

Small Sample Gear Fault Diagnosis Method Based on Transfer Learning

Han Zhang, Shihao Liu*, Xiyang Wang, Junlong Zhang

School of Aircraft Engineering, Nanchang Hangkong University, Nanchang, China

Email: *523112232@qq.com

How to cite this paper: Zhang, H., Liu, S.H., Wang, X.Y. and Zhang, J.L. (2023) Small Sample Gear Fault Diagnosis Method Based on Transfer Learning. *Open Journal of Applied Sciences*, 13, 2461-2479. <https://doi.org/10.4236/ojapps.2023.1312192>

Received: November 24, 2023

Accepted: December 26, 2023

Published: December 29, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Aiming at the problems of lack of fault diagnosis samples and low model generalization ability of cross-working gear based on deep transfer learning, a fault diagnosis method based on improved deep residual network and transfer learning was proposed. Firstly, one-dimensional signal is transformed into two-dimensional time-frequency image by continuous wavelet transform. Then, a deep learning model based on ResNet50 is constructed. Attention mechanism is introduced into the model to make the model pay more attention to the useful features for the current task. The network parameters trained by ResNet50 network on ImageNet dataset were used to initialize the model and applied to the fault diagnosis field. Finally, to solve the problem of gear fault diagnosis under different working conditions, a small sample training set is proposed for fault diagnosis. The method is applied to gearbox fault diagnosis, and the results show that: The proposed deep model achieves 99.7% accuracy of gear fault diagnosis, which is better than the four models such as VGG19 and MobileNetV2. In the cross-working condition fault diagnosis, only 20% target dataset is used as the training set, and the proposed method achieves 93.5% accuracy.

Keywords

Gear Fault Diagnosis, Transfer Learning, CWT, Deep Residual Network, Deep Learning

1. Preface

Gearbox as a mechanical transmission system is widely used in the key components, its working condition directly determines the normal operation of the equipment, and in the gearbox failure form of gear failure accounted for 60% [1]. Therefore, gear fault diagnosis is important for improving the reliability of

industrial equipment and reducing production costs. And with the rapid development of industrial big data and Internet of Things, deep learning-based fault diagnosis methods have been widely studied [2]. Deep learning methods utilize a deep network consisting of a series of nonlinear layers to adaptively learn complex representations of data in the deep feature space, and extract fault features related to the health state of the equipment through the deep representation of the data, which can be used to achieve end-to-end fault diagnosis if fault diagnosis classification algorithms are then incorporated into the model [3]. SAUFI [4] *et al.* applied stacked sparse selfencoders to gearbox fault diagnosis and achieved high diagnostic accuracy with limited samples. JIA [5] *et al.* constructed a model for fault diagnosis based on a one-dimensional convolutional neural network and validated it in motor bearing fault data. JIANG [6] *et al.* proposed a multi-scale CNN that can automatically learn effective fault features directly from the vibration signals of a device.

Deep learning is widely used in the field of mechanical fault diagnosis, but still suffers from the following two problems:

1) Deep learning relies on data-driven, without more high-quality data to train the model, it will affect the diagnostic efficiency and accuracy of the model, and the lack of fault data is a common problem in the field of mechanical fault diagnosis research [7].

2) Deep learning models extract the deep features of the data by stacking hidden layers, but as the model deepens, more and more data need to be labeled for its training, and the difficulty of training increases, which is prone to overfitting phenomenon and reduces the generalization ability of the model, so most of the existing hidden layers of the fault diagnosis model are less than five layers [8]. However, shallow models are limited by their structure and are unable to characterize the deep and complex mapping relationships between signals and equipment health states, resulting in models that cannot meet the needs of machinery big data diagnosis [9].

Migration learning is dedicated to migrating the knowledge acquired in the source domain to the target domain to achieve efficient training of models in the target domain, which can solve the above problems of deep learning [10]. ZHAO [11] *et al.* proposed a multi-scale CNN-based transfer learning framework for fault diagnosis under variable operating conditions and across devices. HE [12] *et al.* used multi-channel signals to construct an integrated migration CNN fault diagnosis model, in which the proposed decision fusion strategy based on the ideas of weight assignment and majority voting flexibly fuses the results of each channel, and the constructed model is validated in gearbox fault diagnosis. WEN [13] *et al.* applied sparse self-encoders to migration learning to construct an intelligent fault diagnosis model to realize fault diagnosis under multiple operating conditions. LU [14] *et al.* proposed a migration learning-based fault diagnosis method that achieved fast and accurate diagnostic results on two datasets with different operating conditions. However, the above existing migration learning methods for fault diagnosis still need to train a deep learning model from scratch

in a source domain fault data set, and thus need to spend a lot of resources on training and tuning of the deep learning model. Moreover, in the actual application of fault diagnosis engineering, different working conditions may cause perturbation of mechanical signals, resulting in changes in signal fault characteristics, which may lead to a drastic performance degradation of the originally trained classification prediction model during deployment and application, resulting in model failure.

Aiming at the above problems, this paper proposes a gearbox fault diagnosis method based on improved deep residual network with migration learning. Firstly, the vibration signals are converted into 2D time-frequency maps by CWT; then a deep learning model based on ResNet50 network is constructed, and the structure and training parameters of the shallow layer of ResNet50 network trained on ImageNet image dataset are migrated to the proposed model for initializing the model parameters through the migration learning method so that the shallow layer of the model will have a mature feature extraction performance at the beginning of training. Aiming at the problem of fault diagnosis in real production environments where gears are under different operating conditions and lack of fault data, we propose a migration learning method based on a small-sample training set, which enables features useful for the task to receive better attention by embedding an attention mechanism in the model. In cross-case fault diagnosis using a small scale target domain dataset as a training set, the proposed model has better fault diagnosis performance in gear fault diagnosis compared to other deep models.

2. Theoretical Foundations

2.1. Image Generation Methods

One-dimensional vibration signals are often analyzed when deep learning methods are used in the field of fault diagnosis, but compared to one-dimensional signals, image data are two-dimensional matrices that can carry more information, and therefore can be used to characterize the distribution of more complex data structures. And in the field of machine vision, mostly RGB three-channel color images are used as inputs, and the conversion of one-dimensional signals into two-dimensional images is necessary if migration learning is done using well-trained models in the field of machine vision. There are two main types of methods in existing research to convert one-dimensional vibration signals into two-dimensional images: One is to segment the acquired signals at equal intervals, and then splice them into a data matrix to obtain a two-dimensional grayscale image, which has the advantage of simplicity and speed, but lacks frequency domain information; Secondly, two-dimensional time-frequency images are acquired by time-frequency imaging methods, such as STFT, WVD, HHT and CWT. Among them, STFT has low time-frequency resolution and is not effective in matching multicomponent time-varying signals; WVD is not robust to noise and has cross-interference terms for multicomponent time-varying signals;

and HHT suffers from end-point effects and modal aliasing; CWT is an effective multi-resolution signal analysis technique, which can effectively explore the signal frequency domain fault information, so this paper uses the CWT method to obtain the time-frequency map.

