machines

MDPI

Article

Multi-Scale Recursive Semi-Supervised Deep Learning Fault Diagnosis Method with Attention Gate

Shanjie Tang, Chaoge Wang, Funa Zhou *, Xiong Hu and Tianzhen Wang

School of Logistic Engineering, Shanghai Maritime University, Shanghai 201306, China
* Correspondence: zhoufn@shmtu.edu.cn

Abstract: The efficiency of deep learning-based fault diagnosis methods for bearings is affected by the sample size of the labeled data, which might be insufficient in the engineering field. Self-training is a commonly used semi-supervised method, which is usually limited by the accuracy of features for unlabeled data screening. It is significant to design an efficient training mechanism to extract accurate features and a novel feature fusion mechanism to ensure that the fused feature is capable of screening. A novel training mechanism of multi-scale recursion (MRAE) is designed for Autoencoder in this article, which can be used for accurate feature extraction with a small amount of labeled data. An attention gate-based fusion mechanism was constructed to make full use of all useful features in the sense that it can incorporate distinguishing features on different scales. Utilizing large numbers of unlabeled data, the proposed multi-scale recursive semi-supervised deep learning fault diagnosis method with attention gate (MRAE-AG) can efficiently improve the fault diagnosis performance of DNNs trained by a small number of labeled data. A benchmark dataset from the Case Western Reserve University bearing data center was used to validate this novel method which shows that 7.76% accuracy improvement can be achieved in the case when only 10 labeled samples was available for supervised training of the DNN-based fault diagnosis model.

Keywords: deep learning; semi-supervised learning; multi-scale recursion; attention gate; fault diagnosis

Citation: Tang, S.; Wang, C.; Zhou, F.; Hu, X.; Wang, T. Multi-Scale Recursive Semi-Supervised Deep Learning Fault Diagnosis Method with Attention Gate. *Machines* **2023**, *11*, 153. https://doi.org/10.3390/machines11020153

Academic Editors: Xiang Li and Jie Liu

Received: 3 January 2023
Revised: 17 January 2023
Accepted: 18 January 2023
Published: 23 January 2023

1. Introduction

As a key component of electromechanical equipment, rolling bearings play a crucial role in electric motors. Fault diagnosis of bearings is important for healthy operation of the motor [1–3]. Data-driven fault diagnosis methods can mine data features without precise mathematical models and expert knowledge [4]. Deep learning is widely used in the field of fault diagnosis for its powerful capability in feature extraction [5–7]. However, it requires huge amounts of labeled data to train a well-performing model, which is always unavailable in the engineering field. Semi-supervised deep learning methods can use massive unlabeled data to optimize the supervised model trained by a small size of labeled data [8–10]. Existing semi-supervised deep learning methods can be mainly classified into deep generative methods, consistency regularization methods, graph-based methods, and pseudo label-based self-training methods [11–13]. The graph-based methods treat samples as nodes in a graph and use the graph structure to reflect the similarity between samples to complete the propagation of labels [14–16]. Chen et al. [14] constructed a graph based on the weighted sparse adjacency matrix through the k-nearest neighbor and Gaussian kernel weighting algorithm, then the labels were propagated to unlabeled data. Yi et al. [15] constructed a graph model for label propagation after removing redundant information from the data using feature screening based on mutual information. Wang et al. [16] transformed the diagnosis task into a hierarchical fault attribute representation task, which reduced the complexity of the classification problem in a semi-supervised model. These methods mentioned above used unlabeled data to improve the fault diagnosis accuracy

of the model. However, the graph-based, semi-supervised methods assumed that the connected nodes in the graph had the same labels and ignored the additional information of neighboring nodes, which greatly limited the model performance. Liang et al. [17] used wavelet transform to achieve time–frequency domain conversion of the vibration signal and fed it to adversarial learning model, which achieved the purpose of combining masses of unlabeled data with little labeled data for fault diagnosis. However, it was difficult to train a good generator and semi-supervised classifier at the same time. Laine et al. [18] introduced self-ensembling to achieve consistent prediction of unlabeled data through data augmentation and dropout. However, this method relied too much on regularization and data augmentation techniques.

Compared with the methods mentioned above, the semi-supervised methods based on pseudo-labeling for self-training are easier to train the model [19–22]. Lee et al. [19] utilized the assumption of low-density separation in semi-supervised learning to select labels with maximum probability for unlabeled data, which improved fault diagnosis accuracy. Yu et al. [20] performed the K-means method to obtain the clustering centers of labeled data and constructed a loss function to minimize the distribution discrepancy by Kullback–Leibler (KL) divergence between the features of unlabeled data and their corresponding clustering centers. Sohn et al. [21] combined consistency regularization and pseudo labels to predict label information for weakly enhanced unlabeled data. The obtained pseudo labels would be used to optimize the model as supervised information of strongly enhanced unlabeled data. Berthelot et al. [22] used random augmentation multiple times on the same unlabeled data to obtain multiple predicted labels, then mixed the labeled and unlabeled data to calculate the loss function separately. The above-mentioned self-training methods can enhance the robustness of the model, but it is limited by inaccurate features for unlabeled data screening.

Accurate features can better represent the original data and improve the accuracy of semi-supervised models [23–25]. Tao et al. [23] used density peak clustering to generate pseudo labels for unlabeled data, and then regularized the inter-class scatter and intra-class scatter based on the pseudo labels to improve the discriminative performance of the extracted feature. Zhang et al. [24] used a variational Autoencoder (VAE) to mine features in the data and then used the features of the labeled data to train the classifier and that of unlabeled data to reconstruct the original data, which improved the fault diagnosis accuracy of semi-supervised model. To fully mine the unlabeled data features, Tang et al. [25] built an unsupervised network to extract the unlabeled data feature and improved fault diagnostic accuracy by jointly fine-tuning with a supervised network. These methods mentioned above improved the accuracy of fault diagnosis by making full use of the features contained in the data, but did not consider the problem of layer-by-layer information loss and error accumulation in deep learning. On the other hand, the accuracy of the pseudo labels used for model optimization also affects the final performance of the semi-supervised model [26–28]. To reduce the error rate of pseudo label screening, Long et al. [26] proposed a stepwise utilization mechanism that assigned nearest labels to each unlabeled data based on distance. Hu et al. [27] designed an unsupervised loss function based on the statistical distance between similar unlabeled data, then combined it with the method proposed in Ref. [22] in order to improve the classification performance. Liu et al. [28] alternately optimized a teacher model and pseudo labels, then jointly trained a classifier using labeled data and unlabeled data in the student model training phase. The above-mentioned methods improved the accuracy of semi-supervised fault diagnosis by selectively utilizing pseudo labels of unlabeled data to reduce the influence of erroneous pseudo labels in the fault diagnosis model. But only single-scale features were used in the screening process, which could not guarantee the adequacy of the information utilized and inevitably caused errors in pseudo label screening.

To solve the problem of inadequate information utilization by single-scale features, some scholars have conducted research on multi-scale feature fusion methods [29–32]. Jiang et al. [29] used a deep belief network (DBN) for feature extraction of vibration

signals, and then used locality preserving projection (LPP) to fuse different scale features, improved the accuracy of fault diagnosis. Shao et al. [30] designed a new deep Autoencoder (AE) model to extract different scale features of the data, and then used LLP for feature fusion, which finally improved the robustness and effectiveness of the diagnostic model. Zhuang et al. [31] stitched the local features extracted from the convolutional layer and the down sampled features from the pooling layer into a multi-scale feature matrix, which was connected to the fully connected layer to broaden and deepen the neural network and obtain more robust and accurate diagnosis results. Zhou et al. [32] used deep neural networks (DNNs) to extract different scale features and then designed a feature fusion network to fuse the multi-scale features to improve the sufficiency of information utilization and fault diagnosis accuracy. However, the methods mentioned above fused the features at different scales directly without considering the degradation of model performance caused by redundant features. The attention mechanism can guide the model to assign more attention to important features, thus reduce the information redundancy in the feature fusion process. Yao et al. [33] used an attention mechanism to design a feature fusion network, which reduced feature redundancy in the fusion process and improved fault diagnosis accuracy. Xie et al. [34] used an attention mechanism to design an adaptive feature fusion layer combined with graph-based, semi-supervised learning to achieve cross-scale information fusion in different neighborhoods. The methods mentioned above fused features through the attention mechanism, which can guide the model to focus on the key information, thus improved the fault diagnosis accuracy. It is significant to use this idea of information screening to improve the adequacy of feature utilization in the unlabeled data screening process of self-training semi-supervised methods.

