processes

MDPI

Article

A Model-Driven Approach to Extract Multi-Source Fault Features of a Screw Pump

Weigang Wen [1], Jingqi Qin [1,*], Xiangru Xu [2], Kaifu Mi [2] and Meng Zhou [1]

[1] School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China; wgwen@bjtu.edu.cn (W.W.); zmetyee@163.com (M.Z.)
[2] Branch of Industry, Beijing Petroleum Machinery Co., Ltd., Beijing 102206, China; xuxrdr@cnpc.com.cn (X.X.); mikfdr@cnpc.com.cn (K.M.)
* Correspondence: qjq8787@163.com

Abstract: Screw pumps' faulty working conditions affect the stability of oil production. At project sites, different sensors are used simultaneously to collect multi-dimensional signals; the data fault labels and location are not clear, and how to comprehensively use multi-source information in effective fault feature extraction has become an urgent issue. Existing diagnostic methods use a single signal or part of a signal and do not fully utilize the acquired signal, which makes it difficult to achieve the required accuracy of diagnostic results. This paper focuses on the model-driven approach to extract multi-source fault features of screw pumps. Firstly, it constructs a fault data model (FDM) by analyzing the fault mechanism of the screw pump. Secondly, it uses the FDM to select an effective data set. Thirdly, it constructs a multi-dimensional fault feature extraction model (MDFEM) to extract featured signal features and data features, for which we also comprehensively used multi-source signals in effective fault feature extraction, while other traditional methods only use one or two signals. Finally, after feature selection, unsupervised fault diagnosis was achieved by using the k-means method. After experimental verification, the method can comprehensively use multi-source information to construct an effective data set and extract multi-dimensional, effective fault features for screw pump fault diagnosis.

Keywords: feature extraction; model-driven; multi-source information; screw pump; fault diagnosis

Citation: Wen, W.; Qin, J.; Xu, X.; Mi, K.; Zhou, M. A Model-Driven Approach to Extract Multi-Source Fault Features of a Screw Pump. *Processes* **2024**, *12*, 2571. https://doi.org/10.3390/pr12112571

Academic Editors: Yuhe Wang and Chunhui Zhao

Received: 24 September 2024
Revised: 5 November 2024
Accepted: 15 November 2024
Published: 17 November 2024

1. Introduction

The screw pump has become one of the widely used oil lifting methods due to the advantages of small size of system equipment, easy maintenance and management, smooth liquid flow, and high pumping efficiency. With the development of oilfield exploration, the lifting height is increasing. The working conditions of the screw pump are affected by various factors when it works in a complex environment, and the faults, such as rod and pipe breakage, pump leakage, sand jamming, and waxing of tubing, occur from time to time [1,2], which seriously restricts the production efficiency of the oil wells. Therefore, screw pump fault diagnosis is of great significance. The traditional screw pump fault diagnosis methods require a lot of expert experience, and using single current or torque and other signals for diagnosis is ineffective.

In recent years, scholars from various countries have conducted extensive research on the fault diagnosis of screw pumps. Aiming at the screw pump fault diagnosis problems of inefficiency and lack of precision, one proposal to improve the wavelet packet transform is by introducing the idea of power spectrum refinement combined with cuckoo search (CS) to optimize the diagnosis method of the back propagation (BP) neural network [3]. Based on the statistical process control, extended discriminant criterion, and the multi-parameter rule, the field fault diagnosis model of electric submersible screw pump unit is constructed, and a multi-parameter process control fault diagnosis method is proposed [4]. A supervised

neural network was established using Bayes classification decision theory to build a PNN model for fault classification of a certain type of screw pump [5]. Scholars proposed that the current method and holding pressure method can be used to comprehensively diagnose the pumping condition of screw pump wells, combined with production, liquid level, and other parameters to synthesize the research and judgment, in order to accurately determine the specific operation of downhole pumps [6]. One study selected rod torque and axial force as parameters for comparing working condition status, stating that using a diagnosis model to calculate the reasonable area of two parameters and comparing them with the actual test data can diagnose the various situations, determine whether there is a fault and analyzing failure [7]. Another paper proposed an unsupervised fault diagnosis methodology to leverage readily available dynamometer cards (DCs) to diagnose collected unlabeled MPCs, and a mathematical model of the SRPS was presented to convert actual DCs to MPCs [8]. Ren Weijian proposed using a wavelet packet to filter and eliminate noise from an active power signal and decomposed fault signal, then they used an Elman neural network to identify the decomposed fault feature [9]. Wavelet packet theory was used to decompose and reconstruct the active power signal of a submersible screw pump, extracted the main fault information contained in the power signal, and constructed the fault feature vector of a submersible screw pump combined with parameters such as output, oil pressure, casing pressure, and dynamic liquid level [10]. A thesis aimed to accurately and efficiently identify the fault forms of a submersible screw pump, and proposed a fault diagnosis method of the submersible screw pump based on random forest. An HDFS storage system and MapReduce processing system were established based on the Hadoop big data processing platform [11]. Xu Jun developed a screw pump condition-monitoring system based on dual micro-controllers, which uses real-time drive electrical parameters to calculate the rod torque and speed and to make a comparative analysis, and achieved the fault diagnosis of the screw pump [12]. Min Li established a screw pump fault diagnosis expert system based on fuzzy neural network. The test results showed that the diagnosis of this fault diagnosis expert system is operable, and the fuzzy neural network is reliable, which enriched the diagnosis method of the screw pump well [13]. Xue Jianquan proposed a screw pump diagnosis method based on the BP neural network and expert system, and developed the fault diagnosis software with Basic and Matlab nnet toolbox, and the test results of the well plant data proved the feasibility of the diagnosis method [14]. Qu Wentao used the node system method to establish a relationship model between active power and pump energy consumption and analyzed the influence of different faults on the screw operating performance [15]. Chen Shiwen calculated and analyzed the force of the screw pump and rod and proposed the method of dividing the fault feature into different thresholds and diagnosing screw pump working conditions by means of the support vector machine algorithm [16]. Zheng Chunfeng studied the system and operation performance of an electric submersible screw pump, analyzed the correspondence between operation parameters and fault types, and used a BP neural network to diagnose faults of electric submersible screw pump [17].

However, the project site data are multi-dimensional signals collected simultaneously by multiple sensors, including current, voltage, power, torque, rotation speed, load, oil pressure, and casing pressure, and the data type is complex, and the collection period is long. Additionally, the fault label of the data is not clear, and the position of the fault information in the whole signal data is not clear as well, which will lead to a low diagnostic accuracy if using this kind of signal data to carry out the fault diagnosis. Traditional screw pump fault diagnosis methods rely on a large number of expert experiences, and diagnosis accuracy using a single signal is low [18,19]. The current method is affected by the combination of all parts of the system [20]; the torque method has an error between the calculated result and the actual torque; the fluid production method does not directly reflect the type of fault; the pressure method involves holding the wellhead to a higher pressure, which must be used on specific conditions. In the field of fault diagnosis, multi-source information technology comprehensively collects equipment fault state information by

using multiple sensors, taps the coupled complementary information between multi-source sensor data, takes multi-dimensional feature fusion analysis as a way to greatly improve the reliability and accuracy of fault diagnosis, and overcomes the shortcomings such as limited fault information contained in a single segmentation and large uncertainty [21–26].

