frequency. The variance is calculated by the following formula:

\[ \sigma^2 = \frac{\sum (X - \mu)^2}{N} \]  \hspace{1cm} (5)

In the formula, \( X \) is the oil pump frequency sample in the abnormal event period, \( \mu \) is the mean value of this sample, and \( N \) is the total value of this sample. To further reduce the anomaly false-positive rate, anomalies with an oil pump frequency value variance that is not 0 are regarded as false abnormal events.

2.4. Evaluation of the Model’s Performance. In a production environment, managers prefer a method with a high false alarm rate and without missed alarms when choosing a leak identification method because leaks can cause serious economic losses. It is better to make a false alarm than to miss a leak. In such a case, accuracy and recall were more reflective of the actual effectiveness of the model in a production environment than other metrics.

According to the true category and the prediction category, the test samples were divided into true positive (TP), false-positive (FP), true negative (TN), and false-negative (FN) cases [52]. The confusion matrix of the classification results is listed in Table 1.

We evaluated the performance of the model by calculating the accuracy, recall rate, and \( F1 \) value [44]. The precision ratio \( P \), recall ratio \( R \), and \( F1 \) values were calculated as follows:

\[ P = \frac{TP}{TP + FP} \]  \hspace{1cm} (6)

\[ R = \frac{TP}{TP + FN} \]

The precision \( P \) represents the number of actual abnormal samples among the samples with predicted results that are anomalies. The recall ratio \( R \) represents the number of samples with predicted anomalies in actual abnormal samples, which measures the ability of the classifier to identify anomalies. \( F1 \) is the harmonic mean of the
“precision and recall ratio” and reflects a comprehensive score of the precision and recall ratio.

3. Results

This experiment included 19 records of pipeline data, and the data are approximately 22.8 million pressure detection records. We took 80% of the data, approximately 18.24 million records as the test data. The 20%, approximately 4.56 million pieces of data, were used as test data. With regard to verification, to ensure the effectiveness of the model, we used the real-time data of the production environment for direct verification.

3.1. Experiment 1: Analysis of Abnormal Event Detection Results. Before learning and training the pressure data using the OCSVM method, we needed to mark the samples with normal pressure fluctuations. We marked the data with stable pressure fluctuations as normal samples and with abrupt pressure fluctuations as abnormal samples, as shown in Figure 3. Blue represented normal samples and red represents abnormal samples. Normal samples were used as input data for training and learning of the OCSVM.

We conducted experimental tests on different sliding window sizes. The time window sizes were set from 80 to 1200 seconds, and a group of experiments was conducted every 40 seconds. A total of 28 training sessions were conducted, and the experimental results are listed in Table 2.

The results showed that the accuracy and recall rate of the prediction results changed very smoothly when the time window size increases from 80 to 1200 seconds. The prediction accuracy of abnormal samples was stable at 0.90, and the recall rate was stable at 0.87. This indicated that the OCSVM had a good effect on anomaly recognition. At the same time, the data from this month included real leak data, and the abnormal pressure in this leak period was recalled by the OCSVM. For real leaks, the recall rate was 100%. In terms of the timeliness of the anomaly alarm, the experimental results showed that the performance of the OCSVM was very stable for different sizes of time windows ranging from 80 to 1200 seconds. Anomalies were detected within 80 seconds of their occurrence, and the response time of the model to anomalies could be less than 3 minutes, indicating a high level of timeliness in the model.

Based on the same data, other machine learning algorithms were used to distinguish normal samples from abnormal samples, and the results were compared. The comparison algorithms were artificial neural networks (ANNs) and random forests (RFs), which were widely used. The selection of algorithm parameters was based on the existing research. We compared these two algorithms with OCSVM. The comparison results are listed in Table 3. OCSVM was better than the others.

3.2. Experiment 2: Analysis of Anomaly Pattern Recognition of Accidents. Anomalies mainly included leak anomalies and manual operation anomalies. The purpose of anomaly pattern differentiation was to identify the difference between anomalies caused by accidents and manual operations.

