A Fault Diagnosis Method for Drilling Pump Fluid Ends Based on Time–Frequency Transforms

Aimin Tang and Wu Zhao *

School of Mechanical Engineering, Sichuan University, Chengdu 610065, China; tangaiminhh@yeah.net
* Correspondence: zhaowu@scu.edu.cn

Abstract: Drilling pumps are crucial for oil and gas operations. Timely diagnosis and troubleshooting of fluid end faults is crucial to ensure the safe and stable operation of drilling pumps and prevent further deterioration of faults. Hence, from a data-driven perspective, this study proposes a fault diagnosis method for the fluid end of drilling pumps based on the generalized S transform (GST) and convolutional neural networks (CNN), using the vibration signal of the fluid end. To address the issue of noise pollution in the vibration signal resulting in unclear feature information and difficult feature extraction, the vibration signal is transformed into a time–frequency diagram based on GST, which more accurately characterizes the fault characteristics of the vibration signal. An AlexNet model, improved by introducing batch normalization and optimizing the number of neurons in the fully connected layer, is used to analyze the recognition performance of the model for the normal, minor damage, and severe damage states of the fluid end of the drilling pump. Finally, the diagnosis results are compared to other methods, with the results showing that the proposed method has the highest fault diagnosis accuracy. With an average recognition rate of 99.21% for the nine types of fluid end, the method proposed in this study provides a way to accurately diagnose fluid end failures, thus supporting the safe and efficient operation of drilling pumps.

Keywords: drilling pump; fluid end; fault diagnosis; vibration signal; generalized S transform; AlexNet

1. Introduction

The drilling pump as shown in Figure 1 serves as the essential circulation equipment in oil and gas drilling operations. Its primary function is to transport high-pressure drilling fluid to the bottom of the well for circulation, washing the well bottom, breaking rocks, and cooling and lubricating the drill bit while also returning rock cuttings to the surface. The drilling pump consists of a power end and a fluid end, with the fluid end being a critical component responsible for controlling the flow direction and rate of fluids within the pump. The suction valve (SV) and discharge valve (DV) of the fluid end enable the pump to intake and discharge drilling fluid simultaneously. The fluid end of the drilling pump is prone to wear and tear, particularly the SV and DV, under the duress of high-pressure loads and heavy grinding from drilling fluid. However, there have been difficulties in fault diagnosis of this component, for two primary reasons. First, original signals tend to be highly contaminated by noise, making it challenging to extract fault features accurately. Second, traditional fault diagnosis methods exhibit limited accuracy and cater to only a small number of fault types, rendering them unsuitable for engineering applications. As such, undertaking extensive research to enhance fluid end fault diagnosis is of utmost importance in forging ahead with efficient and safe drilling operations.

Machine learning (ML) represents a research hotspot in the field of artificial intelligence and has yielded promising results in terms of pump fault diagnosis applications [1–4]. However, most of these studies focus on the fault diagnosis method of one-dimensional signals. It is often necessary to analyze the signal in the time domain, frequency domain, power spectrum, and other parameter indicators to extract the fault characteristics of
the signal. There are not many diagnostic categories, and the effect is not ideal. Some researchers have also converted one-dimensional signals into images to achieve a variety of fault diagnoses of the fluid end, but the method is not ideal for the identification of severe damage and minor damage to the valve [5]. The standard approach for traditional ML may prove ineffective in processing raw data directly, and researchers frequently face the challenge of having to extract features from vast amounts of data, which can be a time-consuming and arduous process. Deep learning (DL) stands out in contrast to ML as it has the ability to extract richer features and solve more intricate problems.

