applied sciences

MDPI

*Article*

# The Fault Diagnosis of a Plunger Pump Based on the SMOTE + Tomek Link and Dual-Channel Feature Fusion

**Xiwang Yang [1,2], Xiaoyan Xu [2], Yarong Wang [2], Siyuan Liu [2], Xiong Bai [3], Licheng Jing [3], Jiancheng Ma [2] and Jinying Huang [3,*]**

[1] School of Information and Communication Engineering, Shanxi University of Electronic Science and Technology, Linfen 041000, China; yangxw@nuc.edu.cn
[2] School of Computer Science and Technology, North University of China, Taiyuan 030051, China; sz202207034@st.nuc.edu.cn (X.X.); wangyarong0820@foxmail.com (Y.W.); b200706@st.nuc.edu.cn (S.L.); b1907076@st.nuc.edu.cn (J.M.)
[3] School of Mechanical Engineering, North University of China, Taiyuan 030051, China; b20230210@st.nuc.edu.cn (X.B.); s202202021@st.nuc.edu.cn (L.J.)
* Correspondence: jyhuang@nuc.edu.cn

**Abstract:** Mechanical condition monitoring data in real engineering are often severely unbalanced, which can lead to a decrease in the stability and accuracy of intelligent diagnosis methods. In this paper, a fault diagnosis method based on the SMOTE + Tomek Link and dual-channel feature fusion is proposed to improve the performance of the sample imbalance fault diagnosis method, taking the piston pump of a turnout rutting machine as the research object. Combining the data undersampling method and the oversampling method to redistribute the collected normal data and fault data makes the diagnostic model have better diagnostic performance in the case of insufficient fault samples. And, in order to fully utilize the global features and local features, a global–local feature complementary module (GLFC) is proposed. Firstly, the generated data similar to the original data are constructed using the SMOTE + Tomek Link method; secondly, the generated data are input into a GLFC module and BiGRU at the same time, the GLFC module extracts the spatial global features and local features of the original vibration data, and BiGRU extracts the temporal information features of the original vibration data, and fuses the extracted feature information, and inputs the fused features into the attention layer; finally, a GLFC module is proposed by the SMOTE + Tomek Link method to make full use of the global features and local features. The extracted feature information is fused, and the fused features are input to the attention layer; finally, the fault classification is completed by the softmax classifier. In this paper, the accuracy and robustness of the proposed model are demonstrated through experiments.

**Keywords:** fault diagnosis; plunger pump; two-channel feature fusion; SMOTE + Tomek Link



**Citation:** Yang, X.; Xu, X.; Wang, Y.; Liu, S.; Bai, X.; Jing, L.; Ma, J.; Huang, J. The Fault Diagnosis of a Plunger Pump Based on the SMOTE + Tomek Link and Dual-Channel Feature Fusion. *Appl. Sci.* **2024**, *14*, 4785. https://doi.org/10.3390/app14114785

Academic Editor: Jérôme Morio

Received: 30 April 2024
Revised: 25 May 2024
Accepted: 29 May 2024
Published: 31 May 2024



## 1. Introduction

With the rapid development of information technology and intelligent technology, the massive monitoring data of railway systems provides strong support for the intelligent operation and maintenance of railway equipment and health management. Among them, the condition monitoring and health management of a plunger pump, a key component of a railway switch machine, is of great significance. However, in the actual complex industrial process, the frequency of failure is very low, and normal data are greater in number than fault data. Therefore, the industry's big data often present types of unbalanced characteristics. It is a hot topic to classify non-equilibrium data efficiently and precisely. At present, there are two kinds of classification methods for unbalanced data: the first method changes the data distribution to achieve the data level of data balance between classes; the second approach is an algorithm-level approach that focuses on a small number of data samples. In terms of data, scholars mainly use resampling techniques to adjust original

samples to reduce the impact of dataset imbalance. Jeyalakshmi, K. et al. [1] proposed a new oversampling method, the weighted synthetic Minority oversampling technique (WSMOTE). This method generates new synthetic data by randomly selecting some samples of a few classes, and assigns different weights to those data according to the importance of the few classes, which improves the importance of the samples in the classifier. Lin et al. [2] integrated the clustering idea into the undersampling method and dynamically adjusted the undersampling proportion according to the sample distribution and model performance to balance the class distribution more accurately and improve the model performance. Duan F. et al. [3] put forward the Mean Radius-SMOTE oversampling method to solve the problem of data unbalance in mechanical fault diagnosis. This method is used to calculate the neighborhood radius based on the average distance between the sample point and its nearest neighbors of the same class. Xu Y. et al. [4] proposed an improved SMOTE method (i.e., MSMOTE) for processing datasets with imbalanced classifications and applied this to multi-fault diagnosis. Rao S. et al. [5] combined the SMOTE algorithm with an artificial neural network to solve the problem of sample imbalance and classification in power transformer fault diagnosis, which improved the accuracy and reliability of the model. Yang et al. [6] proposed a hybrid classifier architecture which first uses density undersampling to obtain balanced subsets of multiple categories, and then uses a cost-effective classification method to deal with information incompleteness. Bernardo et al. [7] proposed a SMOTEOB method that enables the dataset to be rebalanced in a continuous data stream and the classifier to be trained and updated online, thus making the model more adaptable and robust.

Aimed at the problem that traditional fault diagnosis methods require a lot of manual experience and that feature extraction is incomplete, Fern et al. [8] discussed the current status and applications of SMOTE to commemorate its 15th anniversary, and identified the next set of challenges for SMOTE in solving big data problems. Torgo et al. [9] used sampling methods to solve regression tasks where the target variable is continuous. The SMOTER method proposed can be used in conjunction with any existing regression algorithm, making it a universal tool for solving rare extremum problems in predicting continuous target variables. Blagus et al. [10] studied the properties of SMOTE from both theoretical and empirical perspectives, using simulated and real high-dimensional data to explain the class prediction that affects high-dimensional data. Chawla et al. [11] described a method for constructing classifiers from imbalanced datasets, which combines the oversampling of minority classes with the undersampling of majority classes to achieve better classifier performance.

Liu et al. [12] proposed a continuous learning model based on weight space element representation to apply to the incremental fault diagnosis of point machine plunger pumps. Wei X. et al. [13] proposed a cavitation fault diagnosis method based on the combination of long short-term memory (LSTM) and a one-dimensional convolutional neural network (1D-CNN) to identify the cavitation grade of vibration signals under different inlet pressures. Although the above research methods provide important reference value in various fault diagnosis tasks, some challenging problems still need to be solved in order to further improve the accuracy and stability of diagnosis.

(1) Most of the existing research on intelligent fault diagnosis methods for plunger pumps is based on the premise of balanced data. When the proportion of fault samples is much smaller than that of healthy samples or the size of different types of samples is greatly different, the accuracy and stability of the diagnostic model will be significantly reduced.

(2) The method of the data enhancement of samples by generating generative models, such as adversarial networks, is prone to problems such as training collapse, resulting in the poor quality of the generated samples and their insufficient reliability as training samples [14–16].