The main contributions of the paper are described as follows:

(1) The fault data model (FDM) is proposed and applied to select an effective data set from the original data with no fault labels and unclear fault locations.
(2) A multi-dimensional fault feature extraction model (MDFEM) is proposed and applied to extract featured signal feature and data feature from multi-source information.

Therefore, this paper proposes a multi-source fault feature extraction method for a screw pump based on model-driven data. The chapters of the paper are as follows: Section 2 analyzes the screw pump fault mechanism and constructs the FDM; Section 3 uses the FDM to select the fault data and construct an effective data set, then constructs a MDFEM, and extracts multi-dimensional fault features; Section 4 verifies the validity and accuracy of the proposed method; and Section 5 is the conclusion.

2. Fault Model Construction

2.1. Methodological Framework

The flow of the multi-source fault feature extraction method for a screw pump based on model-driven data is shown in Figure 1. The process is described as follows: (1) analyzing the fault mechanism and establishing the current, load, rotational speed, and oil pressure fault mechanism; (2) revealing the mechanism characterization and establishing the FDM; (3) analyzing the original data and carrying out data pre-processing, including data cleaning, normalization, and slice enhancement; (4) using the FDM to select fault data and constructing an effective data set; (5) studying the multi-dimensional feature extraction method, extracting signal and data features, and obtaining effective multi-dimensional features; (6) using the k-means unsupervised clustering method to carry out experimental validation to realize the diagnosis of screw pump faults.

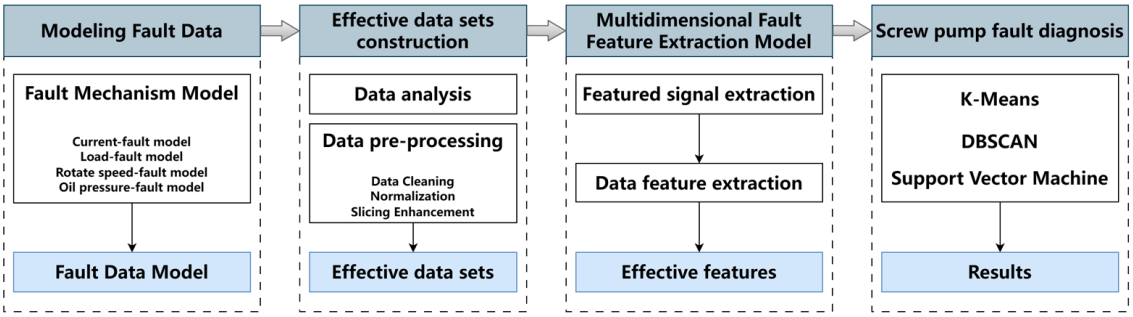


Figure 1. Screw pump fault diagnosis framework.

2.2. Fault Mechanism Model

(1) Current Fault Model

The current method is the quantitative long-time measurement of parameters such as operating current, voltage, and power of the drive motor using specialized instruments. The motor input current is deduced based on the motor output power:

$$I_1 = \sqrt{\frac{N_1^2}{N_e^2} \times \left(I_N^2 - I_0^2\right) + I_0^2} = \sqrt{\frac{M_g^2 n_g^2}{9549^2 \times \eta_t^2 N_e^2} \times \left(I_N^2 - I_0^2\right) + I_0^2} \quad (1)$$

where I_1 is the motor input current, A; N_e is the power under a rated load, kW; I_N is the stator current of the motor under a rated load, A; I_0 is the no-load current of the motor, A. N_1 is the output shaft power of the motor, kW; M_g is the driving torque of the rod, N·m; n_g

is the rotational speed of the rod, r/min; and η_t is the total transmission efficiency of the motor, %.

After analysis, the relationship between electrical parameters and torque can be obtained as shown below:

$$I_1^2 \propto M_g^2 \tag{2}$$

The analysis shows that the square of the motor input current is proportional to the square of the rod torque.

According to Equation (2), it can be deduced that when the current drops to the no-load current, which means the rod torque drops to the no-load torque, the oil rod or oil pipe may break off; when the current is smaller than the lower limit of the reasonable range and higher than the lower limit of the limit range, the pump may leak or the oil pipe may leak; when the current fluctuates in the upper limit of the limit range, the oil pipe may be waxed; when the current increases to the outside of the limit range, the stator may be dissolved or the parameters may be high; when the current is smaller than the lower limit of the reasonable range and fluctuates, the stator may be waxed. When the current fluctuates, the stator may be degumming.

(2) Load Fault Model

The axial load F on the screw pump rod can be expressed by the following formula:

$$F = F_1 + F_2 - F_3 - F_4 \tag{3}$$

where F_1 is the rod's own gravity, N; F_2 is the axial load generated by the pressure difference between the inlet and outlet of the pump, N; F_3 is the friction force between the well fluid and the rod when it flows upward in the pipe, N; and F_4 is the upward buoyancy force that the rod receives in the well fluid, N.

The gravity of the pumping rod can be expressed by the following equation:

$$F_1 = GL \tag{4}$$

where G is the linear density of the rod, N/m; L is the total length of the pumping rod, m.

The axial load generated by the pressure difference between the inlet and outlet of the pump can be expressed by the following equation:

$$F_2 = \left(\pi R^2 + 16eR\right)\Delta p \tag{5}$$

where e is the eccentric moment of the screw pump, m; R is the rotor radius of the screw pump, m; Δp is the differential pressure between the inlet and outlet of the pump, MPa.

The friction force between the well fluid and the rod can be expressed by the following equation:

$$F_3 = 2\pi\mu_l e_l \nu \Delta L \tag{6}$$

where $e_l = \frac{m^2-1}{(m^2+1)\ln m-(m^2-1)}$, $m = \frac{d_t}{d_r}$, ΔL is the length of the section, m; μ_l is the average viscosity of the well fluid, mPa·s; ν is the average flow speed of the well fluid in the section, m/s.

The buoyancy of the rod generated by well fluid can be expressed by the following equation:

$$F_4 = \rho_l g \pi \left(\frac{d}{2}\right)^2 L \tag{7}$$

where g is the acceleration of gravity, N/kg; ρ_l is the density of the fluid in the wellbore, kg/m^3.

After analyzing, it can be obtained that when F_2 decreases, it means the liquid in the screw pump decreases; when F_3 increases, it means the oil velocity increases, i.e., the oil production increases. According to Equations (3)–(7), we can conclude that when F is zero,

there may be a rod break; when F is reduced, there may be an oil pipe leakage, a broken oil pipe, or pump leakage, etc.; when F is increased, there may be waxing of the oil pipe or high parameters.

(3) Rotation Speed Fault Model

The rotate speed of a screw pump is closely related to the oil production, as shown in the following equation:

$$Q = 1440 \times 4nEDT \tag{8}$$

where Q is the theoretical oil production, m^3/d; n is the rotor speed, r/min; E is the eccentricity of the rotor, m; D is the truncated circle diameter of the rotor, m; T is the stator lead, m.