We used the hierarchical clustering algorithm to classify the anomaly patterns. The anomaly samples contained a real leak incident. Since the number of anomaly patterns was unknown, the clustering trend graph method was used to determine the optimal number of clusters. When the number of clusters increased and the number of anomaly samples in the cluster where real leakage anomalies were located tended to be stable, the number of clusters was optimal.

As shown in Figure 4, when the cluster number was 10, the proportion of abnormal samples was stable at 20% of the total number. Thus, 10 was the optimal cluster size.

Figure 5 shows the hierarchical clustering result graph divided into 10 categories, where the red arrow points to a real leak anomaly sample. We retained clusters containing leak exception events because they reflected the leak exception pattern. The other nine types of exceptions were discarded. As a result, we reduced the anomaly false-positive rate and we were able to discard 79.6% of the abnormal samples.

Figure 6 shows the visualization result of anomaly pressure patterns for a period of one month, which shows the occurrence frequencies of 10 types of anomaly patterns within a month. The anomalies in the red box were the real abnormal leak events that were identified. In the production environment, if the real-time data contained the characteristics of a leak anomaly pattern, the monitoring system program would send an alarm. The false anomaly alarm rate was reduced to 20.4% with the use of the hierarchical clustering analysis algorithm.

3.3. Experiment 3: Joint Analysis of Anomaly Pattern Results. We aligned the pressure data with the pump frequency data. The alignment operation mainly involved keeping the time scales of two sets of data consistent, as shown in Figure 6. The figure above presented the pressure monitoring curve, and the figure below presented the oil pump frequency curve. Since manual operation of the oil pump led to abnormal fluctuations in pressure and abrupt changes in the oil pump frequency curve mainly came from manual operations, there was a great correlation between pressure data and abrupt changes in the oil pump frequency curve, as shown in Figure 7.

At this point, we used the abnormal oil pump frequency to eliminate some false anomalies in the pressure data. First, according to the method described in Section 2.3, to identify the oil transfer pump frequency anomalies, the anomalies found in pressure data at the same time as the periods of frequency anomalies were classified as false anomalies, and thus, they were eliminated. The experimental results showed

<table>
<thead>
<tr>
<th>Table 1: Confusion matrix.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real situation</td>
</tr>
<tr>
<td>Anomaly</td>
</tr>
<tr>
<td>Anomaly</td>
</tr>
<tr>
<td>Normal</td>
</tr>
</tbody>
</table>

3. Results

This experiment included 19 records of pipeline data, and the data are approximately 22.8 million pressure detection records. We took 80% of the data, approximately 18.24 million records as the test data. The 20%, approximately 4.56 million pieces of data, were used as test data. With regard to verification, to ensure the effectiveness of the model, we used the real-time data of the production environment for direct verification.

3.1. Experiment 1: Analysis of Abnormal Event Detection Results. Before learning and training the pressure data using the OCSVM method, we needed to mark the samples with normal pressure fluctuations. We marked the data with stable pressure fluctuations as normal samples and with abrupt pressure fluctuations as abnormal samples, as shown in Figure 3. Blue represented normal samples and red represents abnormal samples. Normal samples were used as input data for training and learning of the OCSVM.

We conducted experimental tests on different sliding window sizes. The time window sizes were set from 80 to 1200 seconds, and a group of experiments was conducted every 40 seconds. A total of 28 training sessions were conducted, and the experimental results are listed in Table 2.

The results showed that the accuracy and recall rate of the prediction results changed very smoothly when the time window size increases from 80 to 1200 seconds. The prediction accuracy of abnormal samples was stable at 0.90, and the recall rate was stable at 0.87. This indicated that the OCSVM had a good effect on anomaly recognition. At the same time, the data from this month included real leak data, and the abnormal pressure in this leak period was recalled by the OCSVM. For real leaks, the recall rate was 100%. In terms of the timeliness of the anomaly alarm, the experimental results showed that the performance of the OCSVM was very stable for different sizes of time windows ranging from 80 to 1200 seconds. Anomalies were detected within 80 seconds of their occurrence, and the response time of the model to anomalies could be less than 3 minutes, indicating a high level of timeliness in the model.