2.2. Convolutional Neural Networks

CNN can adaptively extract data features to achieve efficient recognition of targets without human intervention, and are generally composed of a convolutional layer, a pooling layer, and a fully connected layer. In this case, successive convolutional and pooling layers are used for data feature extraction, while the fully connected layer is responsible for integrating the extracted features and outputting the predicted values of the classification results.

Convolutional layer is the core of CNN, mainly consists of a number of convolutional kernel, convolutional kernel with a certain step size traverses the data process that is feature extraction, Equation (1) is the convolutional layer output operation formula:

$$X_i = f(X_{i-1} * W_i + b_i) \quad (1)$$

where: X_i is the feature matrix of layer i ; $f(\cdot)$ is the activation function; X_{i-1} is the input of layer i ; W_i is the weight; b_i is the bias.

ReLU is commonly used as an activation function in CNNs, but GELU is smoother compared to ReLU. Equation (2) is the expression for GELU:

$$\text{GELU}(x) = x\varnothing(x) \quad (2)$$

where: $\varnothing(x)$ is the cumulative probability distribution of a Gaussian normal distribution of x .

2.3. Transfer learning

Transfer learning is widely used in various fields, it is defined as: Given a source domain D_s and a source task T_s , Goal domains D_t and target tasks T_t , transfer learning aims to use the knowledge acquired in the source domain to help optimize the objective mapping function $f(\cdot)$, among them, $D_s \neq D_t$ or $T_s \neq T_t$.

According to Yosinski [15] *et al.* it has been shown that in deep learning models the features extracted from the shallow layers of the network are generalizable and similar, whereas the features extracted from the deeper layers are more abstract compared to the deeper layers, so that the shallow layers of the model are better suited for migration whereas the deeper layers are better suited for dealing with specific tasks.

The transfer learning approach for the fault diagnosis domain is shown in **Figure 1**. A new deep learning model is first trained using the source domain data, while the trained feature extraction layer is later migrated to the target domain, where the classification layer is newly trained in the target domain.

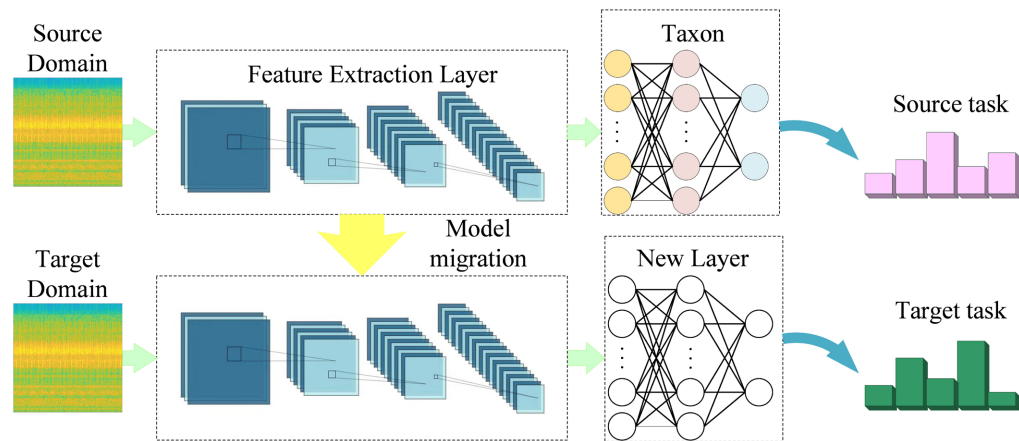


Figure 1. Transfer learning in fault diagnosis.

2.4. Deep Residual Network, Migration Model Construction

2.4.1. Deep Residual Networks

In deep learning, the expressiveness of the network is positively correlated with the depth of the network. Therefore, to design deep learning networks with excellent performance it is necessary to ensure that they have a certain depth. Traditional methods focus on increasing depth by linearly stacking network layers, which does produce some results e.g. VGG networks. However, this method brings exponential growth of training parameters, unstable gradient of the network and network degradation problem, which leads to the need to consume a lot of resources to train the network, and the performance may instead be degraded. In 2016 HE [16] *et al.* proposed ResNet, whose unique residual link structure reduces the impact of the above problems to some extent. As shown in **Figure 2** for the two residual structures in the ResNet50 network, when x is used as the input to the network, assuming that the output fitting function is $H(x)$, the residual mapping $F(x)$ can be denoted as $H(x) - x$, and $H(x)$ can be denoted as $F(x) + x$, where $F(x) + x$ can be interpreted as the summation operation of $F(x)$ with x using shortcut prior to the output of the fitting function in the model building process.

In the residual structure, shortcut branches can bypass some layer links to converge with the main branch without introducing parametric conditions. Therefore, during the training process, the bottom error can be propagated directly upward through the shortcut branch, which can attenuate the phenomenon of gradient vanishing caused by too many layers.

2.4.2. Migration Modeling

Based on transfer learning theory, In this paper, we use pre-training parameters to initialize the ResNet50 shallow network structure and parameters, replacing the last fully connected layer output with 5, and the residual structure was modified in Conv5, as shown in **Figure 3**.

A separate downsampling layer was first constructed before Conv5, Replaces the downsampling operation in the original residual block by setting the step

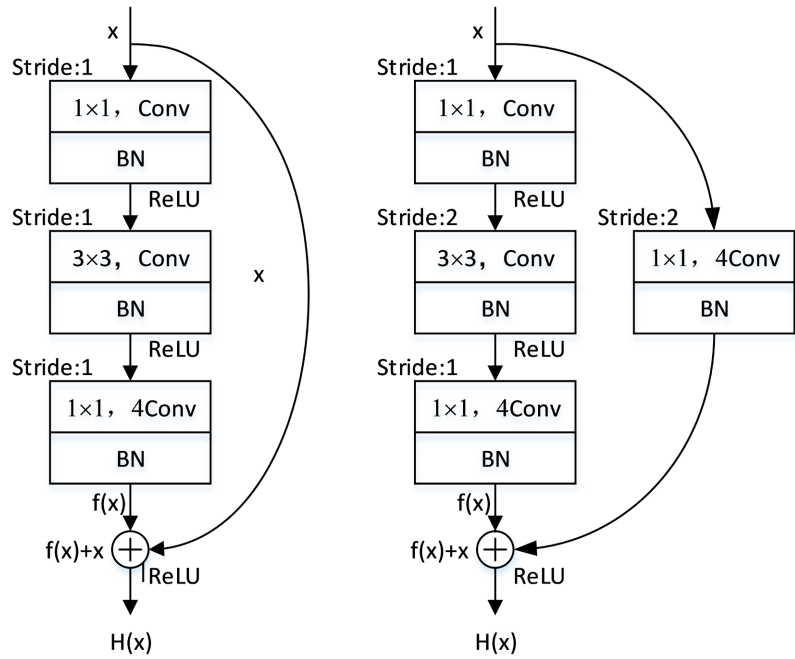


Figure 2. Residual structure in ResNet50.