The actual oil production of a screw pump can be expressed by the following equation:

$$Q' = Q\eta_v = \eta_v 1440 \times 4EDTn \tag{9}$$

where η_v is the volumetric efficiency of the screw pump; Q' is the actual oil production, m^3/d.

It can be seen that after the structural parameters E, D, and T of the screw pump are determined, the oil production is only related to the rotational speed n and the volumetric efficiency η_v, and the rotational speed needs to be increased in order to achieve higher oil production.

Increasing the rotational speed of the screw pump can improve oil production, but extremely high rotational speed will lead to an increase in the centrifugal force of the rod, which will cause vibration and decrease the oil lifting height. At the same time, high speed rotation will also accelerate the wear of stator rubber.

After analyzing, it can be concluded from Equation (9) that when n is zero, the pump may be jammed; when n is reduced, the rod and pipe may show biased wear, or the oil pipe may be waxed, etc.; when n increases, there may be an oil rod break, oil pipe breakage, oil pipe leakage, or pump leakage, and so on.

(4) Oil Pressure Fault Model

Letting the oil pressure be P at the moment of starting pumping t, the relationship between pressure and volume is:

$$\beta_m V_t \Delta P = V_p \Delta t \tag{10}$$

where β_m is the compression coefficient of the gas–liquid mixture in the oil pipe; Δt is the amount of time change, s; ΔP is the amount of pressure change, MPa; and V_t is the pumping volume flow rate m^3/s.

When pipe leakage occurs, the relationship between the leakage flow rate and the pressure difference between the inside and outside of the oil pipe at the leaking place is:

$$V_{le} = \varepsilon \varphi A \sqrt{2gh} \tag{11}$$

where V_{le} is the leakage flow rate, m^3/s; ε is the shrinkage coefficient of the leakage section due to the inertia of the liquid; φ is the flow coefficient associated with the liquid; A is the size of the cross-sectional area at the orifice of the liquid leakage, m^2; g is the acceleration of gravity, m/s^2; h is the difference in liquid pressure between the inside of the tubing at the location of the leakage and the outside of the tubing, m.

Since the leakage flow rate V_{le} is a function of pressure P, and increases nonlinearly with P, the relationship between pressure and time is broken when leakage occurs, resulting in a slower rate of pressure increase.

2.3. Fault Data Model

After the mechanism analysis, a FDM can be established to reveal the mechanism characterization from various aspects, as shown in Figure 2.

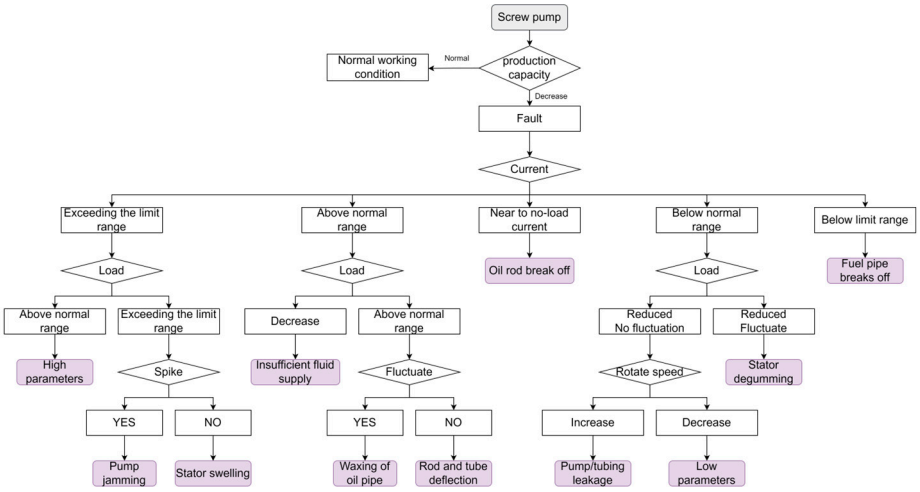


Figure 2. Screw pump fault data model.

The screw pump FDM uses the mechanism analysis of the screw pump and historical data to hierarchically reveal the mechanistic characterization of the faults, such as pump jamming, stator swelling, and pipe waxing, to classify faults and obtain the corresponding fault labels.

3. Effective Data and Feature

3.1. Data Preprocessing

By analyzing the actual working data for pump wells of an oil field in Xinjiang province, China, it can be concluded that the actual signals are collected by a variety of sensors simultaneously, the data types are complex, and the collection cycle is long, with a collection length of about 1150 sampling points. The signals start from about the 100th sampling point, showing a gradual upward trend, and continue to about the 300th sampling point, and fault data are concentrated in the area between about the 100th and 500th sampling points; the abnormal data of the oil pressure signals are also found in the area between about the 100th and 400th sampling points. From the above analysis, it can be seen that most of the signal fault data are concentrated in a small range, and do not cover the entire signal collection length. If the original data are used for fault diagnosis, the diagnostic accuracy will be adversely affected.

(1) Data cleaning

Voltage and current signals contain noise because of the non-stationarity of the system; and pressure and load signals contain random fluctuation caused by the oil pump and environment. This may lead to inaccurate features because some features are sensitive to small fluctuations, while others are the opposite.

The actual data collected from the sensors contain abnormal values. The 3σ method is considered according to the data characteristics.

The 3σ method default is that data obey normal distribution, the probability of data distributed within the interval $(\mu - 3\sigma,\ \mu + 3\sigma)$ is 99.73%, and data distributed outside the interval is considered outlier data. In this section, the sliding window method is used for outlier determination. The standard deviation of the data within the sliding window is calculated as follows:

$$avg = \frac{x_1 + x_2 + x_3 + \cdots + x_l}{l} \tag{12}$$

$$\sigma = \sqrt{\frac{1}{l}\sum_{i=1}^{l} x_i - avg^2} \tag{13}$$

where avg is the mean value of each parameter within the sliding window; σ is the standard deviation of the group of data.

According to Equations (12) and (13), if the distance between a data value and the avg is greater than 3σ, the data will be rounded off, and the rounded data will be filled in using the mean of the neighbouring numbers.

(2) Data normalization

There are different unit dimension features that cannot provide an evaluation in such a multidimensional system. The purpose of normalization is to make data be limited to a certain range (e.g., [0, 1] or [−1, 1]), thus eliminating the adverse effects caused by singular sample data. As the sample data does not involve distance measures or covariance calculation, and the data does not meet the normal distribution, the maximum–minimum normalization (min-max normalization) can be used, and the linear function will convert the original data to the range of [0 1]. The equation is as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{14}$$

where x' is the data after normalization; x is the data before normalization.

(3) Data slicing enhancements

Using the proposed FDM, the collected signals with unequal lengths are cut into signal segments with a length of 500 sampling points. Data slicing can reduce the data volume of a piece of input signal and improve the model computation efficiency; and the data segments containing fault information can be intercepted, so that the fault information features are more obvious, which is conducive to improving the accuracy of the results.