Based on the same data, other machine learning algorithms were used to distinguish normal samples from abnormal samples, and the results were compared. The comparison algorithms were artificial neural networks (ANNs) and random forests (RFs), which were widely used. The selection of algorithm parameters was based on the existing research. We compared these two algorithms with OCSVM. The comparison results are listed in Table 3. OCSVM was better than the others.

3.2. Experiment 2: Analysis of Anomaly Pattern Recognition of Accidents. Anomalies mainly included leak anomalies and manual operation anomalies. The purpose of anomaly pattern differentiation was to identify the difference between anomalies caused by accidents and manual operations.

We used the hierarchical clustering algorithm to classify the anomaly patterns. The anomaly samples contained a real leak incident. Since the number of anomaly patterns was unknown, the clustering trend graph method was used to determine the optimal number of clusters. When the number of clusters increased and the number of anomaly samples in the cluster where real leakage anomalies were located tended to be stable, the number of clusters was optimal.

As shown in Figure 4, when the cluster number was 10, the proportion of abnormal samples was stable at 20% of the total number. Thus, 10 was the optimal cluster size.

Figure 5 shows the hierarchical clustering result graph divided into 10 categories, where the red arrow points to a real leak anomaly sample. We retained clusters containing leak exception events because they reflected the leak exception pattern. The other nine types of exceptions were discarded. As a result, we reduced the anomaly false-positive rate and we were able to discard 79.6% of the abnormal samples.

Figure 6 shows the visualization result of anomaly pressure patterns for a period of one month, which shows the occurrence frequencies of 10 types of anomaly patterns within a month. The anomalies in the red box were the real abnormal leak events that were identified. In the production environment, if the real-time data contained the characteristics of a leak anomaly pattern, the monitoring system program would send an alarm. The false anomaly alarm rate was reduced to 20.4% with the use of the hierarchical clustering analysis algorithm.

3.3. Experiment 3: Joint Analysis of Anomaly Pattern Results. We aligned the pressure data with the pump frequency data. The alignment operation mainly involved keeping the time scales of two sets of data consistent, as shown in Figure 6. The figure above presented the pressure monitoring curve, and the figure below presented the oil pump frequency curve. Since manual operation of the oil pump led to abnormal fluctuations in pressure and abrupt changes in the oil pump frequency curve mainly came from manual operations, there was a great correlation between pressure data and abrupt changes in the oil pump frequency curve, as shown in Figure 7.

At this point, we used the abnormal oil pump frequency to eliminate some false anomalies in the pressure data. First, according to the method described in Section 2.3, to identify the oil transfer pump frequency anomalies, the anomalies found in pressure data at the same time as the periods of frequency anomalies were classified as false anomalies, and thus, they were eliminated. The experimental results showed
Table 2: OCSVM results in an 80-second experimental window.

<table>
<thead>
<tr>
<th>Size of window (s)</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>0.90</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td>200</td>
<td>0.90</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td>400</td>
<td>0.91</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>600</td>
<td>0.91</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td>800</td>
<td>0.90</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td>1000</td>
<td>0.90</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td>1200</td>
<td>0.90</td>
<td>0.86</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 3: Comparison results of various algorithms.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>0.89</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>RF</td>
<td>0.90</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td>OCSVM</td>
<td>0.91</td>
<td>0.88</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Figure 3: Data marking results.

Figure 4: Clustering trend diagram.

Figure 5: Hierarchical clustering result diagram.

Figure 6: Classification of abnormal events.
that this method effectively eliminated artificial false anomalies caused by the operation of the oil transfer pump, reducing the content of false anomalies to 19.04% and retaining the real leak events. The anomaly recognition recall rate was 100%.

In conclusion, the method of anomaly identification based on true and false anomaly differentiation proposed in this study could identify anomalies related to leak events in complex production environments and could achieve a good application effect. The leak anomaly recall rate reached 100%, and the false anomaly exclusion rate reached 83.49%.