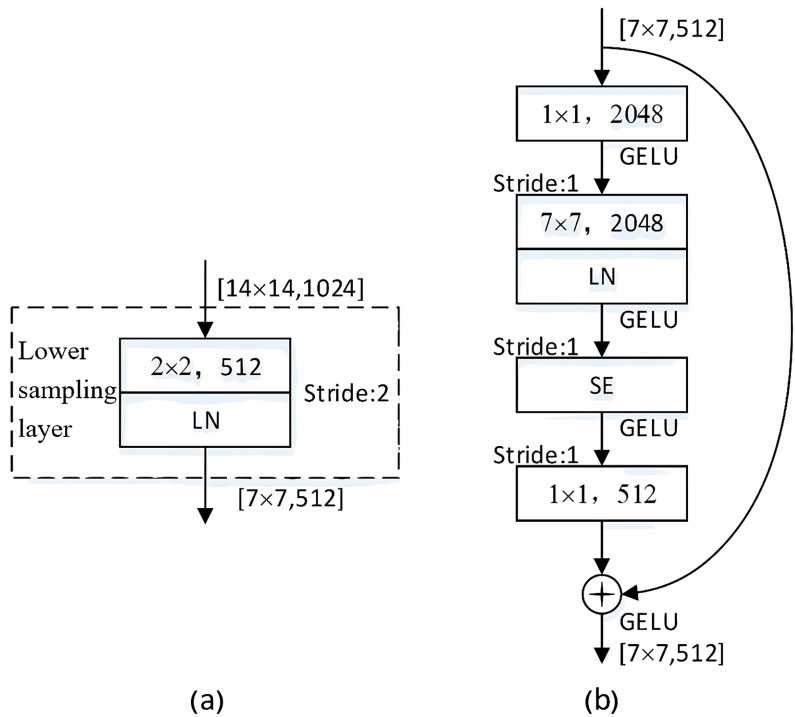


Figure 3. Residual structure and downsampling layer proposed in this paper.

size of the second convolution kernel on the main branch of the first residual structure to 2.

Second, it has been shown that replacing the residual structure in the ResNet model with an inverted residual structure can effectively improve the performance of the model [17], therefore, in this paper, we adjusted the number of

convolution kernels of each convolutional layer on the main branch of the residual structure, so as to present an inverted residual structure with thin top and bottom ends and thick center.

Third, the continuous use of small convolutional kernels has been widely used in deep learning models since the VGG network was proposed, as has the ResNet network. However, small convolutional kernels are more likely to lead to loss of global features compared to larger convolutional kernels, and successive use of small convolutional kernels leads to loss of more detailed features. To this end, we modify the kernel size of the second convolutional layer on the main branch of the residual structure to 77, and for the resulting parameter proliferation, we introduce the depthwise convolution. Assuming that the number of input feature maps is N , the number of parameters required to use DW will be about $1/N$ of the normal convolution.

Fourth, the BN layer can overcome the gradient dispersion problem caused by network deepening in ResNet, but BN is to do standardization in the sample batch dimension, so it is more dependent on the size of the batch size, and the research is often limited by the equipment can not be trained to train a larger batch size samples, so as to fail to play the advantages of BN. LN is another normalization method, which does layer normalization within each sample, and thus can effectively get rid of the batch size limitation. In this paper, the BN layer is replaced with an LN layer and the use of the normalization layer is reduced in order to retain more detailed features extracted by the network.

Fifth, the shallow level of the model extracts generic features, and in order to enhance the importance of features useful for fault diagnosis in the model, this paper uses the squeeze and excitation channel attention mechanism as in **Figure 4** to allow the network to perform feature recalibration. Through the SE mechanism, the model can selectively assign the weight values to each feature channel in the global information in an auto-learning manner, so as to achieve the purpose of emphasizing useful features and suppressing useless features.

Finally, GELU, an activation function proposed in research a few years after the proposal of ResNet, is considered a smoother variant of ReLU and is mostly

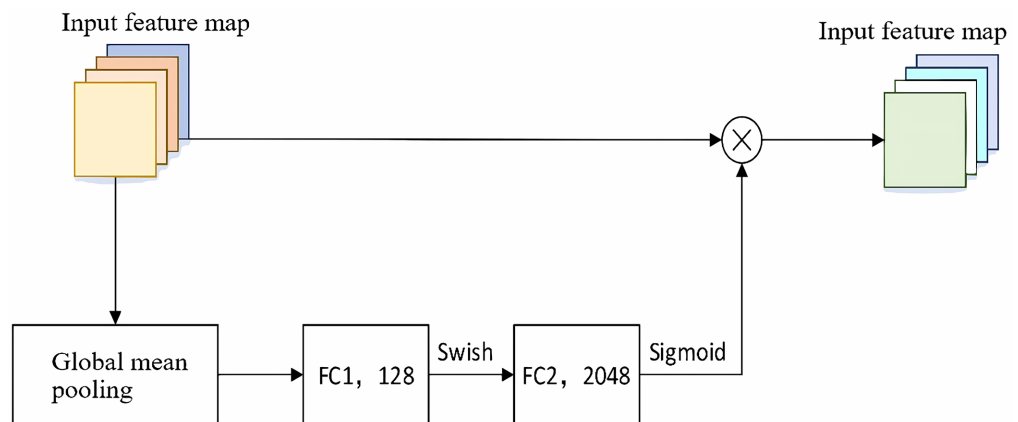


Figure 4. SE layers used in this paper.

used to replace the latter, which is also used as an activation function in the new residual structure. The detailed architecture and parameters of the final migration model are shown in **Table 1**.

3. Algorithm Framework Details

In this paper, we propose a fault diagnosis method based on improved ResNet50 with migration learning for gearbox fault diagnosis, and use a small sample training set to accomplish cross-case diagnosis of gear faults. The specific process is shown in **Figure 5**.

1) The acquired gear failure vibration signals were segmented at equal distances, while a three-channel 2D time-frequency map was generated for each segmented sample using CWT as a sample data set.

2) Constructing a ResNet50-based deep learning model that retains all model structures prior to Conv5 in the ResNet50 network as feature extractors, create a new separate downsampling layer and Conv5 residual layer and change the output of the final fully connected layer to 5.

3) Layers prior to model Conv5 were initialized using ResNet50 network parameters trained on the ImageNet image dataset, train the entire model in a single-case gear dataset to obtain fault diagnosis results, and retain the trained model that can be directly applied to cross-case fault diagnosis.

