RESEARCH ARTICLE

GAN-SAE based fault diagnosis method for electrically driven feed pumps

Hui Han^{1,2}, Lina Hao^{1*}, Dequan Cheng², He Xu²

1 School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China, **2** School of Mechanical Engineering, Shenyang Ligong University, Shenyang, China

* haolina@me.neu.edu.cn

Abstract

The running of high-speed electrically driven feed pump has a direct impact on the safety of personnel equipment and economic benefits of power plant, as the result, intelligent condition monitoring and fault diagnosis of electrically driven feed pump becomes an urgent need. In the practical process of electrically driven feed pump fault diagnosis, the running of the equipment is in normal state for a long time, occasionally, with faults, which makes the fault data very rare in a large number of monitoring data, and makes it difficult to extract the internal fault features behind the original time series data. When the deep learning theory is used in practice, the imbalance between the fault data and the normal data occurs in the operation data set. In order to solve the problem of data imbalance, this paper proposes a fault diagnosis method of GAN-SAE. This method first makes compensation for the imbalance of sample data based on the Generative Adversarial Network (GAN), and then uses the Stacked Auto Encoder (SAE) method to extract the signal features. By designing the fault diagnosis program, compared with only using SAE, back propagation neural networks (BP) and multi-hidden layer neural networks (MNN) method, the GAN-SAE method can offer better capability of extracting features, and the accuracy of fault diagnosis of electrically driven feed pump could be improved to 98.89%.

OPEN ACCESS

Citation: Han H, Hao L, Cheng D, Xu H (2020) GAN-SAE based fault diagnosis method for electrically driven feed pumps. PLoS ONE 15(10): e0239070. <https://doi.org/10.1371/journal.pone.0239070>

Editor: Yanzheng Zhu, National Huaqiao University, CHINA

Received: May 4, 2020

Accepted: August 28, 2020

Published: October 22, 2020

Copyright: © 2020 Han et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was partly supported by the National Natural Science Foundation of China under Grant U1613205. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

High-speed electrically driven feed pump is an essential equipment of utility boiler, which plays a very important role in the normal running and safety of power utility. However, compared with large-scale equipment such as boiler, steam turbine and generator, electrically driven feed pump unit has not been given enough attention, which is the weak link of equipment condition monitoring as well as the major reasons for the unexpected shutdown of the generator set [1]. So it is necessary to monitor and diagnose running condition of electrically driven feed pump unit. At present, due to many monitoring points, high sampling rate, long time of monitoring data collection and other reasons, the electrically driven feed pump monitoring system obtains a large amount of health monitoring data, which on the one hand promotes the field of equipment health monitoring into the "big data era", on the other hand, it also puts forward higher requirements for fault information mining and diagnosis [2]. Taking the electrically driven feed pump of power utility as an example, although there is a large

amount of monitoring data, when the mechanical failures occurring occasionally, there is a small amount of fault state data, which presents the imbalance between fault data and normal data, which makes it very difficult to diagnose the fault based on data-driven methods.

Aiming at the problem of imbalanced data sample, most of the research is carried out from the data and algorithm level [3–5]. From the data level, the training set is usually reconstructed, mainly by oversampling the minority samples and undersampling the majority samples. The most typical is the synthetic minority oversampling technique method (SMOTE) [6], which increases the number of minority samples by randomly interpolating virtual samples to achieve a balanced sample data set. Mao et al. [7] introduced the master curve to extract the characteristics of the flow distribution, and used the SMOTE algorithm to oversample and undersample the minority and majority samples to balance the fault sample data. Finally, the online sequence extreme learning machine (OSELM) classifier was used for bearing fault diagnosis. Tao Xinmin et al. [8] used the SMOTE algorithm to generate minority samples, then used the original samples and newly generated samples to train SVM, and realized bearing fault diagnosis. E. Ramentol et al. [9] combined fuzzy rough set theory with SMOTE, and used two different thresholds to process the minority class and the majority class respectively to achieve a balanced data set, and finally used the C4.5 classifier to perform the fault of the high-voltage circuit breaker diagnosis. From the perspective of algorithm, it mainly improves the diagnostic performance by improving the classification algorithm, including cost-sensitive learning method and integrated learning method, etc. The cost-sensitive learning method mainly includes CS decision trees and neural networks, and integrated learning method includes Bagging algorithm and Adaboost algorithm. Duan et al. [10] studied a support vector data description method based on cross-tree. This method drew a binary tree structure of multiple clusters in the order of the distance from the cluster to the cluster calculated by the Mahalanobis distance. Then used the PSO algorithm to optimize the parameters of the support vector data description (SVDD), and finally used the binary tree network to complete the fault diagnosis of the rotor in the rotor system under the imbalanced sample size. Fan et al. [11] proposed an entropy-based fuzzy support vector machine (EFSVM), which guarantees the importance of the minority class by assigning relatively large fuzzy memberships to the minority class, thereby realizing various imbalanced sample set patterns recognition. Guo et al. [12] first used BPSO for feature selection, then used the Adaboost method to train multiple k-NN classifiers, and finally performed pattern recognition under various imbalanced sample sets. Currently, researchers apply the deep neural network learning method to the problem of sample imbalance in fault diagnosis. Lei et al. [13] studied a classifier loss strategy with adjustable weight coefficients combined with a deep normalized convolutional neural network (DNCNN). In this method, a small number of samples were given a large weight coefficient according to the degree of imbalance, so as to complete the fault diagnosis of bearings. Zhang et al. [14] designed a weighted minor-class oversampling self-coding network (WMODAE), in which the weighted minor-class oversampling technique was used to obtain the balanced sample data set, then the effective features were extracted with an automatic encoder, and finally the bearing fault diagnosis was realized with decision tree C4.5 algorithm.