3.4. Experiment 4: Sensitivity Analysis and the Minimum Leak Detectable Rate. To improve the practicability of this method for general application, the modified Morris screening method was used to analyze the parameter sensitivity of this method. The false anomaly exclusion rate was used as the objective function. We perturbed the model parameters with the convergence accuracy (0.1, 0.01, 0.001, 0.0001), the number of sliding windows (80, 200, 400, 600), the number of clusters (5, 10, 15, 20, 25), and the number of leaf nodes (2, 3, 4, 5, 6) of one-class support vector machines and obtained the sensitivity distribution of the model.

As shown in Figure 8 and Table 4, the most sensitive parameters to the false anomaly elimination rate were the number of clusters, while the convergence accuracy of OCSVM, the number of sliding windows, and the number of hierarchical clustering leaf nodes were less sensitive. Therefore, in practical applications, it was necessary to focus on the selection of the number of clusters.

We monitored the minimum leak detectable rate of the method. In the experimental environment, the identification ability of this method was tested by artificially changing the size of pipeline leakage holes, and the minimum leakage rate of leakage events identified by this method was calculated. First of all, small leaks in the pipeline were made artificially, and liquid was transmitted normally. Besides, whether this method can identify leakage events normally was tested. If it could identify the abnormal leakage, the leakage rate at this time should be calculated; if it cannot identify, the size of the leakage hole should be enlarged and the identification ability of the method should be tested again. This cycle was carried out to find the minimum leakage rate that can be identified by this method. The leak hole was tested by the weighing method. When the ambient temperature was constant (20°C), the initial value $m_0$ of the standard leak hole was tested, and the end value $m_1$ of the leak hole was tested for a certain time $t$ (after 24 h), then the leakage rate $Q$ of the leak hole was calculated as follows:

$$ Q = \frac{m_0 - m_1}{t}. $$

(7)

During the test, the error was eliminated by taking the average value of multiple measurements. In order to reduce the error, the ambient temperature (generally 20°C) needed to be controlled during the test, and the test space was a closed atmospheric static environment. The measurement results are recorded in Table 5.

As can be seen from Table 5, the minimum leak detectable rate by this method was 0.43.

4. Discussion

In anomaly detection, it is often difficult to collect anomaly data, and the acquisition costs may be too high, resulting in imbalanced data categories, which makes the two major classification methods prone to deviations and difficult to apply effectively. In this paper, a classifier is introduced and combined with time series pressure monitoring data to
construct an abnormal pressure monitoring model. This solves the problem of sample imbalance caused by rare abnormal samples in anomaly detection and can achieve real-time and efficient abnormal pressure monitoring of oil pipelines.

Considering the diversity of anomaly patterns, this study divides the anomaly patterns of oil pipelines into two types according to the characteristics of the anomaly generation mechanism: accident anomalies and artificial anomalies. Two sets of anomaly pattern recognition models are constructed to identify accident anomaly patterns and the artificial anomaly model, which further reduces the anomaly false-positive rate.

The prediction accuracy and the recall rate of the OCSVM model are found to be stable. The prediction accuracy of abnormal samples is stable at 0.96, and the recall rate is stable at 0.85. The OCSVM model can capture most pressure anomalies in a short period of time. Through cluster analysis, anomaly patterns related to leak events are obtained, and false abnormal events without leaks are excluded to ensure that the alarm rate is reduced to 21.4%. By adding the oil pump frequency, pressure anomalies caused by manual operation of the oil pump are excluded, reducing the alarm rate of false anomalies.

Considering the characteristics and processing difficulties associated with engineering data (nonexperimental data) from oil pipelines, this method makes full use of the advantages of machine learning methods, such as the support vector machine and hierarchical clustering, to effectively identify abnormal events in pipelines in the production environment. The identification results obtained with this method are stable, and the operation and maintenance costs are low in practical applications. The method is applied to an oilfield production line. Machine learning method innovation has always been difficult in practical applications mainly because practical application environments are more complex than laboratory environments, and general methods cannot be directly applied to multistate event prediction. Because many states are similar in terms of data representation but have different causes in practice, it is difficult to identify anomalies using data representation differences. This method solves this problem by integrating the advantages of multiple machine learning methods and provides ideas for the combination of machine learning method innovations and industrial applications.