To overcome the problems mentioned above in traditional self-training methods, a multi-scale recursive semi-supervised deep learning fault diagnosis method with an attention gate is proposed to ensure more accurate feature extraction and more adequate multi-scale feature utilization. A novel training mechanism is designed to ensure the accuracy of feature extraction by performing multi-scale reconstruction of deep features during pre-training. An attention gate designed to fuse features at different scales enables fuller use of the information contained in the data, which helps the model to screen unlabeled data with high quality pseudo labels. The method proposed is able to improve the fault diagnostic accuracy of the model using a large amount of unlabeled data when the amount of labeled data is small. The innovations of the research are as follows:

1. A multi-scale, recursive, semi-supervised, deep learning fault diagnosis method with an attention gate is proposed to enhance the performance of a semi-supervised model by improving the accuracy of feature extraction and the adequacy of feature utilization.
2. The feature extraction network is trained in a scale recursive manner by designing loss functions using multi-scale features to obtain deep features with good representability of the original data. The adequacy of information utilization can be secured by constructing an attention gate to fuse multi-scale features, thus the features used to screen unlabeled data are more comprehensive.
3. The proposed method can still achieve a satisfying accuracy of fault diagnosis even when there is a very small number of labeled training samples available, which is common in the field of practical engineering diagnoses.

The remainder of this article is structured as follows. Section 2 introduces the theory related to the DNN and attention mechanism. Section 3 details the specific improvement measures of the proposed method. Section 4 verifies the effectiveness and superiority of the designed method using the bearing dataset. Finally, the conclusions are summarized in Section 5.

2. Related Theories

2.1. Deep Neural Network Based on Stacked Autoencoder

The DNN consists of multiple Autoencoders stacked on top of each other. The structure of an AE is shown in Figure 1. The DNN is able to extract the deeper features of the original

data, and its training process includes unsupervised feature extraction and supervised global fine-tuning.

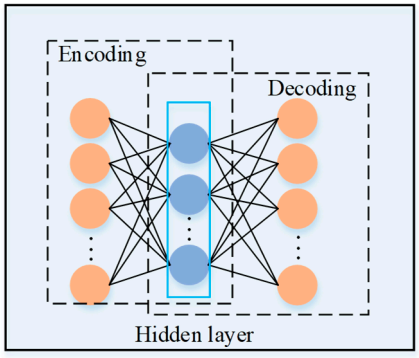

Figure 1. Schematic of AE.

2.2. Attention Mechanism

An attention mechanism is a special structure that can be embedded into a deep learning model to weight the importance of different parts of features, allowing the model to focus on certain parts while ignoring others. An attention mechanism generates a weighted sum of the v based on the similarity between the q and the k. The calculation process is shown in Figure 2.

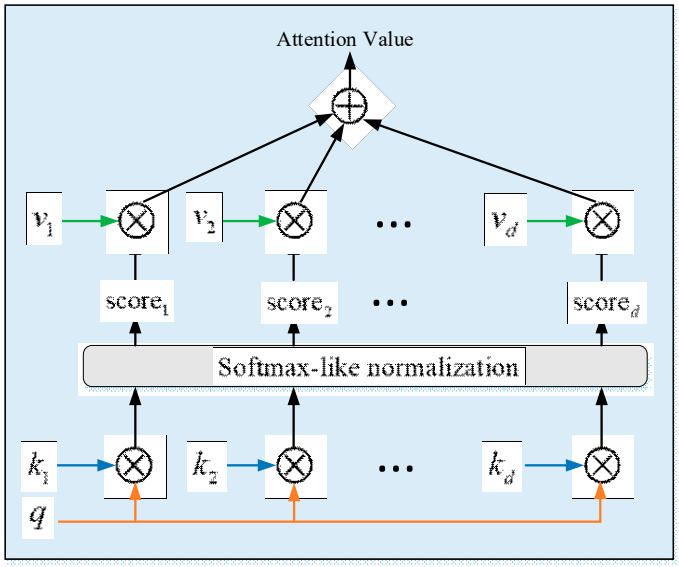

Figure 2. Schematic of an attention mechanism.

3. Multi-Scale Recursive Semi-Supervised Deep Learning Fault Diagnosis Method with Attention Gate

The traditional DNN does not consider the ability of features to reconstruction the original data in the pre-training stage of the network, which makes it difficult to ensure the accuracy of feature extraction. In addition, features at the most abstract scale are difficult to provide comprehensive characteristics of the data because of information loss in the process of layer-by-layer feature extraction, which leads to insufficient information utilization when only single-scale features are used to screen unlabeled data. Therefore, MRAE-AG is designed to ensure the accurate extraction of features and the full use of information. The method includes multi-scale recursive feature reconstruction oriented to accurate feature extraction, an attention gate fusion mechanism oriented to full information utilization, and an unlabeled data screening strategy based on fused features. The network structure is shown in Figure 3.