(3) The convolutional neural network (CNN), as a plunger pump-monitoring signal feature extractor, lacks the ability to capture the features of signal timing on a time-scale, so it is necessary to conduct in-depth studies on this aspect [17,18].

To solve the above problems, this paper puts forward a SMOTE + Tomek Link and two-channel feature fusion model. The combination of the data undersampling method and the oversampling method can reduce the number of samples, ensure the balance of training samples, and realize the fault diagnosis of data sample imbalance in the piston pump of the switch machine. The accuracy and robustness of the proposed model are proved through experiments.

## 2. Method Principles

### *2.1. Convolutional Neural Network*

After data input, the convolutional layer uses multiple convolution cores with the same weight to capture the spatial features of the local region and obtains multiple feature mappings which are used as the input data of the next layer. The specific calculation equation is shown in Equation (1):

$$x_i^l = f\left(w_i^l * x^{l-1} + b_i^l\right) \tag{1}$$

where $x_i^l$ is the feature vector of the output value, $w_i^l$ is the weight matrix of the convolution kernel, $b$ is offset, and $f(\cdot)$ is the activation function, $*$ is a convolution operation performed on two elements.

### *2.2. Bidirectional Gated Recursive Unit*

The upper channel extracts the spatial features of the sample data, but the temporal features in the sample data cannot be processed. Therefore, the lower channel adopts a Bidirectional Gated Recurrent Unit (BiGRU) neural network to extract features from data in both forward and reverse directions [19]. Its specific structure is shown in Figure 1.

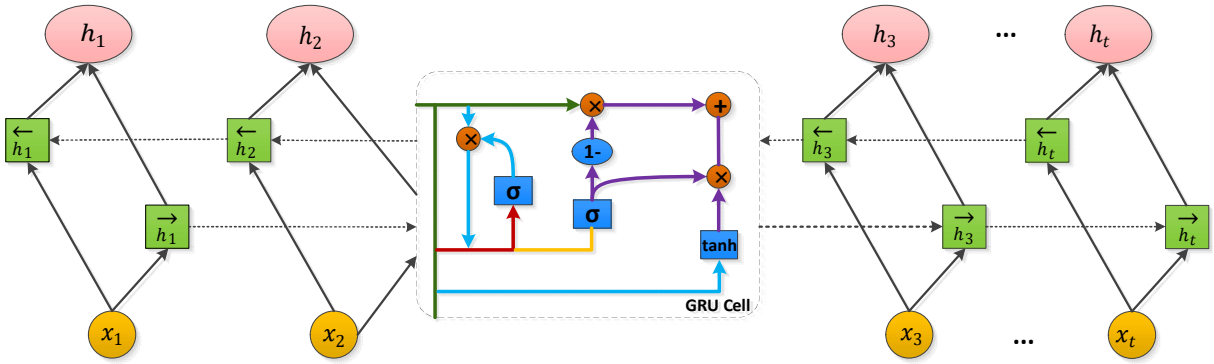


**Figure 1.** A diagram of the BiGRU's structure.

The BiGRU neural network compensates for the GRU's inability to encode information from back to front. In the BiGRU neural network, information has two transmission routes, so you can extract more before and after information, and have a stronger learning ability. The BiGRU is composed of two GRUs stacked on top of each other, and its output depends on the state of the two GRUs. At time t, the hidden state of the BiGRU is obtained by the weighted summing of the forward hidden state $\overrightarrow{h}_t$ and the reverse hidden state $\overleftarrow{h}_t$, as shown in Equations (2)–(5).

$$\overrightarrow{h}_t = GRU(x_t, \overrightarrow{h}_{t-1}) \tag{2}$$

$$\overleftarrow{h}_t = GRU(x_t, \overleftarrow{h}_{t-1}) \tag{3}$$

$$h'_t = w_t \overrightarrow{h_t} + v_t \overleftarrow{h_t} + b_t \tag{4}$$

$$q_t = \sigma(w_O \bullet h'_t) \tag{5}$$

where $x_t$ represents the input information, the $GRU(\cdot)$ function is used to perform a nonlinear transformation on the input vector, thereby encoding the vector into the corresponding GRU hidden state. In this process, $w_t$ and $v_t$ represent the weights corresponding to the forward hidden state and the reverse hidden state of the BiGRU at time $t$, respectively, $b_t$ represents the bias of the corresponding hidden state at time $t$, and $q_t$ representing the output at time $t$.

### *2.3. Global–Local Feature Complementary Module*

In order to make full use of global feature and local feature, a global–local feature complementary module (GLFC) is proposed. In order to make full use of the feature information of different dimensions to supplement the original feature information, the module uses the extended convolution of different expansion rates to extract the feature information of different dimensions. Firstly, local feature information is obtained through dilated convolution layers with smaller dilation rates (d = 1, 3), and global feature information is obtained through dilated convolution layers with larger dilation rates (d = 5, 7). Secondly, dot product operations are performed on the local and global features separately. Finally, the local features, global features, and original input features are added point by point to obtain the output features. The structure of the GLFC module is shown in Figure 2.

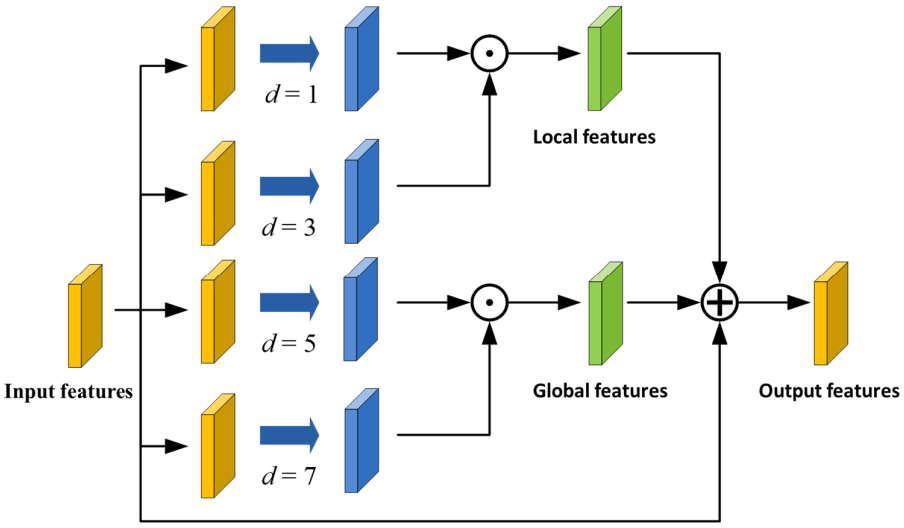


**Figure 2.** The GLFC module's structure.

## 3. Unbalanced Data Processing Method Based on Oversampling and Undersampling

Although SMOTE oversampling can generate new samples, it simply interpolates on a few samples and does not take into account the distribution of the surrounding majority samples, which may cause the minority samples to be misclassified as majority samples, and may also increase the overlap problem within the sample set. In contrast, the data-cleaning undersampling method specifically cleans out overlapping data. Therefore, a mixed sampling method is used in this paper, that is, the dataset is oversampled first to increase the number of a few classes, and then the dataset that has completed the oversampling is cleaned to reduce the numbers of most classes. This approach can both balance the dataset and eliminate overlapping data.

