Research on abnormal data identification of Internal inspection data in natural gas pipelines

Xinjie Du[1] Lei Mou[2,*] Changchao Qi[1] and Yin Qing[3]

[1] *Safety, Environment & Technology Supervision Research Institute, Southwest Oil & Gasfield Company of Petrochina, Chengdu, China*

[2] *Petroleum Engineering School, Southwest Petroleum University, Chengdu, China*

[3] *Mathematics School, Chengdu University of Information Technology, Chengdu, China*

Abstract

Abnormal data often exist in the internal inspection data of natural gas gathering and transmission pipelines, which has a significant impact on the internal corrosion assessment of pipelines. However, there are many different algorithms for identifying abnormal data, and their recognition effects are also different. Therefore, firstly, the abnormal data in the internal detection data of natural gas gathering and transmission pipelines was analyzed. Secondly, several widely used anomaly data recognition algorithms were selected for comparison, namely: Box plot, k-Nearest Neighbor (KNN), Local Outlier Factor (LOF), and Isolation Forest (IForest). These algorithms were applied to identify abnormal data within the internal detection datasets of natural gas gathering and transmission pipelines. A comparative analysis was conducted to determine the optimal recognition performance exhibited by each algorithm, aiming to identify the most effective method for detecting anomalies in this specific domain. The results show that for nearly 80,000 sets of pipeline internal inspection data, the KNN algorithm had the best recognition effect, and was able to effectively identify discrete or abnormal data.

Keywords

pipeline internal inspection, abnormal data, algorithm selection, normal distribution, K-Nearest Neighbor (KNN)

1. Introduction

It is often necessary to analyze the inspection data in natural gas gathering and transportation pipelines to calculate the corrosion rate or distribution of corrosion defects in the pipeline, and to evaluate the corrosion situation in the pipeline. However, due to detection sensor failure, damage or human negligence, there will be abnormal data in the detection data in the natural gas gathering and transportation pipeline. The existence of abnormal data will inevitably lead to an increase in data analysis errors and have an important impact on the assessment of corrosion in pipelines. Therefore, how to identify

ICCIC 2024: International Conference on Computer and Intelligent Control, June 29–30, 2024, Kuala Lumpur, Malaysia

* Corresponding author.

duxinjie@petrochina.com.cn (X. Du); 202121000877@stu.swpu.edu.cn (L. Mou); qichangchao@petrochina.com.cn (C. Qi); qingyin@cuit.edu.cn (Y. Qing)

0009-0009-8552-3364 (X. Du); 0009-0002-0153-4656 (L. Mou); 0009-0005-2977-0628 (C. Qi); 0009-0001-9377-0412 (Y. Qing)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

and remove abnormal data is the primary issue in assessing corrosion in natural gas gathering and transportation pipelines.

Many scholars from various industries have applied various algorithms for identifying and removing abnormal data from sample datasets. Alhussein [1] used the DBSCAN algorithm to identify and detect abnormal data in airport terminals. Abid [2] used the DBSCAN algorithm to identify and detect outliers in sensor detection data. Gu [3] and Osman [4] applied box plots to identify abnormal data in sensor data. Mohiuddin [5] monitored abnormal values in daily stock trading information and found that the LOF algorithm performed the best. Li [6] and Mansoor [7] et al. used an improved SM-Iforest algorithm to identify and detect outliers in machine monitoring data and IoT sensor data. He [8] combined the Grubbs's criterion with the KNN algorithm and proposed a method for detecting network traffic outliers with high accuracy.

The research of abnormal data in pipeline detection is also an active research field in the world, especially in the improvement and application of algorithms. For example, Smith [9] proposed an anomaly detection method based on reinforcement learning, which improved the accuracy and robustness of detection by adaptively adjusting the parameters of the detection model. Jones [10] applied big data analysis technology to process and analyze massive pipeline inspection data in real time, which significantly improved the speed and accuracy of outlier recognition. Mohamed [11] studied the anomaly detection method based on convolutional neural network (CNN) and applied it to industrial pipeline data, achieving good results. Garcia [12] combined genetic algorithm with fuzzy logic to propose a new abnormal data detection method, which effectively improved the sensitivity and specificity of detection. Zhang [13] proposed a novel anomaly detection model by introducing deep generative adversarial networks (GANs), which was successfully applied to actual pipeline monitoring systems.

In general, there are various algorithms for identifying abnormal data, but their recognition effects are also different. To address this issue, firstly, the abnormal data in the internal inspection data of the pipeline were analyzed. Secondly, based on four algorithms: Box plot, KNN, LOF, and IForest, abnormal data in the internal detection data of natural gas gathering and transmission pipelines were identified, and the recognition effects of the four algorithms were compared. Finally, the abnormal data in the pipeline internal inspection data were identified and removed, effectively ensuring the accuracy of pipeline corrosion assessment.

