processes

MDPI

Article

Intelligent Monitoring Model for Lost Circulation Based on Unsupervised Time Series Autoencoder

Liwei Wu [1,2], Xiaopeng Wang [1,2], Ziyue Zhang [3,*], Guowei Zhu [1,2], Qilong Zhang [1,2], Pinghua Dong [1,2], Jiangtao Wang [4] and Zhaopeng Zhu [3]

[1] State Key Laboratory of Offshore Oil and Gas Exploitation, Beijing 102209, China; wulw2@cnooc.com.cn (L.W.); wangxp5@cnooc.com.cn (X.W.); zhugw@cnooc.com.cn (G.Z.); zhangql40@cnooc.com.cn (Q.Z.); dongph@cnooc.com.cn (P.D.)
[2] CNOOC China Limited, Tianjin Branch, Tianjin 300459, China
[3] School of Petroleum Engineering, China University of Petroleum, Beijing 102249, China; zhuzp@cup.edu.cn
[4] China France Bohai Geoservices Co., Ltd., Tianjin 300456, China; wangjiangt@cfbgc.com
* Correspondence: 2023215177@student.cup.edu.cn; Tel.: +86-135-1287-2028

Abstract: Lost circulation, a common risk during the drilling process, significantly impacts drilling safety and efficiency. The presence of data noise and temporal evolution characteristics pose significant challenges to the accurate monitoring of lost circulation. Traditional supervised intelligent monitoring methods rely on large amounts of labeled data, which often do not consider temporal fluctuations in data, leading to insufficient accuracy and transferability. To address these issues, this paper proposes an unsupervised time series autoencoder (BiLSTM-AE) intelligent monitoring model for lost circulation, aiming to overcome the limitations of supervised algorithms. The BiLSTM-AE model employs BiLSTM for both the encoder and decoder, enabling it to comprehensively capture the temporal features and dynamic changes in the data. It learns the patterns of normal data sequences, thereby automatically identifying anomalous risk data points that deviate from the normal patterns during testing. Results show that the proposed model can efficiently identify and monitor lost circulation risks, achieving an accuracy of 92.51%, a missed alarm rate of 6.87%, and a false alarm rate of 7.71% on the test set. Compared to other models, the BiLSTM-AE model has higher accuracy and better timeliness, which is of great significance for improving drilling efficiency and ensuring drilling safety.

Keywords: lost circulation monitoring; unsupervised; time series autoencoder; BiLSTM-AE

check for updates

Citation: Wu, L.; Wang, X.; Zhang, Z.; Zhu, G.; Zhang, Q.; Dong, P.; Wang, J.; Zhu, Z. Intelligent Monitoring Model for Lost Circulation Based on Unsupervised Time Series Autoencoder. *Processes* **2024**, *12*, 1297. https://doi.org/10.3390/pr12071297

Academic Editor: Qingbang Meng

Received: 18 May 2024
Revised: 17 June 2024
Accepted: 19 June 2024
Published: 22 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As the development of conventional oil and gas resources continues to deepen, the oil and gas exploration and development industry is gradually expanding into complex areas, including unconventional, deep, and deep-sea resources. As a crucial step in oil and gas exploration and development, drilling consistently emphasizes safe and efficient drilling, a primary concern for domain experts. Complex oil and gas formations, characterized by high temperatures, high pressures, intricate layers, and instability, lead to high risks and a concealed nature of lost circulation. Identifying lost circulation risks is further complicated by data noise, real-time fluctuations, and nonlinear mapping relationships. If not detected in time, delays in handling can easily result in severe loss and blowout accidents. Therefore, accurate and effective monitoring of lost circulation risks is significant for safe and efficient drilling.

Lost circulation monitoring methods are mainly divided into traditional and intelligent methods. Traditional methods primarily utilize sensor equipment to monitor various parameters during the drilling process, with engineers then determining whether lost circulation has occurred based on real-time changes in these parameters. For example, Shoujun Liu [1] used the mud pit page method and ultrasonic sensors to monitor changes in

the mud pit page to determine well leakage; Dong Tan [2] chose the flow difference method and proposed a combination of surface flow monitoring and downhole pressure prediction to monitor and predict drilling overflow in pressure-sensitive formations. Limited by the sensors' precision and engineers' experience, traditional methods have drawbacks such as intense subjectivity, low accuracy, and poor timeliness.

In recent years, with the rapid development of machine learning technology, intelligent monitoring methods have gradually become a research hotspot in drilling risk monitoring. For instance, Liu Biao et al. [3] analyzed various parameters and formation characteristics during the drilling process and used support vector regression to build an intelligent prediction model for lost circulation, achieving early identification of lost circulation risks. Yingzhuo X et al. [4] developed an innovative lost circulation prediction model based on deep learning and conducted a quantitative analysis of the impact of each feature on the model's prediction results. Song Yan [5] proposed an intelligent recognition method for lost circulation risk status based on extreme learning machines, achieving high-accuracy identification of lost circulation risks. Baek S et al. [6] established a lost circulation correlation model by analyzing formation conditions, engineering parameters, and operational dynamic parameters related to lost circulation, using this model to achieve a lost circulation risk response and early warning for data anomalies with different weights. Li Changhua et al. [7] employed convolutional neural networks and multidimensional data fusion algorithms to establish a lost circulation prediction model, enhancing the reliability of lost circulation identification and prediction. Zheng Zhuo et al. [8] used the XGBoost algorithm to construct a real-time lost circulation early warning model tailored to the specific geological conditions of the Bohai Bay Basin, improving the accuracy of formation loss judgments. Sun Weifeng et al. [9] effectively realized accurate monitoring of minor drilling fluid losses by combining dilated causal convolutional networks and long short-term memory networks. Luo Ming et al. [10] proposed a lost circulation prediction method based on deep convolutional feature reconstruction by extracting key characterization parameters of lost circulation and constructing a temporal feature matrix, improving the prediction accuracy of lost circulation risks. For time series learning methods, Li et al. [11] used LSTM and Bi-LSTM for leakage monitoring in 2023, achieving an identification accuracy between 88.93% and 98.38%.

Currently, traditional methods for intelligently monitoring lost circulation mainly focus on supervised learning algorithms, which are extremely dependent on a large amount of well-leakage risk data during the model training process; in addition, they show apparent limitations in parsing well leakage data with complex and nonlinear characteristics. Moreover, accessing a large amount of labeled risk data is often tricky in practical applications. Secondly, since the traditional supervised model relies on training data, if the training data does not adequately cover all possible lost circulation scenarios, the model may exhibit poor generalization ability in practical applications and fail to accurately identify new lost circulation risks. Therefore, an unsupervised learning approach can overcome the limitations of traditional models that rely on a large amount of risk data.