Currently, some popular variants of the mixed resampling method are a combination of SMOTE [20,21] and ENN [22], and a combination of SMOTE and Tomek Link undersampling. While the Tomek Link and ENN are both algorithms for dealing with unbalanced datasets, their ideas and processings are different [23]. The ENN algorithm eliminates noise

and boundary samples by comparing the Euclidean distance between each sample and its *k*-nearest neighbor samples. However, in some cases, this approach may over-eliminate noise and boundary samples, resulting in the reduction of a small number of class samples, affecting the performance and accuracy of the classifier. In addition, the effect of the ENN algorithm largely depends on the choice of *k* value. If the selected *k* value is too small, then the ENN algorithm may ignore some important minority class samples, and if the selected *k* value is too large, then the ENN algorithm may include some majority class samples, resulting in the unbalance of the dataset not being solved. On the one hand, the Tomek Link algorithm can retain a few important class samples because the Tomek Link algorithm finds the nearest sample pairs that belong to different classes and then removes most class samples in these sample pairs. On the other hand, the Tomek Link algorithm does not need to explicitly select the *k* value because it is a distance-based down-sampling algorithm rather than a neighbor-based algorithm, and it does not need to specify the number of neighbors.

To sum up, the Tomek Link algorithm has solved the above two problems perfectly, so this chapter adopts the combination of SMOTE and Tomek Link undersampling to solve the problem of piston pump sample imbalance. In the following sections, the process of the hybrid sampling method is described in detail, and its working principle is illustrated by Figure 3.

(1) Randomly select a sample from a minority class and determine *k* as the number of nearest neighbor samples. If there is no certainty, then *k* = 5.
(2) Select a neighbor of the sample, generate a new composite sample by interpolation, and then add it to a few categories.
(3) Repeat step (2) until enough synthetic samples are generated.
(4) The Tomek link algorithm is used to detect the distance between two samples to calculate whether they belong to the same category.
(5) If there is no other sample of the same class between two samples of different classes, then there is a Tomek link between the two samples. Delete the Tomek link pair.
(6) Combine the generated synthetic sample with the original sample retained to obtain a new balanced dataset.

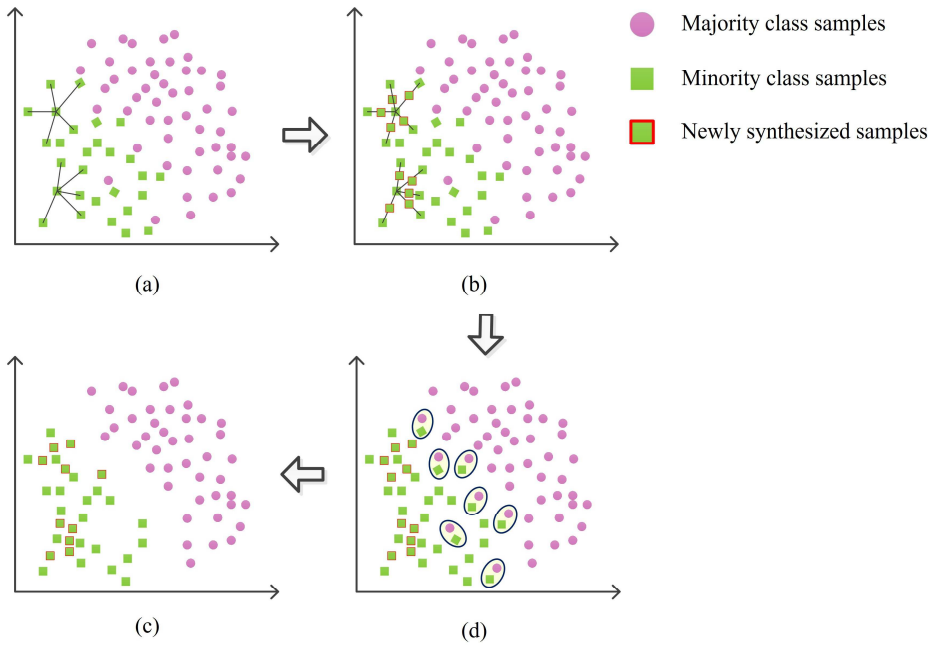


**Figure 3.** The SMOTE + Tomek Link sampling flow chart. (**a**) Identify nearest neighbor samples with *k* = 5 around a few class samples, (**b**) synthesize new minority class samples (marked in red box), (**c**) delete the Tomek Link pair to get balanced data, (**d**) and select the Tomek Link pair.

## 4. CNN-BiGRU-Attention Unbalanced Classifier for the Fault Diagnosis of a Plunger Pump

In this paper, a two-channel feature fusion fault diagnosis model of a plunger pump is proposed, which is composed of three modules, namely, a feature extraction layer, a feature fusion layer, and a fault classification layer. The feature extraction layer is composed of the upper channel GLFC module and the lower channel BiGRU. The upper channel GLFC module can extract the local and global spatial features contained in the sample data, while the lower channel BiGRU can fully extract the temporal features of the sample data from both forward and reverse directions in the time dimension. The feature fusion layer fuses the spatial feature and the time feature. The self-attention mechanism assigns different weights to the extracted features, enhances the features related to the signal, and suppresses the features unrelated to the signal. Finally, the softmax classifier is used in the fault classification layer to classify feature vectors. The overall structure of the fault diagnosis model is shown in Figure 4.

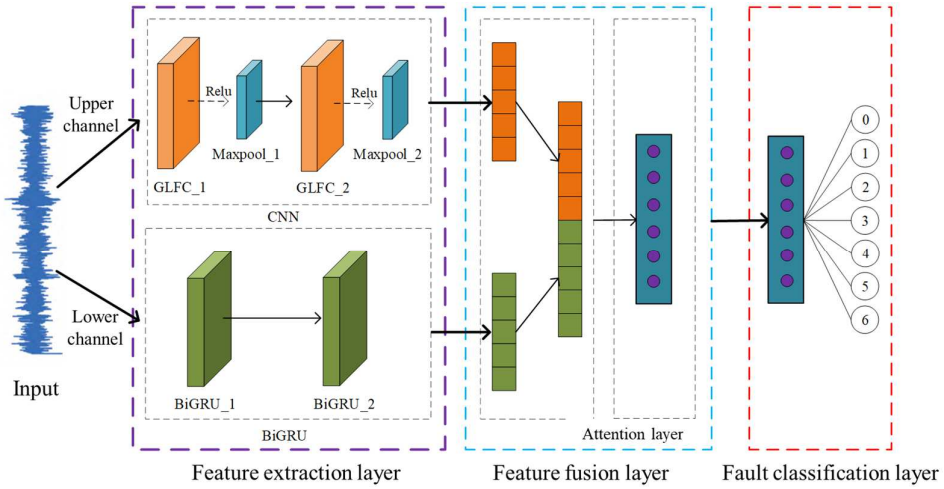


**Figure 4.** The structure of the CNN-BiGRU-Attention fault diagnosis model based on two-channel feature fusion for a rutting machine.