This method mainly studies and analyzes data in a single region, and multiregion joint analysis is also very important, as it allows for the analysis of common characteristics of abnormal events. At the same time, for multisource data joint analysis, more data sources mean that more anomaly information is involved, and a fuller understanding of abnormal events is achieved.

However, the use of multisource data also means that more interference information is involved, which leads to more difficult data analysis and modeling. We plan to expand multiregion joint analysis and multisource data joint analysis methods in the future to make the anomaly recognition model more stable, robust, and extensive.

### 5. Conclusions

To upgrade the stock technology of well sites in oil fields, the focus of this study is to distinguish true anomalies from false ones of tubing leak anomalies. Because there are few leak anomalies in the production environment, traditional methods cannot directly identify leak anomaly patterns with generalizability, while the identified normal patterns lack adaptability to dynamic changes in the environment. Moreover, the sudden changes in monitoring data caused by artificial adjustment of pump frequency and instrument calibration, such as false anomaly, are difficult to distinguish from true anomaly. This leads to a lower recall rate and a higher false alarm rate of leak anomaly identification. To solve this problem, this study proposes a leak anomaly identification method based on the distinction between true and false anomalies. One-class SVM is used to learn the normal working mode of the oil pipeline, and this mode is used to screen out the suspected anomalies of the pipeline, or the true and false anomalies. The curve morphological differences of true and false anomalies are increased by superimposing multisource data, and anomaly patterns of leak events are found by similarity clustering. This method has been validated in an oil transportation production environment, and the results show that the recall rate of leak anomaly identification is 100%, while the removal rate of false anomalies is 83.49%. This method realizes real-time and efficient monitoring of abnormal oil pipeline leak events in complex production environments and provides practical ideas for the application of machine learning methods in production environments.

This method is mainly used in the old industrial oil production well site with old and insufficient monitoring equipment, which is its advantages and disadvantages as well. In factories with more sophisticated and well-distributed monitoring instruments, our method does not have an advantage over other methods. However, in factories where equipment is in poor conditions and there is no money to buy new equipment, our method is very applicable. For example, the negative pressure wave method needs monitoring data at both ends of the pipeline, and the data should come in high accuracy, so that leakage identification can be effective. However, this method is not limited by this and can deal with the situation of insufficient monitoring equipment, low data accuracy, and lack, which provides ideas for upgrading existing technology.

There are still some deficiencies in this study. For example, this method lacks the ability to identify abnormal oil pipe dripping events. The main reason for this is that dripping events may not cause sudden changes in the pressure curve, except for small changes in the pressure curve. Therefore, abnormal events are mixed with normal

<table>
<thead>
<tr>
<th>No.</th>
<th>$M_0$ (g)</th>
<th>$M_1$ (g)</th>
<th>Time (h)</th>
<th>Q (g/24 h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>380.23</td>
<td>370.81</td>
<td>24</td>
<td>0.39</td>
</tr>
<tr>
<td>2</td>
<td>370.37</td>
<td>359.90</td>
<td>24</td>
<td>0.43</td>
</tr>
<tr>
<td>3</td>
<td>364.05</td>
<td>353.01</td>
<td>24</td>
<td>0.46</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td>0.43</td>
</tr>
</tbody>
</table>
events, making it difficult to be caught. At the same time, the amount of data does not support the classification of leak causes due to the low number of leak events in the production environment. In the future, we can study multi-source fusion methods and improve the sensitivity of traditional machine learning algorithms to the identification of oil pipe dripping events by observing oil pipe dripping events in multiple dimensions to realize the identification of abnormal oil pipe dripping events [53].

Data Availability
The data are freely available at https://github.com/Oshima-Ryu/oil-data.

Conflicts of Interest
The authors declare that they have no conflicts of interest.

Authors’ Contributions
L.C. conceptualized the study. L.C. and D.W. designed the methodology. H.W. and T.J. performed the software analysis. L.C., M.F., and H.W. performed the validation. L.C., C.L., and T.J. wrote the original draft. X.W. and F.X. reviewed and edited the manuscript. L.G., Z.Y., and F.M. performed the visualization.

Acknowledgments
This research was supported by the Geological Survey Project of China Geological Survey (no. DD20190392).

References