Due to the limited number of actual fault data samples at the site, the data set needs to be enhanced in order to increase the generalization ability of the model. Since the collected samples involving fault data are generally longer than the required input signal, the data enhancement method proposed in this paper is the sliding window sampling (SWS) method. Sliding window sampling takes a time window of the same size as the standard sample, with a step size smaller than the standard sample. The sampling method is shown in Figure 3. The blue irregular curve represents the timing signal. The green rectangular box represents the time window with a width of "sample length". After the first window captures the signal fragment, it moves a distance of "sliding step" to the right and become the second window. And so on, after moving to the right for n − 1 "sliding steps", the nth time window is obtained. In this way, a total of n signal segments are obtained. The data set after data enhancement contains a total of 13,000 samples.

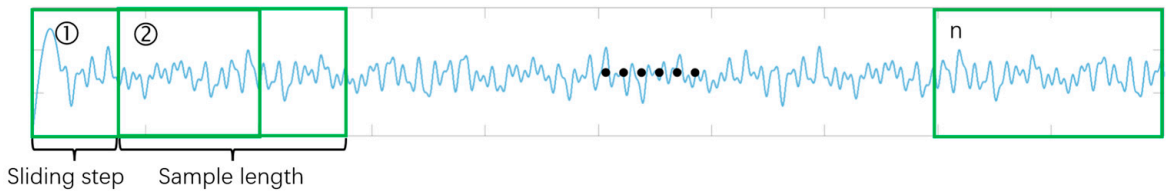


Figure 3. Slide sampling method.

3.2. Effective Data Set

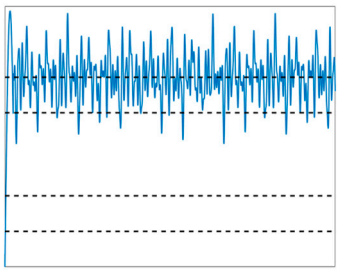
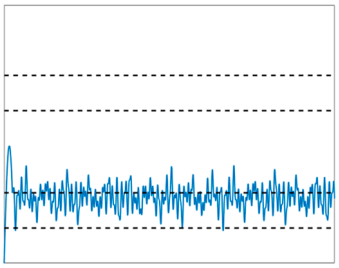
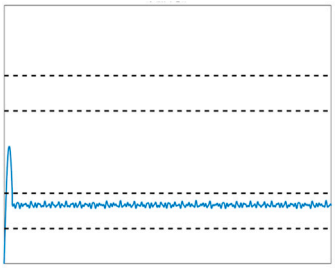
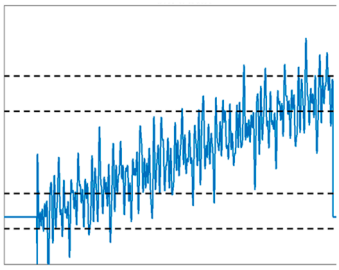
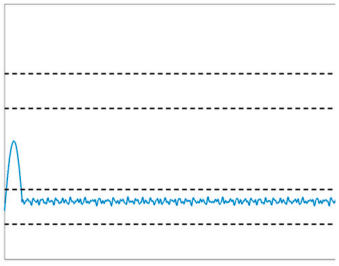
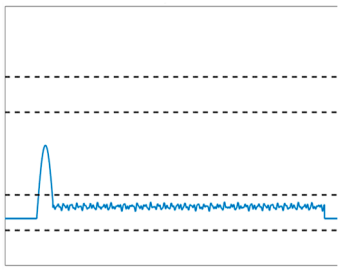
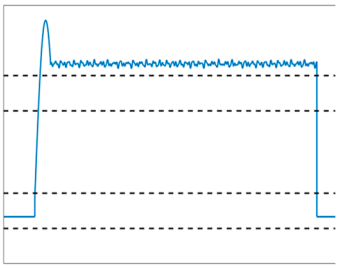
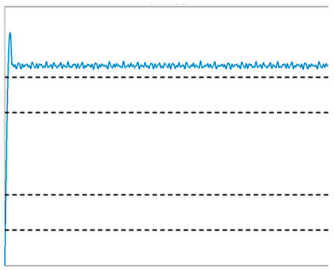
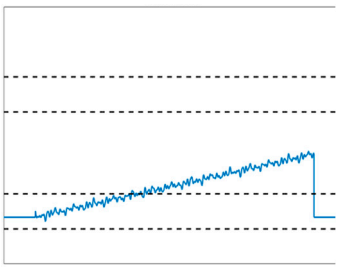
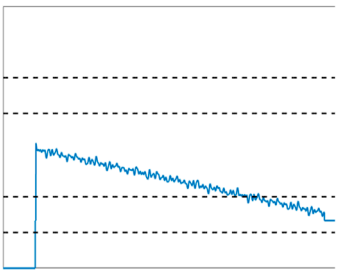
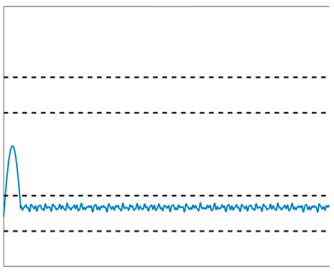
After data pre-processing, the effective data set can be constructed by using FDM to select data containing fault information from original signals.

The FDM provides rules and principles for selecting effective data from a long-period signal. FDM describes the change modality of current, load, rotate speed, oil pressure, etc. on the fault working conditions, including average value, fluctuation, increase or decrease, and so on. Taking oil pipe waxing as an example, when it happens, the current exceeds the maximum limit and wildly fluctuates; the load increases from a low limit to a nearly maximum limit, and wildly fluctuates; the rotate speed steadily stays below the normal

range and larger than the minimum limit; the oil pressure increases stably, with small fluctuations.

It is important for the improvement of the accuracy and efficiency of the screw pump fault diagnosis. Table 1 shows some fault data.

Table 1. Some fault data.

	Waxing of Oil Pipe	**Stator Degumming**	**Pump Leakage**
Current-time diagram			
Load-time diagram			
Rotate speed-time diagram			
Oil pressure-time diagram			

3.3. Multidimensional Fault Feature Extraction

According to the fluctuation of the featured signal we can extract data features that reflect the changes in the featured signal, that is, reflecting the changes in the production data, and then we can judge the fault types.

3.3.1. Extraction of Signal Feature

Since the collected original signals contain 17-dimensional signals such as voltage, current, power, load, torque, and oil pressure, which are a large amount of data and have considerable redundancy, the 17-dimensional data will be analyzed by correlation.

The Pearson correlation coefficient method is defined as the covariance product of two continuously distributed parameters x and y divided by their standard deviation, as in the following equation:

$$\rho_{x,y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \tag{15}$$

where σ_x σ_y is the standard deviation of variables x and y, respectively, and $\mathrm{cov}(x,y)$ denotes the covariance of the two variables. The covariance formula is provided in the following equation:

$$\mathrm{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} \tag{16}$$

The correlation coefficient reflects the strength and direction of correlation between variables. Table 2 shows the general pattern of correlation coefficients and strength of correlation [27–30].

Table 2. Correlation table.