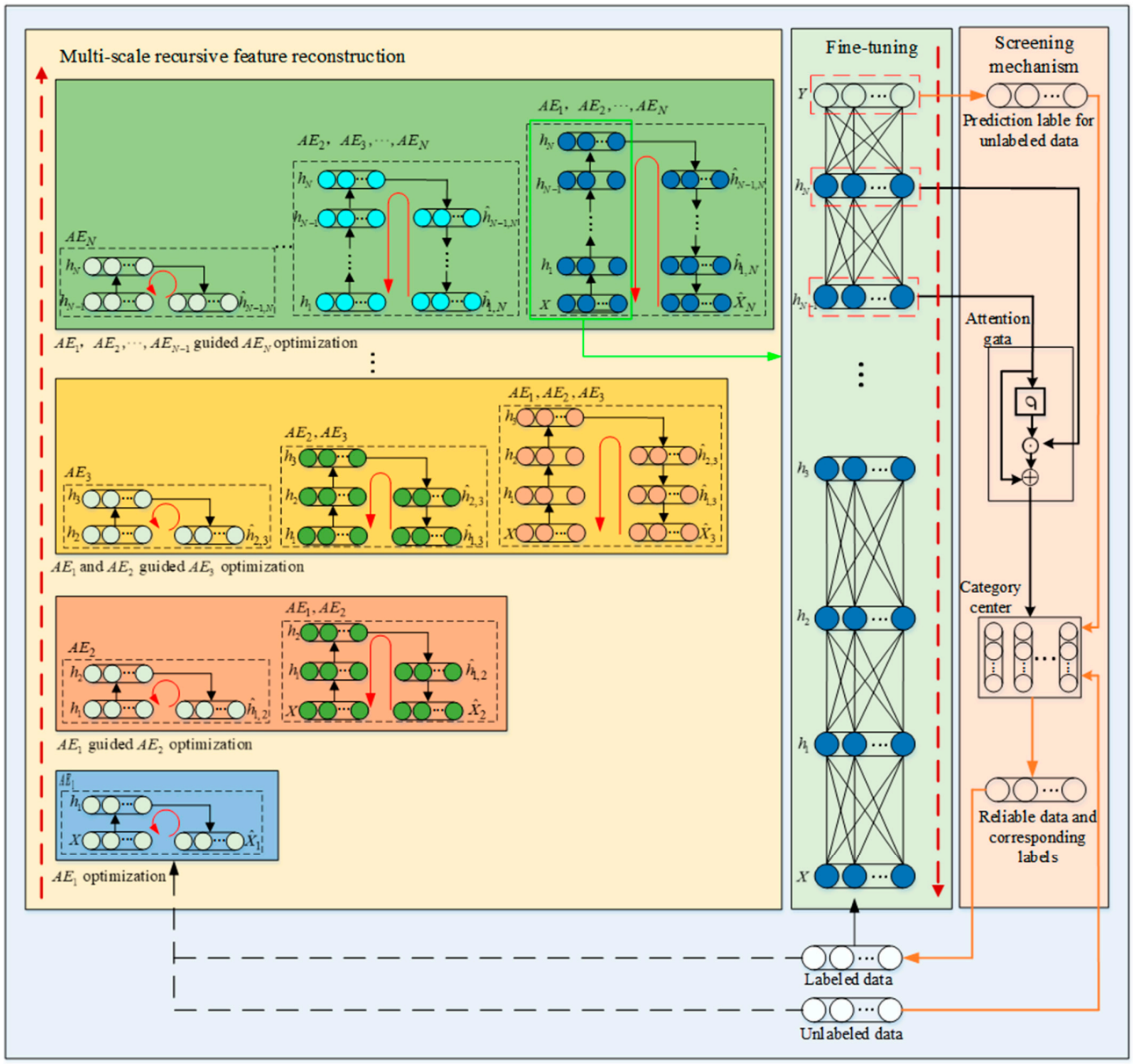

Figure 3. Schematic of the MRAE-AG.

3.1. Multi-Scale Recursive Feature Reconstruction Oriented to Accurate Feature Extraction

Accurate features play a crucial role in the final fault diagnosis performance of the semi-supervised model. Traditional DNNs obtain more abstract features by means of layer-by-layer extraction of features, but the decoder of each AE aims to reconstruct only the input of the current layer rather than the previous layers, which will result in inaccurate feature representation. Therefore, a learning mechanism is proposed in this section to make the AEs more powerful in accurate feature representation by designing a new loss function taking both the reconstruction error of the current layer and the previous layers into account.

The labeled and unlabeled data constitute the dataset used for model pre-training, which is shown in Equation (1).

$$X = [x_l, x_u] \tag{1}$$

where x_l represents labeled data and x_u represents unlabeled data.

The 1*st* AE is trained by using the loss function corresponding to the error between the original data X and the output of decoder. The i AE ($i \geq 2$) is trained by using the following steps.

3.1.1. Preliminary Training of the Current Layer

Using the traditional loss function shown in Equation (2) to train the AE on the i layer:

$$Loss_{AE_i} = \frac{1}{M}\sum_{m=1}^{M}\left(h_{i-1} - \hat{h}_{i-1,i}\right)^2 \tag{2}$$

where M is the number of samples used for pre-training, h_{i-1} is the feature of the $(i-1)$ layer, and $\hat{h}_{i-1,i}$ is the reconstruction result of h_{i-1} by using the feature of the i layer.

3.1.2. Joint Update of the Current Layer and the Previous Layer

In order to reduce information loss from the previous layer to the current layer, the current AE_i and the previous AE_{i-1} are jointly trained in Equations (3)–(5).

$$\begin{cases} h_i = \sigma(h_{i-1}W_{i,1} + b_{i,1}) \\ \hat{h}_{i-1,i} = \sigma(h_i W_{i,2} + b_{i,2}) \end{cases} \tag{3}$$

$$\begin{cases} h_i = \sigma(\sigma(h_{i-2}W_{i-1,1} + b_{i-1,1})W_{i,1} + b_{i,1}) \\ \hat{h}_{i-2,i} = \sigma(\sigma(h_i W_{i,2} + b_{i,2})W_{i-1,2} + b_{i-1,2}) \end{cases} \tag{4}$$

$$Loss_{AE_i,AE_{i-1}} = \frac{1}{M}\sum_{m=1}^{M}\left(h_{i-1} - \hat{h}_{i-1,i}\right)^2 + \frac{1}{M}\sum_{m=1}^{M}\left(h_{i-2} - \hat{h}_{i-2,i}\right)^2 \tag{5}$$

where Equation (5) is the designed loss function for joint training of AE_i and AE_{i-1}. By minimizing $Loss_{AE_i,\ AE_{i-1}}$, the extracted feature on the current layer can reconstruct the input of the previous layer as well as the current layer. In such a way, a more accurate feature extraction process can be secured and information loss between adjacent layers can be effectively reduced. $W_{i,1}$ and $b_{i,1}$ are the weight and bias matrix of the encoder of the i AE, respectively. $W_{i,2}$ and $b_{i,2}$ are the weight and bias matrix of the decoder of the i AE, respectively.

3.1.3. Recursively Update of All Previous Layers

To ensure that h_i is a good representation of the original data, the current AE_i and previous $AE_{i-1}, \ldots, AE_1$ are recursively trained in Equations (6)–(9).

$$\begin{cases} h_i = \sigma(h_{i-1}W_{i,1} + b_{i,1}) \\ \hat{h}_{i-1,i} = \sigma(h_i W_{i,2} + b_{i,2}) \end{cases} \tag{6}$$

$$\begin{cases} h_i = \sigma(\sigma(h_{i-2}W_{i-1,1} + b_{i-1,1})W_{i,1} + b_{i,1}) \\ \hat{h}_{i-2,i} = \sigma(\sigma(h_i W_{i,2} + b_{i,2})W_{i-1,2} + b_{i-1,2}) \end{cases} \tag{7}$$

$$\begin{cases} h_i = \sigma(\sigma(\sigma(\cdots\sigma(XW_{1,1} + b_{1,1}))W_{i-1,1} + b_{i-1,1})W_{i,1} + b_{i,1}) \\ \hat{X}_i = \sigma(\sigma(\sigma(\cdots\sigma(h_i W_{i,2} + b_{i,2}))W_{2,2} + b_{2,2})W_{1,2} + b_{1,2}) \end{cases} \tag{8}$$

$$Loss_{AE_i-AE_1} = \frac{1}{M}\sum_{m=1}^{M}\left(h_{i-1} - \hat{h}_{i-1,i}\right)^2 + \frac{1}{M}\sum_{m=1}^{M}\left(h_{i-2} - \hat{h}_{i-2,i}\right)^2 + \cdots + \frac{1}{M}\sum_{m=1}^{M}\left(X - \hat{X}_i\right)^2 \tag{9}$$

where Equation (9) is the designed loss function $Loss_{AE_i-AE_1}$ based on the reconstruction results of the features of each layer by h_i. By minimizing $Loss_{AE_i\ -\ AE_1}$, we can recursively optimize $AE_{i-1}, \cdots, AE_1$ according to the ability of h_i to reconstruct the features of each layer, and then comprehensively ensure the accuracy of deep features. $\hat{X}_i$ is the reconstruction result of X by using the feature of the ith layer.

3.2. Attention Gate Fusion Mechanism Oriented to Full Information Utilization

Due to information loss, it is difficult to ensure the adequacy of information utilization by using only the most abstract features for unlabeled data screening, which will inevitably lead to unlabeled data screening errors. Traditional feature fusion methods often simply splice features of different scales directly, which cannot adaptively focus on key features. In this section, an information fusion mechanism based on an attention gate is designed to enable useful features at different scales to participate in the model building for unlabeled data screening.

The features of the last two layers in the DNN are the final feature representations of the model and can better describe the data, thus the last two scale features extracted from the network are fused in this article. For a DNN with I hidden layers, the features used for fusion are h_I, h_{I-1}. The specific implementation steps for feature fusion are as follows.

3.2.1. Multi-Scale Feature Fusion Based on Attention Gate

The fusion of features at different scales of the DNN can ensure the comprehensiveness of the information, which can be shown in Equation (10).

$$Att = (1 - G_I) \odot h_{I-1} + G_I \odot h_I \tag{10}$$

where Att is a more comprehensive fused feature of the data, $\odot$ is the Hadamard product of the matrix, and G_I is the fusion strategy of the attention gate for features at different scales:

$$G_I = \sigma(h_{I-1} W_{gate} + b_{gate}) \tag{11}$$

where W_{gate} and b_{gate} are the weight and bias matrix of the attention gate, respectively.