Table 1. Detailed model architecture.

network layer	Number of output channels	Structural parameters	quantity of participants
Conv1	112×112	7×7.64 , stride2	9536
Maxpool	56×56	3×3 maxpool, stride2	0
Conv2	56×56	$\begin{bmatrix} 1 \times 1.64 \\ 3 \times 3.64 \\ 1 \times 1.256 \end{bmatrix} \times 3$	215,808
Conv3	28×28	$\begin{bmatrix} 1 \times 1.128 \\ 3 \times 3.128 \\ 1 \times 1.512 \end{bmatrix} \times 4$	1,219,584
Conv4	14×14	$\begin{bmatrix} 1 \times 1.256 \\ 3 \times 3.256 \\ 1 \times 1.1024 \end{bmatrix} \times 6$	7,098,368
Downsample	7×7	2×2.512 , stride2	2,097,664
Conv5	7×7	$\begin{bmatrix} 1 \times 1.2048 \\ 7 \times 7.2048 \\ 1 \times 1.512 \end{bmatrix} \times 3$	8,185,728
Avapool	7×7	7×7 average pool, 512	0
FC	5	1×1.512	2565
Total model parameters			18,829,253

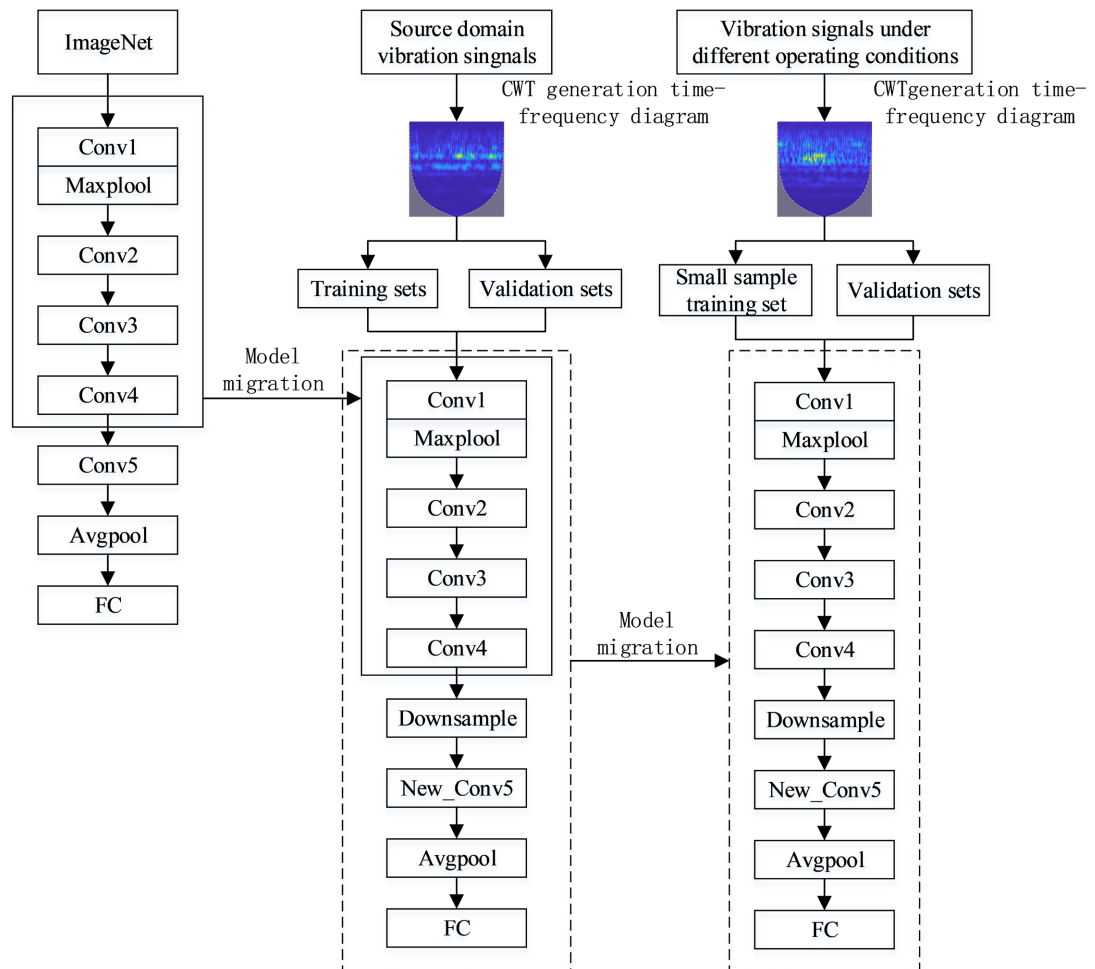


Figure 5. Schematic diagram of the methodological framework proposed in this paper.

4) The gear failure dataset under different operating conditions is partitioned in a small percentage to form the training set and the rest is used as the validation set. The model trained in step 3 is fine-tuned using the training set, while the performance of the fine-tuned model is later tested using the validation set.

4. Experimental Verification and Analysis

The experiments were done on a computer configured with i5-8300H, 8G RAM, and GTX1050Ti, based on the Pytorch deep learning framework of Pytorch 3.6. The training parameters for the experiment are as follows: training batch size of 16, initial learning rate of 0.0005, learning rate decay final multiplicity of 0.001, L2 regularization factor of 0.05, Adan optimizer and cross-entropy loss function.

4.1. Description of the Experimental Data Set

The data used in the experiment were collected from the DPS fault diagnosis experimental platform of SQI, and its structure is shown in **Figure 6**. The subject of the experiment is the involute spur pinion gear on the intermediate shaft labeled in the perspective view of the parallel gearbox in **Figure 7**, artificially ar-

ranged for a single point of failure, and the types of failures are shown in **Figure 8** as wear, broken teeth, root cracks, and missing teeth, respectively.

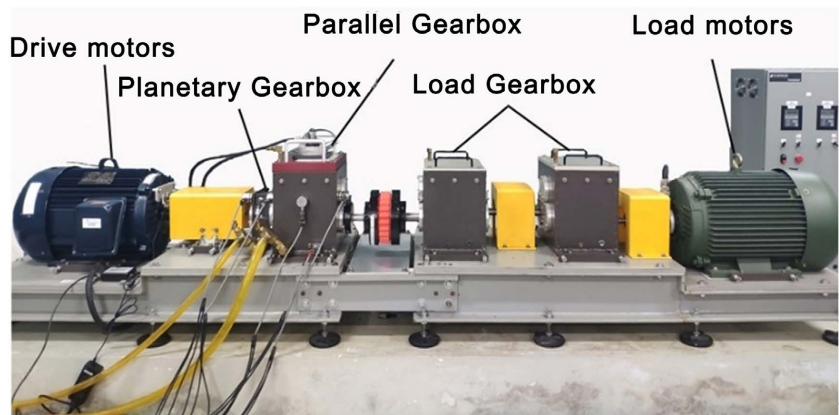


Figure 6. Lab bench.

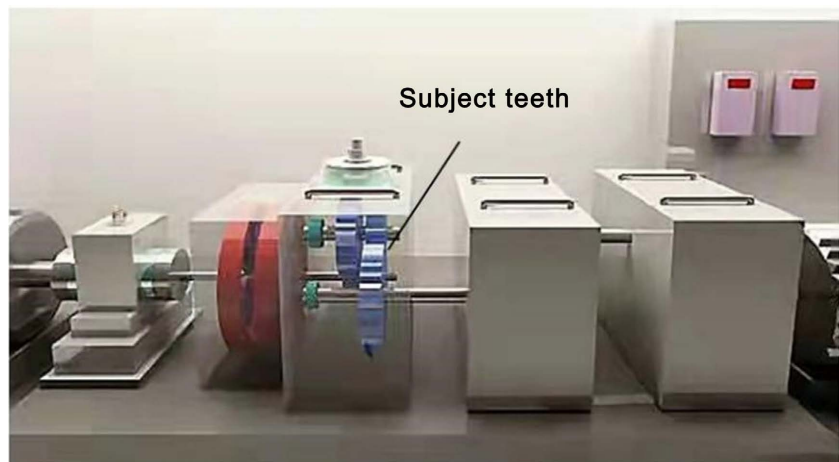


Figure 7. Perspective view of parallel gearbox.