### *4.1. Feature Extraction Layer*

The original signal is intercepted with a fixed data length and labeled as the input of the model. The pre-processed data are input into the upper and lower channels at the same time, and then the spatial and temporal features contained in the signal are mined from the two channels, respectively, and the extracted features are input into the feature fusion layer.

### *4.2. Feature Fusion Layer*

In order to improve the performance of the model, the feature fusion layer fuses the features of multiple inputs to obtain a more comprehensive feature representation. In this paper, the concatenate layer is used for feature fusion, which can concatenate multiple feature vectors of the same dimension according to the specified dimension to obtain a feature vector with a larger dimension. In the case of there being no loss of information, multiple different sources and different types of feature vectors are integrated to obtain a more comprehensive and comprehensive feature representation. This plays an important role in processing multi-modal data or processing multiple tasks.

Specifically, in the two-channel feature fusion CNN-BiGRU-Attention plunger pump fault diagnosis model, Maxpool-2 and BiGRU-2 are the two-channel convolutional pooling layer output and the BiGRU layer output, respectively. The concatenate layer is obtained through the feature fusion operation of Maxpool-2 and BiGRU-2. Then, the fused feature is input to the attention layer, the feature vector is re-weighted, and finally, the $N$-dimensional feature vector is output.

### 4.3. Fault Classification Layer

The softmax classifier classifies the $N$-dimensional feature vectors output by the feature fusion layer, and the softmax model can be represented by Equation (6) as follows:

$$Y = soft\ \max(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{N} e^{x_k}} \tag{6}$$

where $x_i$ is the $i$-th element of the input vector, $Y$ is the probability of each class, and the softmax function converts each element in the input vector into its form as a probability value.

### 4.4. Model Training

The specific process of the CNN-BiGRU-Attention plunger pump fault diagnosis method based on dual-channel feature fusion is shown in Figure 5. The specific fault diagnosis steps are as follows.

(1) Obtain the original vibration data of the plunger pump and randomly divide the data samples into a training set and a test set according to a 7:3 ratio.
(2) Initialize the CNN and BiGRU weights and bias items and input the data of the training set into the CNN and the BiGRU in batches at the same time for fault feature learning.
(3) A CNN and BiGRU were used to extract the spatial and temporal characteristics of vibration data and then feature fusion was carried out, and the fused features were input into the attention layer and, finally, into the softmax layer.
(4) Use softmax to classify the faults of the plunger pump and fine-tune the model parameters according to the changes in loss value and accuracy.
(5) Input the test set into the trained model and output the diagnostic results.

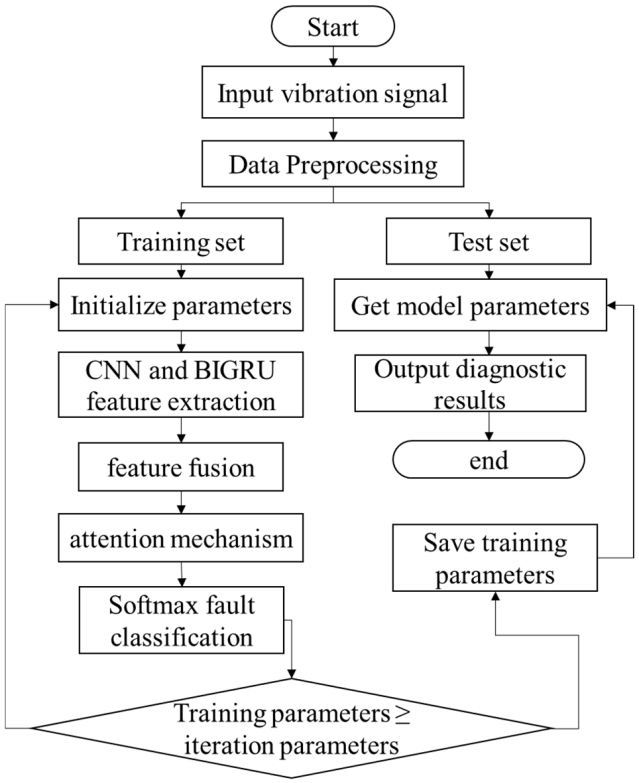


**Figure 5.** A flowchart of the CNN-BiGRU-Attention plunger pump fault diagnosis based on two-channel feature fusion.

## 5. Experiment and Analysis

The experimental running environment includes the following: the server model is a Dell Powe Edge C6145, the operating system is a 64-bit Windows server 2012 R2, the programming language is Python (3.7.9), the editor is PyCharm (v.2023.1), and the integration platform is Anaconda 2020.3.

### *5.1. Experimental Platform and Dataset*

The fault diagnosis experimental platform for the plunger pump is shown in Figure 6. The ZT axial plunger pump, which is a specialized electro-hydraulic pump, is selected as the test object. After being powered on, the motor drives the plunger pump to work, the cylinder body rotates, and the plunger moves back and forth. The high-pressure oil discharged is connected to the oil return pipeline through the overflow valve and flows back to the oil tank. The vibration signal of the pump casing is collected through an accelerometer, and the data acquisition system and supporting upper computer software are used to complete the data acquisition work.

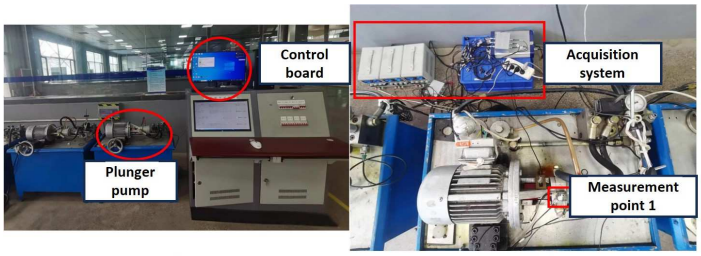


**Figure 6.** The experimental platform. (**a**) The equipment and (**b**) the measurement point layout.

This experiment selected a sampling frequency of 5120 Hz, and the experimental platform collected data from seven working conditions: the normal operation of the axial piston pump, the condition where plunger ball head wear occurs, that where there is triangular hole-plugging, that where valve plate wear occurs, that where cylinder block wear occurs, that where bowl wear occurs, and that where plunger wall pitting occurs. The sensor measuring points are arranged on the pump casing. The specific fault mode is shown in Figure 7.