Convolutional neural networks, as one of the typical DL algorithms, are widely used in various fields. Li et al. [6] analyzed the seismic signal based on a short-time Fourier transform (STFT), converted it into a time–frequency image, and then converted the time–frequency image into a grayscale image as the input of CNN. The average accuracy of recognition was 97.39%, which was higher than the traditional ML method. Cai et al. [7] converted a one-dimensional voltage disturbance signal into a two-dimensional image based on Wigner–Ville distribution (WVD) and introduced CNN to realize the classification of the power quality display signal, which can achieve high-precision feature extraction and classification. Miao et al. [8] implemented the continuous wavelet transform (CWT) to process signals and obtained two-dimensional time–frequency images. These images were then fed into a CNN model for training and prediction, thus enabling them to recognize and identify weld defects with greater accuracy than via traditional methods—the process was approximately 10% more accurate. Luo et al. [9] proposed a feature-coding matrix based on the generalized S transform (GST) and two-dimensional non-negative matrix factorization and input it into a deep belief network (DBN), which solved the problem of difficult feature extraction of one-way valve vibration signals under strong noise but only realized the identification of three states of one-way valve.

The time–frequency analysis methods used by the above researchers include STFT, WVD, the continuous wavelet transform (CWT), and GST [10,11]. The S transform (ST) combines the advantages of STFT and WT. It has a strong time–frequency analysis ability and is not easily affected by noise, but it has the disadvantage of a fixed window function. To solve this problem, researchers have improved ST and proposed GST, but currently GST is mostly used for seismic signal processing [12].

As the vibration signal is a non-stationary signal with strong randomness, traditional machine learning has limited fault identification types and accuracy. To overcome this issue, fault diagnosis based on CNN provides an effective solution. This study converts the vibration signal into a time–frequency image using GST, which is then input into an optimized CNN to automatically extract feature information and mine the hidden
correlations between input data. Through this approach, the problem of traditional single-layer neural networks having difficulty in feature extraction and pattern recognition is better resolved and a more rapid and accurate diagnosis of fluid end faults of drilling pumps is provided.

2. Preparations

2.1. Time–Frequency Transform Method

2.1.1. Generalized S Transform

ST is a reversible time–frequency analysis method proposed by Stockwell et al. [13]. It has good time–frequency characteristics and is a powerful tool for analyzing non-stationary signals. The ST of signal \( h(t) \) is shown in Equation (1):

\[
S(\tau, f) = \int_{-\infty}^{+\infty} h(t)w(\tau - t, f) \exp(j2\pi ft)dt
\]

where \( \tau \) is the time-shift parameter, \( f \) is continuous frequency, \( t \) is time, and \( j \) is complex unit.

When processing the original signal, ST introduces a Gaussian window function [14], as shown in Equation (2), so that its time–frequency resolution changes with frequency. However, due to the constant speed of window size changing with frequency, the actual use is limited [15]. In order to solve this problem, the researchers improved ST and obtained GST [16], which is defined as shown in Equation (3):

\[
w(\tau - t, f) = \frac{|f|}{\sqrt{2\pi}} \exp\left(-\frac{(\tau - t)^2 f^2}{2}\right)
\]

\[
GST_x(\tau, f, \lambda) = \int_{-\infty}^{+\infty} x(t) \frac{|f|^\lambda}{\sqrt{2\pi}} \exp\left(-\frac{(\tau - t)^2 f^2}{2}\right) \exp(-i2\pi ft)dt
\]

where \( w(\tau - t, f) \) is Gaussian window function and \( i \) is an imaginary number; usually, \( \lambda \in [0.6, 1] \).

2.1.2. Short-Time Fourier Transform

The Fourier transform is a commonly used method of signal time–frequency transform. After years of development, the Fourier transform has developed into STFT [17], also known as windowed Fourier transform, which is an effective time–frequency analysis method. The fundamental concept involves dividing the time domain signal into multiple fixed time intervals on the time axis, assuming that the signal is stationary within each window. The Fourier transform is then utilized on the signal within each window to capture the time-varying spectrum. Continuously sliding the time window enables repeated application of the Fourier transform until the end of the signal, thereby enabling the capturing of the time-varying spectrum of the entire signal. The STFT can be represented by Equation (4):

\[
STFT_x(t, f) = \sum_{n=0}^{L'-1} x(n)k(n - t)e^{-2\pi fn}
\]

where \( x(n) \) denotes the original signal at time \( n \), \( k(n - t) \) denotes the sliding window function of the original signal at time \( n \), and \( L' \) denotes the length of the window function. For a given time \( t \), \( STFT(t, f) \) can be regarded as the spectrum of that time.