3.2.2. Training of the Attention Gate for Multi-Scale Feature Fusion

The goal of feature fusion based on an attention gate is to obtain key features that help to perform unlabeled sample screening by comprehensive different scale features. A loss function is designed using the labeled data:

$$Loss_{gate} = \sum_{c=1}^{C} d_c + \frac{1}{\sum_{c_1=1}^{C} \sum_{\substack{c_2=1 \\ c_2 \neq c_1}}^{C} d_{c_1,c_2}} \tag{12}$$

where C is the total number of classes, d_c is the intra-class distance of the fused features of the c class, and d_{c_1,c_2} is the inter-class distance of the fused features of the c_1 class and c_2 class. The Euclidean distance is used in the algorithm design, which can be calculated using Equation (13).

$$d_{(a,b)} = \sqrt{\sum_{k=1}^{n} (a_k - b_k)^2} \tag{13}$$

By minimizing $Loss_{gate}$, the attention gate can be well trained to guide the feature fusion strategy to focus more attention on the important features and less attention on other features with less contribution.

3.3. Unlabeled Data Screening Strategy Based on Fused Feature

In the semi-supervised methods based on self-training, the means of screening pseudo label affects the final diagnosis accuracy of the semi-supervised method. Once the fused feature on a multi-scale is obtained, unlabeled samples can be well screened as long as the screening criterion with a higher confidence is used. The detailed steps are as follows.

3.3.1. Determination of the Center of Each Class According to Labeled Data

The fused feature obtained by the attention gate can be used to calculate the center of each class with labeled data:

$$\mathrm{Cen}_c = \frac{1}{M_c}\sum_{m=1}^{M_c} Att_{c,m} \tag{14}$$

where Cen_c is the center of the labeled data in the c class, $Att_{c,m}$ is the feature for the mth labeled sample corresponding to the c class, and M_c is the total number of the labeled data in the c class.

3.3.2. Attention Gate-Based Fusion of Multi-Scale Features for Unlabeled Data

The multi-scale features of the unlabeled data are fused by using an attention gate, which is shown in Equation (15).

$$\begin{cases} G_{I,u} = \sigma(h_{I-1,u}W_{gate} + b_{gate}) \\ Att_u = (1 - G_{I,u}) \odot h_{I-1,u} + G_{I,u} \odot h_{I,u} \end{cases} \tag{15}$$

3.3.3. Criteria Designing for Screening Unlabeled Data

To improve the correct rate of screened pseudo labels, screening criteria are established based on the inter-class distance and intra-class distance. The screened unlabeled data and their corresponding pseudo labels will be applied to optimize the semi-supervised model.

For a certain unlabeled data, it belongs to the p class once Equation (16) is satisfied.

$$d_{(Cen_p, Att_u)} < \min\left\{ d_{(Cen_c, Att_u)} \middle| c \in [1, C], c \neq p \right\} \tag{16}$$

The labeled dataset is expanded by splicing the labeled data and the screened unlabeled data, which can be shown in Equation (17).

$$\begin{cases} x_{l,new} = [x_l, x_{u,reliable}] \\ y_{l,new} = [y_l, y_{u,reliable}] \end{cases} \tag{17}$$

where $x_{u,reliable}$ is the screened unlabeled sample by using the designing criterion and $y_{u,reliable}$ is the corresponding pseudo labels.

The screening criterion proposed in this section is based on the similarity between the fused features rather than the affiliation of the fault diagnosis results, which can effectively avoid the problem of unlabeled data screening errors caused by insufficient information utilization. The flowchart of MRAE-AG is shown in Figure 4.

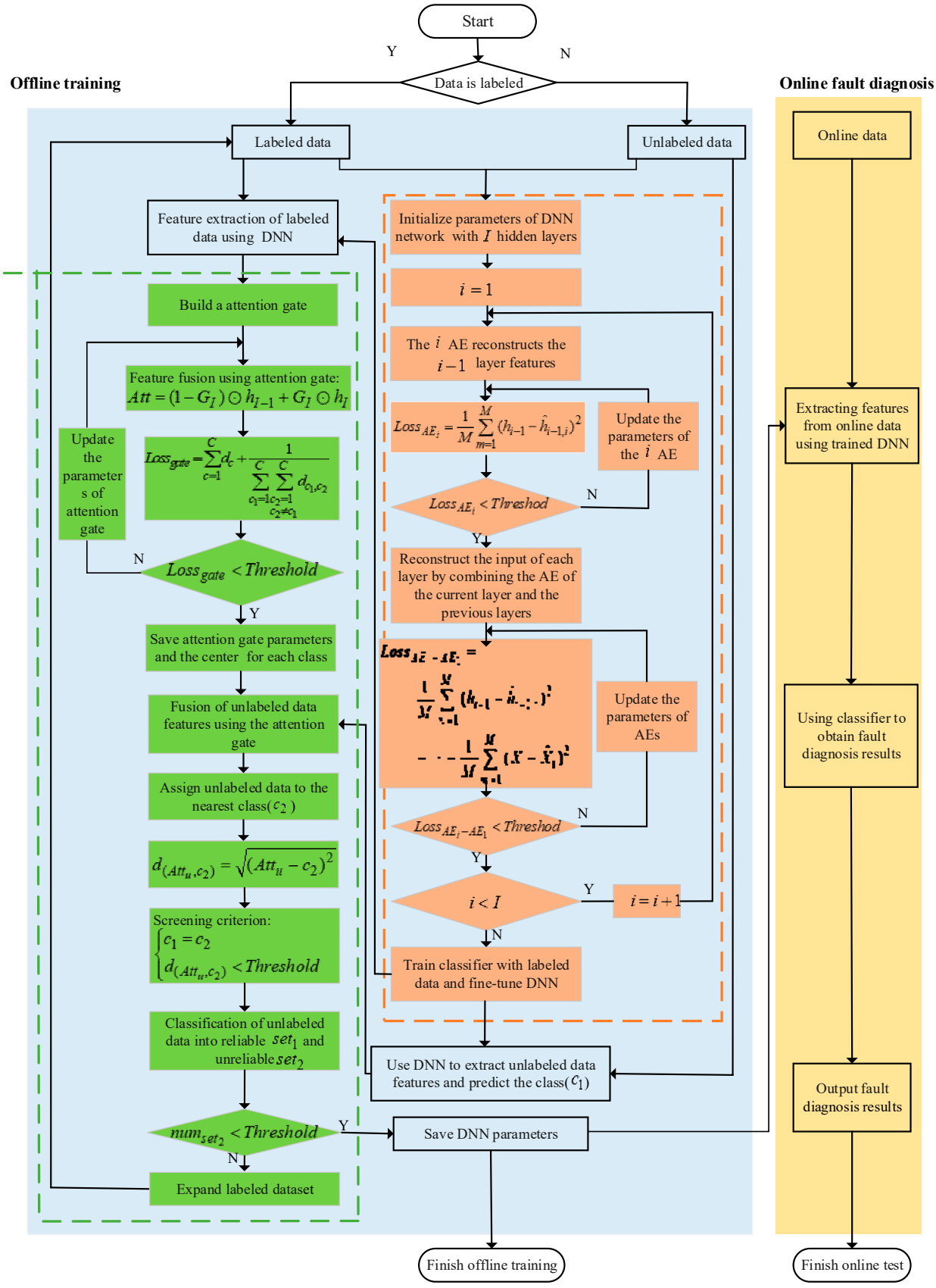

Figure 4. Flowchart of MRAE-AG.

4. Experiment Analysis

Bearings are a crucial part of electric motors. In this section, experiments are conducted using the Case Western Reserve University rolling bearing dataset to validate the effectiveness of the proposed method. The presented model is programmed in the Tensorflow framework.