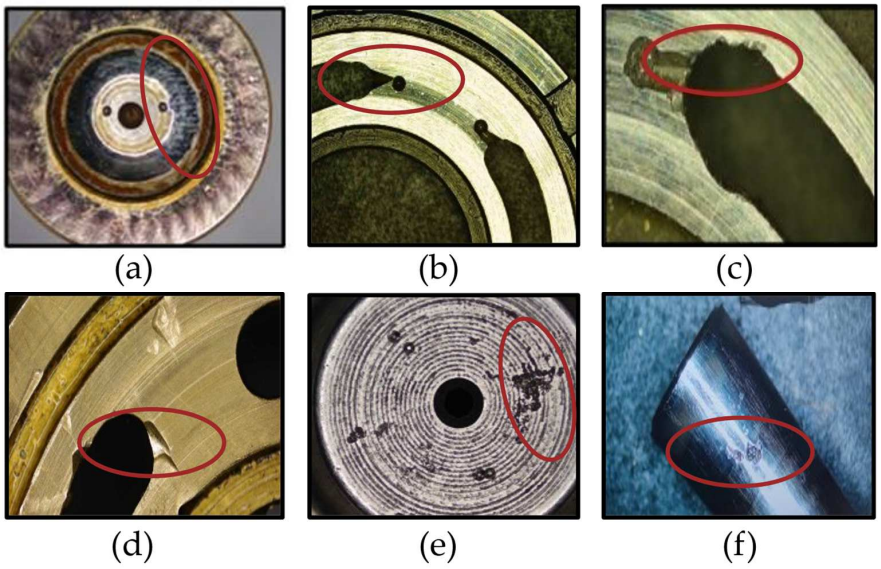


**Figure 7.** The plunger pump failure mode. (**a**) Plunger ball head wear, (**b**) triangular hole plugging, (**c**) the valve plate wear, (**d**) cylinder block wear, (**e**) bowl wear, (**f**) plunger wall pitting.

### *5.2. Performance Analysis of an Unbalanced Dataset in the Dual-Channel Feature Fusion Model*

#### 5.2.1. Reconstruction of Unbalanced Datasets

In order to fully simulate the data collection of mechanical equipment in the real environment and further reduce the samples of different working conditions, the unbalanced dataset of different working conditions of the artificially constructed plunger pump is further balanced. The unbalanced dataset of different working conditions of the artificially constructed plunger pump is generated, as shown in Table 1:

**Table 1.** A comparison of the imbalanced dataset with the multi-classification of the dataset processed based on the SMOTE + Tomek Link integration algorithm.

| Type | Label | Raw Dataset | SMOTE + Tomek Link Post-Equilibrium Dataset |
|---|---|---|---|
| Normal | 0 | 4600 | 4495 |
| Plunger ball head wear | 1 | 129 | 4253 |
| Triangular hole plugging | 2 | 200 | 4060 |
| Valve plate wear | 3 | 270 | 4423 |
| Cylinder block wear | 4 | 217 | 4227 |
| Bowl wear | 5 | 281 | 3836 |
| Plunger wall pitting | 6 | 242 | 4432 |

#### 5.2.2. Experimental Results

As can be seen from the results in Figure 8, when unbalanced data are used as the training set, the classification effect of plunger ball head wear and triangle hole blockage on the test set is somewhat different from that of other categories of data, decreasing by 31% and 25%, respectively. The main reason for this is that the real samples of these two types of faults are small, and the information learned during the training of the diagnosis algorithm is very limited, so it is easy to cause the problem of misclassification, which also confirms the problem of data imbalance restricting the performance of the diagnosis model mentioned above.

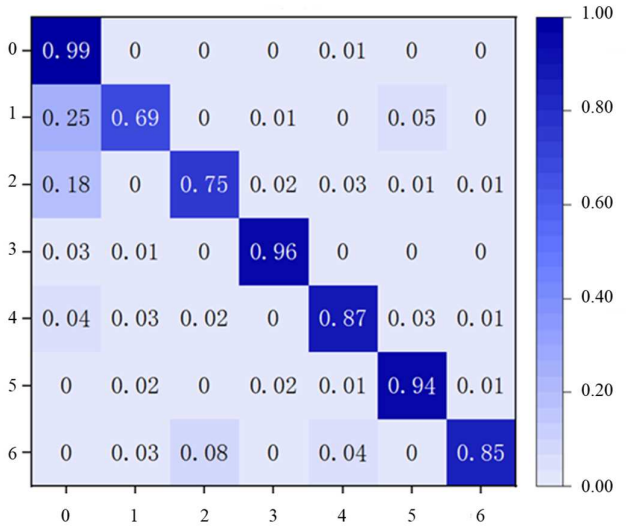


**Figure 8.** Test results based on the unbalanced dataset.

### *5.3. SMOTE + Tomek Link Mixed Sampling Performance Analysis*

#### 5.3.1. Resampling Data

Resampling was performed using SMOTE + Tomek Link mixed sampling. The unbalance dataset used in the experiment is a multi-classification dataset, which is selected to

analyze six kinds of samples, specifically, plunger ball head wear, triangle hole blockage, and valve plate wear, cylinder block wear, bowl wear, and plunger wall pitting. Oversampling was conducted, and then repeated data points in the balanced samples were deleted using the Tomek link method. Tables 1 and 2, respectively, show the comparison of the original training set of binary and multi-classification plunger pump data and the training set with the SMOTE + Tomek Link integrated algorithm. It can be seen that after using the integrated algorithm to balance the training set among the various categories, they have reached a relatively balanced state.

**Table 2.** A comparison of the imbalanced dataset with the binary classification of the dataset processed based on the SMOTE + Tomek Link integration algorithm.

| Datasets | Normal Working Condition | Abnormal Condition | Scale |
|---|---|---|---|
| Raw datasets | 4600 | 1339 | 3.4:1 |
| SMOTE + Tomek Link post-equilibrium datasets | 4531 | 4400 | 1.03:1 |

5.3.2. Evaluation Indicators

In this experiment, the precision $P$, recall $R$, $F1$-score, and confusion matrix were selected as evaluation indexes to evaluate the fault diagnosis model of the CNN-BiGRU-Attention plunger pump with dual-channel feature fusion. The following is a detailed introduction to the evaluation indicators.

Samples can first be divided according to the combination of their true class and the model-predicted class, respectively: a false negative ($FN$) is a positive sample that is incorrectly determined to be a negative sample. A false positive ($FP$) occurs when a negative sample is incorrectly interpreted as a positive sample. A true negative ($TN$), that is, a truly negative sample, is accurately identified as a true negative. A positive sample is considered a true positive ($TP$) and is a true positive sample.

(1) Precision

The precision rate measures the proportion of successfully predicted operating conditions to all operating cases. It can be expressed by Equation (7).

$$P = \frac{TP}{TP + FP} \tag{7}$$

(2) Recall

The recall rate, also known as detection rate, is the ratio of all successfully detected instances to all true instances. It can be expressed by Equation (8).

$$R = \frac{TP}{TP + FN} \tag{8}$$

(3) $F1$-score

In general, there is a contradictory relationship between the accuracy rate and the recall rate: when the accuracy rate is high, the recall rate is often low, and when the recall rate is high, the accuracy rate is often low. To balance the precision and recall, you can use the $F1$-score indicator, which is the weighted harmonic average of the precision and recall. The $F1$-score determines the accuracy and robustness of a classifier and is often used to evaluate the performance of binary classification models. It can be represented by Equation (9).

$$F1 = \frac{(\alpha^2 + 1)P * R}{\alpha^2 P + R} \tag{9}$$

where the purpose of the $F1$ is to merge the scores of the precision and recall into one score, and in the process of merging, the weight of the recall is based on precision $\alpha$ times.

When the parameter $\alpha$ = 1, this is the most common $F1$-score, which is a way to evaluate the performance of a system by looking at its recall rate and accuracy. It can be represented by Equation (10).

$$F1 = \frac{2 * P * R}{P + R} = \frac{2TP}{N + TP - TN} \tag{10}$$

where $N$ is total number of examples.