For the non-temporal learning method, it is often difficult to effectively reveal the changing rule of parameters over time, and the temporal continuity and sequence of data are neglected when analyzing the data. Therefore, this study focuses on considering time-sequential data in the drilling process to explore the relationship of parameter changes over time.

To address the shortcomings of the supervised learning and non-time series autoencoder model algorithms mentioned above in lost circulation monitoring studies, an unsupervised bi-directional long and short time series autoencoder model based on bi-directional long and short time series memory networks (BiLSTM-AE) is used in this study. The encoder and decoder of the model are BiLSTM, which can capture temporal correlations. When the time series data is input into the model, it encodes, compresses and decodes the data, then calculates the reconstruction error between the output data and the input data. If the reconstruction error is larger than a threshold, it is risky data, and vice versa; it is normal

and risk-free data. BiLSTM-AE is more comprehensive in the contextual understanding of sequence data and can extract the characteristics of temporal variations. This method can identify lost circulation risk in real-time in practical applications, make up for the lack of relying on a large amount of lost circulation risk data, mine the changing law of data, improve real-time monitoring, and provide an efficient and accurate method for lost circulation monitoring.

2. Methodology

2.1. Bidirectional Long Short-Term Memory Network (BiLSTM)

Long Short-Term Memory (LSTM) networks are a particular type of recurrent neural network (RNN) specifically designed to address the long-term dependency issues encountered by traditional RNNs when processing long sequence data. LSTM was proposed by Jürgen Schmidhuber [12] and optimized by Alex Graves et al. in 1999. Through subsequent improvements and optimizations by various researchers, LSTM has become one of the mainstream methods for handling sequential data tasks, such as language modeling and text generation, speech recognition, and machine translation.

Bidirectional Long Short-Term Memory (BiLSTM) networks are an extension of LSTM that better capture the dependencies in sequence data in both forward and backward directions [13]. BiLSTM was introduced by Schuster and Paliwal [14] in 1997. In a standard LSTM, the state transmission of each unit is unidirectional from front to back, meaning the output at the next moment can only be predicted based on the temporal information of the previous moment. Therefore, LSTM models can only learn the features of risk characterization parameters from past moments and cannot learn features from future moments.

BiLSTM, on the other hand, has two state transmission paths: one from front to back and the other from back to front. By placing two independent LSTM layers in parallel on the same output, one processing forward sequence information and the other processing backward sequence information, and then merging the information from both directions, the network can simultaneously consider contextual information from both past and future moments. BiLSTM outperforms unidirectional LSTM and other traditional RNN variants in many tasks because it can capture more contextual information.

Similar to h_t for the network state of the backward LSTM hidden layer at moment t, the calculation formula is shown as Equation (1).

$$h_t = LSTM(x_t, h_{(t-1)}) \tag{1}$$

where h_t is the state of the backward LSTM network's hidden layer at moment t; x_t is the input at moment t; $h_{(t-1)}$ is the state of the backward LSTM network's hidden layer at moment $t-1$. The output of the BiLSTM network is the combination of two parts of hidden layer states h_t and h_t, thus forming the overall network's hidden state h_t.

2.2. Autoencoders (AE)

Autoencoders (AE) are an important deep learning architecture in the field of unsupervised learning [15]. They can automatically learn effective features from a large amount of unlabeled data. The effective compression and reconstruction of data are mainly achieved through two parts: the encoder and decoder. The main structure is shown in Figure 1. This structure can not only learn low-dimensional feature representations from high-dimensional data to achieve data dimensionality reduction and feature learning but also improve the model's resistance to noise by denoising the self-encoder. This improvement enhances the efficiency of data processing and analysis, providing a highly efficient tool for tasks such as data dimensionality reduction, feature extraction, denoising, and more.

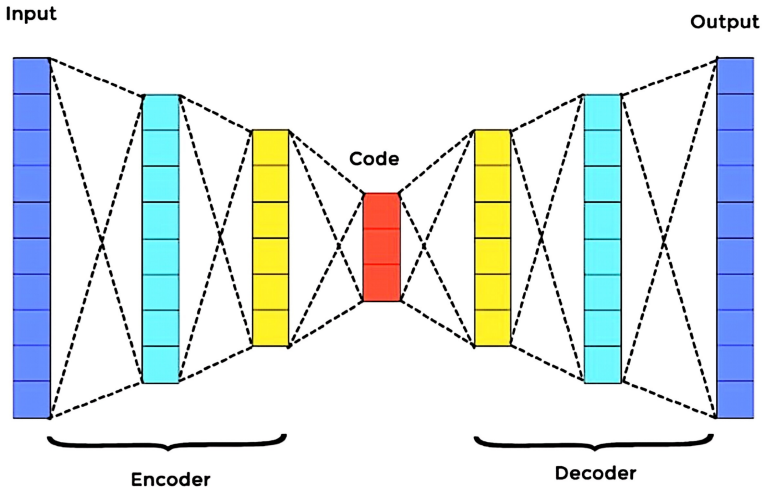


Figure 1. Time Series Autoencoder Structure.

The auto-encoder consists of two functions, the encoder and the decoder, denoted as ϕ and ψ, respectively. The goal of the auto-encoder is to approximate the input data by minimizing a constant function as shown in Equation (2):

$$\text{argmin}_{\phi,\psi}\|X - \psi(\phi(X))\|^2 \tag{2}$$

In the encoding phase, the autoencoder maps the input vector X to the intermediate vector V through Equation (3):

$$V = \sigma(WX + b) \tag{3}$$

W is the weight matrix, b is the bias vector, and σ is the activation function of the coding layer.

In the decoding stage, the autoencoder maps the intermediate vector V to the output vector $\hat{X}$ by Equation (4), here $m > n$. Then, the reconstruction error between the reconstructed vector $\hat{X}$ and the original input vector X is calculated using Equation (5).

$$\hat{X} = \hat{\sigma}(\hat{W}V + \hat{b}) \tag{4}$$

$$E(X, \hat{X}) = \sum_{i=1}^{m} \|\hat{\sigma}(\hat{W}\sigma(Wx_i + b) + \hat{b}) - x_i\|^2 \tag{5}$$

where $\hat{W}$ is the weight matrix of the decoder, $\hat{b}$ is the bias vector, and $\hat{\sigma}$ is the activation function of the decoding layer.