2.1.3. Wigner–Ville Distribution

Wigner–Ville distribution [18] is directly defined as the time–frequency joint distribution function of the signal. It has no local stability limitation on the signal, and the time domain resolution is independent of the frequency resolution, which can be selected sepa-
rately. Wigner–Ville distribution is essentially the Fourier transform of the instantaneous correlation function of the signal. Assuming that there is a continuous-time signal \( x(t) \), its WVD can be expressed as Equation (5):

\[
WVD_{z}(t, f) = \int_{-\infty}^{\infty} z(t + \frac{\tau}{2})z^{*}(t - \frac{\tau}{2})e^{-j2\pi f \tau} d\tau
\]  

(5)

where \( z(t) \) is the analytic signal of signal \( x(t) \) and \( z(t + \frac{\tau}{2})z^{*}(t - \frac{\tau}{2}) \) is the instantaneous correlation function of the signal. The superscript ‘\( H \)’ represents conjugate transpose; \( t \) is time; \( f \) is the frequency; \( \tau \) is time delay; \( j \) is an imaginary unit. WVD can reflect the time–frequency characteristics of the signal and has good time–frequency aggregation.

2.1.4. Continuous Wavelet Transform

The continuous wavelet transform [19] decomposes the signal into wavelet coefficients at different scales through the scaling and translation of the wavelet function. CWT can adjust the window size for multi-resolution analysis according to the characteristics of different signals and is widely used in time–frequency analysis of signals. For any signal \( s(t) \) with finite energy, the formula for the continuous wavelet transform is shown in Equation (6):

\[
CWT = \langle s(t), \psi_{a,b}(t) \rangle = \frac{1}{\sqrt{c}} \int_{-\infty}^{\infty} s(t)\psi^{*}(\frac{t-b}{a})dt
\]

(6)

where \( CWT \) is the wavelet coefficient, \( c \) is the scale factor, \( b \) is the translation amount, \( \psi_{a,b}(t) \) is the wavelet function after scaling and translation, \( \psi^{*}(t) \) is the conjugate complex number of \( \psi(t) \), and \( \psi(t) \) is the mother wavelet function.

2.2. Convolutional Neural Networks

Convolutional neural networks originated from the study of biological neural vision [20]. CNN is composed of feature extractor and feature classifier. According to the input dimension, CNN has different types [21,22].

Generally, CNN architecture is primarily comprised of an input layer, convolutional layer, activation function, pooling layer, fully connected layer, and output layer [23]. In CNN, the kth convolution feature of the lth layer is expressed as Equation (7):

\[
I_{l}^{k}[x, y] = f \left( \sum_{c} I_{l-1}^{c}[x, y] \ast w_{ck}[x, y] + b_{k} \right)
\]

(7)

where \( l \) is the layer index, \( c \) is the channel index in the previous layers, and \( k \) is the channel index in the current layers; \( I_{l-1}^{c}[x, y] \) is the output from the previous layer, \( w_{ck}[x, y] \) is the convolution filter, \( b_{k} \) is the bias term, and \( f(\cdot) \) is the activation function. This paper uses ReLU as the activation function to ensure the convergence speed of the network [24].