2. Abnormal data

The magnetic flux leakage detector is used to detect the growth and distribution of corrosion defects in the pipeline. Magnetic flux leakage testing mainly utilizes the high permeability characteristics of ferromagnetic materials, as well as the fact that the permeability of ferromagnetic materials is greatly affected by material defects under magnetic saturation conditions. If the material is free of defects, the magnetic field lines only exist inside the material, otherwise there is a leakage magnetic field. Therefore, the size and shape of defects can be detected through the leakage magnetic field signal and the Hall effect of the sensor. Draw a magnetic flux leakage detection signal curve based on the internal magnetic flux leakage signal data of the pipeline, and this is expressed in Eqs. (1).

$$y = NDy + Yn \tag{1}$$

where N is the number of signal channels, Dy is the distance between two adjacent channels on the Y-axis, Yn is the data pulse value corresponding to a certain data point. The X-axis of the horizontal axis of the curve image is defined as the pipeline mileage axis, and the Y-axis is defined as the data pulse Value.

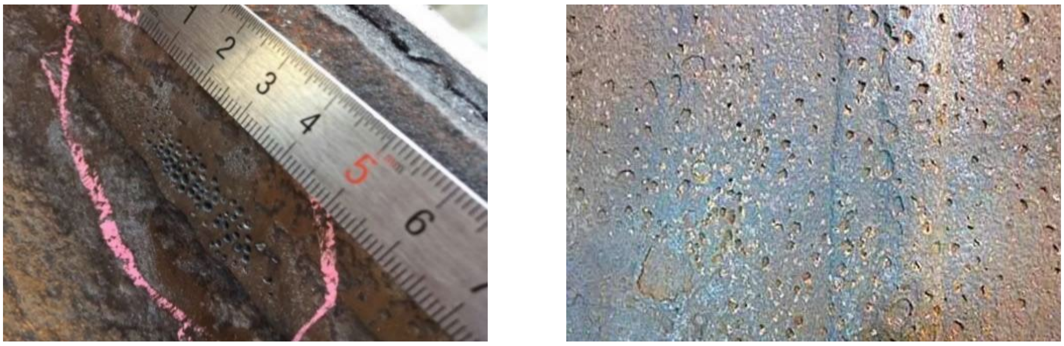


Figure 1: Distribution of internal corrosion defects of pipeline.

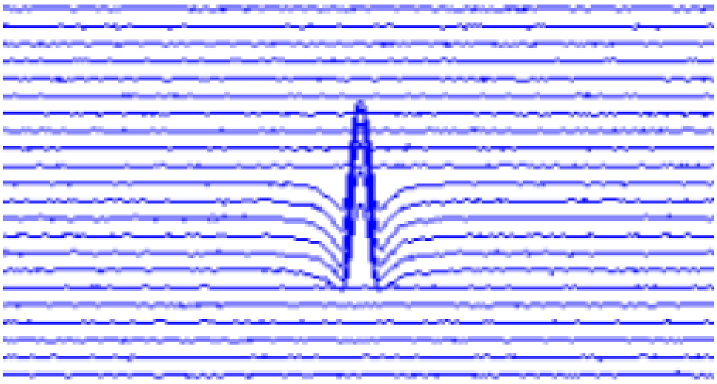


Figure 2: Corrosion defect signal curve of magnetic flux leakage detection.

As shown in Table 1, based on the magnetic flux leakage detection signal curve, the length, width, clock orientation, and depth of corrosion defects in the pipeline are identified, with 80,000 sets for each feature. These data may contain some abnormal data, which typically deviates from the main body of the dataset and appears in a discrete state. Compared with normal data, abnormal data is often difficult to identify and remove directly through manual observation. In addition, because of the abnormal data detected in the pipeline is usually caused by equipment failure, the existence of abnormal data will directly affect the authenticity of the statistical results of pipeline corrosion data. Based on the collected internal inspection data of the pipeline, scatter plots were generated for the distribution of corrosion defect depth percentage, length, width, and clock orientation data.

Table 1
Statistical Results of Corrosion Defects in Pipelines

Pipe Materials	Types of Natural Gas	Number of Internal Corrosion Defects
L360QB	Sulfur-containing wet natural gas	7
L245NCS	Sulfur-containing wet natural gas	7
L360QS	Sulfur-containing wet natural gas	433
L360	Sulfur-containing dry natural gas	871
	Sulfur-containing wet natural gas	59875
20#	Sulfur-containing dry natural gas	7126
	Sulfur-containing wet natural gas	10183
	Total	79149

As shown in Figure 3 and Figure 4, it can be seen that there are obvious scattered abnormal data in the depth percentage, length, and width of corrosion defects, while the clock orientation data is evenly distributed. Therefore, it is necessary to adopt effective methods for identifying abnormal data to handle these outliers.