4.1. Dataset Description and Experiment Design

The dataset provided by the Case Western Reserve University bearing data center website is widely used in the field of fault diagnosis [35]. The experimental platform is shown in Figure 5, which is mainly composed a 2HP motor (left), torque transducer (center), and dynamometer (right).

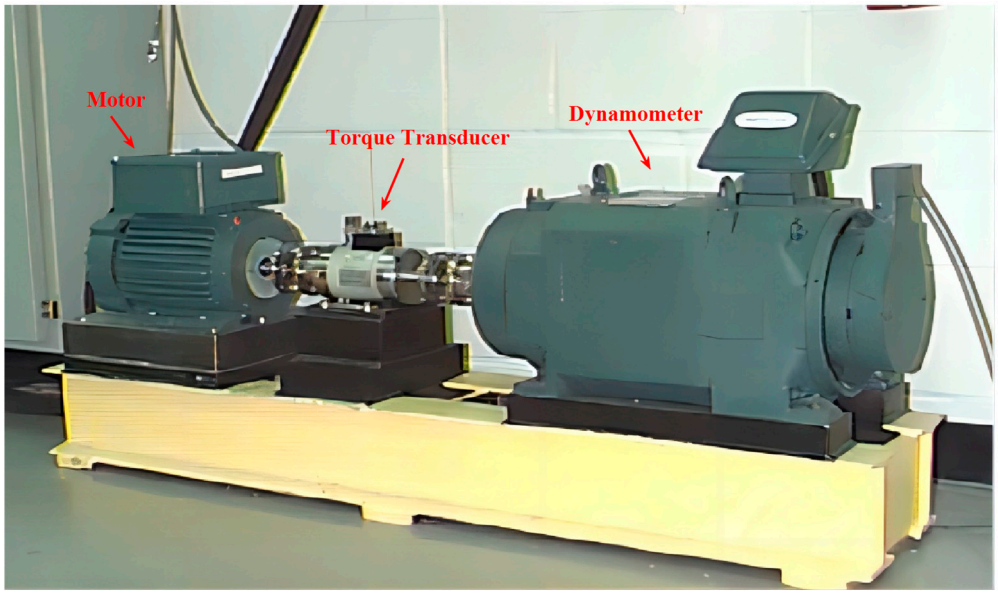

Figure 5. Case Western Reserve University bearing dataset collection platform.

The data used for experiment verification in this section comes from the vibration data collected by the accelerometer at the drive end of the Case Western Reserve University bearing dataset when the motor load is 0 hp and the speed is 1797 rpm, with a sampling frequency of 48 KHz. The three faults were measured for the inner race, ball, and outer race of the rolling bearings at 0.021 inches, which constitute the four health states of the rolling bearings together with the normal operation. The data applied in the experiments can be obtained through a sliding window with a size of 400 and a step size of 30. The details of the data are shown in Table 1.

Table 1. Details of the experiment data.

Bearing Health Condition	Fault Size (Inches)	Label
Normal	0	0
Inner race fault	0.021	1
Outer race fault	0.021	2
Roller fault	0.021	3

The DNN is used as the base network model in the experiment validation process, the number of neurons in each hidden layer is 600/200/101/50/30/30/9, the learning rates is 0.005, and the Adam optimizer was used to train the model. To verify the superiority of the proposed method, it was compared to four other semi-supervised methods, which is shown in Table 2.

Table 2. The related semi-supervised models for comparison.

Semi-Supervised Model	Model Description
SAE-SSL	Combines unlabeled data with labeled data for pre-training of SAE
π-Model [18]	Achieves prediction of unlabeled data through data augmentation and dropout
DNN-SSL [25]	Designs an unsupervised network to extract the features of unlabeled data
VAE-M1 [24]	Uses VAE to extract feature of both labeled and unlabeled data
MRAE	Designs training mechanism with multi-scale recursive for SAE
MRAE-AG	The method proposed in this article

Experiments 1–6 are designed to verify the superiority of the proposed method using different sizes of unlabeled data. Experiments 7–16 are designed to compare the improvement of fault diagnosis accuracy corresponding to different sizes of labeled data. A total of 4236 samples were utilized to test the fault diagnosis accuracy of the semi-supervised model. The detail of 16 experiments is shown in Table 3.

Table 3. Experiment design for semi-supervised fault diagnosis.

	Number of Labeled Data for Training	Number of Unlabeled Data for Training
Experiment 1	4×10	0
Experiment 2	4×10	4×25
Experiment 3	4×10	4×100
Experiment 4	4×10	4×250
Experiment 5	4×10	4×2000
Experiment 6	4×10	4×4000
Experiment 7	4×5	0
Experiment 8	4×5	4×4000
Experiment 9	4×25	0
Experiment 10	4×25	4×4000
Experiment 11	4×35	0
Experiment 12	4×35	4×4000
Experiment 13	4×45	0
Experiment 14	4×45	4×4000
Experiment 15	4×1000	0
Experiment 16	4×1000	4×4000

4.2. Results and Analysis

The features extracted by traditional SAE and MRAE are visualized using t-SNE (t-distributed Stochastic Neighbor Embedding), which is shown in Figure 6. The t-SNE1 and t-SNE2 represent the two principal components obtained by t-SNE. The features extracted by MRAE outperformed the features extracted by traditional methods, because the fault classes represented by yellow and gray in Figure 6a are more prone to be confused, while the distinguishability of these two classes is higher in Figure 6b. It can be seen from Figure 6 that the class features extracted by MRAE are easier to distinguish since they are well clustered.

The six experiments for the different numbers of unlabeled data used are designed to test the efficiency of the proposed method. The fault diagnosis results are shown in Table 4.

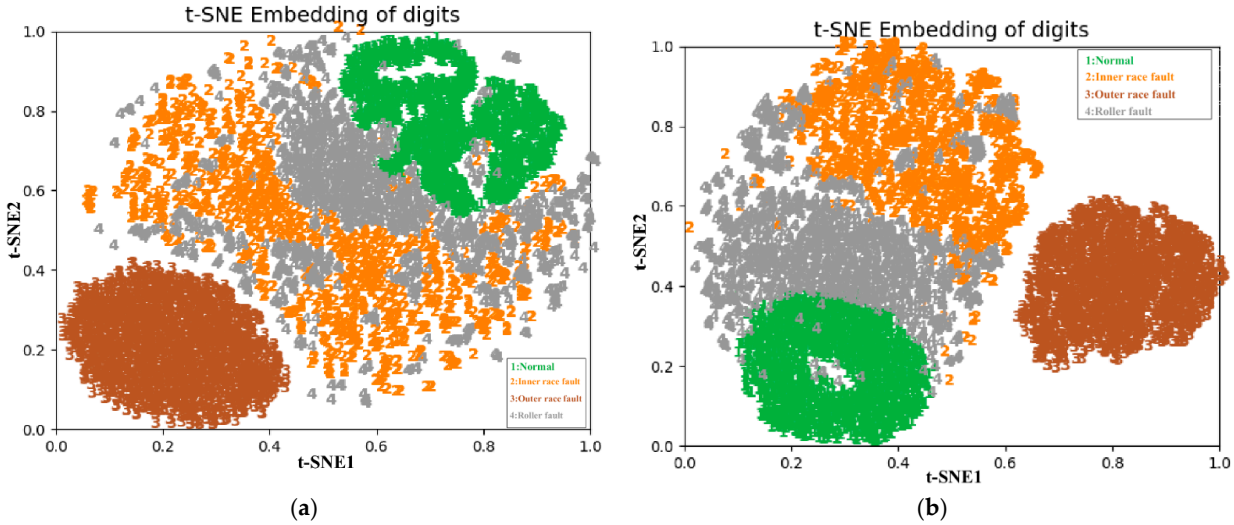

Figure 6. Feature visualization. (**a**) Features extracted by SAE; (**b**) Features extracted by MRAE.

Table 4. Fault diagnosis results of experiments 1–6.