Relevance	Negative Value	Positive Value
Irrelevant	−0.09~0.00	0.00~0.09
Low relevance	−0.50~0.10	0.10~0.50
High relevance	−0.80~−0.50	0.50~0.80
Significant relevance	−0.90~−10	0.9~1

A heat map of correlation coefficients using data from the normal operating conditions of well number z70-01-3060601 in a certain oil field is shown in Figure 4. According to the correlation coefficient, it can be seen that the voltage, current, and power correlation is higher than 0.91, and the correlation of current and torque is 0.95, so the voltage, power, and torque signals are discarded; the correlation of current and load is 0.20; the correlation of current and oil pressure is 0.49; and the correlation of load and speed is 0.44. In the case that voltage, power, torque, and oil pressure are discarded, the correlation of current, load, speed, and oil pressure is the lowest, that is, 3.48. It can be concluded that the 4-dimensional signals of current, load, speed, and oil pressure can be used as the signal feature.

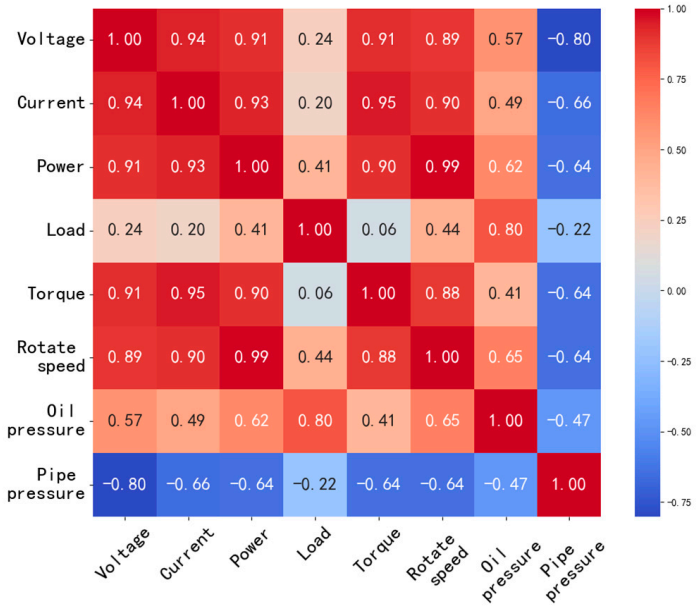


Figure 4. Heat map of signal correlation coefficients.

3.3.2. Extraction Data Feature

Statistical features can comprehensively describe the original signal state from different perspectives, which can quantitatively reflect the degree of signal dispersion, change, and asymmetry, and play an important role in fault diagnosis research.

Based on the fault mechanistic analysis, we chose the mean, variance, and peak-to-peak value (as shown in Table 3) as data features:

Table 3. Characteristics of statistics.

Feature Name	Formula
Mean	$\overline{x} = \frac{1}{N_s}\sum_{i=1}^{N_s} x(i)$
Variance	$\frac{1}{N_s-1}\sum_{i=1}^{N_s}(x(i)-\overline{x})^2$
Peak-to-peak value	$\max(x(i)) - \min(x(i))$

(1) The mean monitors the signal center trend, which can be the most intuitive representation of the signal changes. For example, an extremely low load mean value could mean oil rod break off; an extremely high value could mean pump jamming.
(2) The variance monitors the distribution trend of the signal around the mean value, which can help us better understand the fluctuation of the parameters and trend changes. For example, a high variance could mean waxing of oil pipes.
(3) The peak-to-peak value can be used to describe the magnitude of change of a parameter over a period of time, and can better reflect the fluctuation of the parameter. For example, a high peak-to-peak could mean stator degumming.

The effective features extracted by the multidimensional feature extraction model are shown in Table 4.

Table 4. Effective characteristics.

Current	Load	Rotate Speed	Oil Pressure	
Mean	Cm	Lm	Sm	Pm
Variance	Cv	Lv	Sv	Pv
Peak-to-peak value	Cp	Lp	Sp	Pp

4. Experiments

We chose 10 common types of faults, shown in Table 5, and we set the number of clusters to 10.

Table 5. Common types of faults.

Fault Number	Type of Fault	Reason
0	Oil rod break-off	Excessive torque/tension
1	Oil pipe leakage	Oil pipe corrosion
2	Oil pipe break-off	Anti-rotation anchor damage
3	Waxing of oil pipe	High wax content in oil wells
4	Stator swelling	Liquid-absorbing, expanding
5	Stator Degumming	Low bonding strength
6	Pump leakage	Stator wear and aging
7	Pump jamming	Excessive surplus
8	High parameters	Displacement is greater than the fluid supply capacity
9	Low parameters	Insufficient liquid supply capacity

(1) Comparative experiment I

In order to verify the effectiveness of the effective feature, comparison experiments are conducted using the extracted features (Feature Set-1, Table 6) and remaining features

(Feature Set-2, Table 7) except for the extracted features, respectively. First, the dimensions of the two feature sets are reduced to 2 dimensions using principal component analysis (PCA), and then the K-means algorithm is used to cluster.

Table 6. Feature Set-1.

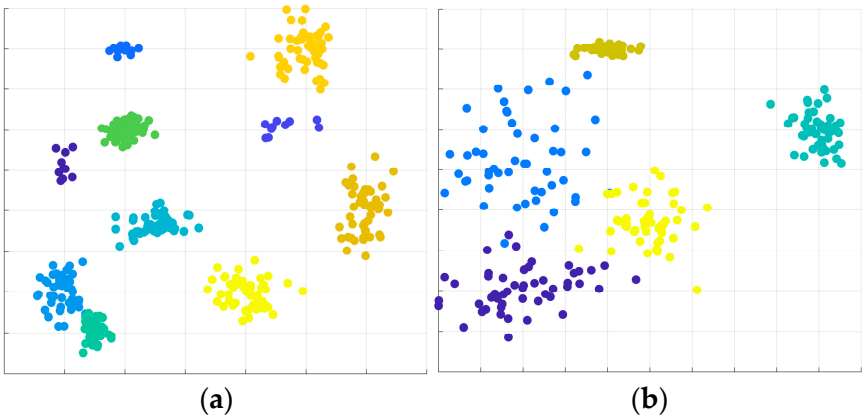
Value Features	Signal Features			
	Current	**Load**	**Rotate Speed**	**Oil Pressure**
Mean	Cm	Lm	Sm	Pm
Variance	Cv	Lv	Sv	Pv
Peak-to-peak value	Cp	Lp	Sp	Pp

Table 7. Feature Set-2.

Value Features	Signal Features			
	Current	**Load**	**Rotate Speed**	**Oil Pressure**
Kurtosis	Ck	Lk	Sk	Pk
Impulse	Ci	Li	Si	Pi

K-means is computationally efficient, especially when the number of clusters and dimensions are not too large. This makes it suitable for large datasets. The output of K-means, i.e., the cluster centroids, is straightforward to interpret. Each centroid represents the average of all points in that cluster, which can help explain the result of each cluster.

The clustering results are shown in Figure 5. It shows that using the Feature Set-1, the data can be effectively classified into 10 classes. Points clustered into one category represent data for one type of fault. The smaller the intra-class distance and the larger the inter-class distance, the better the clustering is, and accordingly, the clearer the distinction among different faults. Different colors correspond to different types of faults. The pump leakage (green cluster) and the oil pipe leakage (light-blue cluster) are similar, so they are closer in the clustering results. However, using Feature Set-2, the data can be classified into only 5 classes, and the intra-class spacing is large and the inter-class spacing is small, so that the clustering results are not effective. Therefore, it can be concluded that the effective features can be extracted using the proposed multidimensional feature extraction model.