(4) Confusion matrix

In order to better describe the classification results of the proposed model, the confusion matrix is usually presented in four forms: $FN$, $FP$, $TN$, and $TP$. If there are $N$ classes, the confusion matrix is an $N \times N$ matrix, with the left axis representing the true class and the upper axis representing the predicted classification corresponding to that true class. Each element $Q$ of the matrix represents the number of elements in class $A$ that are actually classified as class $B$. The confusion matrix is shown in Table 3.

**Table 3.** The confusion matrix.

| Real Situation | Forecast Result ($Q$) | |
|---|---|---|
| | **Positive Example** | **Counter Example** |
| Positive example | TP | FN |
| Counter example | FP | TN |

5.3.3. Comparative Experiment of Different Mixed Sampling Methods

The accuracy of the SMOTE + ENN algorithm and of the SMOTE + Tomek Link algorithm after classification on an unbalanced dataset are compared, as shown in Table 4.

**Table 4.** An accuracy table of different sampling algorithms on dual-channel feature fusion model.

| Method | SMOTE + ENN | SMOTE + Tomek Link |
|---|---|---|
| Accuracy rate | 94.82% | 96.93% |

*5.4. Comparative Analysis of Different Datasets in Different Models*

The experimental results are mainly divided into three groups. The first group presents the different evaluation metrics of imbalanced and balanced datasets on the CBA model. The second group compares the classification performance of the imbalanced dataset and the balanced dataset after using the SMOTE + Tomek Link mixed sampling method on the model proposed in this paper. The third group analyzes the multi-classification results of the balanced dataset on the model proposed in reference [13] and the model proposed in this paper. All the charts in the experimental results, based on the attention mechanism of the CNN-LSTM plunger pump fault diagnosis model, are uniformly represented by CLA (CNN-LSTM-Attention). The fault diagnosis model for plunger pumps based on dual-channel feature fusion is uniformly represented by CBA (CNN-BiGRU-Attention). The plunger pump fault diagnosis model based on SMOTE + Tomek Link mixed sampling and dual-channel feature fusion is uniformly represented by ST-CBA (SMOTE + Tomek Link).

5.4.1. Different Evaluation Index Results of an Unbalanced Dataset and a Balanced Dataset on the CBA Model

Below, different evaluation indexes of the CBA model and the ST-CBA model on multi-classification datasets are evaluated by referencing their recall rates, precision rates, and $F1$-scores.

Tables 5 and 6 compare the experimental results of CBA and ST-CBA on multi-classification datasets, and it can be seen that ST-CBA has improved the evaluation indexes for each working condition. Below is a bar chart for a more intuitive comparison.

**Table 5.** An evaluation of the effectiveness of CBA on multi-categorical datasets.

| Working Condition | Recall | Precision | *F*1-Score |
|---|---|---|---|
| Normal | 0.992 | 0.975 | 0.984 |
| Plunger ball head wear | 0.695 | 0.868 | 0.772 |
| Triangular hole plugging | 0.752 | 0.874 | 0.809 |
| Valve plate wear | 0.965 | 0.932 | 0.945 |
| Cylinder block wear | 0.869 | 0.915 | 0.891 |
| Bowl wear | 0.941 | 0.918 | 0.929 |
| Plunger wall pitting | 0.854 | 0.911 | 0.882 |

**Table 6.** An evaluation of the effect of ST-CBA on multi-classification datasets.

| Working Condition | Recall | Precision | *F*1-Score |
|---|---|---|---|
| Normal | 0.996 | 0.995 | 0.996 |
| Plunger ball head wear | 0.879 | 0.961 | 0.918 |
| Triangular hole plugging | 0.965 | 0.943 | 0.953 |
| Valve plate wear | 0.983 | 0.966 | 0.974 |
| Cylinder block wear | 0.957 | 0.957 | 0.957 |
| Bowl wear | 0.975 | 0.951 | 0.963 |
| Plunger wall pitting | 0.957 | 0.978 | 0.967 |

The comparison of experimental results between CBA and ST-CBA on multi-classification datasets is shown in Figures 9–11. It can be seen that the classification performance of the CBA model is not ideal when using the original data as training data. After expanding the minority class samples through the SMOTE + Tomek Link algorithm, the different indicator values of the model have significantly increased, showing good accuracy. The specific indicator improvement rate is shown in Table 7.