	Number of Labeled Data	Number of Unlabeled Data	SAE-SSL	π-Model	DNN-SSL	VAE-M1	MRAE	MRAE-AG
Experiment 1	4 × 10	0	36.53%	45.98%	46.08%	45.90%	46.50%	46.50%
Experiment 2	4 × 10	4 × 25	57.64%	58.66%	58.61%	59.99%	63.10%	65.62%
Experiment 3	4 × 10	4 × 100	58.29%	60.24%	60.67%	66.90%	66.97%	67.98%
Experiment 4	4 × 10	4 × 250	60.65%	64.06%	61.23%	66.98%	71.76%	75.49%
Experiment 5	4 × 10	4 × 2000	60.86%	65.76%	61.83%	69.21%	72.20%	76.61%
Experiment 6	4 × 10	4 × 4000	61.59%	67.46%	69.27%	69.84%	72.49%	77.60%

Comparison of the different rows of Table 4 shows that in the case when only 10 labeled samples are available, the unlabeled data can be used to optimize the supervised model. By increasing the number of unlabeled data, the fault diagnosis accuracy can be enhanced. It can be seen from column 4 and column 5 of Table 4 that the π-model is superior to SAE-SSL since it designs an unsupervised loss function for the model during training, but it is difficult to balance the unsupervised loss function and the supervised loss function. It can be seen from column 5 and column 6 of Table 4 that the features extracted by DNN-SSL will be more accurate due to the unsupervised feature extraction network designed separately for unlabeled data. However, the error rate of pseudo labels was higher because of the direct use of fault diagnosis result affiliation for unlabeled data screening. It can be seen from columns 4–7 that VAE-M1 improves the accuracy of feature extraction and fault diagnosis over other conventional methods due to the powerful data generation capability. However, its final fault diagnosis accuracy is limited since the unlabeled data is not referred to in global fine-tuning. Column 7 and column 8 of Table 4 indicate that the feature extraction ability of MRAE was better than that of VAE-M1 because each AE's decoder of MRAE aims to reconstruct the input of all previous layers rather than only the current layer. However, it is still difficult to train a classifier with satisfying diagnosis performance since a unique layer of feature is to referred for screening unlabeled data. Column 8 and column 9 show that the fault diagnosis accuracy of MRAE-AG was significantly improved because an attention gate was designed to make more adequate use of useful features of different layers to ensure that sample screening is based on comprehensive features, which is helpful for improving the correct rate of pseudo labels.

The confusion matrix is shown in Figure 7 to make the fault diagnosis result clearer. Figure 7 shows the same conclusion as Table 4.

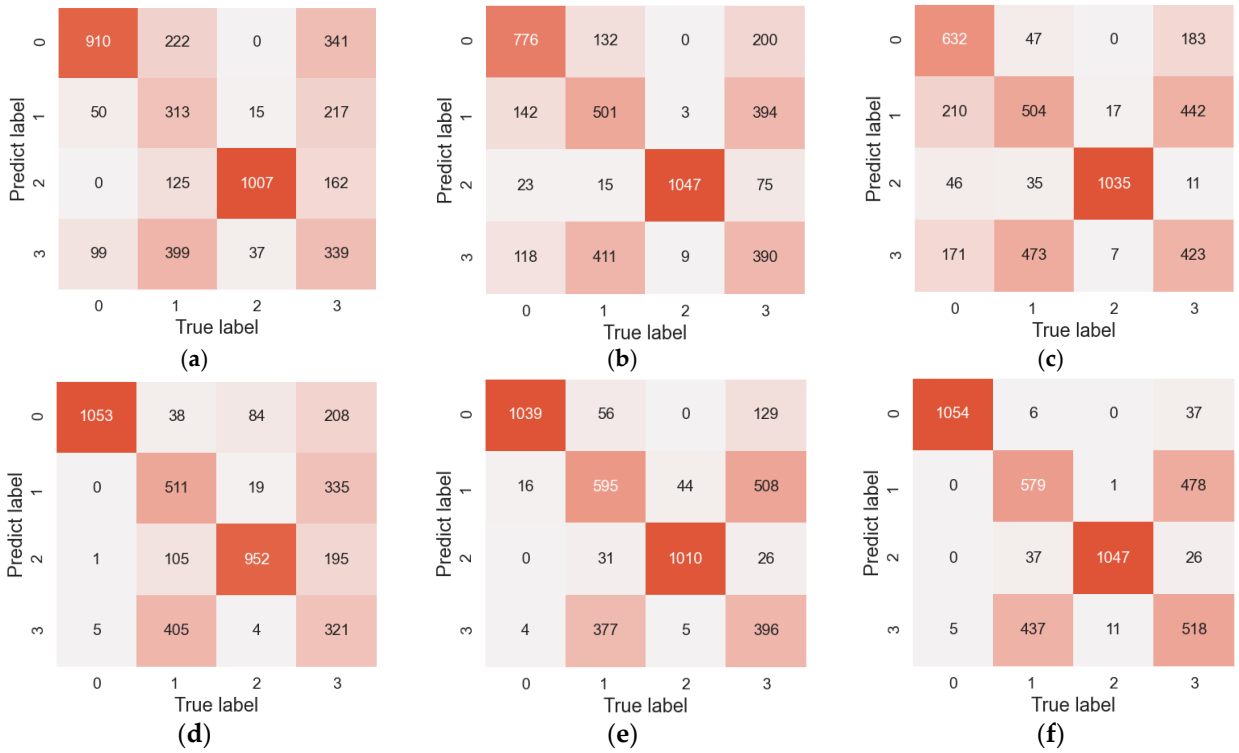

Figure 7. Confusion matrix of experiment 4. (**a**) SAE-SSL; (**b**) π-Model; (**c**) DNN-SSL; (**d**) VAE-M1; (**e**) MRAE; (**f**) MRAE-AG.

The 10 total experiments for the different number of labeled data used showed that the proposed MRAE-AG method was significantly superior to the existing methods in the case when there is only a very small number of labeled data available. The fault diagnosis results for these ten experiments are shown in Table 5.

Table 5. Fault diagnosis results of experiments 7–16.

	Number of Labeled Data	Number of Unlabeled Data	SAE-SSL	π-Model	DNN-SSL	VAE-M1	MRAE	MRAE-AG
Experiment 7	4 × 5	0	31.56%	36.30%	34.04%	36.04%	39.07%	39.07%
Experiment 8	4 × 5	4 × 4000	47.59%	54.91%	30.14%	62.25%	66.92%	70.15%
Experiment 1	4 × 10	0	36.53%	45.98%	46.08%	45.90%	46.50%	46.50%
Experiment 6	4 × 10	4 × 4000	61.59%	67.46%	69.27%	69.84%	72.49%	77.60%
Experiment 9	4 × 25	0	50.99%	56.30%	57.78%	58.09%	60.10%	60.10%
Experiment 10	4 × 25	4 × 4000	68.38%	70.30%	71.17%	71.34%	72.82%	77.75%
Experiment 11	4 × 35	0	57.12%	62.46%	58.59%	60.12%	63.83%	63.83%
Experiment 12	4 × 35	4 × 4000	64.56%	71.34%	71.53%	72.07%	73.79%	78.23%
Experiment 13	4 × 45	0	63.87%	66.43%	63.98%	64.87%	66.74%	66.74%
Experiment 14	4 × 45	4 × 4000	71.19%	71.57%	72.22%	73.04%	74.12%	81.72%
Experiment 15	4 × 1000	0	93.88%	94.40%	93.15%	93.22%	95.91%	95.91%
Experiment 16	4 × 1000	4 × 4000	94.16%	95.84%	93.83%	95.63%	97.85%	98.01%

Analysis of different rows of the last column shows that the proposed method can achieve higher diagnosis accuracy with the increase in the number of labeled data. Comparison of the different columns in Table 5 arrives at the same conclusion as Table 4.

The confusion matrix is shown in Figure 8 to make the fault diagnosis result clearer. Figure 8 shows the same conclusion as Table 5. The histogram is shown in Figure 9 to indicate the comprehensive experiment results of the corresponding methods.