(**a**) (**b**)

Figure 5. Results of comparison of Experiment I. (**a**) Clustering results of Feature Set-1; (**b**) clustering results of Feature Set-2.

The Calinski–Harabasz Index (CHI) measures how good the clustering is by the ratio of interclass scatter to intraclass scatter. A higher CHI represents better clustering, as it means that there is more variability between classes as well as more compactness within

classes. Figure 6 shows the CHI of different feature sets with the number of clusters ranging from 4 to 10. It can be seen that Feature Set-1 determined the right cluster number, which means Feature Set-1 is effective for fault diagnosis.

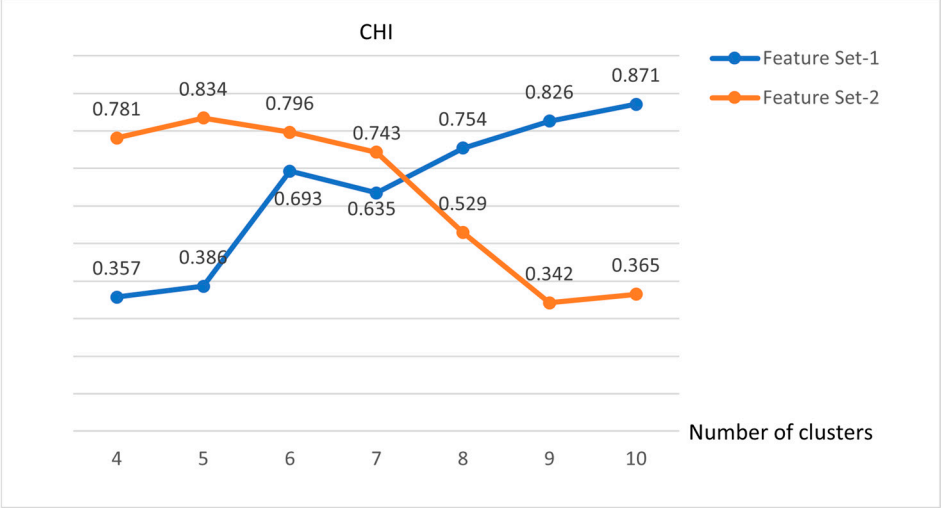
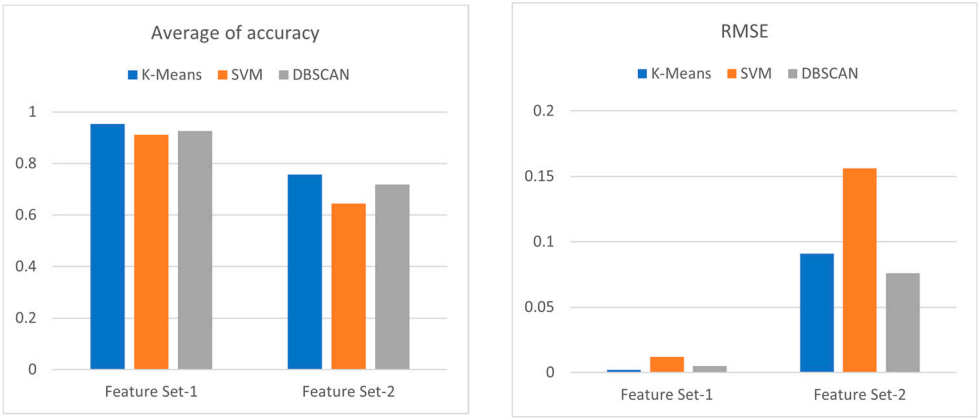


Figure 6. CHI of different feature sets with the number of clusters.

In order to further verify the validity of the model, we reran k-means with 8 different initial random seeds, SVM with 8 different ε-SVR values, and (DBSCAN) with 8 different bandwidths and summarized the results with the test results of K-means, and the average of accuracy and root-mean-square error (RMSE) for different datasets after clustering are shown in Figure 7 and Table 8. It can be seen that, for Feature Set-1, all three methods have high average accuracy and low RMSE for diagnosis results. Therefore, using FDM and MDFEM, we can select effective fault data and extract effective fault features, which have high-quality clustering performance for fault diagnosis.



(**a**) (**b**)

Figure 7. Diagnosis results for different datasets after clustering: (**a**) average of accuracy; (**b**) RMSE.

Table 8. The average of accuracy and RMSE for different datasets.

Datasets	K-Means		SVM		DBSCAN	
	Average of Accuracy (%)	RMSE	Average of Accuracy (%)	RMSE	Average of Accuracy (%)	RMSE
Feature Set-1	95.3	0.002	91.2	0.012	92.6	0.005
Feature Set-2	75.7	0.091	64.5	0.156	71.9	0.076

(2) Comparative experiment II

In order to verify the validity of multi-source signals, this section uses a single signal to extract features for experiments. The mean, variance, and peak-to-peak value of current, load, rotational speed, and oil pressure are extracted, respectively, constituting Feature Set-3 to Feature Set-6 (Tables 9–12), and then the dimensions of the four feature sets are reduced to 2 dimensions using principal component analysis (PCA); then clustering analysis is carried out using the K-means algorithm, and the clustering results are shown in Figure 8.

Table 9. Feature Set-3.

Value Features	Current
Mean	Cm
Variance	Cv
Peak-to-peak value	Cp

Table 10. Feature Set-4.

Value Features	Load
Mean	Lm
Variance	Lv
Peak-to-peak value	Lp

Table 11. Feature Set-5.

Value Features	Rotate Speed
Mean	Sm
Variance	Sv
Peak-to-peak value	Sp

Table 12. Feature Set-6.

Value Features	Oil Pressure
Mean	Pm
Variance	Pv
Peak-to-peak value	Pp

Different colors correspond to different types of faults. It shows that the data can only be classified into 4 classes using current or rotational speed signals, respectively, and the intraclass spacing is large, the interclass spacing is small; using a load or oil pressure signal can only divide the data into 3 classes and 2 classes, respectively, and the clustering is not effective. Therefore, it can be concluded that the clustering effect of the features extracted by using a single signal is not effective, and there is a considerable gap with the clustering effect of the effective features extracted in this paper.

Figure 9 shows the CHI of different feature sets with the number of clusters ranging from 4 to 10. We already set the number of clusters to 10. It can be seen that none of these 4 datasets determined the right cluster number, which means using a single signal to extract fault features has low quality for diagnosis.