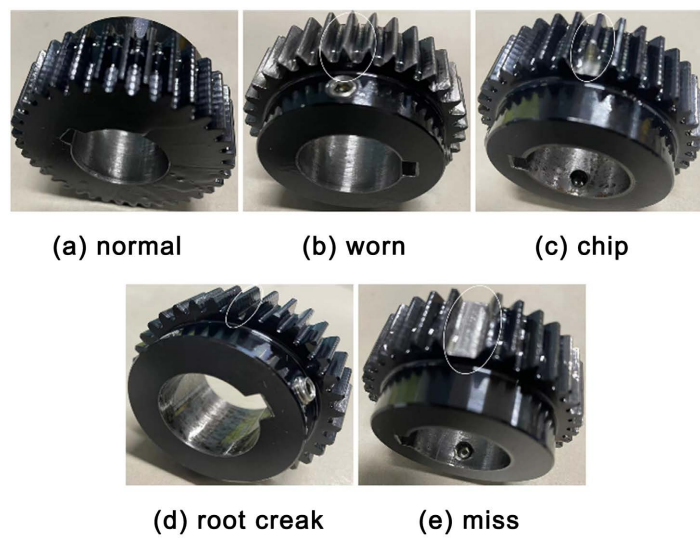


Figure 8. Gear health.

The vibration signals of five different gears under four load conditions were collected with a sampling frequency of 20,480 Hz using a triaxial acceleration sensor (PCB 604B31) under the condition that the driving motor rotational frequency was 30 Hz, and the single acquisition time was 32 seconds. The samples are intercepted with a sliding window of 2048 length without overlapping, and CWT is done for each sample to obtain a time-frequency map of 224,224 pixels in size, to obtain the gear failure dataset under different loading conditions, and the detailed information of the dataset is shown in Table 2. Figure 9 presents some of the data.

4.2. Analysis of Model Optimization Results

To verify the impact of each improvement on the model performance, the model was gradually optimized starting from the base ResNet50, and pre-training parameters were used to initialize the model after each model optimization. Using 20% of the samples from dataset A as the training set and the rest as the validation set, the model is trained 10 times for 40 rounds for each optimization, and the results are shown in Table 3 and Figure 10 (where ID is independent downsampling and IR is inverted residual).

From the experimental results, the use of independent downsampling, inverted residual structure, LN and SE optimization models all play a positive role in the improvement of fault diagnosis accuracy, the addition of GELU had essentially

Table 2. Details of the dataset.

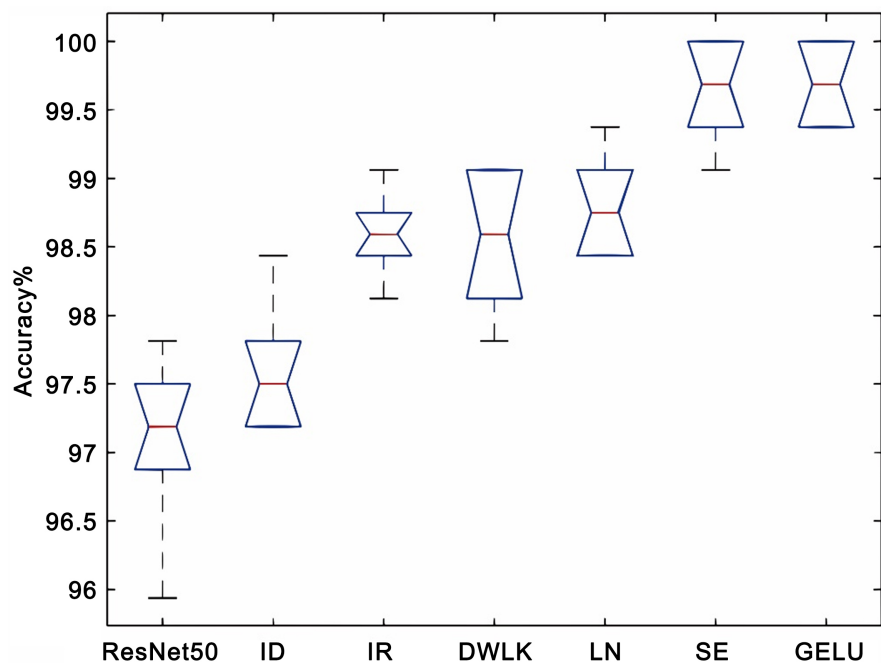
Data set	Load (0.45 N/m)	Fault type	Sample size
A	0	Normalcy	1600 (320 × 5)
B	1	wear and tear	1600 (320 × 5)
C	3	broken tooth	1600 (320 × 5)
D	0~4	tooth root crack missing tooth	1600 (320 × 5)

Legend	Channel1 Tacho1	Channel2 Tacho2	Channel3 Tacho3	Channel4	Channel5	Channel6	
On/Off	ON OFF	ON OFF	ON OFF	ON	ON	ON	
Volts/Unit	1	1	1	1	1	1	
Data	0	0.075487	-0.01348	0.03437	0.044246	-0.01486	0.00282
	0.000049	0.075471	0.005622	0.0325	0.022401	-0.00234	0.009322
	0.000098	0.075547	0.01753	0.025077	0.024816	0.012292	0.017278
	0.000146	0.075357	0.025734	0.01645	0.055363	0.02946	0.024885
	0.000195	0.075614	0.035236	0.010588	0.065739	0.052344	0.03209
	0.000244	0.075196	0.052404	-0.00019	0.046037	0.072135	0.032907
	0.000293	0.075809	0.062802	-0.01168	0.025752	0.075406	0.024652
	0.000342	0.074747	0.073127	-0.0282	0.004998	0.084605	0.029378
	0.000391	0.076226	0.088547	-0.02396	0.005516	0.084327	0.026062
	0.000439	0.074472	0.095663	-0.03351	-0.00566	0.070164	0.013065
	0.000488	0.076988	0.101587	-0.03913	-0.02286	0.068969	0.013914

Figure 9. Partial data presentation.

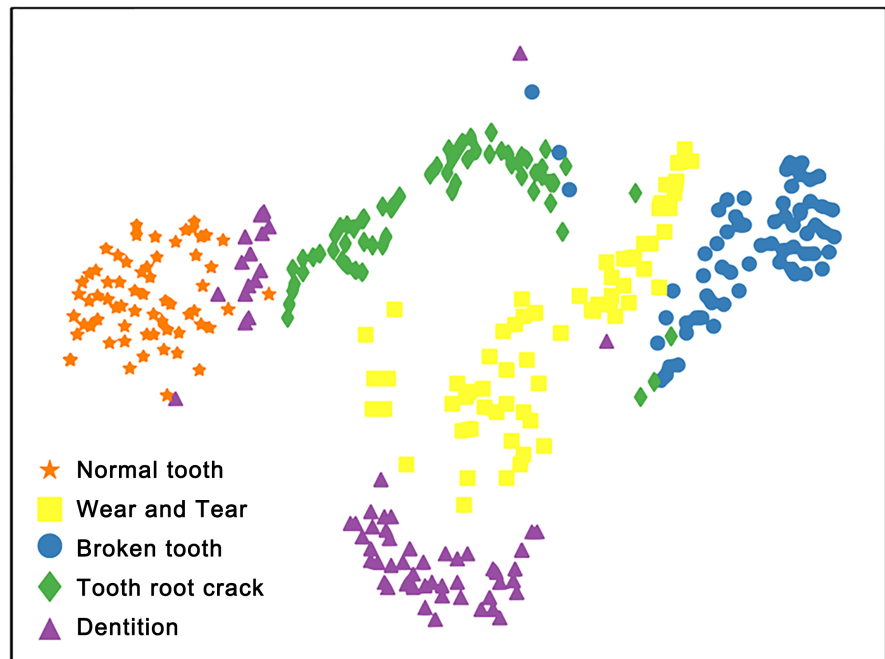
Table 3. Impact of each improvement on the model.