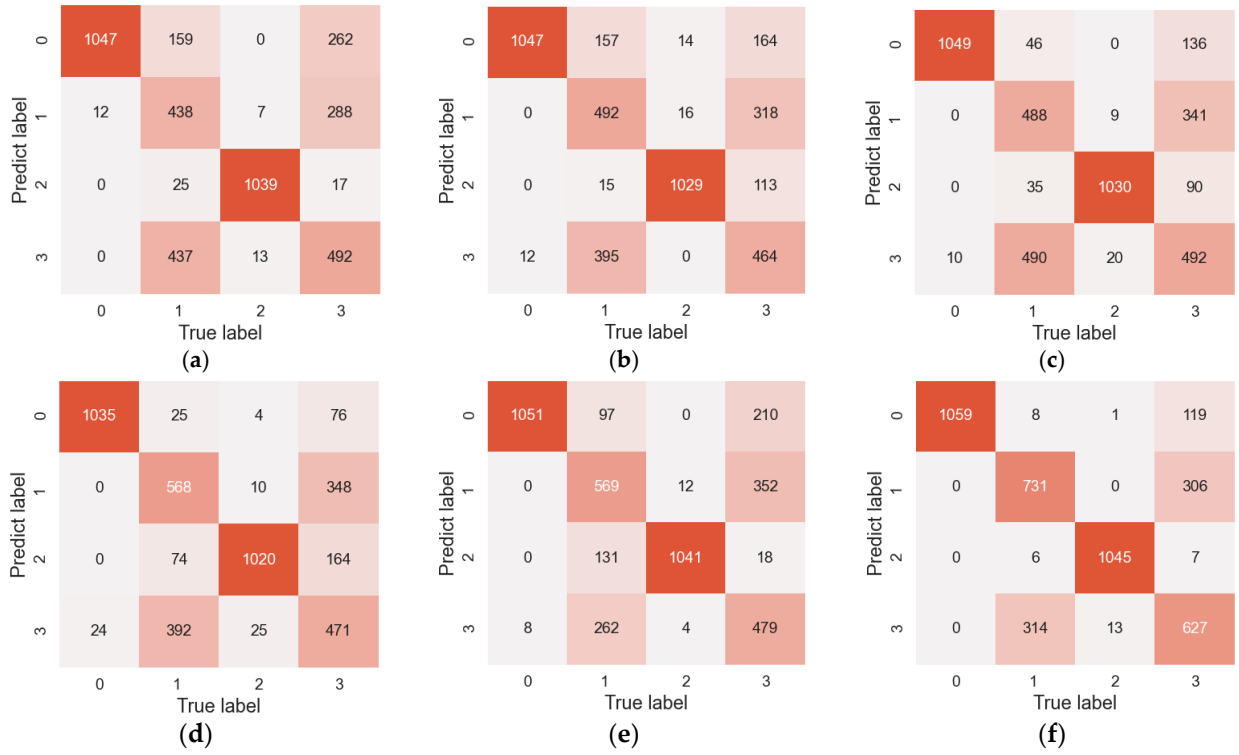

Figure 8. Confusion matrix of experiment 14. (**a**) SAE-SSL; (**b**) π-Model; (**c**) DNN-SSL; (**d**) VAE-M1; (**e**) MRAE; (**f**) MRAE-AG.

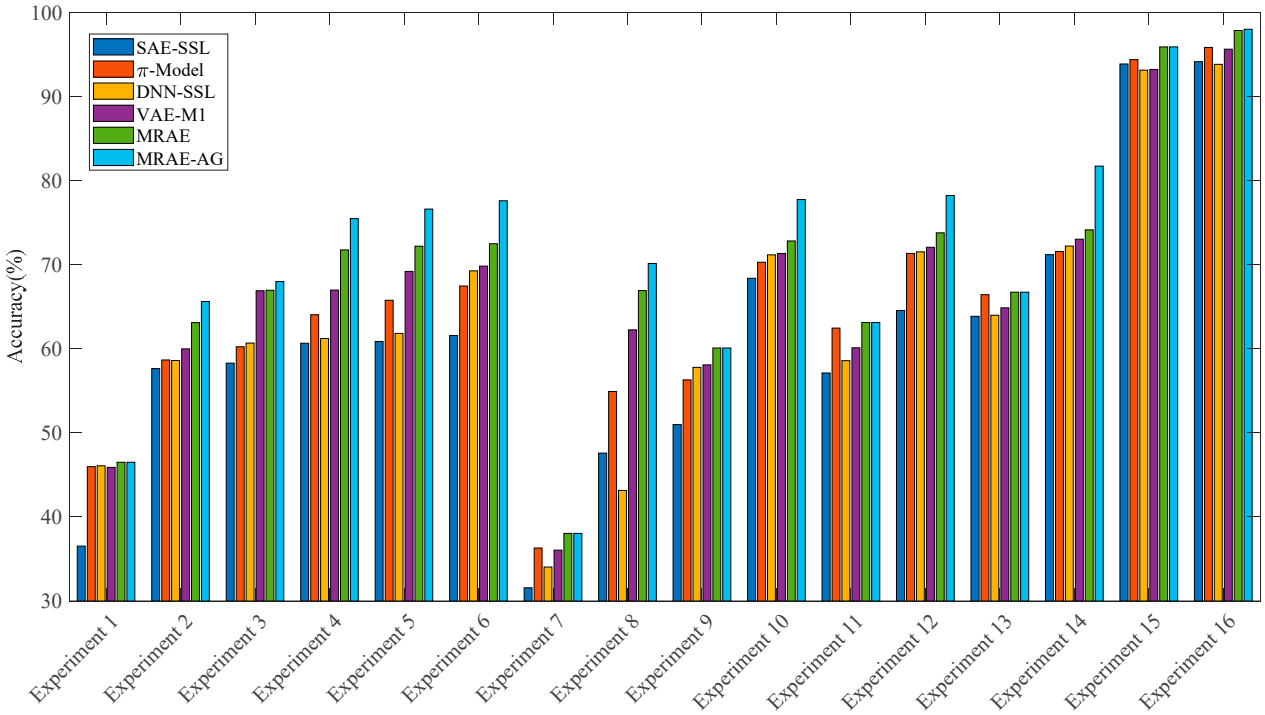

Figure 9. Histogram of all experiment results.

5. Conclusions

Existing semi-supervised methods cannot obtain satisfying fault diagnosis accuracy in the case when there is only a small amount of labeled data available due to inaccurate feature extraction and inadequate information utilization. It is crucial to develop a training

mechanism and a fusion strategy in order to train a well-performing, semi-supervised model. A new loss function was designed to take both the reconstruction error of the current layer and the previous layers into account in the pre-training stage of the network, which makes the model more powerful in feature representation. A fusion mechanism based on an attention gate was proposed to make the fusion feature more comprehensive, which is prone to screen unlabeled data. The features obtained by the designed method were more distinguishable than traditional methods in the sense that features were extracted more accurately and the information was more fully utilized. The experiment results showed that the proposed method outperformed existing semi-supervised methods and a more than 7% accuracy improvement can be obtained in the case when only a small size of labeled data is available. It was obvious that much less labeled data is sufficient for the proposed method to train a satisfying semi-supervised model.

Although the proposed method outperforms existing methods, it cannot achieve satisfying diagnostic accuracy in the case when there is less unlabeled data or no labeled data available. We will build a joint learning model using data from different clients in future work.

Author Contributions: Writing—original draft, S.T.; Writing—review & editing, C.W. Methodology, F.Z.; Resources, X.H.; Validation, T.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62073213) and the National Natural Science Foundation of China Youth Fund (52205111).