The goal of the autoencoder is to minimize the reconstruction error. Error minimization is achieved by adjusting the weight matrices W and $\hat{W}$ and the bias vectors b and $\hat{b}$, by means of gradient descent and backpropagation algorithms.

3. Data Processing

3.1. Data Cleaning

This study utilizes time series data collected from 13 wells in real drilling sites. The data collection interval is 5 s, comprising over 300,000 normal state data points and more than 5000 lost circulation state data points. Upon reviewing the raw data, it was found that there were anomalies and missing values. Anomalies can distort data trends, and missing values can lead to incorrect data formats, both of which can affect the normal construction and training of intelligent models. An improved sliding window 3σ method was used to filter out anomalies. By setting an appropriate sliding window size, this method avoids misjudging normal data as anomalies. The anomalies in the data were replaced with missing values, and spline interpolation was used to fill both the originally missing values

and those generated by anomaly replacement. This method preserves the overall trend of the data and performs well when there are continuous missing data.

To further reduce noise and improve data quality, a smoothing filter was applied to denoise the time series data. Each smoothed point is the average of the current point and the previous 99 data points (a total of 100 data points). The aim is to reduce random fluctuations in the data, facilitating subsequent calculations, analysis, and visualization. Additionally, differences in dimensions between different parameters can affect data analysis and model training effectiveness. Therefore, the min-max normalization method was used to process the data dimensionlessly, which helps shorten the model training time and reduce overfitting [16]. The relevant calculation formula is as follows (6):

$$x_{scale} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{6}$$

where x is the data sample, x_{scale} is the normalized value of x, $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the variable x.

3.2. Feature Selection

To reduce irrelevant parameter inputs and computational redundancy, it is necessary to optimize the input features of the model. This study employs the Spearman correlation coefficient method [17] combined with the knowledge of lost circulation identification mechanisms for feature selection. Spearman's correlation coefficient can be an effective measure of non-linear relationships between variables because it is based on rankings rather than specific values and requires less data distribution. Multiple parameter relationships in lost circulation monitoring are usually non-linear, so the Spearman correlation coefficient can more accurately reflect such complex relationships, thus optimizing model input features and improving model accuracy and robustness. First, the Spearman correlation coefficient method is used to calculate the correlation between each parameter and the lost circulation labels, optimizing feature parameters from a data perspective. A Spearman correlation coefficient greater than 0 indicates a positive correlation, while a Spearman correlation coefficient less than 0 indicates a negative correlation. The closer the value is to 1 or -1, the stronger the correlation. A Spearman correlation coefficient equal to 0 indicates no correlation between the two variables. The calculation formula is shown as (7):

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{7}$$

where r_s is the correlation coefficient between the two variables, n is the number of data points in the variables, and d_i is the rank difference between the data points in the two variables. The absolute value of the Spearman correlation coefficient is taken; the larger this value, the stronger the correlation. The correlation between the feature parameters and the lost circulation labels is shown in Figure 2. As can be seen from the figure, parameters such as standpipe pressure, total pit volume, outlet flow rate, weight on bit, and torque have a strong correlation with the lost circulation labels.

Considering that the core of accurately identifying lost circulation is monitoring the trends in sequence variation of parameters, this study analyzes the dynamic changes in parameters in multiple typical lost circulation cases [18]. Through data visualization, as shown in Figure 3, it is found that in different lost circulation cases, three parameters outlet flow rate, total pit volume, and standpipe pressure exhibit a more noticeable downward trend and form a general pattern of change. In contrast, the variation patterns in parameters such as weight on bit and torque are more influenced by the specific lost circulation case itself.

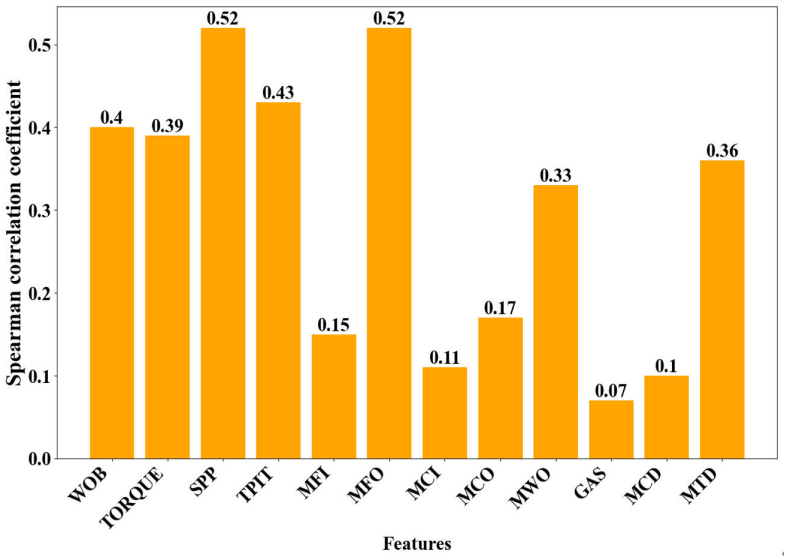


Figure 2. Feature Selection Using Spearman Correlation Coefficient Method.

When experts at the drilling site assess lost circulation risks, they often analyze changes in outlet flow rate, total pit volume, and standpipe pressure. When lost circulation occurs, drilling fluid flows from the wellbore into the formation through the lost circulation path, causing a decrease in the return speed of the drilling fluid in the annulus. This decrease reduces annular friction and consequently lowers the standpipe pressure [19]. Additionally, if there are no significant changes in the inlet flow rate, the outlet flow rate decreases. Finally, as some drilling fluid flows into the formation, the volume of drilling fluid returning to the surface decreases, resulting in a reduction in the total pit volume [20].

Therefore, considering both the correlation of parameters and the mechanistic knowledge, this study uses total pit volume (TPIT), outlet flow rate (MFO), and standpipe pressure (SPP) as the input parameters for the subsequent intelligent model.