The pooling layer is ordinarily situated after the convolutional layer, which serves to decrease the volume of data to be processed through methods like maximum pooling and mean pooling. Following this, the fully connected layer is leveraged to compile the features derived from the obtained feature map into a fixed-length feature column vector. Lastly, the output layer is composed of an n-class softmax layer that can generate an n-class label distribution. This distribution can be represented as Equation (8):

\[
\text{softmax}(P)_{o} = \frac{\exp(p_{o})}{\sum_{v=1}^{n} \exp(p_{v})}, \quad o = 1, 2, \ldots, n \text{ and } p = [p_{1}, p_{2}, \ldots, p_{n}]^{T} \in \mathbb{R}^{n}
\]

(8)

where \( \text{softmax}(P)_{o} \) is the output of the last fully connected layer, \( o \) is the target class index, and \( n \) is the number of categories. In multi-class problems, softmax normalizes the output value and converts the output value into probability.
As one of the classic CNN models, AlexNet [25] can perform parallel processing of image data by multiple GPUs for information interaction and can realize multi-feature learning.

3. Drilling Pump Fluid End Fault Diagnosis

3.1. Time–Frequency Image Generation

The poor working conditions of the drilling pump often result in serious noise pollution in the fluid end signal collected during experiments, which makes it challenging to extract the corresponding features efficiently and accurately using traditional methods. Hence, this study proposes the conversion of the vibration signal into a time–frequency image based on GST to preserve the signal’s time–frequency characteristics. Figure 2 illustrates the time–frequency image of the fluid end vibration signal generated through GST transformation, wherein signal segments with large fluctuations are revealed to possess more distinct features in the image.

![Time–frequency image of the fluid end vibration signal generated by GST.](image)

Figure 2. Time–frequency image of the fluid end vibration signal generated by GST.

Figure 3 depicts the process of transforming the fluid end vibration signal into a time–frequency image utilizing GST. In this process, L signal segments of length N are randomly selected from the collected vibration signal to generate L pieces.

3.2. Network Structure Optimization

By utilizing GST to create a time–frequency image of the fluid end vibration signal, an image dataset is formed and subsequently leveraged for training and prediction using CNN. This paper outlines a fault diagnosis model for the drilling pump fluid end that is based on the AlexNet architecture. In this model, a batch normalization (BN) layer [26] is introduced to optimize AlexNet and the number of neurons in the fully connected layer is modified to enhance the network’s ability to extract features. When the BN layer is introduced into the convolution layer, assuming the input of the first BN layer is 

$$y^l = (y_{1}^l, \ldots, y_{c}^l, \ldots, y_{d}^l)$$

the BN can be represented by Equation (9):

$$z_{ck}^l = \gamma_{ck}^l \frac{y_{ck}^l - \frac{1}{bd} \sum_{k=1}^{b} \sum_{j=1}^{d} y_{ck}^j}{\sqrt{\frac{1}{bd} \sum_{k=1}^{b} \sum_{j=1}^{d} (y_{ck}^j - \frac{1}{bd} \sum_{k=1}^{b} \sum_{j=1}^{d} y_{ck}^j)^2} + \beta} + \alpha_{ck}^l$$

where \(\gamma_{ck}^l\) is the scaling factor of the BN layer, \(\alpha_{ck}^l\) is bias, \(z_{ck}^l\) is the output, \(\beta\) is constant, and \(b\) is batch size.

The BN layer is a useful tool that can accelerate the convergence speed of the network, improve gradient dispersion, and enhance the generalization of the model. Figure 4
presents the parameter details and model structure of the optimized AlexNet network model with BN layers incorporated.

**Figure 3.** GST time–frequency image generation process. (a) Random selection of signals; (b) generation of time–frequency image datasets.

**Figure 4.** The AlexNet model in this paper. (a) Model details; (b) model structure.
3.3. Fluid End Fault Diagnosis Process

Figure 5 depicts the fault diagnosis process of the fluid end, which involves the following implementation steps:

1. Collecting vibration signals from the fluid end valve box of the drilling pump under different damage degrees of the SV and the DV at the fluid end.
2. Converting the vibration signals of valve boxes in different states into time–frequency images, using GST to form image datasets.
3. Dividing the dataset into training and test sets, inputting them into the optimized AlexNet model for training and prediction, calculating the accuracy rate, outputting the confusion matrix, and visualizing the diagnosis results.