Data Availability Statement: The data involved in this article have been presented in the article.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (62073213) and the National Natural Science Foundation of China Youth Fund (52205111). The authors would like to thank Case Western Reserve University for providing the motor bearing vibration data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Toma, R.N.; Prosvirin, A.E.; Kim, J.M. Bearing Fault Diagnosis of Induction Motors Using a Genetic Algorithm and Machine Learning Classifiers. *Sensors* **2020**, *20*, 1884. [CrossRef] [PubMed]
2. Alabsi, M.; Liao, Y.B.; Nabulsi, A.A. Bearing fault diagnosis using deep learning techniques coupled with handcrafted feature extraction: A comparative study. *J. Vib. Control* **2021**, *27*, 404–414. [CrossRef]
3. Zhao, X.; Qin, Y.; Fu, H.; Jia, L.; Zhang, X. Blind source extraction based on EMD and temporal correlation for rolling element bearing fault diagnosis. *Smart Resilient Transp.* **2021**, *3*, 52–65. [CrossRef]
4. Kankar, P.K.; Sharma, S.C.; Harsha, S.P. Fault diagnosis of ball bearings using machine learning methods. *Expert Syst. Appl.* **2011**, *38*, 1876–1886. [CrossRef]
5. Li, X.; Zhang, W.; Ding, Q. Cross-Domain Fault Diagnosis of Rolling Element Bearings Using Deep Generative Neural Networks. *IEEE Trans. Ind. Electron.* **2019**, *66*, 5525–5534. [CrossRef]
6. Qiao, Z.J.; Shu, X.D. Stochastic Resonance Induced by Asymmetric Potentials Enhanced Mechanical Repetitive Transient Extraction. *J. Mech. Eng.* **2021**, *57*, 160–168.
7. Wang, X.; Mao, D.; Li, X. Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network. *Measurement* **2021**, *173*, 108518. [CrossRef]
8. Yi, C.A. Bearing Fault Diagnosis with Deep Learning Models. In Proceedings of the 2020 International Conference on Image Processing and Robotics (ICIP), Negombo, Sri Lanka, 6–8 March 2020.
9. Farajzadeh-Zanjani, M.; Hallaji, E.; Razavi-Far, R.; Saif, M.; Parvania, M. Adversarial Semi-Supervised Learning for Diagnosing Faults and Attacks in Power Grids. *IEEE Trans. Smart Grid.* **2021**, *12*, 3468–3478. [CrossRef]
10. Li, X.; Jiang, H.; Zhao, K.; Wang, R. A Deep Transfer Nonnegativity-Constraint Sparse Autoencoder for Rolling Bearing Fault Diagnosis with Few Labeled Data. *IEEE Access* **2019**, *7*, 91216–91224. [CrossRef]
11. Van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
12. Li, Y.F.; Liang, D.M. Safe semi-supervised learning: A brief introduction. *Front. Comput. Sci.* **2019**, *13*, 669–676. [CrossRef]
13. Ge, Z.Q. Semi-supervised data modeling and analytics in the process industry: Current research status and challenges. *IFAC J. Syst. Control* **2021**, *16*, 100150. [CrossRef]
14. Chen, X.; Wang, Z.; Zhang, Z.; Jia, L.; Qin, Y. A Semi-Supervised Approach to Bearing Fault Diagnosis under Variable Conditions towards Imbalanced Unlabeled Data. *Sensors* **2018**, *18*, 2097. [CrossRef] [PubMed]

15. Yi, H.K.; Jiang, Q.C. Graph-based semi-supervised learning for icing fault detection of wind turbine blade. *Meas. Sci. Technol.* **2020**, *32*, 035117. [CrossRef]
16. Wang, X.; Wang, T.; Ming, A.; Zhang, W.; Li, A.; Chu, F. Semi-supervised hierarchical attribute representation learning via multi-layer matrix factorization for machinery fault diagnosis. *Mech. Mach. Theory* **2022**, *167*, 104445. [CrossRef]
17. Liang, P.; Deng, C.; Wu, J.; Yang, Z.; Zhu, J.; Zhang, Z. Single and simultaneous fault diagnosis of gearbox via a semi-supervised and high-accuracy adversarial learning framework. *Knowl. Based Syst.* **2020**, *198*, 105895. [CrossRef]
18. Laine, S.; Aila, T. Temporal Ensembling for Semi-Supervised Learning. *arXiv* **2016**, arXiv:1610.02242.
19. Lee, D.H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In Proceedings of the 2013 International Conference on Machine Learning (ICML), Miami, FL, USA, 16–21 June 2013; p. 896.
20. Yu, K.; Lin, T.R.; Ma, H.; Li, X.; Li, X. A multi-stage semi-supervised learning approach for intelligent fault diagnosis of rolling bearing using data augmentation and metric learning. *Mech. Syst. Signal Process.* **2021**, *146*, 107043. [CrossRef]
21. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Li, C.L. Fixmatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 596–608.
22. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. MixMatch: A Holistic Approach to Semi-Supervised Learning. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5050–5060.
23. Tao, X.; Ren, C.; Li, Q.; Guo, W.; Liu, R.; He, Q.; Zou, J. Bearing defect diagnosis based on semi-supervised kernel Local Fisher Discriminant Analysis using pseudo labels. *ISA Trans.* **2021**, *110*, 394412. [CrossRef] [PubMed]
24. Zhang, S.; Ye, F.; Wang, B.; Habetler, T.G. Semi-Supervised Bearing Fault Diagnosis and Classification Using Variational Autoencoder-Based Deep Generative Models. *IEEE Sens. J.* **2020**, *21*, 6476–6486. [CrossRef]
25. Tang, S.J.; Zhou, F.N.; Liu, W. Semi-supervised bearing fault diagnosis based on Deep neural network joint optimization. In Proceedings of the 2021 China Automation Congress(CAC), Beijing, China, 22–24 October 2021; pp. 6508–6513.
26. Long, J.; Chen, Y.; Yang, Z.; Huang, Y.; Li, C. A novel self-training semi-supervised deep learning approach for machinery fault diagnosis. *Int. J. Prod. Res.* **2022**, 1–14. [CrossRef]
27. Hu, Z.; Yang, Z.; Hu, X.; Nevatia, R. Simple: Similar Pseudo Label Exploitation for Semi-Supervised Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 21–24 June 2021; pp. 15099–15108.
28. Liu, T.; Ye, W. A semi-supervised learning method for surface defect classification of magnetic tiles. *Mach. Vis. Appl.* **2022**, *33*, 1–14. [CrossRef]
29. Jiang, H.; Shao, H.; Chen, X.; Huang, J. A feature fusion deep belief network method for intelligent fault diagnosis of rotating machinery. *J. Intell. Fuzzy Syst.* **2018**, *34*, 3513–3521. [CrossRef]
30. Shao, H.; Jiang, H.; Wang, F.; Zhao, H. An enhancement deep feature fusion method for rotating machinery fault diagnosis. *Knowl. Based Syst.* **2017**, *119*, 200–220. [CrossRef]
31. Zhuang, Z.; Wei, Q. Intelligent fault diagnosis of rolling bearing using one-dimensional multi-scale deep convolutional neural network based health state classification. In Proceedings of the 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, China, 27–29 March 2018; pp. 1–6.
32. Zhou, F.N.; Zhang, Z.Q.; Chen, D.M. Bearing fault diagnosis based on DNN using multi-scale feature fusion. In Proceedings of the 2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC), Zhenjiang, China, 16–18 October 2020; pp. 150–155.
33. Yao, D.C.; Liu, H.C.; Yang, J.W.; Zhang, J. Implementation of a novel algorithm of wheelset and axle box concurrent fault identification based on an efficient neural network with the attention mechanism. *J. Intell. Manuf.* **2021**, *32*, 729–743. [CrossRef]
34. Xie, Z.L.; Chen, J.L.; Feng, Y.; He, S.L. Semi-supervised multi-scale attention-aware graph convolution network for intelligent fault diagnosis of machine under extremely-limited labeled samples. *J. Manuf. Syst.* **2022**, *64*, 561–577. [CrossRef]
35. Loparo, K.A. Case Western Reserve University Bearing Data Center. Available online: http://csegroups.case.edu/bearingdatacenter/pages/download-data-file (accessed on 10 May 2022).