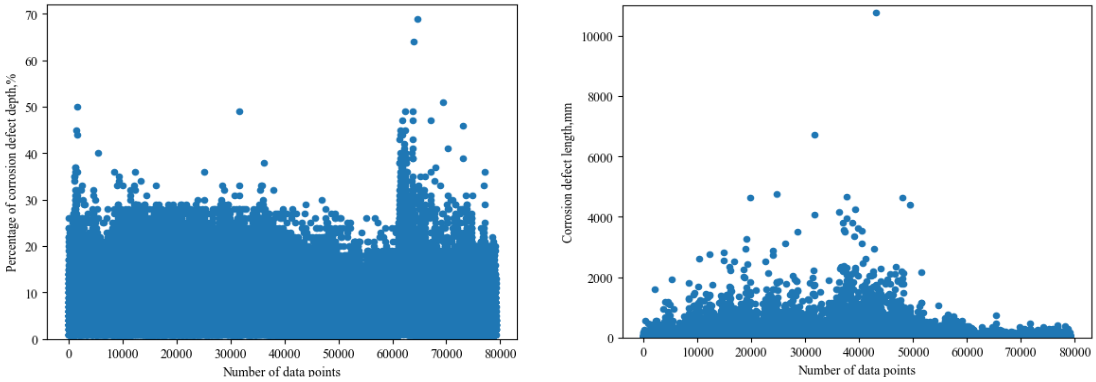


Figure 3: Scatter plot of corrosion defect depth percentage data distribution (left) and scatter plot of corrosion defect length data distribution (right).

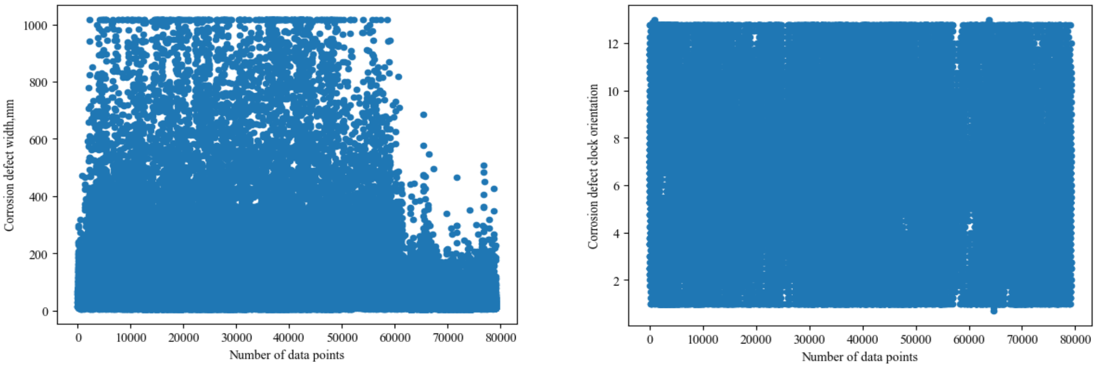


Figure 4: Scatter plot of corrosion defect width data distribution (left) and scatter plot of corrosion defect clock orientation distribution (right).

3. Selecting an algorithm for identifying abnormal data

Different algorithms for identifying abnormal data have different applicable scopes: The pauta criterion [14] is suitable for single-dimensional data that follows a normal or approximately normal distribution; the Box plot [15] is suitable for single-dimensional data without requiring a normal or approximately normal distribution; LOF [16] is suitable for medium-to-high-dimensional datasets where the densities of different clusters are significantly diverse. In addition, six algorithms for identifying abnormal data were also studied, including the Median Absolute Deviation (MAD) [17], Grubbs's criterion [18], K-means [19], DBSCAN [20], KNN [21], and IForest [22]. The applicability of most algorithms for identifying abnormal data can be classified into three categories: sample size, data dimension, and normality distribution.

The sample size and data dimension of the data can be directly observed and determined. However, there are various methods for testing normality, such as the Kolmogorov-Smirnov test, Shapiro-Wilk test, D'Agostino and Pearson omnibus normality test, kurtosis and skewness distribution, Q-Q plot, P-P plot, histogram, etc. Among them, the Shapiro-Wilk test is suitable for small sample sizes, while the Kolmogorov-Smirnov test is suitable for large sample sizes. Since there is no clear boundary for sample size, here we use 1000 as the boundary to distinguish between small and large sample sizes. The above normality tests will ultimately return a p-value, and when $p > 0.05$, it indicates that the dataset follows a normal distribution. However, it is often difficult to achieve an absolute normal distribution, so when the skewness and kurtosis values of the data are within ±2 [23-25], it can be considered as approximately normally distributed.

Therefore, a method for selecting an algorithm for identifying abnormal data is proposed, as shown in Figure 5, which selects the appropriate algorithm for identifying abnormal data in the sample data by determining the sample size, normality, and data dimension of the sample data.