Although the above-mentioned traditional imbalanced sample processing methods have improved the classification accuracy of imbalanced samples to a certain extent, they still have several shortcomings: (1) Virtual sample generation methods such as SMOTE are prone to generate noisy samples or invalid samples, especially in the case of "non-convex" fault sample space and large change conditions. (2) The classification performance of the existing classification network is reduced under the data set after sample expansion. (3) When the sample size imbalance ratio between the majority class and the minority class is too large, the effect of a large number of improved algorithms will become smaller and smaller.

Generative adversarial networks (GAN) is a new unsupervised learning model, which was first proposed by Goodfellow etc in 2014 [15]. GAN is widely used to solve the problem of data imbalance because of its powerful data generation capability. Ref. [16] used GAN generator to capture the real distribution of the original samples, and then generated new samples with similar distribution to the original samples, expanded the number of training fault samples, and then improved the accuracy of planetary gearbox fault diagnosis. Ref. [17] proposed a new method of selective GAN, solving the problem of high data imbalance in the fault diagnosis of reciprocating machinery. Ref. [18] proposed a deep convolutional GAN (DCGAN) model to simulate the original distribution from minority classes and generate new data to solve the data imbalance problem, improving the accuracy of fault diagnosis. In addition to the above studies, there are many studies around deep learning methods such as variational auto encoders (VAE) [19] and auto encoders (AE) [20], combined with the advantages of generative adversarial networks, aiming to improve the accuracy and authenticity of samples generated by GAN. For the research of pump fault diagnosis, there is no relevant literature on the application of the GAN method.

Therefore, in order to the problem of data sample imbalance in the fault diagnosis of electrically driven feed pump, this paper combines the powerful data generation capabilities of the GAN and deep learning technology of the stacked auto encoder (SAE) with feature learning and extraction capabilities [21–23], proposes a fault diagnosis method of GAN-SAE for the first time. This method firstly makes compensation for the imbalance of sample data based on the GAN, uses the SAE method to extract the signal features, integrates GAN with SAE networks to design the fault diagnosis program, and applies it in the fault diagnosis system of the electrically driven feed pump, so as to improve the diagnosis accuracy. This paper mainly consists of four parts. Firstly, it introduces the related theoretical basis. Secondly, it focuses on the GAN-SAE fault diagnosis algorithm. Thirdly, it gives the fault diagnosis experiment and analysis. Finally, it draws the conclusion and analysis.

1 Introduction to GAN and SAE

1.1 Generative adversarial networks (GAN)

Generative adversarial network is a new type of deep learning network model, which directly generate similar distributions from sample data. In Ref. [13], this network model is structurally inspired by the minimax two-player game theory, and in game theory, this model is composed of a generator (G) and a discriminator (D). The generator captures the potential distribution of real data samples and generates new data samples; the discriminator is a binary classifier, its objective is to determine whether the input sample is real or generated by the generator. Therefore in the training process of network model, the minimax two-player game composed by generator and discriminator. The goal of the training is to get the game to Nash Equilibrium [24]. The general structure of GAN is shown in Fig 1. $G(z)$ represents the generated sample

Fig 1. General GAN structure diagram.

<https://doi.org/10.1371/journal.pone.0239070.g001>

output by the generator, and $D(G(z))$ represents the discriminator's evaluation value of the generated sample. x is the real sample and $D(\bullet)$ is the discriminator output.

In the network model, the generator G receives a set of random noise z as input, and outputs a generated sample $G(z)$. The input of the discriminator D includes the real samples x from the real data distribution and the fake samples $G(z)$ generated by the generator G . The output is a probability value from 0 to 1. The value represents that the probability which discriminator D considers the input data to be real. At this time, the discriminator D is equivalent to a two-classifier **to determine** whether the input data comes from the real data distribution or the generated data distribution [25]. For example, when D considers the input data to be the real sample x , it outputs 1; otherwise, when D considers the input data to be the samples $G(z)$, it outputs 0.

During the training process, the discriminator D discriminates the fake samples generated by the generator G , and G updates its network parameters according to the discriminant results of D , and continuously improves the generation ability, so that the generated samples gradually approach the real data distribution, thereby they can cheat D . In order not to be deceived by G , D tries to find the difference between the real sample and the fake sample generated by G , and continuously improves its discriminate ability. Specifically, for the fake sample, the goal of the discriminator D is to make the output close to 0, and the goal of the generator G is to make it close to 1. The generator G and the discriminator D continuously optimize the respective networks through their iterative training process until they reach a balanced state.

The design of the generative adversarial network loss function is based on the Nash equilibrium of the two-player game. The calculation formula of the loss function of the generator used in this paper is shown in Eq (1).

$$G_{loss} = \log(D(G(z))) \quad (1)$$

The goal of the generator parameter update is to make $D(G(z))$ approach 1, that is, the discriminator will judge the generated sample as a true sample, and take the logarithm of the value as the loss function, so that the target can be converted to make the loss function approach 0.

The calculation formula of the discriminator's loss function is shown in Eq (2).

$$D_{loss} = -\log(D(x)) + \log(1 - D(G(z))) \quad (2)$$

The purpose of the discriminator parameter update is to make the evaluation value of the real sample approach 1, and the evaluation value of the generated sample tend to 0, which maximizes the discrimination ability between the generated sample and the real sample. By taking the logarithm of $D(x)$ and $1 - D(G(z))$, and summing them as the discriminator's loss function, the goal can be converted to make the loss function approach zero. This function embodies a major characteristic of GAN, that is, the minimax two-player game is formed by the loss function of the generator and discriminator, and the convergence of the network is promoted through the process of the game. At the same time, due to the existence of Nash equilibrium in game theory, network convergence is guaranteed and network reliability is improved.