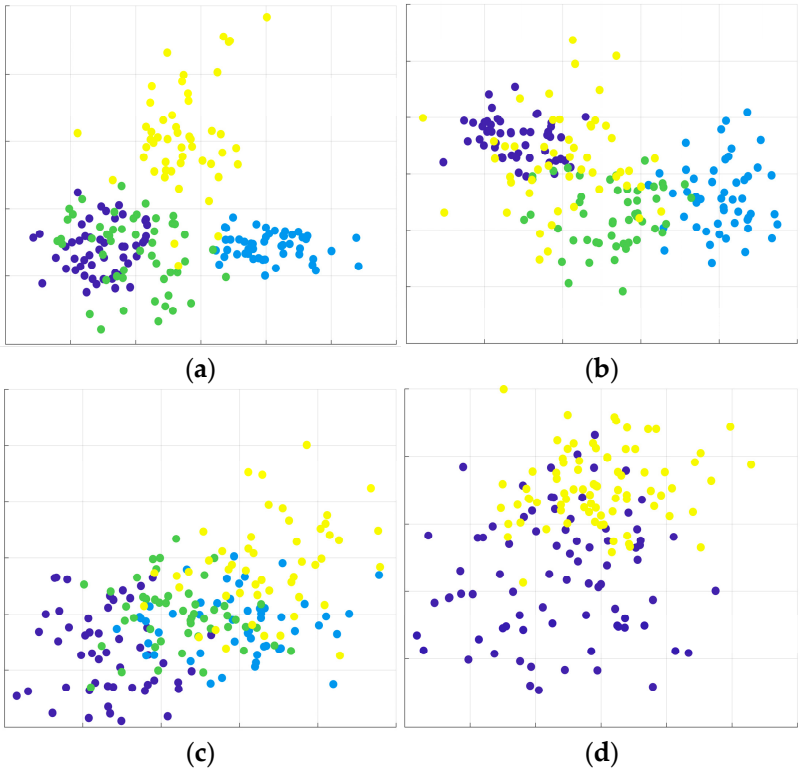


Figure 8. Results of comparison of Experiment II. (**a**) Clustering results of Feature Set-3; (**b**) clustering results of Feature Set-4; (**c**) clustering results of Feature Set-5; (**d**) clustering results of Feature Set-6.

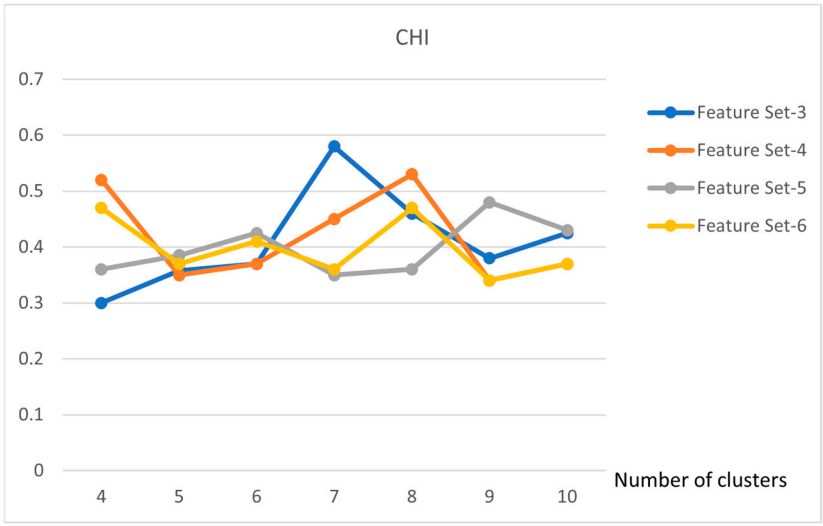
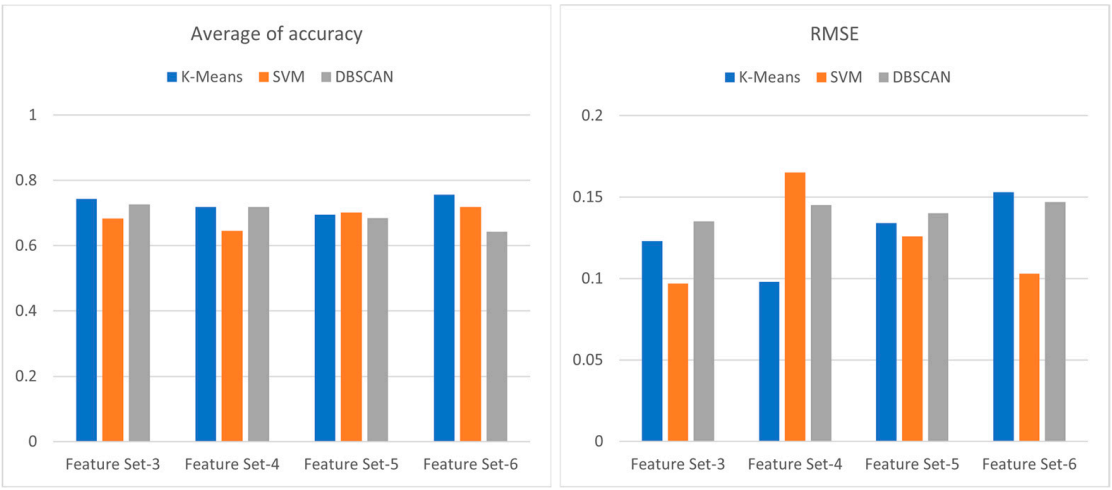


Figure 9. CHI of different feature sets with the number of clusters.

The averages of accuracy and RMSE for different datasets after clustering using different methods are shown in Figure 10 and Table 13. It can be seen that when we use a single signal to select an effective fault and extract fault features for diagnosis, we will get a low average of accuracy and a high RMSE, which means we will get a low quality of fault diagnosis.



(**a**) (**b**)

Figure 10. Diagnosis results for different datasets after clustering: (**a**) average of accuracy; (**b**) RMSE.

Table 13. The average of accuracy and RMSE for different datasets.

Datasets	K-Means		SVM		DBSCAN	
	Average of Accuracy (%)	RMSE	Average of Accuracy (%)	RMSE	Average of Accuracy (%)	RMSE
Feature Set-3	74.3	0.123	68.3	0.097	72.6	0.135
Feature Set-4	71.8	0.098	64.5	0.165	71.9	0.145
Feature Set-5	69.5	0.134	70.2	0.126	68.5	0.140
Feature Set-6	75.6	0.153	71.8	0.103	64.3	0.147

5. Conclusions

This paper proposed a model-driven approach to extract multi-source fault features of a screw pump. Firstly, the screw pump fault mechanism model and the FDM were constructed; then the original multi-dimensional data were cleaned, normalized, sliced, and enhanced, and the FDM was used for data selection to establish an effective data set; then the multi-dimensional fault feature extraction model (MDFEM) was constructed to extract the 4-dimensional featured signal and the 3-dimensional data feature, and the 12-dimensional effective fault features were extracted; finally, experiments were carried out to verify the effectiveness and accuracy of the proposed method.

We draw the following conclusions: (1) By analyzing the different fault mechanisms of the screw pump, the FDM can be constructed to select the effective data and solve the problem of unclear fault labels and locations; (2) By using MDFEM, we can determine that the 4-dimensional featured signals of the original multi-source signals, such as current, load, rotational speed, and oil pressure, are fault-related, and the 3-dimensional statistic feature of the mean, variance, and peak-to-peak are fault-related; (3) The method proposed in this paper can be used to extract 12-dimensional effective fault features and can achieve a multi-source informational fault diagnosis of a screw pump. After experimental verification, the proposed method can comprehensively use the multi-source information collected at the project site and can accurately and efficiently identify the types of screw pump faults, which has high application value. Due to the limited fault sample data collected in this paper, the next step is to obtain more comprehensive fault sample data as much as possible to improve the accuracy of the method.