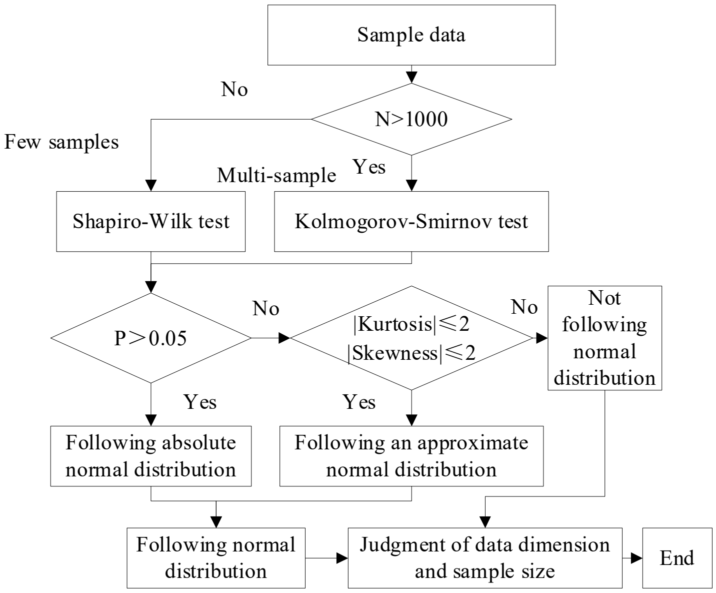

Sample data
No
N>1000
Few samples
Multi-sample
Yes
Shapiro-Wilk test
Kolmogorov-Smirnov test
P>0.05
No
|Kurtosis|≤2
|Skewness|≤2
No
Not following normal distribution
Yes
Yes
Following absolute normal distribution
Following an approximate normal distribution
Following normal distribution
Judgment of data dimension and sample size
End

Figure 5: The process of selecting an algorithm for identifying abnormal data.

4. Select the instance application of the process

Based on observation, it can be inferred that the data type of the corrosion defect depth percentage, defect length, and width data is a large sample single-dimension dataset. Using the Kolmogorov-Smirnov test method and combining the skewness and kurtosis changes of the data, we can judge the normality of the data. The results are shown in Table 2.

Table 2
Sample Data Normality Test Results

Characteristics of Corrosion Defects	Statistical Significance: P	Skewness	Kurtosis
Depth percentage	0.00	1.236	2.620
Length	0.00	14.208	483.815
Width	0.00	3.242	12.112

As shown in Table 2, the Kolmogorov-Smirnov test for the percentage of depth of corrosion defects in pipelines revealed a significant P value of 0.00, far below 0.05, with a skewness of 1.239 and a kurtosis of 2.620, both greater than 2.0, which is inconsistent with normal distribution. The significance P of the length data of corrosion defects in the pipeline is 0.00, which is much smaller than 0.05. The skewness is 14.208 and the kurtosis is 483.815, which is greater than 2.0, and does not conform to the normal distribution. The Kolmogorov-Smirnov test on the data of the width of corrosion defects in the pipeline shows that the significance P is 0.00, far less than 0.05, the skewness is 3.242, and the kurtosis is 12.112, which is greater than 2.0, and does not conform to the normal distribution.

In general, the data of the depth percentage, length, and width of the corrosion defects in the pipeline do not conform to the normal distribution, so the data type is large sample size, single dimension, and non-normal distribution data. Based on algorithm research, four abnormal data identification algorithms were qualitatively selected: Boxplot, KNN, LOF, and IForest.

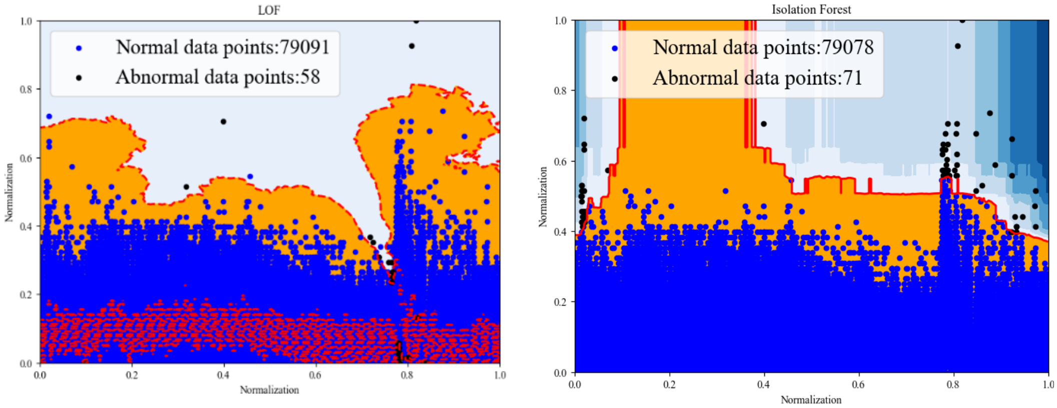
One such method is the Box plot algorithm, which relies on the interquartile range (IQR) to quantify the dispersion of data points. Another algorithm, KNN, classifies data by calculating the distances between distinct values, effectively grouping similar points together. The LOF algorithm, on the other hand, identifies outliers by assessing the density differences among data points, flagging those that deviate significantly from their neighbors. Lastly, the IForest algorithm identifies outliers based on both the numerical structure and the density of the data, utilizing an ensemble of isolation trees to efficiently separate anomalies from the rest of the dataset.