model optimization	average accuracy/%	training time/min
ResNet	97.0625	24.1
Independent downsampling	97.5938	24.6
inverted residual	98.5938	54.8
DW large convolutional kernel	98.5625	25.1
LN	98.8438	25.3
SE	99.6875	26
GELU	99.7188	25.9

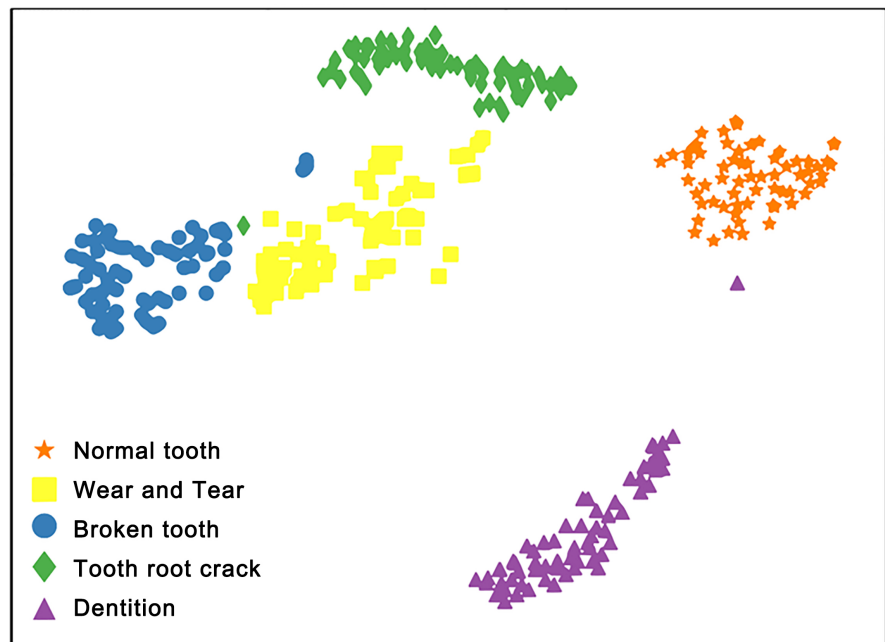
**Figure 10.** Box plots of the effect of each improvement on the model.

no effect, while the use of the 77-size convolutional kernel and DW had a negative effect, due to the fact that while the larger convolutional kernel captures more global information, the DW convolution ignores the information interactions between the channels, which ultimately leads to a slight decrease in accuracy. As can be seen from the training time in **Table 3**, the training time required by the model is significantly increased after the inclusion of the inverted residual structure, viewing the number of Conv5 parameters optimized by the inverted residual structure is 119,565,312, which is more than 91.8% of the total parameters of the model. While the model is optimized by DW with expanded convolutional kernel although there is a small decrease in classification accuracy, the number of Conv5 parameters is 6,626,304, which is a decrease of about 94.5%, so the overall performance improvement of the model by DW and expanded convolutional kernel can be considered as a positive effect. The second convolutional layer in the last residual block of Conv5 and the subsequent SE

layer output features are visualized by the t-distributed stochastic domain embedding algorithm, and the results are shown in **Figure 11**. In comparing the scatter plots of the two fault features, it can be found that the inter- and intra-class spacing of the fault features are significantly reduced by the SE layer, which will undoubtedly provide a positive help to the feature classification, and intuitively explains the effectiveness of the SE layer on the model enhancement.



(a) Second convolutional layer



(b) SE Layer

Figure 11. Scatter plot of fault characteristics of convolutional and SE layers.

From the above experiments, it can be concluded that the model optimization scheme proposed in this paper can effectively improve the performance of Res-Net50 in migration learning.

4.3. Comparative Analysis of Different Models

In order to verify the effectiveness of the method proposed in this paper, four models (AlexNet, VGG19, inception-V3 and MobileNetV2) commonly used in the field of machine learning are selected for comparison experiments. All models used were models that completed pretraining on the ImageNet dataset and performed fault diagnosis under the same training conditions.

The training and validation sets are randomly assigned in a 1:4 ratio under working condition A. Each model is trained 10 times and each iteration is traversed for 40 rounds, and the experimental results are shown in **Figure 12**.

As can be seen in the figure, the method in this paper has the highest fault diagnosis accuracy of 99.72% compared to the other four models under the same training conditions. AlexNet achieved the lowest average accuracy of 92.94% due to its relatively simple structure and limited mining of deeper features of the data. The average accuracy of VGG19 and inception-V3 fault diagnosis is 97% and 96.13%, respectively, which is attributed to the complexity of the model, but also the consequent drastic increase in parameters. MobileNetV2 benefits from depth-separable convolution, allowing it to obtain far more depth than the VGG19 and inception-V3 models while having fewer total parameters relative to the former two, and achieving an average accuracy of 97.63%.

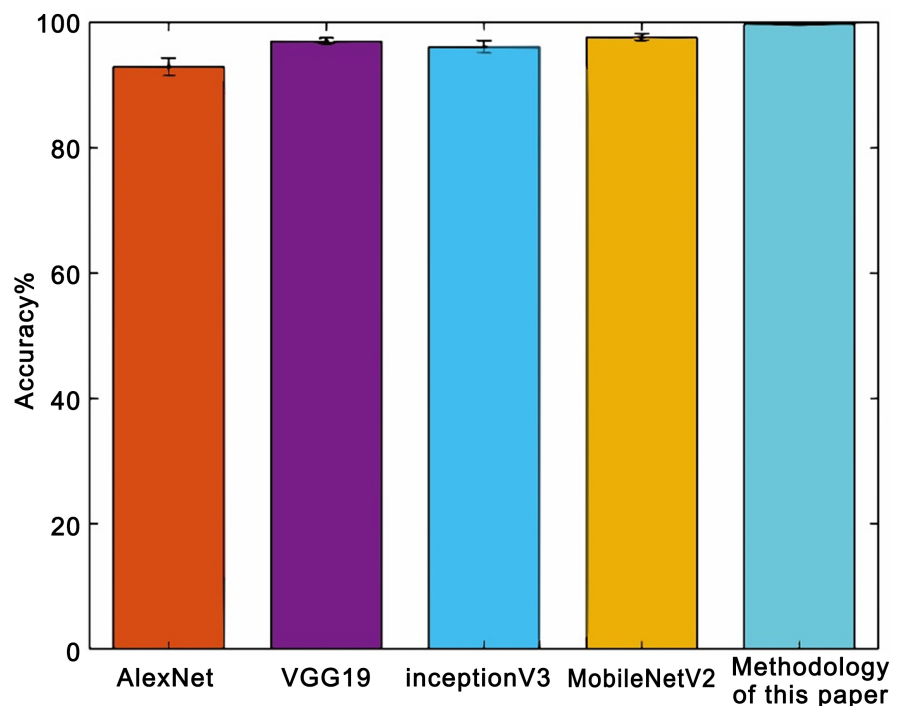


Figure 12. Comparison of fault diagnosis accuracy of different models.