1.2 Stacked Auto Encoder (SAE)

The Auto Encoder (AE) is a typical single hidden layer neural network proposed by Rumelhart in Ref. [26] 1986. Since this network keeps the input and output as consistent as possible. In this case, the hidden layer feature extraction and parameter learning are implemented in an

Fig 2. Schematic diagram of AE structure.

<https://doi.org/10.1371/journal.pone.0239070.g002>

unsupervised manner. Then it's called AE. In the structure of AE neurons possess the same amount of data in the input and output layers. The process from input to hidden layer is called encoding, and that from hidden layer to output is called decoding respectively. The schematic diagram of the Auto Encoder network structure is shown in Fig 2.

Each neuron of Auto Encoder network uses a fully connected method to weight the sum of the input vector $X = (x_1, x_2, \dots, x_n)$, then add the bias and go through the activation function to obtain the encoding vector $Y = (y_1, y_2, \dots, y_m)$ of the hidden layer. The encoding vector is remapped to the output vector $\hat{X} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n)$ through the decoding network. Finally, it is necessary to ensure that the input and output layers are as similar as possible, that is, the weights and offsets should be adjusted and optimized with the goal of minimizing the reconstruction error. The mathematical expression of the Auto Encoder network is as follows:

$$Y = \sigma_a(W_a \bullet X + b_a) \quad (3)$$

$$\hat{X} = \sigma_s(W_s \bullet Y + b_s) \quad (4)$$

$$\min L(\theta) = \|\hat{X} - X\|_2^2 \quad (5)$$

$$\theta = [W_a, b_a; W_s, b_s] \quad (6)$$

Fig 3. Schematic diagram of SAE structure.

<https://doi.org/10.1371/journal.pone.0239070.g003>

$W_n \in R^{n \times m}$, $W_c \in R^{m \times n}$, $b_n \in R^m$, $b_c \in R^n$ are the weight and bias that need to be optimized; $\sigma_a(\bullet)$, $\sigma_s(\bullet)$ are the activation function.

The encoding layer parts of the Auto Encoder are stacked, that is, the hidden layer of the previous Auto Encoder is used as the input layer of the next Auto Encoder network. Finally, the Softmax classifier is added to the last layer of the network constitutes the Stacked Auto Encoder (SAE) [27]. SAE trains layer by layer in the order from the bottom to the top. The SAE structure diagram is shown in Fig 3.

The encoding parts of multiple Auto Encoder networks (AE) are superimposed, the input vector is received at the first layer of the AE, and sent to its hidden layer after training. The raw data is transformed layer by layer to the top layer, so that the feature of the input vector is extracted multiple times to form a more representative feature vector, and then the feature vector is classified by a Softmax classifier. This Stacked Auto Encoder (SAE) network can well extract the features in the input data, and ensure that important information in the data is not lost.

2 Based on GAN-SAE fault diagnosis method

Due to the complex structure of the monitoring equipment and the harsh and variable operation environment, the massive data collected by the equipment operation monitoring system contains rich and variable fault information of mechanical equipment components [28]. Based on the data-driven fault diagnosis method, a large number of data are usually classified labeled and input to the already built deep learning network, and then the diagnosis results are got. However, in the actual production process, the equipment is in normal operation state for a long time, occasionally there are faults, resulting in a small amount of fault state data in a large

Fig 4. Flow chart of GAN-SAE method.

<https://doi.org/10.1371/journal.pone.0239070.g004>

number of monitoring data, then leading to data imbalance in the actual operation data set, which makes it very difficult for fault diagnosis based on data-driven methods. Since GAN is a new type of deep learning network model with strong data generation ability according to Fig 1, this paper intends to use GAN, sending random Gaussian noise variable Z into generator model G to generate a batch of new samples, which will be input into discriminator D together with real fault data samples for judgment. The generator network parameters will be updated according to the gradient of loss function. The generated samples are closer to the real fault data samples after iteration, and the loss function is formed according to the output of the discriminator. In this way, we can mine the features of data, generate data with the same feature distribution, make up for the data imbalance of sample, and improve the classification accuracy of the classifier. SAE can complete the task of dimensionality reduction and feature extraction, which can effectively avoid local optimum and improve the convergence speed of network. Based on this idea, this paper designs the fault diagnosis method integrating GAN with SAE, which is called GAN-SAE method, to diagnose the high-speed electrically driven feed pump. The flowchart of this method is shown as Fig 4.

The steps of the whole process can be summarized as follows: Firstly, obtain the detected signals during monitoring process, without feature extraction, take the detected signals directly

Fig 5. Schematic diagram of data collection system for electrically driven feed pump.

<https://doi.org/10.1371/journal.pone.0239070.g005>

as training samples. Secondly, build GAN network, input the training samples into the GAN network, and complete training for n times, until reaching the set goal, get the pseudo-samples generated by GAN combined with the real data to form a new data sample and input into the SAE network. Finally, complete the training and output the diagnosis result.

3 Experiment and analysis

3.1 Data acquisition system of electrically driven feed pump

The electrically driven feed pump unit is mainly composed of feed pump, hydraulic coupling and motor. As the main driving component, the motor drives the feed pump through the hydraulic coupling. The schematic diagram of the data acquisition system of the electrically driven feed pump shown in Fig 5. During the production process, many parameters are monitored by sensors. Such as, CD-21 vibration sensor is used to monitor the horizontal and vertical vibration of the bearings at the suction side and discharge side, FR1312 sensor to monitor the rotation speed of feed pump, PT100 temperature sensor to monitor the bearing temperature, YN100 pressure sensor to monitor the pressure of the feed pump, AVMN flowmeter to monitor water flow of feed pump and etc parameters. The diagram of local sensor distribution in Fig 6, sensors that lay at the site can be seen from the end face of feed pump, such as vibration sensors, temperature sensors and pressure sensors. According to JB/T-8097-1999, which

Fig 6. Diagram of local sensor layout.