Table 3
The Comparison of the Four Algorithms

The Algorithms	Advantages	Disadvantages
Box plot	Simple to us, low computational cost.	Not suitable for multidimensional data, and the effect of non-normal distribution data is not good.
KNN (K-NearestNeighbor)	Simple and intuitive, no need to assume data distribution.	The calculation cost is high and sensitive to the choice of K value.
LOF (Local Outlier Factor)	Suitable for complex data sets and can handle regions with different densities.	High computational cost, sensitive to parameters
IForest (Isolation Forest)	Efficient, suitable for large-scale data without setting parameters	The effect of high-dimensional data is not good. The results are affected by randomness and need to be averaged by multiple runs.

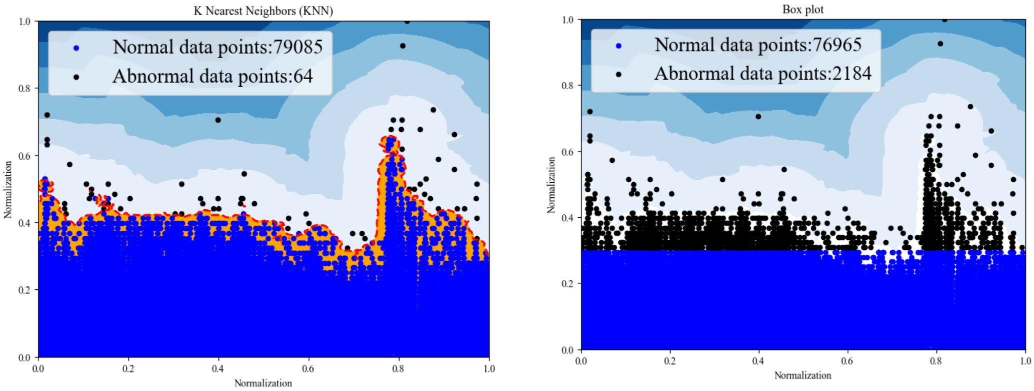
5. Identifying Abnormal Data Application Example

There exist numerous algorithms aimed at identifying abnormal data, each employing unique techniques and approaches. Four abnormal data identification algorithms were used to identify the abnormal data in the depth percentage of corrosion defects, defect length, and width data inside the pipeline. The algorithm code for identifying abnormal data was then crafted based on Python, and the identification results of abnormal data in the percentage data of the depth of corrosion defects in the pipeline are shown in Figure 6.



(a) LOF

(b)IForest

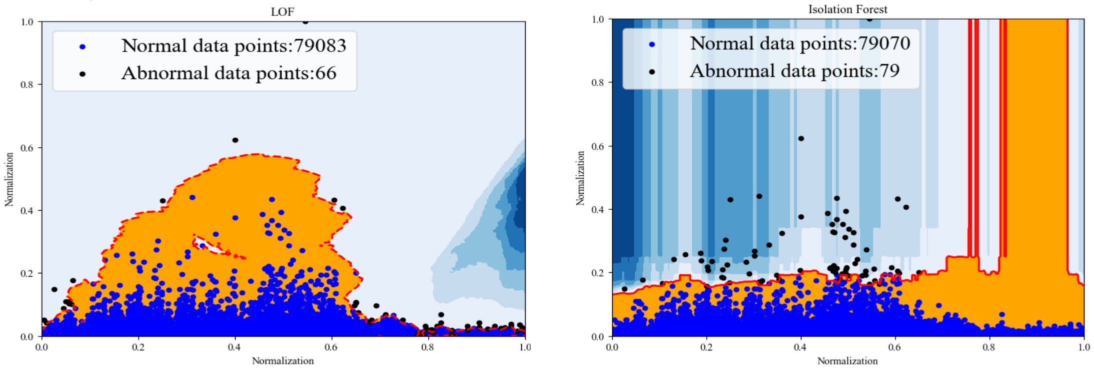


(c) KNN (d) Box plot

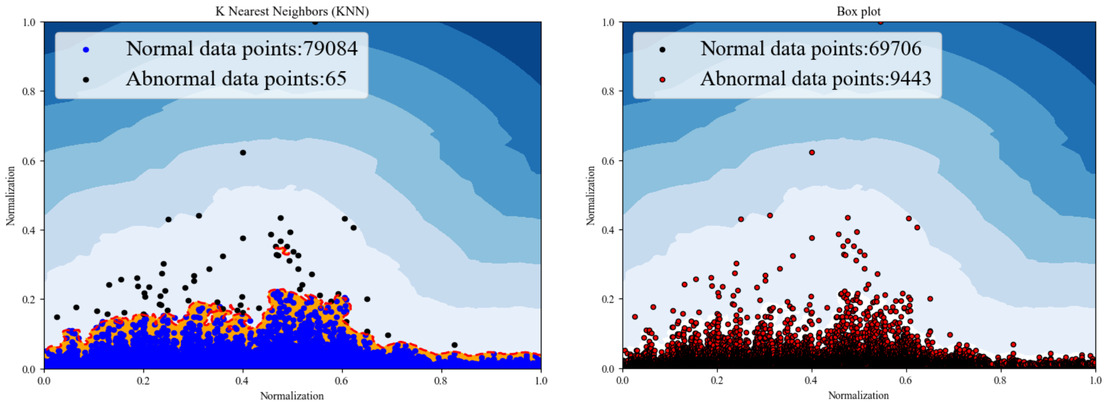
Figure 6: Abnormality identification results of corrosion defect depth percentage data in pipelines under four algorithms.