4. Fluid End Diagnosis Results and Analysis

4.1. Signal Acquisition Experiment

Vibration signals from the fluid end of drilling pumps under different fault types were collected by conducting pump-operation experiments using faulty valve bodies. The experimental setup employed in this study, including the drilling pump, vibration sensor, data acquisition device, and data storage device, are illustrated in Figure 6. The signals were sampled at a frequency of 1 kHz. The suction and DVs used in the experiment were designed to have three states: normal, minor damage, and severe damage, as shown in Figure 7.
Figure 6. Equipment and process of signal acquisition. (a) Drilling pump; (b) fluid end data acquisition point; (c) vibration sensor; (d) acquisition equipment; (e) signals collected.

Figure 7. Operation states of the valve. (a) Normal state; (b) minor damage; (c) severe damage.

In the experiment, C and F represent severe and minor damage, respectively, while N indicates normal state. The combination of the SV and DV forms a total of nine different conditions. SCDC represents severe damage to both the SV and DV; SCDF represents severe damage to the SV and minor damage to the DV; SCDN represents severe damage to the SV and normal state of the DV; SFDC represents minor damage to the SV and severe damage to the DV; SFDF represents minor damage to both the SV and DV; SFDN represents minor damage to the SV and normal state of the DV; SNDC represents normal state of the SV and severe damage to the DV; SNDF represents normal state of the SV and minor damage to the DV; and SNDN represents normal state of both the SV and DV. The nine collected time domain signals are shown in Figure 8.
Figure 8. Nine valve states and their time-domain signals. (a) The nine states’ time domain signals; (b) the nine valve states.

4.2. Experimental Data Process

In the experimental vibration signal for each fault type, L signal segments with a length of N are randomly selected and a time–frequency image dataset is generated using GST. If the fault signal of slight damage at the fluid end contains K points, L vibration signal segments with length N are randomly selected from 1 to K-N, where N is set to be 1400 in this study. Each sampling length N can produce a time–frequency image, resulting in L images, each of which is different. The size of the time–frequency image in this study is $227 \times 227$. Four different datasets are obtained based on GST, STFT, WVD, and CWT methods. Figure 9 shows the images generated by different methods. Specifically, Figure 9a illustrates a time–frequency image generated using GST, Figure 9b depicts a time–frequency image based on STFT, Figure 9c presents a time–frequency image based on WVD, and Figure 9d demonstrates a time–frequency image generated using CWT.

Figure 9. Four methods to generate time–frequency images. (a) GST time–frequency image; (b) STFT time–frequency image; (c) WVD time–frequency image; (d) CWT time–frequency image.

4.3. Comparison of Diagnostic Results

For each generated dataset using the four different methods, there are nine states in the dataset and 2000 images are generated for each state, resulting in a total of 18,000 images which are further divided into training and test sets, with the training set and the test set accounting for 80% and 20%, respectively. The generated time–frequency images are fed
into the model for training and testing. In this study, the AlexNet model is optimized by introducing the BN layer and changing the number of neurons in the first fully connected layer, and the performance of the model in identifying the type of fluid end faults is analyzed. The FC6 is varied across 128, 256, 512, 1024, 2048, 3072, 4096, and 5120, leading to a total of eight models. The training samples are randomly selected from the dataset, and the training set and the test set are different.

The four generated datasets are utilized to evaluate the original AlexNet model and the AlexNet model with the BN layer, using labels assigned for the fault types of the fluid end. In order to minimize discrepancies, the number of neurons in the first fully connected layer of each model is tested five times, and the average value is reported for diagnostic purposes. Table 1 presents the diagnostic results of the original AlexNet model, while Table 2 shows the diagnostic results of the AlexNet model introduced with the BN layer.