In order to verify the effectiveness of the migration strategy proposed in this paper, each model is trained from scratch, and the experimental conditions are the same as those of the appeal training using pre-trained models, and the results are shown in **Table 4**.

Comparing **Figure 11** and **Table 4**, it can be seen that the fault diagnosis accuracy of each model is significantly improved by using the proposed transfer learning method compared to training the model from scratch, in which VGG19 cannot even converge under the condition of training the model from scratch, while the proposed model still achieves the highest diagnostic accuracy among all models. Therefore, it can be determined that the use of migration learning methods can effectively improve the fault diagnosis accuracy, and the deep model constructed in this paper has the advantages of better robustness and higher fault diagnosis accuracy compared to the other four commonly used deep models.

4.4. Small Sample Migration between Different Operating Conditions

In order to verify the fault diagnosis performance of this paper's method in small samples and cross-case conditions, the model trained under Case A dataset is used as a pre-training model, which is migrated and applied to the datasets of Case B, C, and D, respectively, to accomplish the fault diagnosis task without changing the structure of the model and the training parameters.

In the migration process, each dataset is randomly assigned training and validation sets using a 1:4 ratio, and the model is trained on each dataset 10 times, with each iteration traversing 10 rounds. Comparison tests were conducted using the pre-trained model directly traversing the data sets with different operating conditions, and the results are shown in **Figure 13**.

As can be seen from **Figure 12**, the diagnostic results obtained by the fault diagnosis method taking small sample migration are superior to those of the direct application method in each migration task. The overall average accuracy of fault diagnosis methods using small sample migration was 93.5%. In the A to D migration task, the fault diagnosis accuracy is relatively low at 84.26% because the D condition is a complex condition with time-varying loads, while the dataset conditions applied in the A to B and A to C migration tasks are both time-invariant loads, and the diagnosis accuracies are 98.34% and 97.9%, respectively.

Table 4. Comparative experimental results of training from scratch.

model	average accuracy/%	standard deviation/%
AlexNet	76.5625	1.0925
VGG19	19.3065	2.1938
inception-V3	90.7813	0.7254
MobileNetV2	94.5	0.7540
methodology of this paper	97.0938	0.5902

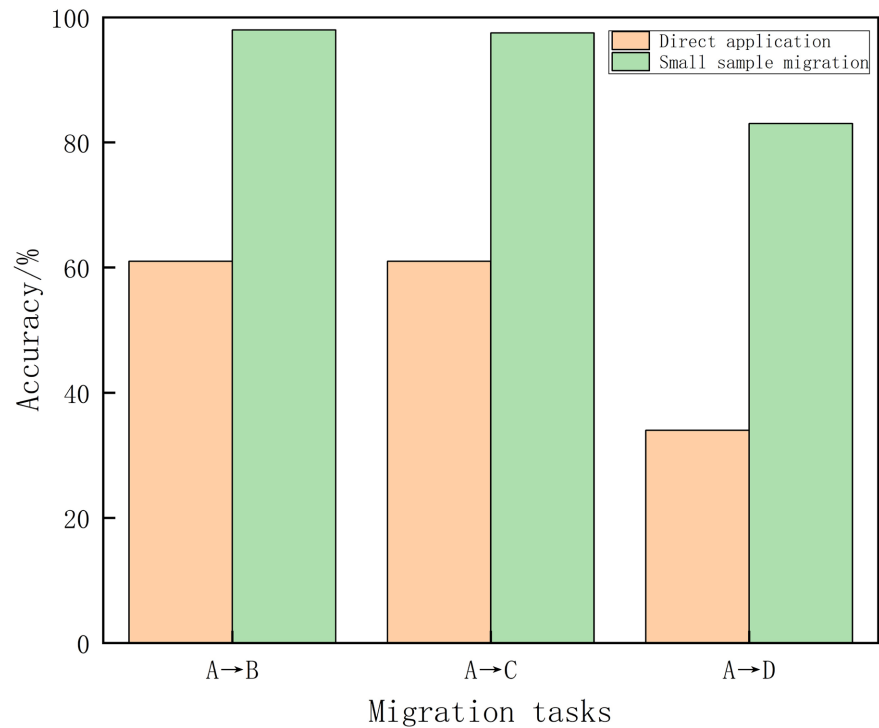


Figure 13. Small sample migration results between different operating conditions.

From the experimental results, it can be concluded that the method proposed in this paper can obtain high fault diagnosis accuracy in cross-case fault diagnosis with small training rounds by using only a small sample training set.

4.5. Effect of Small Sample Size on Migration Results

In deep learning, the size of the number of samples in the training set directly affects the performance of the model after it has been trained, in order to explore the effect of the proportion of small sample training set to the total data set on the training results, respectively, using 5%, 10%, 15%, 20%, 25% and 30% proportion of the division of the small sample training set, to complete the migration task with the A condition as the source domain, the other three conditions as the target domain, each task migration for 10 repetitions of the test, each time 10 rounds of the diagnostic accuracy of the average value of the results are shown in **Figure 14**, which is a direct application of the method for the proportion of the 0% training set.

From the experimental results, it can be seen that as the proportion of the training set to the dataset increases, the fault diagnosis accuracy for each task increases. Among them, the D dataset is more complex, and the model is better fine-tuned for the D task when the small samples are increased, which is particularly significant in improving the accuracy of the migration task. For the migration tasks A to B and A to C, the diagnostic accuracy is only about 60% when the method is applied directly, while the accuracy improves to nearly 90% when the small sample is 5%, stabilizes above 95% when the small sample is 10%, and