<https://doi.org/10.1371/journal.pone.0239070.g006>

is the national machinery industry standard, the fault diagnosis experiment was carried out for feed pump vibration.

3.2 Prepare experimental data

The experimental data come from the monitoring data of the electrically driven feed pump in one power utility. Generally, the running speed of the electrically driven feed pump is in the range of 3000r/min to 5600r/min. The failure rate of feed pump is relatively small when it runs in low speed. the measured data are not enough to be used as the input of pump fault diagnosis, so the data are excluded in the experiment when its running speed is less than 3000r/min. In the experiment, the monitored data for a period of time after feed pump operation are selected to generate diagnosis result, which include the normal operation data of the feed pump before the fault occurs, and the fault data (such as the vibration fault), and the operation data after troubleshooting (reducing vibration by field dynamic balance treatment of electrically driven feed pump). Experimental data consist 32,000 items, including 13 variables, such as generator power, hydraulic coupling opening of feed pump, rotating speed of feed pump, outlet feed-water pressure, current, horizontal vibration value of suction side bearing, vertical vibration value of suction side bearing, horizontal vibration value of discharge side bearing, vertical vibration value of discharge side bearing, radial temperature of suction side bearing, radial bearing temperature of discharge side bearing, as well as outlet flow rate 1 of booster pump and outlet flow rate 2 of booster pump. The part of the captured field data is shown in Fig 7.

3.2.1 Data processing based on GAN. Due to the large numerical differences in collected data in the field with different types of variables, in order to reduce the calculation error caused by the differences of data, to ensure the original data structure is relatively unchanged, all variables are normalized [29]. At first, the data used in deep learning usually needs to be input in the range of 0 ~ 1. Otherwise, the weight, adder calculation and final classification results will be affected by different orders of magnitude of the input value. The range method is used for normalization.

$$y = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (7)$$

Date	Generator power (MW)	Hydraulic coupling opening value (%)	Rotation speed of feed pump (r/min)	Outlet water pressure (MPa)	Motor current value (A)	Suction-side bearing horizontal vibration 1 (μm)	Suction-side bearing vertical vibration 2 (μm)	Discharge side bearing horizontal vibration value 1 (μm)	Discharge side bearing vertical vibration value 2 (μm)	Suction-side radial bearing temperature (°C)	Discharge side radial bearing temperature (°C)	Booster pump outlet flow 1 (t/h)	Booster pump outlet flow 2 (t/h)
2019/7/10 21:45	265.59	36.58	4572.37	17.5	400.07	29.66	14.55	22.44	5.93	42.68	40.59	347.7	347.33
2019/7/10 21:46	263.68	37.95	4581.47	17.5	417.86	31.8	16.05	24.91	6.35	42.9	40.86	379.54	378.48
2019/7/10 21:47	261.78	38.42	4666.1	17.51	436.34	38.29	16.81	28.12	6.44	43.11	41.13	425.63	424.32
2019/7/10 21:48	259.87	38.79	4676.29	17.52	436.63	37.49	16.73	27.63	6.52	43.32	41.41	426.71	427.16
2019/7/10 21:49	257.96	39.09	4679.98	17.53	437.21	35.8	16.43	27.66	6.61	43.53	41.68	420.55	422.23
2019/7/10 21:50	256.05	39.25	4679.9	17.54	433.99	36.89	16.14	29.1	6.69	43.68	41.89	414.55	414.73
2019/7/10 21:51	258.87	39.41	4678.79	17.55	431.4	35.71	15.97	29.42	6.81	43.8	41.99	405.49	405.83
2019/7/10 21:52	262.74	39.57	4737.03	17.56	435.98	40.77	17.44	32.51	7.31	43.91	42.1	420.48	420.33
2019/7/10 21:53	266.62	39.73	4744.59	17.57	440.17	40.91	16.25	32.14	6.76	44.02	42.21	410.04	410.44
2019/7/10 21:54	263.28	39.89	4742.71	17.58	437.15	39.57	16.49	30.41	7.62	44.13	42.32	399.12	399.25
2019/7/10 21:55	259.15	40.98	4755.93	17.59	447.81	44.85	18.24	34.95	7.1	44.24	42.43	419.46	418.07
2019/7/10 21:56	255.01	41.08	4803.4	17.59	450.09	47.02	18.63	34.91	7.15	44.35	42.54	426.87	427.3
2019/7/10 21:57	250.87	39.86	4784.06	17.6	439.12	43.06	17.07	33.11	7.54	44.46	42.64	411.29	411.82
2019/7/10 21:58	251.58	39.91	4764.32	17.61	439.43	42.71	17.78	32.64	6.78	44.58	42.75	409.9	411.66
2019/7/10 21:59	254.9	39.96	4764.88	17.62	438.56	43.19	16.93	32.64	6.7	44.69	42.86	417.96	417.64
2019/7/10 22:00	258.21	40.01	4760.2	17.63	438.77	44.64	18.86	33.48	6.69	44.8	42.97	423.95	423.35
2019/7/10 22:01	260.94	40.06	4764.47	17.64	443.41	45.25	18.87	33.74	7.15	44.91	43.08	433.33	433.81
2019/7/10 22:02	261.76	40.1	4764.56	17.65	449.14	44.05	19.29	34.34	6.85	44.98	43.19	444.75	443.88
2019/7/10 22:03	262.57	41.15	4805.55	17.56	459.34	48.7	19.41	37	6.85	45.01	43.3	476.72	476.02

Fig 7. Collected field data screenshot.