<table>
<thead>
<tr>
<th>No</th>
<th>CNN-128</th>
<th>CNN-256</th>
<th>CNN-512</th>
<th>CNN-1024</th>
<th>CNN-2048</th>
<th>CNN-3072</th>
<th>CNN-4096</th>
<th>CNN-5120</th>
</tr>
</thead>
<tbody>
<tr>
<td>GST</td>
<td>97.65</td>
<td>97.82</td>
<td>98.21</td>
<td>98.46</td>
<td>98.43</td>
<td>98.33</td>
<td>98.46</td>
<td>98.30</td>
</tr>
<tr>
<td>STFT</td>
<td>97.30</td>
<td>97.81</td>
<td>97.86</td>
<td>98.09</td>
<td>98.25</td>
<td>98.39</td>
<td>98.17</td>
<td>98.17</td>
</tr>
<tr>
<td>WVD</td>
<td>95.23</td>
<td>96.12</td>
<td>96.03</td>
<td>96.65</td>
<td>96.64</td>
<td>96.44</td>
<td>96.74</td>
<td>96.76</td>
</tr>
<tr>
<td>CWT</td>
<td>94.25</td>
<td>94.59</td>
<td>95.11</td>
<td>95.34</td>
<td>95.61</td>
<td>95.87</td>
<td>95.88</td>
<td>95.65</td>
</tr>
</tbody>
</table>

Table 2. AlexNet model results with BN layer (%).

<table>
<thead>
<tr>
<th>No</th>
<th>CNN-128</th>
<th>CNN-256</th>
<th>CNN-512</th>
<th>CNN-1024</th>
<th>CNN-2048</th>
<th>CNN-3072</th>
<th>CNN-4096</th>
<th>CNN-5120</th>
</tr>
</thead>
<tbody>
<tr>
<td>GST</td>
<td>98.97</td>
<td>99.10</td>
<td>99.15</td>
<td>99.21</td>
<td>99.10</td>
<td>98.91</td>
<td>99.00</td>
<td>99.05</td>
</tr>
<tr>
<td>STFT</td>
<td>99.02</td>
<td>99.02</td>
<td>99.01</td>
<td>99.02</td>
<td>99.14</td>
<td>99.10</td>
<td>99.03</td>
<td>99.04</td>
</tr>
<tr>
<td>WVD</td>
<td>97.55</td>
<td>97.63</td>
<td>97.47</td>
<td>97.41</td>
<td>97.44</td>
<td>97.48</td>
<td>97.51</td>
<td>97.23</td>
</tr>
<tr>
<td>CWT</td>
<td>96.14</td>
<td>96.34</td>
<td>96.63</td>
<td>96.69</td>
<td>96.63</td>
<td>96.54</td>
<td>96.80</td>
<td>96.70</td>
</tr>
</tbody>
</table>

The experimental results of each model demonstrate that the time–frequency image dataset generated using the GST method achieved the highest diagnostic accuracy and that the diagnostic accuracy of the model utilizing the BN layer is approximately 1% higher than the model without the BN layer. Specifically, the AlexNet model with the BN layer and the first fully connected layer (FC6) containing 1024 neurons achieved the best classification performance, with an average accuracy of 99.21%. This suggests that an appropriate number of fully connected layer neurons can improve the accuracy of the diagnosis.

To validate the performance of the model, T-SNE visualization technology was employed in this study to visualize the features of key network layers in the optimized AlexNet. This is depicted in Figure 10. Initially, the first, third, and last convolutional layers were selected for visualization, followed by the first fully connected layer (FC6) and the last fully connected layer FC8. A total of five features were visualized using this process. By utilizing T-SNE visualization, the study was able to better understand and interpret the performance of the AlexNet model and the effectiveness of the proposed fault diagnosis approach.