As shown in Figure 7, LOF identified 58 abnormal data, IForest identified 71 abnormal data, KNN identified 64 abnormal data, and Box plot identified 2184 abnormal data. Similarly, as shown in Figure 7, the results of identifying abnormal data for the length data of corrosion defects in pipelines are similar, using the above four algorithms. LOF identified 66 abnormal data, IForest identified 79 abnormal data, KNN identified 65 abnormal data, and Box plot identified 9443 abnormal data.When the above four algorithms were used to identify abnormal data of pipeline corrosion defect width data, LOF identified 67 abnormal data, IForest identified 80 abnormal data, KNN identified 52 abnormal data, and Box plot identified 8367 abnormal data.

It can be clearly seen that KNN can effectively identify outliers that are far away from the main body of the dataset, while LOF and IForest have poor identification performance and did not effectively identify the abnormal data. At the same time, Box plot identified too many abnormal data, which is not suitable for the dataset of internal corrosion defect depth percentage.



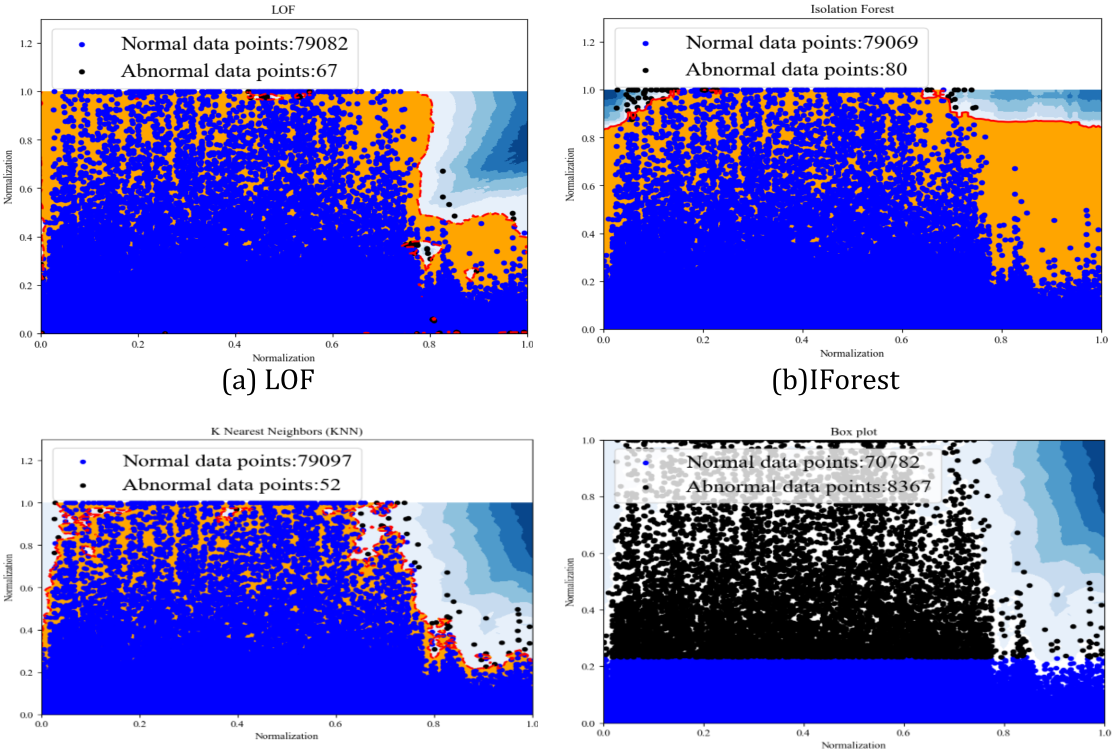
(a) LOF (b)IForest



(c) KNN

(d) Box plot

Figure 7: Abnormality identification results of corrosion defect length data in pipelines under four algorithms.



(a) LOF

(b)IForest

(c) KNN

(d) Box plot

Figure 8: Abnormality identification results of corrosion defect width data in pipelines under four algorithms.

In general, the KNN algorithm has a good effect on the identification of abnormal data in pipeline detection data, and the abnormal data finally identified is more consistent with the real situation. KNN algorithm is used to identify and remove abnormal data, as shown in Figure 9 to Figure 11.