<https://doi.org/10.1371/journal.pone.0239070.g007>

#1 Operation Parameters of the Electrically Driven Feed Pump

Date	Generator power (MW)	Hydraulic coupling opening value (%)	Rotation speed of feed pump (r/min)	Outlet water pressure (MPa)	Motor current value (A)	Suction-side bearing horizontal vibration 1 (um)	Suction-side bearing vertical vibration 2 (um)	Discharge side bearing horizontal vibration value 1 (um)	Discharge side bearing vertical vibration value 2 (um)	Suction-side radial bearing temperature (°C)	Discharge side radial bearing temperature (°C)	Booster pump outlet flow 1 (t/h)	Booster pump outlet flow 2 (t/h)
2019/7/10 21:45	0.6276383	0.3059871	0.50772857	0.677028	0.346988	0.260093168	0.138648937	0.28202987	0.3025887	0.669143666	0.668674699	0.185141961	0.189629
2019/7/10 21:46	0.6140188	0.3319439	0.512220389	0.677028	0.393394	0.282545745	0.153785284	0.31593783	0.3267498	0.676928521	0.679518072	0.215352181	0.219076
2019/7/10 21:47	0.6004706	0.3408488	0.553994306	0.67821	0.4416	0.350637905	0.161454366	0.36000445	0.3319272	0.684359519	0.690361446	0.259082991	0.262409
2019/7/10 21:48	0.5868511	0.347859	0.559024156	0.679391	0.442356	0.342244418	0.160647095	0.35327777	0.3365293	0.691790517	0.701606426	0.260107709	0.265093
2019/7/10 21:49	0.5732316	0.353543	0.560845564	0.680573	0.443869	0.324513178	0.157619825	0.3536896	0.3417067	0.699221515	0.712449799	0.254263019	0.260433
2019/7/10 21:50	0.5596121	0.3565745	0.560806075	0.681754	0.43547	0.335949303	0.154693465	0.37345781	0.3463088	0.70452937	0.720883534	0.248570138	0.253343
2019/7/10 21:51	0.5797205	0.3596059	0.560258172	0.682936	0.428713	0.323568911	0.152978012	0.37785074	0.3532119	0.708775655	0.724899598	0.239973889	0.24493
2019/7/10 21:52	0.607316	0.3626374	0.589005814	0.684118	0.440661	0.376657714	0.167811632	0.42027001	0.3819751	0.712668082	0.729317269	0.254196602	0.258637
2019/7/10 21:53	0.6349829	0.3656688	0.592737479	0.685299	0.45159	0.378126574	0.155803464	0.41519068	0.3503356	0.71656051	0.73373494	0.24429099	0.249288
2019/7/10 21:54	0.6111666	0.3687003	0.591809499	0.686481	0.443713	0.364067484	0.158225279	0.39144138	0.3998083	0.720452937	0.73815261	0.233929947	0.23871
2019/7/10 21:55	0.5817171	0.389352	0.598334976	0.687662	0.47152	0.419464496	0.17588435	0.45376613	0.3698946	0.724345364	0.742570281	0.253228812	0.2565
2019/7/10 21:56	0.5521962	0.3912467	0.621766476	0.687662	0.477467	0.442231828	0.179819801	0.45321701	0.3727709	0.728237792	0.746987952	0.260259519	0.265226
2019/7/10 21:57	0.5226754	0.3681319	0.612220127	0.688844	0.448851	0.400684069	0.164078	0.42850676	0.3952062	0.732130219	0.751004016	0.245477006	0.250592
2019/7/10 21:58	0.5277382	0.3690792	0.602476335	0.690025	0.44966	0.397011919	0.171242537	0.42205464	0.3514861	0.736376504	0.755421687	0.244158156	0.250441
2019/7/10 21:59	0.5514119	0.3700265	0.602752755	0.691207	0.447391	0.402048011	0.162665274	0.42205464	0.346884	0.740268931	0.759839357	0.251805592	0.256094
2019/7/10 22:00	0.5750143	0.3709739	0.600442676	0.692388	0.447938	0.417261205	0.182140707	0.43358609	0.3463088	0.744161359	0.764257028	0.257488984	0.261492
2019/7/10 22:01	0.5944809	0.3719212	0.602550376	0.69357	0.460042	0.423661239	0.182241616	0.43715535	0.3727709	0.748053786	0.768674699	0.266388854	0.271138
2019/7/10 22:02	0.600328	0.372679	0.6025948	0.694751	0.474989	0.411071009	0.186479793	0.4453921	0.355513	0.750530786	0.773092369	0.277224303	0.280899
2019/7/10 22:03	0.6061038	0.3925729	0.62282773	0.684118	0.501596	0.45985815	0.187690701	0.48190836	0.355513	0.751592357	0.77751004	0.307557868	0.311281

Fig 8. Normalized data screenshot.

<https://doi.org/10.1371/journal.pone.0239070.g008>

Where, x_i is the sample value, y is the value after normalization, $x_{\max}$ and $x_{\min}$ are the maximum and minimum value in the sample set.

For example, the data in the first column (power of the generator) in Fig 7, normalization processing is as follows. The maximum value is 317.81029MW and the minimum value is 177.57MW from the first column, the first value 265.59MW is normalized by the above formula (8).

$$\frac{265.59 - 177.57}{317.81029 - 177.57} = 0.627638334 \quad (8)$$

Fig 8 shows a partial data screenshot after the data shown in Fig 7 is normalized by the method.

According to JB/T-8097-1999, which is the national machinery industry standard in China, during the experiments, it is a fault state if the amplitude of bearing horizontal vibration on the suction side is greater than 35um, the amplitude of bearing vertical vibration on the suction side is greater than 15um, the amplitude of bearing horizontal vibration on the discharge side is greater than 35um, or the amplitude of bearing vertical vibration on the discharge side is greater than 10um. Table 1 is the original data classification table, which gives the symbol representation in the experiment and the number of various data under various faults and normal conditions in the experiment. The symbols A represents the suction side of feed pump, and symbols B represents the discharge side of feed pump. The digital angle mark 1 represents the horizontal direction and 2 represents the vertical direction. For example, A1 indicates horizontal fault on suction side, and A1B2 indicates that horizontal fault on suction side and vertical fault on discharge side occur simultaneously. According to the statistics, the number of

Table 1. Original data classification table.