As illustrated in Figure 10, the network model appears to contain all the relevant feature information regarding the input data. Initially, the features of the first convolutional layer and the third convolutional layer were undistinguished. However, after passing through the BN and fourth convolutional layers, the features of the last convolutional layer became increasingly distinguishable. This effect was particularly noticeable in the FC6 and FC8 layers, where their features were well-differentiated. This observation indicates that the optimized AlexNet model is effective in extracting and processing the necessary feature information to accurately diagnose faults in the fluid end of a drilling pump.
To further assess the optimized AlexNet model’s performance, GST-, WVD-, STFT-, and CWT-generated datasets are provided as input to ResNet-18 [27], SqueezeNet [28], and LeNet [29] for training and prediction. Next, the one-dimensional vibration signals of nine states at the fluid end are utilized as the input data. A one-dimensional convolutional neural network paired with the long short-term memory [30] network (CNN-LSTM) is applied for this purpose. The signal processed by the correlated kurtosis deconvolution (MCKD) [31] is input into the method in 1D CNN (MCKD-CNN). Compared with AlexNet network, the innovation of ResNet-18 network structure is that it proposes the concept of residual connection, which directly establishes shortcuts between different layers of the network, making the transmission of information more efficient. After ResNet-18 uses two layers of convolutional local blocks, it uses nine residual block layers and finally uses a fully connected layer to complete the image classification task. Compared with AlexNet, SqueezeNet improves the operation speed to the greatest extent without reducing the accuracy of the model. It can perform more efficient distributed training and has higher expansion efficiency. The main feature of LeNet is the combination of convolutional and downsampling layers as the basic structure of the network, which includes three convolutional layers, two downsampling layers, and two fully connected layers. LeNet was originally designed to identify handwritten characters and print characters, and it achieved good results in this regard. CNN-LSTM network is a neural model that combines CNN and LSTM. It has strong feature extraction ability, memory ability, and characteristics suitable for different data types. It can extract and predict the characteristics of one-dimensional signals well. The MCKD-CNN network used in this paper is a method that combines signal processing algorithms with 1D CNN. Firstly, the MCKD is used to process the one-dimensional signal to highlight the fault characteristics of the signal, and then the 1D CNN is used to train and predict the processed signal. When the type of input signal is a one-dimensional signal, in order to ensure the fairness of the comparison the length of the signal used is also 1400 points; 2000 in each condition. Table 3 shows the comparison results.

Figure 10. Visualization of features in the optimized AlexNet.
Table 3. Results of different fault diagnosis methods.

<table>
<thead>
<tr>
<th>Image Datasets</th>
<th>Methods</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GST</td>
<td>AlexNet-1024</td>
<td>99.21</td>
</tr>
<tr>
<td></td>
<td>ResNet-18</td>
<td>98.56</td>
</tr>
<tr>
<td></td>
<td>SqueezeNet</td>
<td>92.97</td>
</tr>
<tr>
<td></td>
<td>LeNet</td>
<td>86.01</td>
</tr>
<tr>
<td>WVD</td>
<td>AlexNet-256</td>
<td>97.63</td>
</tr>
<tr>
<td></td>
<td>ResNet-18</td>
<td>98.42</td>
</tr>
<tr>
<td></td>
<td>SqueezeNet</td>
<td>90.64</td>
</tr>
<tr>
<td></td>
<td>LeNet</td>
<td>84.17</td>
</tr>
<tr>
<td>STFT</td>
<td>AlexNet-2048</td>
<td>99.14</td>
</tr>
<tr>
<td></td>
<td>ResNet-18</td>
<td>99.09</td>
</tr>
<tr>
<td></td>
<td>SqueezeNet</td>
<td>92.33</td>
</tr>
<tr>
<td></td>
<td>LeNet</td>
<td>78.42</td>
</tr>
<tr>
<td>CWT</td>
<td>AlexNet-4096</td>
<td>96.80</td>
</tr>
<tr>
<td></td>
<td>ResNet-18</td>
<td>96.71</td>
</tr>
<tr>
<td></td>
<td>SqueezeNet</td>
<td>82.78</td>
</tr>
<tr>
<td></td>
<td>LeNet</td>
<td>68.67</td>
</tr>
<tr>
<td>One-dimensional signals</td>
<td>CNN-LSTM</td>
<td>70.32</td>
</tr>
<tr>
<td></td>
<td>MCKD-CNN</td>
<td>92.36</td>
</tr>
</tbody>
</table>