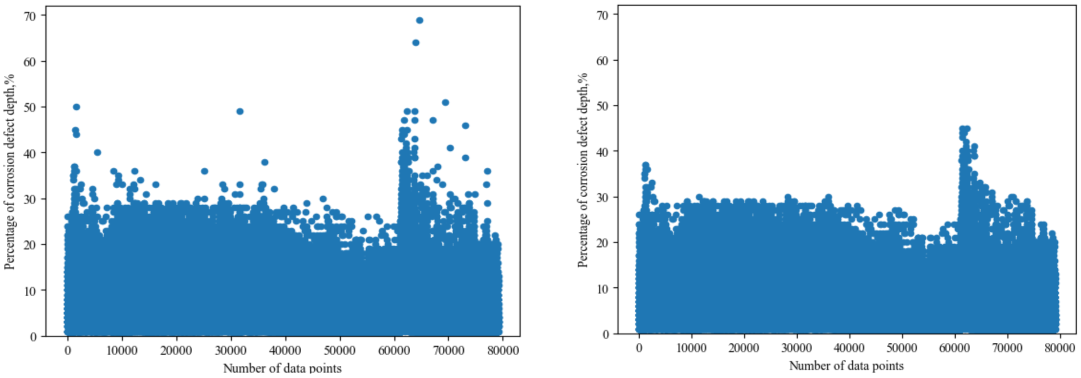


Figure 9: Before processing corrosion defect depth percentage data (left) and after processing corrosion defect depth percentage data (right).

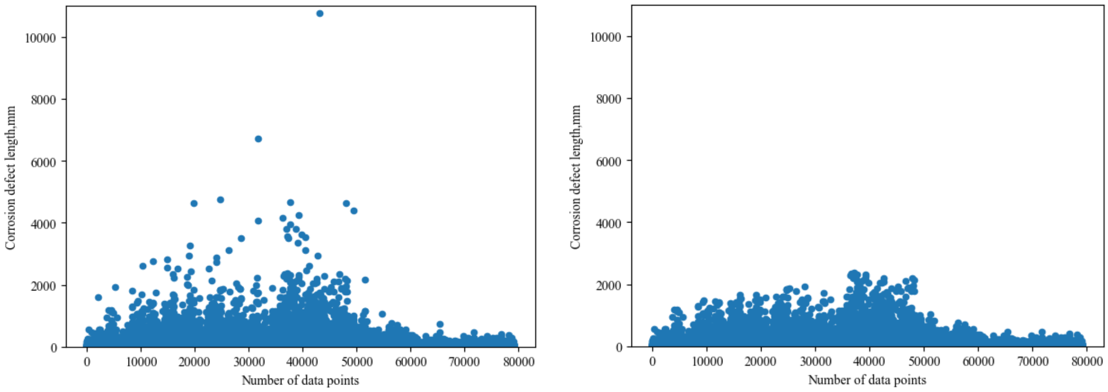


Figure 10: Before processing corrosion defect length data (left) and after processing corrosion defect depth percentage data (right).

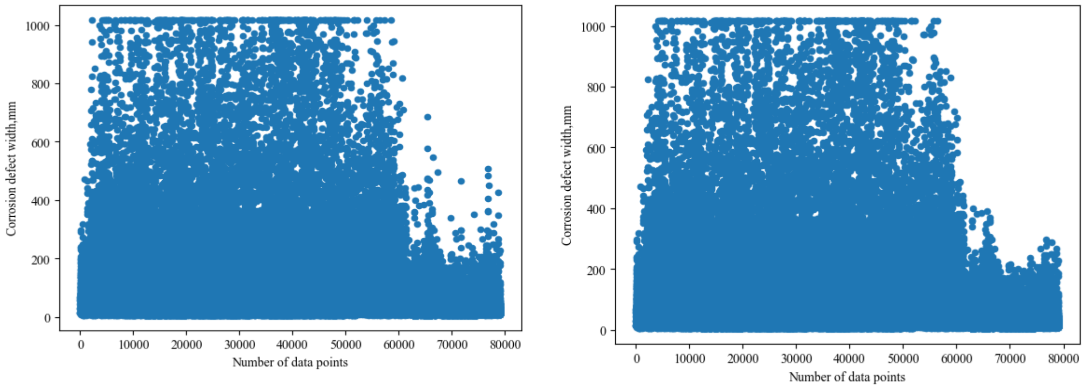


Figure 11: Before processing corrosion defect width data (left) and after processing corrosion defect depth percentage data (right).

6. Conclusion

Natural gas gathering pipeline inspection data includes four characteristics of corrosion defects, which are depth percentage data, length data, width data, and clock orientation data. If there are abnormal data in the data, it may have a significant impact on the corrosion assessment of the pipeline. Therefore, it is necessary to select suitable algorithms for identifying abnormal data and apply them to identify abnormal data in pipeline inspection data. The main conclusions are as follows:

1. Nearly 80,000 sets of pipeline inspection data were collected. By plotting the distribution scatter plot of the depth percentage of corrosion defects, length, width, and clock orientation data, it was found that there are obvious scattered abnormal data in the depth percentage of corrosion defects, defect length, and width data.
2. A method for selecting algorithms for identifying abnormal data based on the sample size, normality, and data dimension of the sample data is proposed. This method can select algorithms that are suitable for identifying abnormal data in the sample data.
3. Based on pipeline inspection data, it is judged that the data type is large sample size, single dimension, and non-normal distribution data. Four algorithms for identifying abnormal data, including Box plot, KNN, LOF, and IForest, are selected.
4. The selected algorithms are applied to pipeline inspection data, and it is found that the KNN algorithm has the best identification performance and can effectively identify scattered or abnormal data.