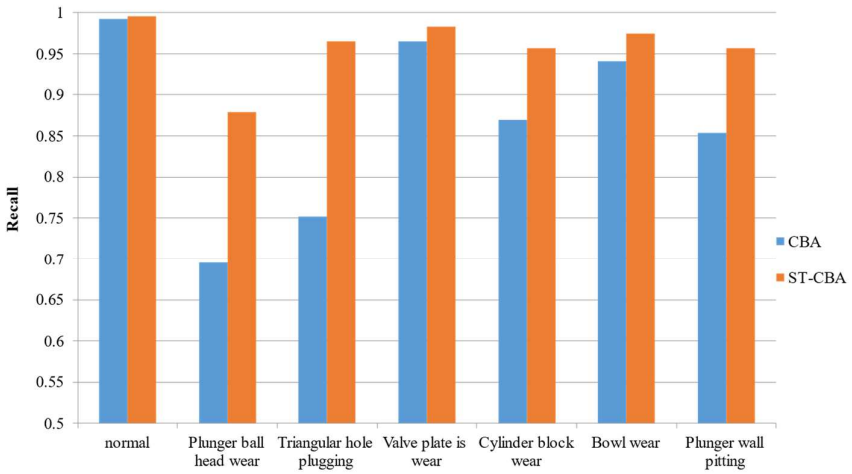


**Figure 9.** A comparison of multiple classification recalls between the CBA and ST-CBA models.

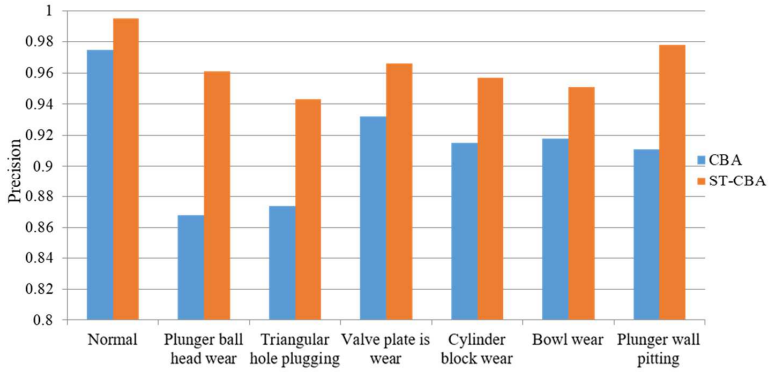


**Figure 10.** A comparison of multiple classification precision values between the CBA and ST-CBA models.

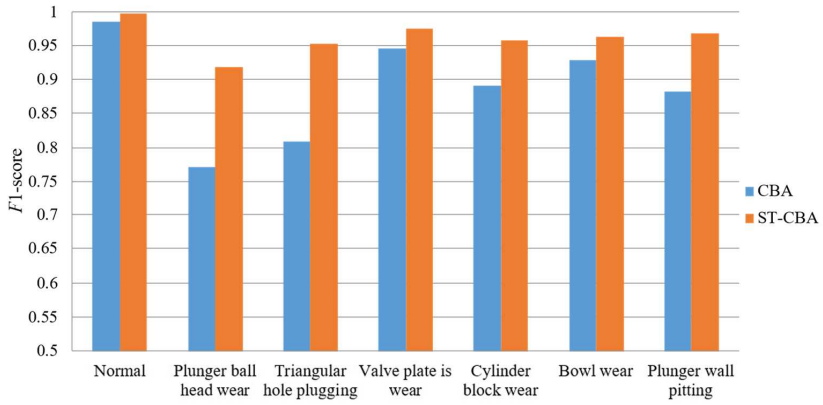


**Figure 11.** A comparison of the CBA and ST-CBA models' multi-classification *F*1-scores.

**Table 7.** Average improvement rate of evaluation indicators.

| Evaluation Indicators | Recall | Precision | *F*1-Score |
|---|---|---|---|
| Average improvement rate | 9.2% | 5.1% | 7.3% |

5.4.2. Classification Performance of Imbalanced and Balanced Datasets on CBA Models

As shown in Figure 12, the method proposed in this paper shows good classification results in the balanced dataset based on the SMOTE + Tomek Link mixed sampling method, with a classification accuracy of 95.86% and an accuracy of 86.43% in the imbalanced dataset. The results show that using the SMOTE + Tomek Link method to construct generated data similar to the original data can significantly improve the accuracy of fault diagnosis.

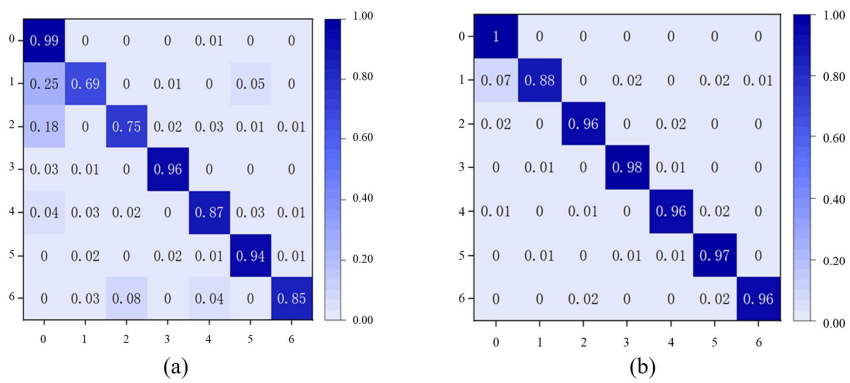


**Figure 12.** The classification performance of the CBA model. (**a**) The imbalanced dataset, (**b**) the balanced dataset.

5.4.3. The Classification Performance of Balanced Datasets on Different Models

Figure 13a shows the diagnostic results of the method proposed in reference [13] under balanced data, with an average accuracy of 91.68% for each category. Figure 13b shows the diagnostic results of the method proposed in this paper, with an accuracy of 95.86%. Figure 13c shows the diagnostic results using only the GLFC method proposed in this paper, with an average accuracy of 89.29%. Figure 13d shows the diagnostic results using the BiGRU method, with an average accuracy of 88.43%. Figure 13e shows the diagnostic results using the CLA method, with an average accuracy of 81.00%. Figure 13f shows the diagnostic results using the CNN method, where the average accuracy is 82.14%. For details, please refer to Table 8. This indicates that the diagnostic performance of the ST-CBA model proposed in this article is superior to that of other methods. The method proposed in this article can be used for the fault diagnosis of plunger pumps, with high diagnostic efficiency and accuracy.

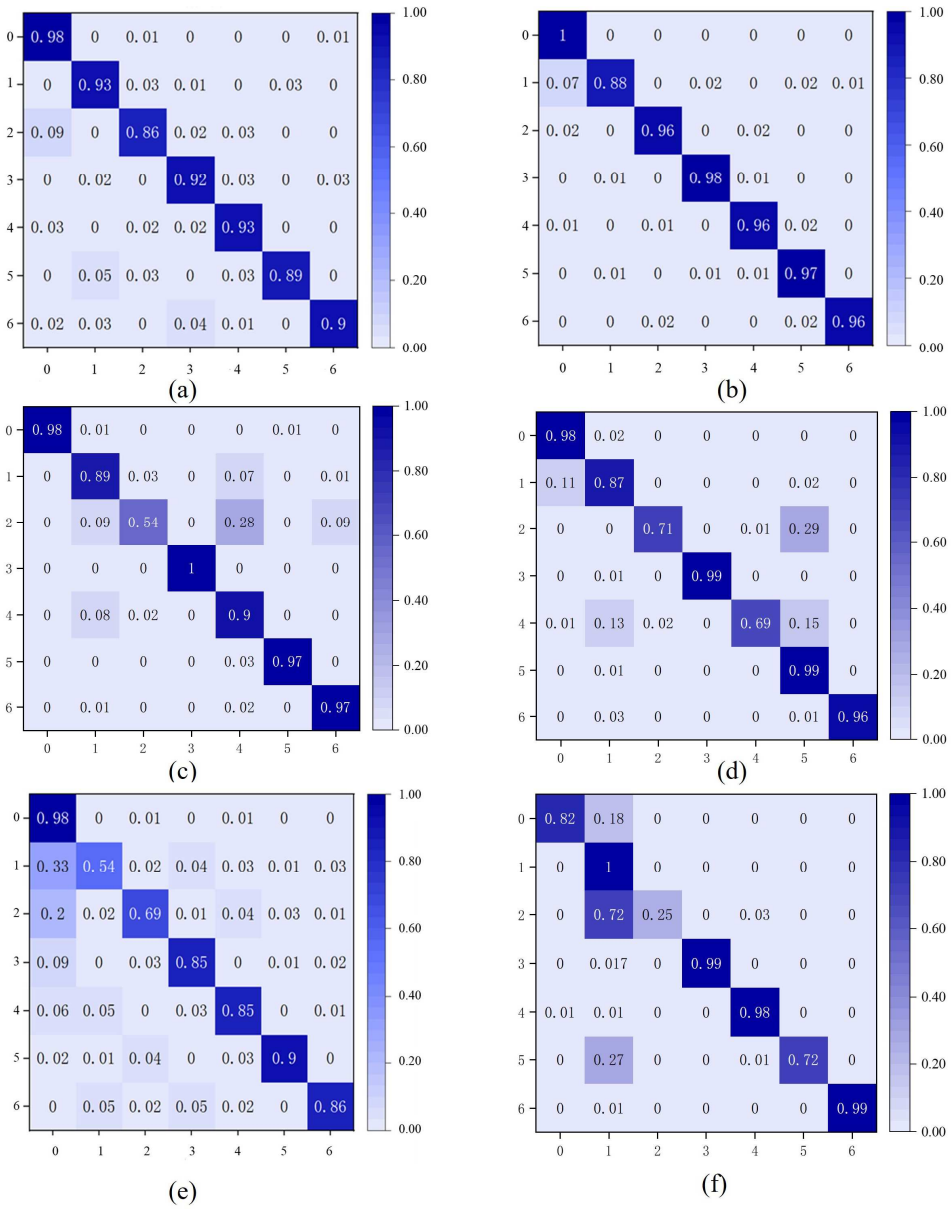


**Figure 13.** Test results after equalizing data based on the SMOTE-Tomek Link. (**a**) Reference [13], (**b**) the ST-CBA, (**c**) GLFC, (**d**) BiGRU, (**e**) CLA, (**f**) and CNN methods.