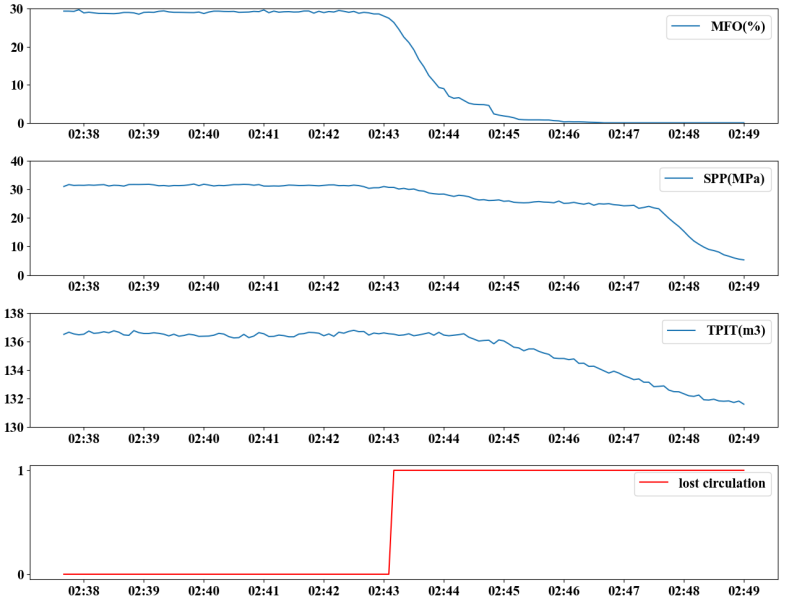


Figure 3. Key Parameter Trends after Lost Circulation.

3.3. Time Series Sample Construction

In lost circulation monitoring, there is much more normal state data than lost circulation state data. This data imbalance limits the training of supervised learning algorithms. This is because they rely on a large number of well-labeled positive/negative labeled

samples. Unsupervised learning algorithms require only normal state data as training data and are able to overcome the data imbalance.

To meet the input requirements of the algorithm model and conform to the identification mechanism of lost circulation risks, the model input must be a continuous data sequence over a period of time rather than independent data samples at single time points. This study uses the sliding window method to construct fixed-length time series samples from the normal state data in the time series dataset. It is constructed as shown in Figure 4.

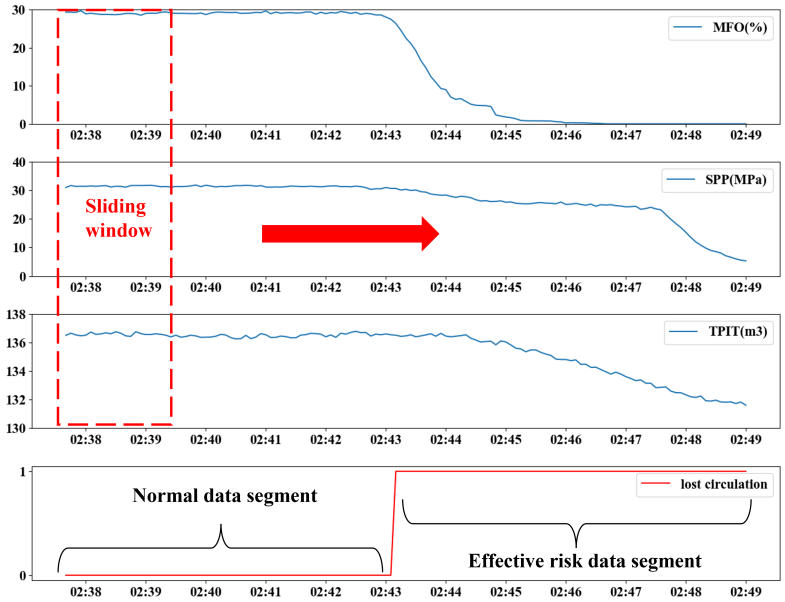


Figure 4. Constructing Sample Methods Using a Sliding Window.

In this paper, the sliding step of the window is set to 1, moving one time step at a time. The time step of the window is set to 30, and the data collection frequency is 5 s. The final time span of each time-series sample is 150 s. For each window, data from 30 consecutive time points starting from the current time step are extracted to form a time-series sample. This setup ensures that each sample contains enough information to capture the trend of the parameters over time.

4. Model Construction

4.1. Evaluation Metrics

An important part of applying machine learning methods to solve engineering problems is evaluating the performance of the algorithm models. This study uses four evaluation metrics to assess the performance of different models in identifying lost circulation risks: accuracy, recall, missing alarm rate, and false alarm rate [21]. The calculation formula for each evaluation metric is shown in Table 1 and Formulas (8)–(11).

Table 1. True Result vs. Forecast Result.

	Positive Class	Negative Class
Positive Class	TP	FN
Negative Class	FP	TN

TP is the number of samples that predict the true positive class as a positive class;
FN is the number of samples that predict the true positive class as a negative class;
FP is the number of samples that predict the true negative class as a positive class;
TN is the number of samples to predict the true negative class as a negative class.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{8}$$

Accuracy: Equation (8) is the ratio of the number of correctly predicted samples inside the classification model to the overall number of predicted samples.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

Recall: Equation (9) is the ratio of predicted positive samples to actual positive samples.

$$MissingAlarm = \frac{FN}{TP + FN} \tag{10}$$

MissingAlarm: Equation (10) is the proportion of actual lost circulation samples that are predicted to be non-lost circulation samples in the classification model, i.e., the proportion of lost circulation samples that are missed.

$$FalseAlarm = \frac{FP}{TP + FP} \tag{11}$$

FalseAlarm: Equation (11) is the proportion of samples that are actually non-lost circulation but are predicted as lost circulation by the classification model. In other words, it is the ratio of samples that are predicted to be lost circulation but are actually non-lost circulation within the total samples predicted as lost circulation.

4.2. Model Structure

The structure of the proposed BiLSTM-AE model is shown in Figure 5. The model structure consists of two parts: the encoder part and the decoder part. The BiLSTM-AE model uses BiLSTM networks as both the encoder and the decoder.

First, the input layer receives the time series data, which then passes through an encoder that contains a BiLSTM layer. Next, a fully connected layer maps the encoded vectors to a low-dimensional encoding. The encoded data are then expanded through a repeat vector layer to match the input requirements of the decoder.

The decoder consists of two BiLSTM layers, which aim to reconstruct the original input sequence into a high-dimensional space, thus completing the process of data encoding compression and decoding reconstruction.