The fault types	Suction-side fault (A)		Discharge side fault (B)		On both sides fault				Normal
	Horizontal fault	Vertical fault	Horizontal fault	Vertical fault	A1B1	A1B2	A2B1	A2B2	
Classification	A1	A2	B1	B2	A1B1	A1B2	A2B1	A2B2	ZC
Classification principle	>35	>15	>35	>10					
Number of the fault data	1716	203	164	454	154	133	0	836	28340

<https://doi.org/10.1371/journal.pone.0239070.t001>

Fig 9. GAN structure diagram for data processing.

<https://doi.org/10.1371/journal.pone.0239070.g009>

fault data of different categories is very different, which leads to the problem of data imbalance.

According to the GAN principle shown in Fig 1, the GAN structure diagram for data processing is shown in Fig 9. The network includes two parts, which are generator part and discriminator one. The generator part includes three encoder layer. The first encoder layer has 8 nodes, the second encoder layer 10 nodes, and the third encoder layer 13 nodes. Additionally, there are three discriminators in the discriminator part. The number of nodes in each layer is 13, 10 and 1 respectively. Table 2 shows the parameter settings of each encoder and each discriminator of GAN network.

After initializing the network structure and parameters, random noise is input into the generator network composed of 3-layer encoder as shown in Fig 9, and a batch of samples are generated. Next steps are to combine the generated samples with the original fault data samples and input them into the discrimination network composed of 3-layer discriminators. After training, the new generated data have a similar distribution to the original fault data, and the amount of data is compensated, so that the amount of each fault data is relatively balanced,

Table 2. Parameters of the GAN structure.

Name	Numbers of node	Activation function
Encoder-1	8	Sigmoid
Encoder-2	10	Sigmoid
Encoder-3	13	Sigmoid
Discriminator-1	13	Sigmoid
Discriminator-2	10	Sigmoid
Discriminator-3	1	Sigmoid

<https://doi.org/10.1371/journal.pone.0239070.t002>

Table 3. Classification tables after data generation table.

The fault types	Suction-side fault (A)		Discharge side fault (B)		On both sides fault				Normal
	Horizontal fault	Vertical fault	Horizontal fault	Vertical fault	A1B1	A1B2	A2B1	A2B2	
Classification	A1	A2	B1	B2	A1B1	A1B2	A2B1	A2B2	ZC
Classification principle	>35	>15	>35	>10					
Number of the fault data	572	400	400	454	400	400	0	418	409

<https://doi.org/10.1371/journal.pone.0239070.t003>

and the number of various state data shown in Table 3 is obtained. The output of the discriminator forms the generator loss function and the discriminator loss function, as shown in Fig 10. The generator network parameters and the discriminator network parameters are updated according to the gradient of the loss function, so that the generated samples output by the generator are closer to the real fault data samples. Table 4 is the loss function change table in the iteration process. After 5000 iterations, the generator loss function becomes 0.00106, as shown in Fig 10. The loss function tends to be close to 0, indicating that the data generated by GAN has similar distribution to the actual fault data, and the generated sample data can be used as the input of fault diagnosis.

Lastly, remove the A2B1 type without data, combine the generated fault data with the real data, label the 7 types of fault data and the normal data respectively, and then it can be used as the sample data of the fault diagnosis system of the electrically driven feed pump.

3.3 Fault diagnosis experiments

3.3.1 Fault diagnosis based on SAE. The original fault data set collected in the field are used as the input of SAE diagnosis method. Two-thirds of the original fault data set will be used as the training set and the remaining one third as the test set. Using SAE method to construct a multi-layer deep learning network. The structure of the fault diagnosis program is shown in Fig 11.

Program flowchart is divided into two parts. First part is the model training process. The fault samples are preprocessed, a data set consisting of the running data and the labels corresponding to the fault types is formed. Two thirds of the data set were used as training sets and the remaining one third as test sets. The unsupervised training process of AE network only

Fig 10. Change of GAN loss function.

<https://doi.org/10.1371/journal.pone.0239070.g010>

Table 4. Changes in the loss function during iteration.

Numbers of iteration	D Loss function	G Loss function
1000	1.59378	-0.87585
2000	0.34549	-0.09859
3000	0.17131	-0.05585
4000	0.11510	-0.04884
5000	0.01058	-0.00106

<https://doi.org/10.1371/journal.pone.0239070.t004>

needs to run the training set to obtain an encoder, AE network can well complete the feature extraction task and save the encoder. The SAE network belongs to supervised training, whose data set consisting of running data and tags corresponding to its fault type. After reading the encoders saved in AE network, adding the category layer of Softmax as a classifier, the SAE network model can be saved by supervised training. Then the SAE network model is tested with the test set until the test model is successful. The second part is the process of fault diagnosis using SAE network model. In the process of fault diagnosis, after preprocessing the data collected in the field, the SAE model is called as the running data for fault diagnosis, and the label vector corresponding to the fault type is output to achieve the purpose of fault diagnosis. In order to reduce the influence of random factors on the diagnosis, the experiment was repeated 15 times. The SAE method has an average diagnostic accuracy of 94.9% and a maximum diagnostic accuracy of 96.5% for vibration-induced faults.