Author Contributions: Conceptualization, W.W. and J.Q.; methodology, W.W. and J.Q.; software, J.Q.; validation, W.W. and J.Q.; formal analysis, W.W., J.Q., X.X. and M.Z.; investigation, W.W. and J.Q.; resources, W.W., X.X. and K.M.; data curation, J.Q.; writing original draft preparation, J.Q.;

writing review and editing, W.W.; visualization, J.Q. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data and materials supporting the results are included within the article.

Acknowledgments: We are very grateful to the reviewers and editors for their contributions to improving this manuscript.

Conflicts of Interest: Author Xiangru Xu and Kaifu Mi were employed by the company Beijing Petroleum Machinery 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. Li, M.; Xue, J. A review of diagnostic methods for oil recovery system conditions in screw pump wells. *Neijiang Sci. Technol.* **2013**, *4*, 47–48.
2. Zu, S.; Bai, J.; Wang, Y.; Li, Z. Analysis of the working conditions of surface-driven screw pumps. *Oil Drill. Process* **2006**, *1*, 28–31+86–87.
3. Li, B.; Song, W.; Xu, J.; Zhang, B. Ground-driven screw pump fault diagnosis based on improved wavelet packet combined with CS-BP. *Sci. Technol. Eng.* **2023**, *23*, 5641–5646.
4. Xie, J.; Cheng, H.; Chu, Y. Process control fault model and intelligent diagnosis method of electric submersible screw pump. *Pet. Mach.* **2023**, *51*, 116–121.
5. Wu, J.; Zhang, Q.; Liu, M. Research on fault diagnosis of screw pump by PNN network method. *Neijiang Sci. Technol.* **2020**, *41*, 39–40.
6. Han, Y. Comprehensive diagnosis method and application of pumping condition in screw pump wells. *Chem. Eng. Equip.* **2023**, *02*, 96–98.
7. Liang, H.Z.; Qi, M.; Li, W.; Li, W.H. Reseach and application of diagnostic intelligence technology for working conditions in screw pump wells. In Proceedings of the 2011 International Conference on Electrical and Control Engineering, Yichang, China, 16–18 September 2011; pp. 1161–1164.
8. Hao, D.; Gao, X. Multi-Weighted Partial Domain Adaptation for Sucker Rod Pump Fault Diagnosis Using Motor Power Data. *Mathematics* **2022**, *10*, 1519. [CrossRef]
9. Ren, W.J.; Lu, Y.; Xiao, K.X. Fault diagnosis of screw pump wells based on wavelet packet and Elman neural network. *J. Syst. Simul.* **2012**, *24*, 176–179.
10. Dong, K.; Li, Q.; Zhang, Z.; Jiang, M.; Xu, S. Submersible screw pump fault diagnosis method based on a probabilistic neural network. *J. Appl. Sci. Eng.* **2022**, *25*, 1067–1075.
11. Jiang, M.; Cheng, T.; Dong, K.; Xu, S.; Geng, Y. Fault diagnosis method of submersible screw pump based on random forest. *PLoS ONE* **2020**, *15*, e0242458. [CrossRef]
12. Xu, J.; Zhao, X.F.; Zhang, F.J.; Wang, H.J. Embedded progressive cavity pump condition monitoring system. *Control Eng.* **2010**, *17*, 557–560.
13. Li, M.; He, P.; Meng, C. Research on intelligent integrated fault diagnosis expert system for screw pump well. *Electr. Appl.* **2011**, *30*, 72–74.
14. Xue, J.Q.; Li, M.H.; Zhang, G.D. Research on screw pump well fault diagnosis technology based on BP neural network. *J. Xi'an Pet. Univ.* **2013**, *28*, 74–76.
15. Zou, W.; Qu, W.T.; Huang, W. Screw pump well fault diagnosis based on RBF neural network. *China Pet. Chem. Stand. Qual.* **2014**, *24*, 43.
16. Chen, S.W. Research on a New Method of Diagnosing the Working Condition of Oil Extraction Wells with Surface-Driven Progressive Cavity Pump. Master's Thesis, China University of Petroleum (Beijing), Beijing, China, 2016.
17. Zheng, C.F.; Wu, X.; Xu, H.Y. Diagnosis of pumping condition of electric submersible screw pump based on BP neural network. *Digit. Des.* **2018**, *7*, 56–58.
18. Mei, J.; Wen, T. Fault diagnosis system with natural repair function for screw oil pump based on radial basic function network. In Proceedings of the 8th International Conference on Electronic Measurement and Instruments, Melbourne, Australia, 6–8 May 2024.
19. Lv, X.X.; Wang, H.X.; Xin, Z.; Liu, Y.X.; Zhao, P.C. Adaptive fault diagnosis of sucker rod pump systems based on optimal perceptron and simulation data. *Pet. Sci.* **2022**, *19*, 743–760. [CrossRef]
20. Pang, T.Z. Investigation on the application of screw pump oil recovery technology in thick oil development. *Pet. Ind.* **2015**, *7*, 136.
21. Wei, J.; Gao, X. Fault Diagnosis of Sucker Rod Pump Based on Deep-Broad Learning Using Motor Data. *IEEE Access* **2020**, *8*, 222562–222571. [CrossRef]

22. Guo, L.B. Analysis of oil recovery technology and utilization of screw pump in thick oil development. *Petrochem. Technol.* **2018**, *4*, 16–18.
23. Lu, C.; Xian, W.; Xiang, Y. Using the motor power and XGBoost to diagnose working states of a sucker rod pump. *J. Pet. Sci. Eng.* **2021**, *199*, 108329.
24. Nojavanasghari, B.; Gopinath, D.; Koushik, J.; Baltrušaitis, T.; Morency, L.P. Deep multimodal fusion for persuasiveness prediction. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016.
25. Zheng, B.; Gao, X.; Pan, R. Sucker rod pump working state diagnosis using motor data and hidden conditional random fields. *IEEE Trans. Ind. Electron.* **2019**, *67*, 7919–7928. [CrossRef]
26. Han, Y.; Li, K.; Ge, F.; Wang, Y.; Xu, W. Online fault diagnosis for sucker rod pumping well by optimized density peak clustering. *ISA Trans.* **2021**, *120*, 222–234. [CrossRef]
27. Li, L.M.; Wen, Z.Z.; Song, Y.Q. Application research of optimized K-means clustering in redundant feature elimination. *Comput. Digit. Eng.* **2019**, *47*, 2836–2840.
28. Wang, X.F.; Qiu, J.; Liu, G.J. Research on feature selection of mechanical failure based on feature correlation and redundancy analysis. *Chin. J. Mech. Eng.* **2006**, *17*, 379–382.
29. Yi, L.Z.; Liu, Z.L.; Long, X. Neural network wind power prediction based on mutual information redundancy analysis. *Nat. Sci. J. Xiangtan Univ.* **2016**, *38*, 5.
30. Zhao, J.Y.; Zhang, Z.L. Feature selection method based on maximum mutual information and maximum correlation entropy. *Comput. Appl. Res.* **2009**, *26*, 4.