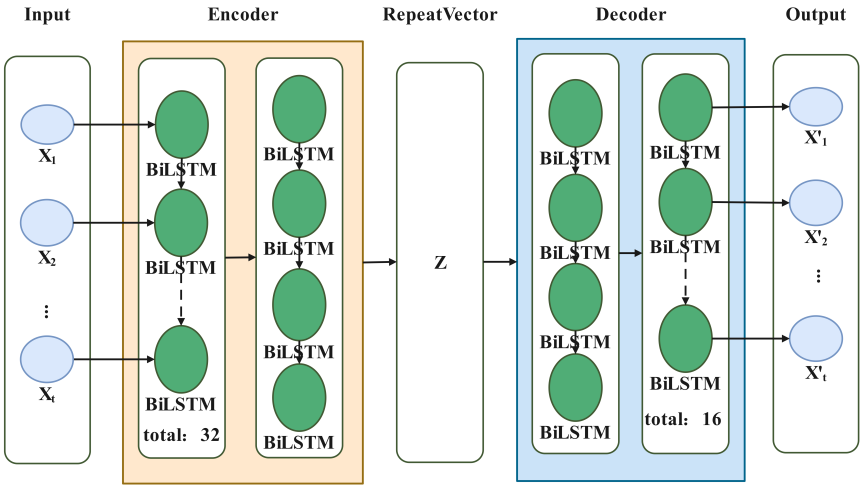


Figure 5. BiLSTM-AE Diagram.

4.3. Parameter Settings

In the proposed BiLSTM-AE model, there are a total of five layers in addition to the input and output layers. The input layer consists of three neurons. The first BiLSTM layer processes each time step with 32 units and returns the full sequence, outputting a 32-dimensional vector for each time step. The second LSTM layer processes the sequence

with four units and returns only the output of the last time step. The third Repeat Vector layer repeats the output of the second layer for the entire sequence length, making its dimension match the sequence length. The fourth BiLSTM layer processes the input from the third layer and returns an eight-dimensional vector sequence. The fifth layer is an LSTM layer that processes the output from the fourth layer and returns a 16-dimensional vector sequence.

The output layer is a TimeDistributed Dense Layer that applies a Dense layer at each time step of the fifth layer, outputting vectors that match the number of input features. The loss function used during the training process is the mean absolute error (MAE), which is the average of the absolute differences between the predicted values and the true values. The formula is shown as (12):

$$\mathrm{MAE}(y, \overline{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \overline{y}_i| \tag{12}$$

1. For each sample i, calculate the absolute error between the true value y_i and the predicted value $\bar{y}_i$, that is $|y_i - \bar{y}_i|$.

2. Sum the absolute errors for all samples.

3. Divide the total error by the total number of $n_{samples}$ to obtain the mean absolute error.

The Adam optimizer was used with a learning rate of 0.001, 50 epochs, and a batch size of 32. The specific model parameters are shown in Table 2.

Table 2. BiLSTM-AE Parameter Settings.

Network Layer	Output Dimension	Param
Input Layer	(None, 10, 3)	0
BiLSTM (L1)	(None, 10, 32)	2560
BiLSTM (L2)	(None, 10, 4)	592
Repeat Vector (L3)	(None, 10, 4)	0
BiLSTM (L4)	(None, 10, 8)	288
BiLSTM (L5)	(None, 10, 16)	1600
Time Distributed Dense Layer	(None, 10, 3)	51

4.4. Model Training

Figure 6 shows the change in the loss value of a training model over epochs. During the first 10 epochs, the training loss decreases rapidly, indicating that the model learns a substantial amount of information during this period. As the epochs increase, the training loss converges to 0.0188 and gradually stabilizes, showing no significant signs of overfitting or underfitting.

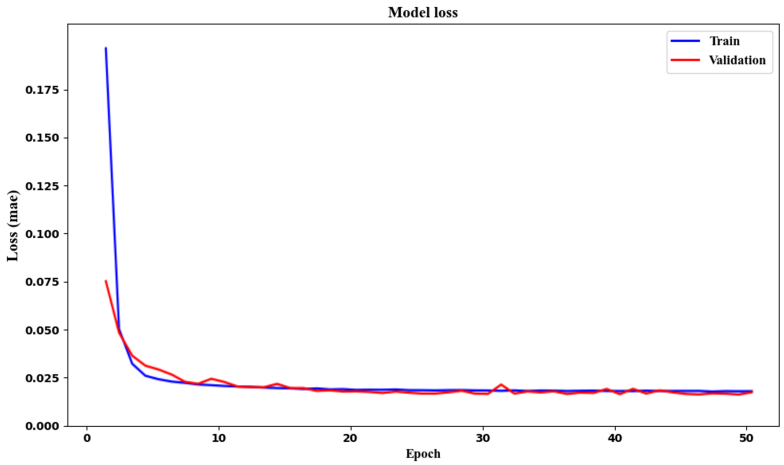


Figure 6. Model loss.

5. Results and Discussion

5.1. Parameter Analysis

The time step length of the time series samples is crucial for the model's performance. An appropriate time step length can improve model accuracy and reduce the risk of overfitting or underfitting. This study conducted comparative experiments with different time step lengths, and the experimental results are shown in Figure 7. When the time step length is set to 30, the model achieves the highest accuracy and recall on the test set, with values of 92% and 93%, respectively. When using smaller or larger time step lengths, the model's performance decreases to varying degrees. This indicates that the overall model performance is optimal with a time step length of 30, resulting in the fewest missed alarms and false alarms.

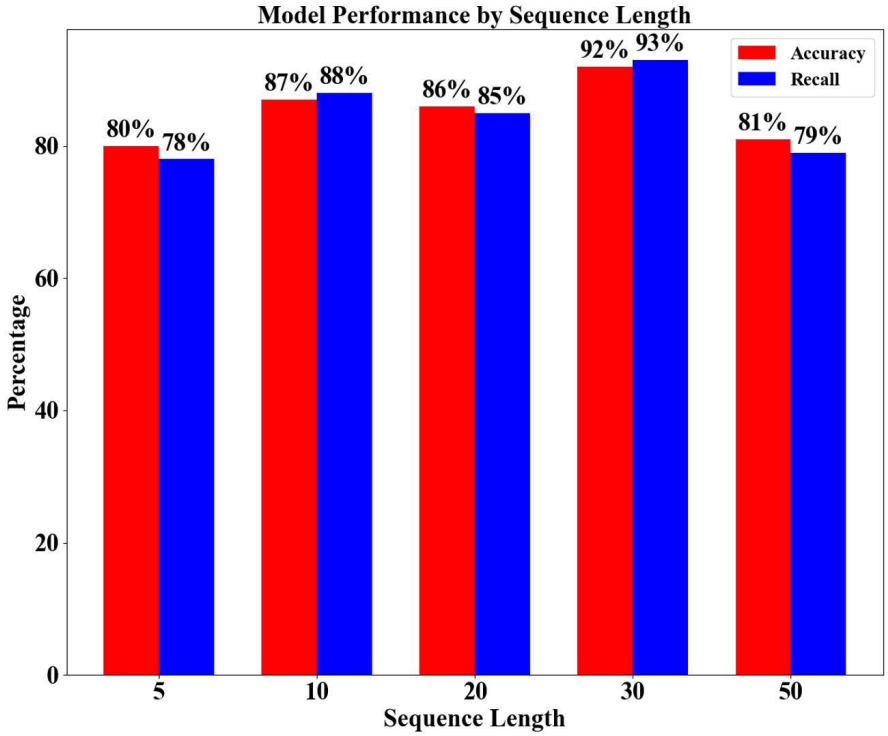


Figure 7. Model Performance and Time Step Length.