For the drilling pump’s fluid end, the comparison results presented in Table 3 indicate that the method proposed in this study is more suitable for fault diagnosis. Figure 11 illustrates the confusion matrix results obtained in an experiment where the BN layer is introduced and the number of FC6 neurons is set to 1024. The rows in Figure 11 correspond to the true label, while the columns correspond to the predicted label. The results reveal that the prediction accuracy rate in this experiment is 99.44%. Labels SFDC, SFDN, SNDC, and SNDN demonstrated the highest recognition accuracy, reaching 100%. However, label SCDF had the lowest identification accuracy at 98.00% since eight SCDF labels are wrongly identified as SCDC. This is because the vibration signal of the severe damage of the DV masks the signal of the minor damage of the DV. The characteristics of the time–frequency images generated by some signal segments are similar, which leads to the wrong identification of the network model. In addition, two SCDC labels are incorrectly identified as SCDF, three SCDC labels are incorrectly identified as SFDC, and three SFDF labels are wrongly identified as SCDC. The misidentification of two SCDC labels as SCDF and three SFDF labels as SCDC may be attributed to the incorrect identification results between severe damage and minor damage. One possible reason for this is that as the pump operates, valves labeled with minor damage continue to deteriorate, resulting in similar signal segments between those with minor and severe damage labels. Specifically, the vibration signal from severe damage of the DV may have masked the signal from minor damage of the same valve, causing some signal segments to generate similar spectrograms. This ultimately led to the misidentification of the network model. One SFDF label is incorrectly identified as SCDN, one SNDF label is inaccurately identified as SCDN, one SNDF label is incorrectly identified as SFDN, and one SCDN label is inaccurately identified as SFDN.
5. Conclusions

This study proposes a fault diagnosis method suitable for the fluid end based on the vibration signal in the actual noise environment, targeting the low diagnostic rate of the fluid end under traditional machine learning. Firstly, the proposed method involves using GST to transform the original vibration signal into a time–frequency image dataset, which preserves the time–frequency characteristics of the vibration signal. Next, the fault identification performance of different AlexNet models for slight and severe damage to the fluid end is explored. Finally, the diagnostic results are compared using multiple diagnostic methods. The main conclusions of this study are as follows.

(1) A GST-CNN-based fault diagnosis method is proposed for the fluid end, which achieves higher fault diagnosis accuracy compared to traditional machine learning methods.

(2) This study explores the optimization of the AlexNet model by incorporating a batch normalization layer (BN) and adjusting the number of neurons. The findings suggest that adding the BN layer to the AlexNet model, along with an appropriate number of fully connected layer neurons, can improve the accuracy of fault diagnosis.

(3) The diagnostic model proposed in this approach has been shown to achieve an impressive average accuracy rate of 99.21% for diagnosing the nine categories of the fluid end. This surpasses the performance of many other available diagnostic methods and provides a fast and reliable alternative in the diagnosis of the fluid end. This development provides a useful tool for ensuring optimal performance and efficiency of fluid ends.

Author Contributions: Conceptualization, W.Z.; methodology, A.T.; software, A.T.; validation, A.T.; investigation, A.T.; resources, W.Z.; writing—original draft, A.T.; writing—review & editing, W.Z.; supervision, W.Z.; project administration, W.Z.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.
Funding: This work was supported by Sichuan Science and Technology Program (Grant Nos. 2022YFG0063 and 2022YFQ0114).

Data Availability Statement: Not applicable.

Acknowledgments: The authors acknowledged the financial support from Sichuan Science and Technology Program (Grant Nos. 2022YFG0063 and 2022YFQ0114).

Conflicts of Interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

References


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