**Table 8.** Balancing the classification accuracy of datasets on different models.

| Model Methods | Reference [13] | ST-CBA | GLFC | BiGRU | CLA | CNN |
|---|---|---|---|---|---|---|
| Accuracy | 91.68% | 95.86% | 89.29% | 88.43% | 81.00% | 82.14% |

## 6. Conclusions

This article proposes a fault diagnosis method based on the SMOTE + Tomek Link and dual-channel feature fusion to address the issue of imbalanced fault samples in the plunger pump of switch machines. Firstly, the SMOTE + Tomek Link algorithm was adopted to solve the problem of imbalanced plunger pump samples. This algorithm combines the advantages of oversampling and undersampling, compensates for the shortcomings of a single mechanism, and not only helps to alleviate the problem of data imbalance, but also achieves more significant feature preservation, thereby improving learning efficiency. Then, the ST-CBA method proposed in this article was experimentally validated by comparing its performance on imbalanced and balanced datasets. The evaluation indicators of the model in this article have significantly improved on the balanced dataset, with a recall rate increase of 9.2%, a precision rate increase of 5.1%, and an $F1$-score increase of 7.3%. This indicates the effectiveness of the ST-CBA method proposed in this article in the diagnosis of plunger pump faults. Finally, by comparing the proposed method with similar excellent methods, the experiments show that the ST-CBA method proposed in this paper is superior to other fault diagnosis methods.

**Author Contributions:** Conceptualization, X.Y. and X.X.; methodology, X.Y. and X.X.; software, X.X.; validation, J.M.; formal analysis, Y.W.; investigation, S.L.; resources, L.J.; data curation, X.B.; writing—original draft preparation, X.X.; writing—review and editing, X.Y.; visualization, S.L.; supervision, J.H.; project administration, J.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Research Project Supported by the Shanxi Scholarship Council of China (2022-141) and the Fundamental Research Program of Shanxi Province (202203021211096).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Jeyalakshmi, K. Weighted synthetic minority over-sampling technique (WSMOTE) algorithm and ensemble classifier for hepatocellular carcinoma (HCC) in liver disease system. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **2021**, *12*, 7473–7487.
2. Lin, W.C.; Tsai, C.F.; Hu, Y.H.; Jhang, J.S. Clustering-based undersampling in class-imbalanced data. *Inf. Sci.* **2017**, *409*, 17–26. [CrossRef]
3. Duan, F.; Zhang, S.; Yan, Y.; Cai, Z. An oversampling method of unbalanced data for mechanical fault diagnosis based on MeanRadius-SMOTE. *Sensors* **2022**, *22*, 5166. [CrossRef] [PubMed]
4. Xu, Y.; Zhao, Y.; Ke, W.; He, Y.L.; Zhu, Q.X.; Zhang, Y.; Cheng, X. A multi-fault diagnosis method based on improved SMOTE for class-imbalanced data. *Can. J. Chem. Eng.* **2023**, *101*, 1986–2001. [CrossRef]
5. Rao, S.; Zou, G.; Yang, S.; Khan, S.A. Fault diagnosis of power transformers using ANN and SMOTE algorithm. *Int. J. Appl. Electromagn. Mech.* **2022**, *70*, 345–355. [CrossRef]
6. Yang, K.; Yu, Z.; Wen, X.; Cao, W.; Chen, C.P.; Wong, H.S.; You, J. Hybrid classifier ensemble for imbalanced data. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 1387–1400. [CrossRef] [PubMed]
7. Bernardo, A.; Della Valle, E. Smote-ob: Combining Smote and Online Bagging for Continuous Rebalancing of Evolving Data Streams. In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December 2021.
8. Fernández, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [CrossRef]
9. Torgo, L.; Ribeiro, R.P.; Pfahringer, B.; Branco, P. Smote for regression. In *Progress in Artificial Intelligence: 16th Portuguese Conference on Artificial Intelligence, Angra do Heroísmo, Portugal, 9–12 September 2013*; Springer: Berlin/Heidelberg, Germany, 2013.

10. Blagus, R.; Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* **2013**, *14*, 106. [CrossRef] [PubMed]
11. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
12. Liu, S.; Huang, J.; Ma, J.; Luo, J. Class-incremental continual learning model for plunger pump faults based on weight space meta-representation. *Mech. Syst. Signal Process.* **2023**, *196*, 110309. [CrossRef]
13. Wei, X.; Chao, Q.; Tao, J. Fault diagnosis of high speed piston pump based on LSTM and CNN. *Acta Aeronaut. Astronaut. Sin.* **2021**, *42*, 435–445.
14. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27, Montreal, QC, Canada, 8–13 December 2014.
15. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In Proceedings of the Advances in Neural Information Processing Systems 29, Barcelona, Spain, 5–10 December 2016.
16. Ye, Y.; Wang, L.; Wu, Y.; Chen, Y.; Tian, Y.; Liu, Z.; Zhang, Z. Gan quality Index (GQI) by Gan-Induced Classifier. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
17. Levent, E. Bearing Fault Detection by One-Dimensional Convolutional Neural Networks. *Math. Probl. Eng.* **2017**, *2017*, 8617315.
18. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. *ImageNet Classification with Deep Convolutional Neural Networks*; NIPS, Curran Associates Inc.: New York, NY, USA, 2012; pp. 1–38.
19. Liu, J.; Yang, Y.; Lv, S.; Wang, J.; Chen, H. Attention-based BiGRU-CNN for Chinese question classification. *J. Ambient. Intell. Humaniz. Comput.* **2019**, 1–12. [CrossRef]
20. Bhagat, R.C.; Patil, S.S. Enhanced SMOTE Algorithm for Classification of Imbalanced Big-Data Using Random Forest. In Proceedings of the 2015 IEEE International Advance Computing Conference (IACC), Banglore, India, 12–13 June 2015.
21. Gu, Q.; Wang, X.M.; Wu, Z.; Ning, B.; Xin, C.S. An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification. *J. Digit. Inf. Manag.* **2016**, *14*, 92–103.
22. Muntasir Nishat, M.; Faisal, F.; Jahan Ratul, I.; Al-Monsur, A.; Ar-Rafi, A.M.; Nasrullah, S.M.; Reza, M.T.; Khan, M.R.H. A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset. *Sci. Program.* **2022**, *2022*, 3649406. [CrossRef]
23. Keller, A.; Pandey, A. SMOTE and ENN Based XGBoost Prediction Model for Parkinson's Disease Detection. In Proceedings of the 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 7–9 October 2021.