The selection of the reconstruction error threshold in the unsupervised model determines the final performance, requiring a balance between reducing false alarms and avoiding missed alarms. After repeated experiments and validation, this study selected a reconstruction error threshold of 1.041. The model's performance with different reconstruction error thresholds is shown in Figure 8. The figure shows that as the threshold is adjusted, the model's performance changes significantly. As the threshold increases, the model's accuracy generally increases, false alarm rates decrease, and missed alarm rates increase, indicating that raising the reconstruction error threshold helps reduce false alarms but makes the model prone to missed alarms. At a threshold of 1.041, the model achieves the highest accuracy of 92.51%. Under this threshold, the false alarm rate is 7.71%, and the missed alarm rate is 6.87%, indicating that this threshold balances the ratio of missed alarms and false alarms while ensuring high accuracy, demonstrating good overall performance.

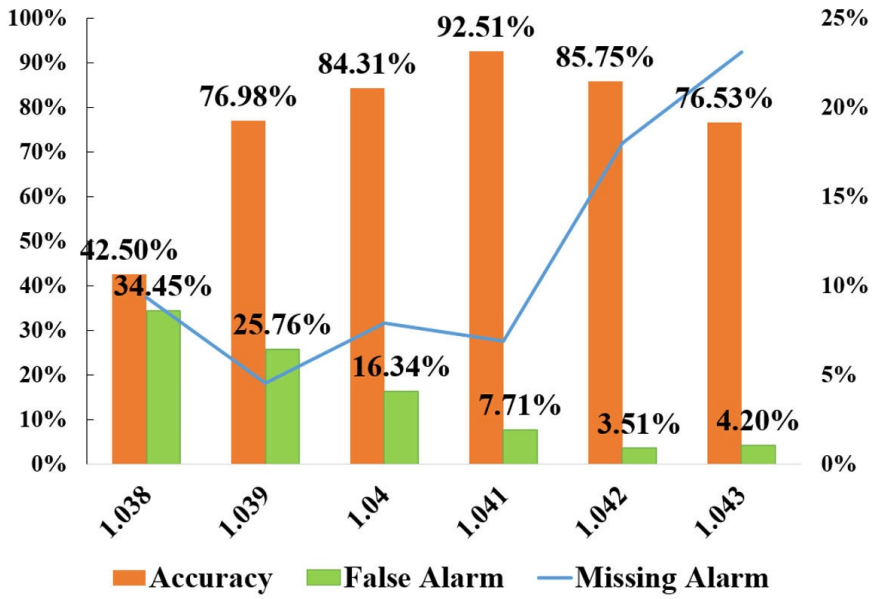


Figure 8. Model Performance and Threshold Selection.

5.2. Algorithm Performance Comparison

To further evaluate the performance of the BiLSTM-AE model, this paper constructs five comparison models: Random Forest, Xgboost, LSTM, BiLSTM, and LSTM-AE. These five models are classic intelligent models in the field of lost circulation and have demonstrated strong and good performance in lost circulation monitoring. Random Forest and Xgboost are non-sequential models, while LSTM, BiLSTM, and LSTM-AE are sequential models. The confusion matrices of different algorithm models on the test set are shown in Figure 9.

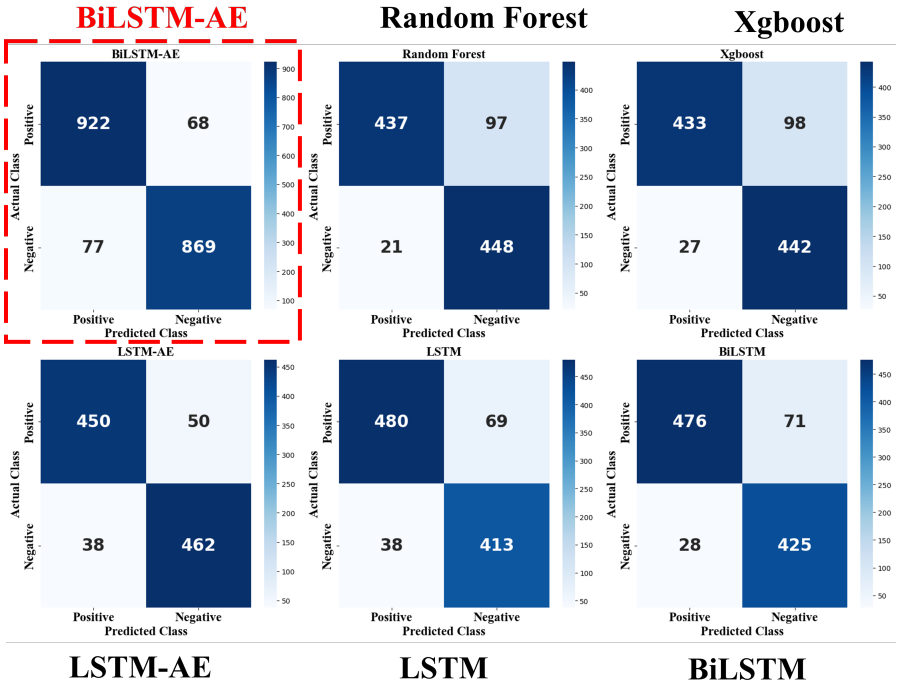


Figure 9. Confusion Matrix.

The comparison results of algorithm performance are shown in Figure 10. From the figure, it can be seen that sequential models perform better overall than non-sequential models, with higher overall accuracy. Among the sequential models, the unsupervised

models LSTM-AE and BiLSTM-AE perform better than the supervised models, with superior accuracy, recall, missing alarm rate, and false alarm rate. The BiLSTM-AE model performs the best, with an accuracy of 92.51%, a missing alarm rate of 6.87%, a false alarm rate of 7.71%, and a recall rate of 93.13%.