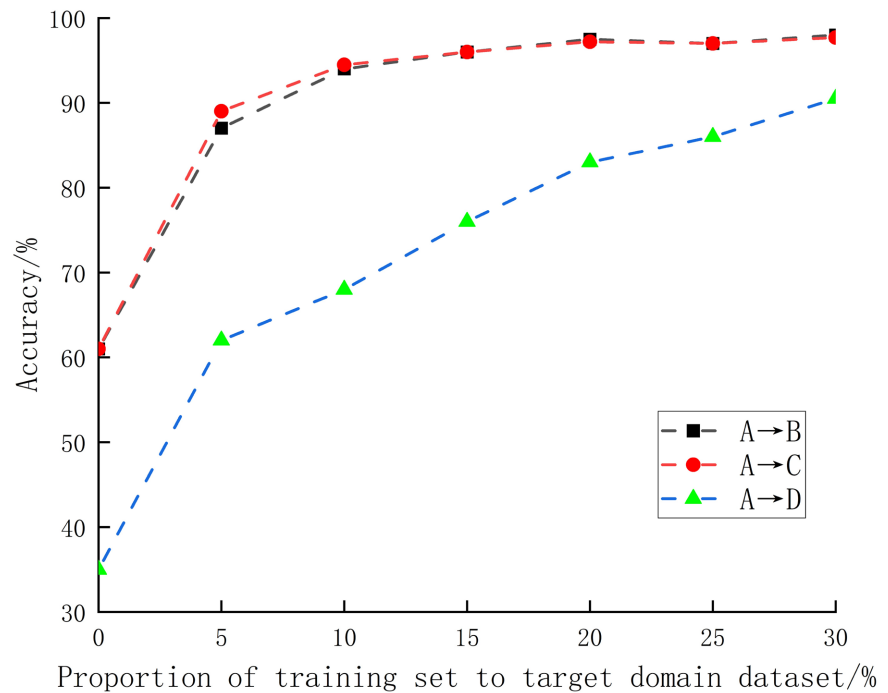


Figure 14. Effect of small sample size on migration results.

then the accuracy improvement slows down as the small sample size increases, and the improvement is not significant when the small sample size is 20% or more relative to 20%.

In summary, in the small-sample migration approach, the fault diagnosis accuracy continues to improve as the size of the target domain dataset accounted for by the small samples increases, and the improvement is particularly significant at ratios of 10% and above. Comparing the relationship between the rate of increase in diagnostic accuracy with small sample size for each task, it was concluded that a training set of 20% proportional size was most suitable for small sample migration.

5. Conclusions

In this paper, a fault diagnosis method based on improved ResNet50 with migration learning is proposed, and the feasibility of the method is verified by utilizing the vibration signals of gears in five different states under different load conditions, and the experimental results show that:

- 1) Applying deep learning models to the field of fault diagnosis through the migration learning method of pre-trained model parameter reuse can effectively improve the overfitting problem of the model on small sample datasets and improve the fault diagnosis performance.
- 2) The proposed improved deep learning model based on ResNet50 achieved 99.7% fault diagnosis accuracy in experimental validation on the gear dataset, outperforming four other commonly used deep learning models and ResNet50.
- 3) The proposed method carries out migration learning fault diagnosis across

operating conditions with 20% of the target domain dataset as the training set and only 10 rounds of traversal, and the overall diagnosis average accuracy is 93.5%, which fully reflects the advantages of this algorithm's strong generalization ability and high accuracy of fault diagnosis, and it can basically satisfy the needs of gear fault diagnosis.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Wang, Y., Yang, S. and Sanchez, R.V. (2018) Gearbox Fault Diagnosis Based on a Novel Hybrid Feature Reduction Method. *IEEE Access*, **6**, 75813-75823. <https://doi.org/10.1109/ACCESS.2018.2882801>
- [2] Li, C., Zhang, S.H., Qin, Y., *et al.* (2020) A Systematic Review of Deep Transfer Learning for Machinery Fault Diagnosis. *Neurocomputing*, **407**, 121-135. <https://doi.org/10.1016/j.neucom.2020.04.045>
- [3] Lecun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. <https://doi.org/10.1038/nature14539>
- [4] Saufi, S.R., Ahmad, Z.A.B., Leong, M.S., *et al.* (2020) Gearbox Fault Diagnosis Using a Deep Learning Model with Limited Data Sample. *IEEE Transactions on Industrial Informatics*, **16**, 6263-6271. <https://doi.org/10.1109/TII.2020.2967822>
- [5] Jia, M.X., Xu, Y.M., Hong, M.Y., *et al.* (2020) Multitask Convolutional Neural Network for Rolling Element Bearing Fault Identification. *Shock and Vibration*, **2020**, Article ID: 1971945. <https://doi.org/10.1155/2020/1971945>
- [6] Jiang, G.Q., He, H.B., Yan, J., *et al.* (2018) Multiscale Convolutional Neural Networks for Fault Diagnosis of Wind Turbine Gearbox. *IEEE Transactions on Industrial Electronics*, **66**, 3196-3207. <https://doi.org/10.1109/TIE.2018.2844805>
- [7] Cao, P., Zhang, S.L. and Tang, J. (2018) Preprocessing-Free Gear Fault Diagnosis Using Small Datasets with Deep Convolutional Neural Network-Based Transfer Learning. *IEEE Access*, **6**, 26241-26253. <https://doi.org/10.1109/ACCESS.2018.2837621>
- [8] Shao, S.Y., Mcaleer, S., Yan, R.Q., *et al.* (2018) Highly Accurate Machine Fault Diagnosis Using Deep Transfer Learning. *IEEE Transactions on Industrial Informatics*, **15**, 2446-2455. <https://doi.org/10.1109/TII.2018.2864759>
- [9] Tang, S.N., Yuan, S.Q. and Zhu, Y. (2019) Deep Learning-Based Intelligent Fault Diagnosis Methods toward Rotating Machinery. *IEEE Access*, **8**, 9335-9346. <https://doi.org/10.1109/ACCESS.2019.2963092>
- [10] Weiss, K., Khoshgoftaar, T.M. and Wang, D.D. (2016) A Survey of Transfer Learning. *Journal of Big Data*, **3**, 1-40. <https://doi.org/10.1186/s40537-016-0043-6>
- [11] Zhao, B., Zhang, X.M., Zhan, Z.H., *et al.* (2020) Deep Multi-Scale Convolutional Transfer Learning Network: A Novel Method for Intelligent Fault Diagnosis of Rolling Bearings under Variable Working Conditions and Domains. *Neurocomputing*, **407**, 24-38. <https://doi.org/10.1016/j.neucom.2020.04.073>
- [12] He, Z.Y., Shao, H.D., Zhong, X., *et al.* (2020) Ensemble Transfer CNNs Driven by Multi-Channel Signals for Fault Diagnosis of Rotating Machinery Cross Working Conditions. *Knowledge-Based Systems*, **207**, Article 106396.

- <https://doi.org/10.1016/j.knosys.2020.106396>
- [13] Wen, L., Gao, L. and Li, X.Y. (2017) A New Deep Transfer Learning Based on Sparse Auto-Encoder for Fault Diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics. Systems*, **49**, 136-144. <https://doi.org/10.1109/TSMC.2017.2754287>
- [14] Lu, T., Yu, F., Han, B.K., *et al.* (2020) A Generic Intelligent Bearing Fault Diagnosis System Using Convolutional Neural Networks with Transfer Learning. *IEEE Access*, **8**, 164807-164814. <https://doi.org/10.1109/ACCESS.2020.3022840>
- [15] Yosinski, J., Clune, J., Bengio, Y., *et al.* (2014) How Transferable Are Features in Deep Neural Networks? *Advances in Neural Information Processing Systems*, **27**, 3320-3328.
- [16] He, K.M., Zhang, X.Y., Ren, S.Q., *et al.* (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. <https://doi.org/10.1109/CVPR.2016.90>
- [17] Sandler, M., Howard, A., Zhu, M.L., *et al.* (2018) Mobilenetv2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4510-4520. <https://doi.org/10.1109/CVPR.2018.00474>