3.3.2 Fault diagnosis based on GAN-SAE. According to Fig 4, the network structure diagram of GAN-SAE method is built as Fig 12. In fault diagnosis of the GAN-SAE method, fault data is collected on site and fault sample data set is generated by GAN network, they are taken together as the input of the method, and two-thirds of the data set is also taken as the training set. It is similar to SAE method for training. The number of nodes in the network is shown in Table 5. There are 6 layers in the network, including one input layer, four encoder layers and one output layer. The number of neuron nodes in each layer is 13, 200, 100, 50, 25 and 8 respectively. GAN-SAE network model is obtained after training. In the process of fault diagnosis, after preprocessing the data collected in the field, the experiment is also repeated for 15 times as the running data. The saved GAN-SAE model is used for fault diagnosis, and the label vector corresponding to the fault type is output. The diagnosis is shown in Fig 12. The GAN-SAE method has an average diagnostic accuracy of 98.16% and a maximum diagnostic accuracy of 98.89% for vibration-induced faults.

Fig 11. Structure of SAE diagnostic program.

<https://doi.org/10.1371/journal.pone.0239070.g011>

Fig 12. GAN-SAE network structure.

<https://doi.org/10.1371/journal.pone.0239070.g012>

3.3.3 Comparison of diagnosis results. In this experiment, compared with diagnostic accuracy for SAE method, GAN-SAE method, and also include other two common fault diagnosis methods: back propagation neural networks (BP) and multi-hidden layer neural networks(MNN). For BP, the number of neurons in each layer is 13, 20, 20 and 8. And for MNN, the number of neurons in each layer is configured as 13, 200, 100, 50, 25 and 8 respectively. After calculating the average accuracy (15 times) for the four methods, we got the mean and maximum accuracy for each diagnosis, and the standard deviation can be calculated. All the results are shown in Table 6. Standard deviation reflects the dispersion degree of diagnostic results. Smaller standard deviation means smaller fluctuations, which proves more stable results can be obtained by that method. Fig 13 shows the accuracy curves of the methods: GAN-SAE, SAE, BP and MNN respectively.

After using the GAN to generate data and using SAE method for training, the GAN-SAE method diagnostic accuracy is at a high level with an average accuracy of 98.16% with a maximum accuracy of 98.89%. Compared with the GAN-SAE method, SAE method with the original imbalance data input gets an average accuracy of 94.9% with a maximum diagnostic accuracy of 96.5% after 15 times training, and its standard deviation is higher than the GAN-SAE method. The average accuracy of MNN method was 92.8%, the maximum accuracy was 94.5% and the average accuracy of BP method was 93.3%, the maximum accuracy was 95.5%. From Fig 13, the fluctuating range of GAN-SAE is smallest among the four methods, proved the smallest deviation value in Table 6. Although average diagnostic accuracy of SAE is higher than MNN and BP, but sometimes it may get worse result than these two methods.

Table 5. GAN-SAE network settings at each layer.

Name	Numbers of node	Activation function
Input layer	13	Leaky_relu
Encoder-layer1	200	Leaky_relu
Encoder-layer2	100	Leaky_relu
Encoder-layer3	50	Leaky_relu
Encoder-layer4	25	Leaky_relu
Softmax layer	8	Softmax

<https://doi.org/10.1371/journal.pone.0239070.t005>

Table 6. Comparison diagnostic accuracy of different methods.

Diagnostic method	Mean diagnostic accuracy	Maximum diagnostic accuracy	Standard deviation
GAN-SAE	98.16%	98.89%	0.3%
SAE	94.9%	96.5%	1.12%
MNN	92.8%	94.5%	1.06%
BP	93.3%	95.5%	1.08%

<https://doi.org/10.1371/journal.pone.0239070.t006>

The higher value for GAN+SAE and SAE are obtained after several repeat, also shows training plays a important role in diagnostic for these methods. By comparison, GAN-SAE method can not only achieve a high diagnostic accuracy, but also get a stable diagnosis result.

4. Conclusion

In order to solve the problem of low diagnosis accuracy caused by the imbalance of the input data for electrically driven feed pump, and overcome the deficiency of traditional imbalanced data processing methods, this paper proposes a fault diagnosis method of GAN-SAE by combining the advantages of GAN and SAE methods, and it is applied for fault diagnosis of electrically driven feed pump.

1. In the proposed method, the powerful data generation ability and data feature mining ability of the GAN method were firstly used to generate data with the same feature distribution to compensate for the imbalanced experimental sample data of electrically driven feed pump, thereby improving the classification accuracy of the classifier.
2. After being processed by GAN, the generated fault data was fused with the real data and used as the fault diagnosis sample data of the electrically driven feed pump. Then, the SAE method, which can complete the data dimensionality reduction and feature extraction tasks, was used for fault identification to avoid the local optima and improve the network convergence speed.

Fig 13. Diagnostic accuracy of the 15 trials.

<https://doi.org/10.1371/journal.pone.0239070.g013>

3. This method was applied for the fault diagnosis of electrically driven feed pump, and it demonstrates superior recognition results through the comparison with other fault identification methods.

In this paper, the GAN method was used to balance experimental data samples, the research on the single fault diagnosis has made some progress, but the compound fault diagnosis method for massive data has some problems such as unstable training, disappearing gradient, and mode collapse. How to solve the above problems more effectively and accurately, further research is needed.

Supporting information

S1 Data.

(XLS)

Author Contributions

Investigation: Dequan Cheng, He Xu.

Software: Dequan Cheng, He Xu.

Writing – original draft: Hui Han.

Writing – review & editing: Hui Han, Lina Hao.