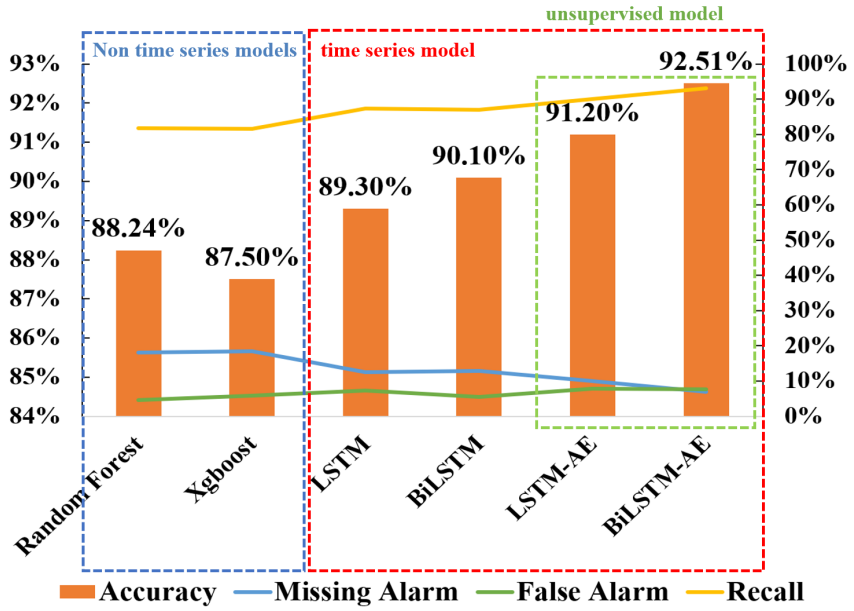


Figure 10. Algorithm Performance Comparison.

The comparison of these models shows that the BiLSTM-AE model demonstrates superior performance in lost circulation risk monitoring, highlighting the rationality and applicability of unsupervised algorithms and the consideration of data temporal fluctuation characteristics in the field of risk monitoring.

5.3. Case Analysis

A typical lost circulation case from an oil field was selected for analysis to visually demonstrate the effectiveness of the BiLSTM-AE model in lost circulation monitoring. The lost circulation monitoring results of different models are shown in Figure 11. The drilling site reported the occurrence of lost circulation after the parameters changed drastically. The alarm time of the LSTM model was slightly earlier than the reported time, while the BiLSTM-AE model issued an alarm at the early stage of the slow parameter decline. The alarm time of the BiLSTM-AE model was close to the corrected actual occurrence time of lost circulation and was about 2 min earlier than the reported time. This highlights the rationality and timeliness of the proposed model, which is significant for timely plugging and reducing drilling fluid leakage.

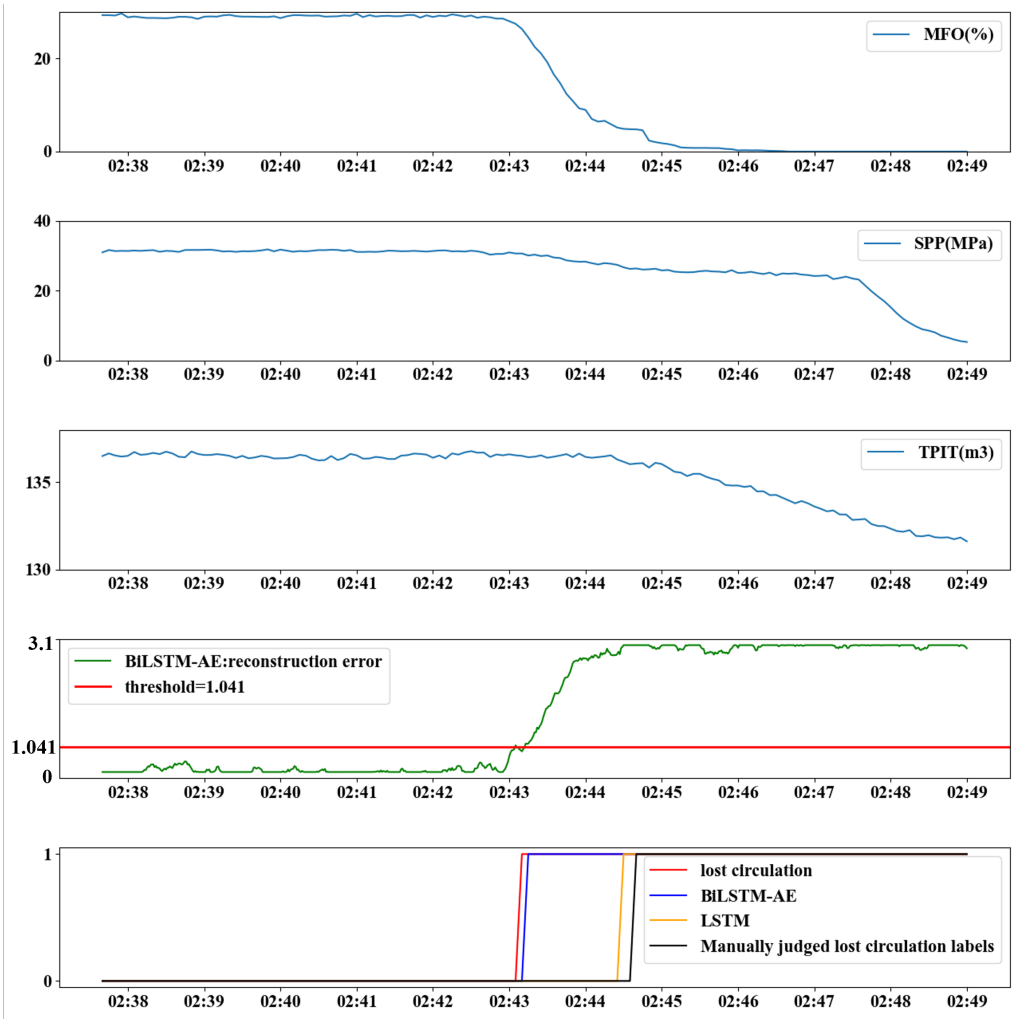


Figure 11. Model Prediction and Lost Circulation Labels.

6. Conclusions

In this paper, a lost circulation monitoring model based on an unsupervised time series autoencoder (BiLSTM-AE) is proposed. This model comprehensively extracts contextual information from time series data. The experimental results show that the model achieves an accuracy of 92.51% on the test set, demonstrating its high accuracy and reliability in practical applications. The relevant patterns and insights obtained during the research can be summarized as follows:

1. Lost circulation exhibits significant temporal evolution characteristics. Compared to non-sequential algorithms, sequential algorithms that consider the temporal variation of sequence data perform better in lost circulation monitoring and diagnosis.