References

[1] I. Alhussein, A. H. Ali, “Application of DBSCAN to anomaly detection in airport terminals,” 2020 3rd International Conference on Engineering Technology and its Applications (IICETA), IEEE, pp. 112-116, 2020.

[2] A. Abid, A. Kachouri, A. Mahfoudhi, “Outlier detection for wireless sensor networks using density-based clustering approach,” IET Wireless Sensor Systems, pp. 83-90, 2017.

[3] G. U. Guo-qing, L. I. Ao-hui, “Exponential weighted smoothing prediction model based on abnormal detection of box-plot,” Computer and Modernization, 2021.

[4] O. Salem, Y. Liu, A. Mehaoua, “Online anomaly detection in wireless body area networks for reliable healthcare monitoring,” IEEE journal of biomedical and health informatics, pp. 1541-1551, 2014.

[5] M. Ahmed, N. Choudhury, S, Uddin, “Anomaly detection on big data in financial markets,” Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 998-1001, 2017.

[6] C. Li, L. Guo, H. Gao, “Similarity-measured isolation forest: anomaly detection method for machine monitoring data,” IEEE Transactions on Instrumentation and Measurement, pp. 1-12, 2021.

[7] M. A. Bhatti, R. Riaz, S. S. Rizvi, “Outlier detection in indoor localization and Internet of Things (IoT) using machine learning,” Journal of Communications and Networks, pp. 236-243, 2020.

[8] G. He, C. Tan, D. Yu, “A real-time network traffic anomaly detection system based on storm,” 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, IEEE, 2015, pp. 153-156.

[9] J. Smith, L. Brown, & H. Wang, “Reinforcement learning-based anomaly detection method for pipeline monitoring,” Journal of Industrial Information Integration, pp. 200-210, 2019.

[10] A. Jones, M. Garcia, & S. Patel, “Big data analytics for real-time pipeline anomaly detection,” IEEE Transactions on Industrial Informatics, pp. 4500-4510, 2020.

[11] S. Mohamed, A. El-Sayed, & M. Hussein, “Convolutional neural networks for anomaly detection in industrial pipelines,” International Journal of Computer Applications, pp. 25-32, 2020.

[12] M. Garcia, R. Lopez, & J. Gonzalez, “A novel anomaly detection approach using genetic algorithms and fuzzy logic,” Expert Systems with Applications, pp. 113750, 2021.

[13] Y. Zhang, B. Liu, & X. Chen, “Deep generative adversarial networks for anomaly detection in pipeline monitoring,” Neural Networks, pp. 78-89, 2022.

[14] L, Zhao, “2021 Prediction model of ecological environmental water demand based on big data analysis,” Environmental Technology & Innovation, 2021.

[15] J. W. Tukey, “Exploratory Data Analysis,” Reading, MA: Addison–Wesley, pp. 131-160， 1977.

[16] M. M. Breunig, H. P. Kriegel, R. T. Ng, & J. Sander, “LOF: identifying density-based local outliers,” In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93-104, 2000.

[17] D. C. Howell, “Median absolute deviation,” Encyclopedia of Statistics in Behavioral Science, 2005.

[18] M. Urvoy, & F. Autrusseau, “Application of Grubbs' test for outliers to the detection of watermarks,” In Proceedings of the 2nd ACM workshop on Information Hiding and Multimedia Security, pp. 49-60, 2014.

[19] J. A. Hartigan, & M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” Journal of the Royal Statistical Society, Series C (applied statistics), pp. 100-108, 1979.

[20] M. Ester, H. P. Kriegel, J. Sander, & X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” In KDD, Vol. 96, No. 34, pp. 226-231, 1996.

[21] L. Che, C. J. Hwang, C. Ni, “Research on SVM Distance based credit scoring model for network management,” Journal of Convergence Information Technology, 2012.

[22] F. T. Liu, K. M. Ting, & Z. H. Zhou, “Isolation Forest,” In 2008 Eighth IEEE International Conference on Data Mining, pp. 413-422, 2008.

[23] W. Trochim, & J. Donnelly, “The research methods knowledge base,” 3 Edition. Mason, Ohio: Atomic Dog Publishing Inc, 2006.

[24] F. J. Gravetter, L. B. Wallnau, L. A. B. Forzano, & J. E. Witnauer, “Essentials of statistics for the behavioral sciences,” Cengage Learning, 2021.

[25] A. Field, Discovering statistics using IBM SPSS statistics, Sage, 2013.