References

1. Xu HW. Research and Application on Condition Assessment and Maintenance Optimization of the Turbine Driven Boiler Feed Water Pump Set. Xi'an Thermal Power Research Institute Co., Ltd. 2018.
2. Lei YG, Jia F, Zhou X, et al. Big data Health Monitoring method of Mechanical equipment based on depth Theory. *Journal of Mechanical Engineering*. 2015; 52 (21), 49–56. <https://doi.org/10.3901/JME.2015.21.049>
3. H He Garcia EA. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 2009, 21(9): 1263–1284. <https://doi.org/10.1109/tkde.2008.239>
4. Quasi-Synchronization of Discrete-Time Lur'e-Type Switched Systems with Parameter Mismatches and Relaxed PDT Constraints, *IEEE Transactions on Cybernetics*, 2020, 50(5): 2026–2037. doi:10.1109/tcyb.2019.2930945.
5. Multiple Lyapunov Functions Analysis Approach for Discrete-Time Switched Piecewise-Affine Systems Under Dwell-Time Constraints, *IEEE Transactions on Automatic Control*, 2020, 65(5): 2177–2184. <https://doi.org/10.1109/TAC.2019.2938302>
6. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 2002, 16(1): 321–357. <https://doi.org/10.1613/jair.953>.
7. Mao WT, He L, Yan YJ, et al. Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine. *Mechanical Systems and Signal Processing*, 2017, 83: 450–473. <https://doi.org/10.1016/j.ymssp.2016.06.024>
8. Tao XM, Liu FR, Tong ZJ, et al. Novel fault detection method based on SVM with unbalanced datasets. *Journal of Vibration and Shock*. 2010, 29(12): 8–12(in Chinese).
9. Ramentol E, Gondres I, Lajes S, et al. Fuzzy-rough imbalanced learning for the diagnosis of High Voltage Circuit Breaker maintenance: The SMOTE-FRST-2T algorithm. *Engineering Applications of Artificial Intelligence*, 2016, 48(C): 134–139. <https://doi.org/10.1016/j.engappai.2015.10.009>
10. Duan LX, Xie MY, Bai TB, et al. A new support vector data description method for machinery fault diagnosis with unbalanced datasets. *Expert Systems with Applications*, 2016 64: 239–246. <https://doi.org/10.1016/j.eswa.2016.07.039>
11. Fan Q, Wang Z, Li DD, et al. Entropy-based fuzzy support vector machine for imbalanced datasets. *Knowledge-Based Systems*, 2017, 115: 87–99. <https://doi.org/10.1016/j.knsys.2016.09.032>
12. Guo HX, Li YJ, et al. BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. *Engineering Applications of Artificial Intelligence*, 2016, 49(C): 176–193. <https://doi.org/10.1016/j.engappai.2015.09.011.s>

13. Jia F, Lei YG, Lu N, et al. Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. *Mechanical Systems and Signal Processing*, 2018, 110: 349–367. <https://doi.org/10.1016/j.ymssp.2018.03.025>
14. Zhang Y, Li X, Gao L, et al. Imbalanced data fault diagnosis of rotating machinery using synthetic over-sampling and feature learning. *Journal of Manufacturing Systems*, 2018, 48: 34–50. <https://doi.org/10.1016/j.jmsy.2018.04.005>
15. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*. 2014; Montreal, QC, Canada, 8–13 December 2014: 2672–2680.
16. Wang Z, Wang J, Wang Y. An intelligent diagnosis scheme based on generative adversarial learning deep neural networks and its application to planetary gearbox fault pattern recognition. *Neurocomputing*. 2018; 310 (2018):213–222. <https://doi.org/10.1016/j.neucom.2018.05.024>
17. Cabrera Diego, Sancho Fernando, Long Jianyu, Sánchez René-Vinicio, Zhang Shaohui, Cerrada Mariela, et al. Generative Adversarial Networks Selection Approach for Extremely Imbalanced Fault Diagnosis of Reciprocating Machinery. *IEEE*. 2019; volume 7:70643–70653. <https://doi.org/10.1109/access.2019.2917604>
18. Xie Yuan, Zhang Tao. Imbalanced Learning for Fault Diagnosis Problem of Rotating Machinery Based on Generative Adversarial Networks. *IEEE*. 2018; 6017–6022. <https://doi.org/10.23919/chicc.2018.8483334>
19. Genevay A, Peyre G, Cuturi M. GAN and VAE from an optimal transport point of view [EB/OL]. <https://arxiv.org/abs/1706.01807>, 2017.
20. Li CL, Chang WC, Cheng Y, et al. Mmd gan: Towards deeper understanding of moment matching network. *Advances in Neural Information Processing Systems*. 2017, 2203–2213.
21. Jia F, Lei Y, Guo L, Lin J, Xing S. A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. *Neurocomputing*. 2018; vol. 272, pp. 619–628. <https://doi.org/10.1016/j.foodchem.2018.06.050> PMID: 30309590
22. Guo L, Gao H, Huang H, He X, Li S. Multifeatures fusion and nonlinear dimension reduction for intelligent bearing condition monitoring. *Shock and Vibration*. 2016; 1–10. <https://doi.org/10.1155/2016/4632562>
23. Liu H, Li L, Ma J. Rolling bearing fault diagnosis based on STFT-deep learning and sound signals. *Shock and Vibration*. 2016; 1–12. <https://doi.org/10.1155/2016/6127479>
24. Sican Cao, Long Wen, Xinyu Li, Liang Gao. Application of Generative Adversarial Networks for Intelligent Fault Diagnosis 2018 IEEE 14th International Conference on Automation Science and Engineering. 2018; 711–715. doi: 10.1109/COASE.2018.8560528.
25. M Mirza, S Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv*. 2014; 1411–1784.
26. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1988; 323(6088): 399–421. doi.org/10.1038/323533a0.
27. Vincent P, Larochelle H, Lajoie I, Bengio Y, et al. Stacked Denoising Autoencoders Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*. 2010; 11(12): 3371–3408.
28. Jardine AKS, Lin D, Banjevic D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*. 2006; 20(7): 1483–1510. <https://doi.org/10.1016/j.ymssp.2005.09.012>
29. Ambika PS, Rajendrakumar PK, Ramchand Rijil. Vibration signal based condition monitoring of mechanical equipment with scattering transform. *Springer-Verlag*. 2019; 33(7): 1–9. <https://doi.org/10.1007/s12206-019-0604-7>