2. Existing data issues severely limit the effectiveness of supervised intelligent algorithms. The proposed unsupervised model effectively overcomes data limitations. One important advantage is that the training set of the model consists of normal drilling time series data. Therefore, the model can detect anomalies by identifying deviations from normal patterns, showing flexibility and adaptability in handling lost circulation risks.

The data used in this study primarily come from normal drilling processes, with relatively fewer well leakage data. Although unsupervised models can partially overcome this issue, the diversity and representativeness of the data remain essential factors affecting the model's generalization ability. In future research, we will further improve and optimize the model to meet the lost circulation monitoring needs under different complex working conditions.

Author Contributions: Conceptualization, L.W. and J.W.; Methodology, L.W. and Z.Z. (Ziyue Zhang); Software, X.W., Z.Z. (Ziyue Zhang) and Z.Z. (Zhaopeng Zhu); Validation, L.W., G.Z. and Q.Z.; Formal analysis, X.W.; Investigation, L.W., J.W. and Z.Z. (Zhaopeng Zhu); Resources, P.D. Data curation, X.W. and Z.Z. (Ziyue Zhang); Writing—original draft, Z.Z. (Ziyue Zhang); Writing—review & editing, G.Z. and J.W.; Visualization, X.W. and P.D. Supervision, Q.Z., P.D. and Z.Z. (Zhaopeng Zhu); Project administration, L.W. and Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key Research and Development Project, grant number 2019YFA07083004, National Science Foundation for Distinguished Young Scholars, grant number 52125401, Science Foundation of China University of Petroleum, Beijing (No. 2462022QNXZ004), the National Natural Science Foundation of China (Grant No. 52274019), Science Foundation of China University of Petroleum, Beijing, grant number 2462022SZBH002.

Data Availability Statement: The data are not publicly available due to involve the information on Chinese oil fields and need to be kept confidential.

Conflicts of Interest: Authors Liwei Wu, Xiaopeng Wang, Guowei Zhu, Qilong Zhang, Pinghua Dong were employed by the company CNOOC China Limited, Jiangtao Wang was employed by the company China France Bohai Geoservices Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Liu, S. Development of Drilling Fluid Level Monitoring and Automatic Grouting Device. *Pet. Mach.* **2006**, *02*, 29–30+78.
2. Tan, D. Research and Application of Overflow Monitoring Methods in Pressure-Sensitive Formation Drilling. Master's Thesis, Southwest Petroleum University, Chengdu, China, 2018.
3. Liu, B.; Li, X.; Li, S.; Tan, J.; Wang, G.; Liu, H. Well Leakage Prediction Based on Support Vector Regression. *Drill. Prod. Technol.* **2019**, *42*, 17–20+1–2.
4. Yingzhuo, X.; Liupeng, W.; Yang, Z.; Mingzhen, L. Real-time monitoring and early warning of well leakage based on big data analysis. *J. Phys. Conf. Ser.* **2021**, *1894*, 012084. [CrossRef]
5. Song, Y. Study on Intelligent Recognition Method of Well Leakage Risk Status Based on Extreme Learning Machine. Master's Thesis, China University of Petroleum (Beijing), Beijing, China, 2022.
6. Baek, S.; Bacon, D.H.; Huerta, N.J. Enabling site-specific well leakage risk estimation during geologic carbon sequestration using a modular deep-learning-based wellbore leakage model. *Int. J. Greenh. Gas Control* **2023**, *126*, 103903. [CrossRef]
7. Li, C.; Zhang, X.; Du, X.; Guo, Z. Well Leakage Prediction Based on Convolutional Neural Networks. *Lab. Res. Explor.* **2023**, *42*, 102–106+142. [CrossRef]
8. Zheng, Z.; Song, Z.; Chen, B.; He, P.; Ji, C.; Xu, T. Research on Well Leakage Early Warning Model Based on XGBoost Algorithm. *Appl. Petrochem.* **2023**, *42*, 112–115.
9. Sun, W.; Bu, S.; Zhang, D.; Li, W.; Liu, K.; Dai, Y. Intelligent Monitoring Method of Drilling Fluid Minor Loss Based on DCC-LSTM. *Nat. Gas Ind.* **2023**, *43*, 141–148.
10. Luo, M.; Li, S.; Peng, W.; Zhou, Z. Well Leakage Accident Prediction Based on Deep Convolutional Feature Reconstruction. *Comput. Simul.* **2023**, *40*, 82–88.
11. Li, H.; Wang, D.; Bu, S.; Kuang, J. Intelligent Lost Circulation Monitoring with Unbalanced Data Samples. In Proceedings of the 2023 5th International Conference on Intelligent Control, Measurement and Signal Processing (ICMSP), Chengdu, China, 19–21 May 2023.
12. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [CrossRef]
13. Li, Q.; Wu, X.; Zhou, X. Tool wear state recognition based on optimized bidirectional long short-term memory network. *Mech. Des. Res.* **2023**, *39*, 147–151. [CrossRef]
14. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
15. Guoliang, Z.; Xiaojun, G. A Review of Network Anomaly Detection Based on Autoencoders. *J. Inf. Secur.* **2023**, *8*, 81–94. [CrossRef]
16. Hanyu, Y.; Xiaoyong, Z.; Lei, W. A Review of Data Normalization Methods. *Comput. Eng. Appl.* **2023**, *59*, 13–22.
17. Hao, C. Analysis of Water Quality Changes in Shanghai Water Bodies Based on Spearman Rank Correlation Coefficient Method. *Environ. Technol.* **2020**, *33*, 28–33. [CrossRef]
18. Haibiao, W. Research on Intelligent Identification and Treatment Decision of Well Leakage. Master's Thesis, Southwest Petroleum University, Chengdu, China, 2017.
19. Yanfeng, L.; Chao, M.; Boreng, D.; Xinhui, Z. Prediction of Drilling Loss Accidents Based on Dynamic Features and Deep Neural Networks. *J. Xi'an Shiyou Univ. (Natural Sci. Ed.)* **2022**, *37*, 64–69.

20. Sun, W.; Feng, J.; Zahng, D.; Li, W.; Liu, K.; Dai, Y. An intelligent lost circulation recognition method using LSTM-autoencoder and ensemble learning. *Pet. Drill. Tech.* **2024**, *52*, 1–7. [CrossRef]
21. Weifeng, S.; Kai, L.; Dezhi, Z.; Weihua, L.; Liming, X.; Yongshou, D. Overflow and Well Leakage Monitoring Method Combining Drilling Conditions and Bi-GRU. *Pet. Drill. Tech.* **2023**, *51*, 37–44.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.