Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P718-cd

A New Approach for Fault Diagnosis of Rolling Bearings Based on Adaptive Batch Normalization and Attention Mechanism

Jingwen Hu,

College of Intelligence Science and Technology, National University of Defense Technology, China. E-mail: hjw1028@163.com

Yashun Wang

College of Intelligence Science and Technology, National University of Defense Technology, China. E-mail: wangyashun@nudt.edu.cn

Xun Chen

College of Intelligence Science and Technology, National University of Defense Technology, China. E-mail: chenxun@nudt.edu.cn

This paper proposes a single branch transfer learning method with the noise reduction attention mechanism for cross-domain fault diagnosis of rolling bearing. First, adaptive batch normalization is added to the model to ensure its domain adaptation capability. Furthermore, to improve the model's ability that suppresses noise-related features in a noisy environment, the noise reduction attention mechanism is introduced. With sufficient experimental verifications carried out, the results support that our proposed method has satisfying performance.

Keywords: Rolling bearing, intelligent fault diagnosis, anti-noise, deep learning.

1. Introduction

Algorithms based on deep learning or machine learning are the prevalent methods in the development of rolling bearing fault diagnosis (Lei et al. 2020). An important premise of most existing methods is to assume that the training and the testing sets follow the same distribution. In practical industrial applications, however, general distribution discrepancies exist between training data and testing data due to variation in working conditions, interference of environmental noise, etc., which leads to a significant decline in diagnostic performance while using traditional methods. Domain shift caused by varying operating conditions brings new challenges to the fault diagnosis of rolling bearings. Recently, transfer learning models have been gradually applied to solve the problem of cross-domain fault diagnosis of rolling bearings. Li et al. (2021) proposed a two-stage transfer adversarial network for failure diagnosis of rotating machinery. This method draws on the idea of generative adversarial networks. Wen et al. (2019) used a sparse autoencoder model along with the maximum mean discrepancy for fault diagnosis of rolling bearing under different loads. Li et al. (2019) used the idea of generative adversarial networks to develop a transfer learning model for bearing fault diagnosis under the lack of data in the target domain. Wang et al. (2020) proposed a multi-scale deep intra-class method for dealing with the fault diagnosis of rolling bearings under different failure modes. The above methods provide a potential solution for solving the cross-domain fault mode diagnosis but are only able to fix a percentage of the problems by far.

Limitations of the existing methods are still in demand to be solved in real-life applications. The above-mentioned transfer learning models are almost all multi-branch network models. An additional branch could greatly complicate the training of a model due to the introduction of new parameters and losses, which is reflected in the cumbersome parameter adjustment and the

convergence and cost of the model. For generative adversarial networks, the model often cannot enter the Nash equilibrium state because of the overfitting of the discriminator, and the performance of the model is very unstable, which is the main problem faced by its promotion and application (Jeong and Shin 2021, Salimans et al. 2016). On the other hand, in one comparative experiment of rolling bearing fault diagnosis, the transfer learning model based on maximum mean discrepancy is proven to have the best performance, but it also costs the most computation (Wang, Michau, and Fink 2019). Another problem that needs attention is the service environment of rolling bearings is often accompanied by various vibrations and interference. Noise will reduce the ability of the model to extract advanced fault features, and this weak point will be amplified when dealing with transfer tasks under different working conditions, resulting in an underwhelming performance. However, the aforesaid methods do not introduce a noise reduction mechanism into the model in a targeted manner.

To further solve the above problems and explore a better algorithm for the fault diagnosis task of rolling bearings under different working conditions in a noisy environment, this paper proposes a new approach. The proposed method firstly adds an adaptive batch normalization to the deep network, which improves the domain adaptive ability of the model. Additionally, a noise reduction attention mechanism is specifically embedded in the model. This ensures the stability of the model so that the model can still learn crucial features in a noisy environment. The input of the proposed approach here is the original vibration signal, which does not require the additional design of hand-crafted features. This paper illustrates how the proposed method acts directly on the raw vibration signal, which can carry out end-to-end fault diagnosis task processing.

2. The Theoretical Framework for The Proposed Methods

2.1 *Problem Formulation*

The mathematical expression for the domain adaptation in cross-domain task is defined as the following. Given a labeled source domain dataset $D_S = \{(x_{s1}, y_{s1}),...,(x_{sn}, y_{sn})\}$, $y_{si} \in Y$, and an unlabeled target domain dataset $D_t = \{x_{t1},...,x_{tk}\}$, where y represents the label of the state, and Y is the set of all class labels {0, ..., M}, M is the fault label. The labels of the target domain are missing during training, so the ground truth labels, for the purpose of this paper, are denoted as $\{y_{t1},...,y_{tk}\}$, $y_{ti} \in Y$. The intention of domain adaptation is to use source domain labeled data D_s to learn a predictive model f on the target domain, which makes $f : x_t \to y_t$ has a smaller prediction error on the target domain, as shown in Eq. (1),

$$f^* = \arg\min_f \mathbb{E}_{(x,y)\in D_s} \epsilon\big(f(x), y\big) \tag{1}$$

2.2 *The Proposed Model*

2.2.1 *Adaptive Batch Normalization*

The Batch Normalization (BN) layer (Ioffe and Szegedy 2015) is used to alleviate the internal covariate shift problem when training deep neural networks. Formally, for a batch of data $B = \{(x_i, y_i)\}_{i=1}^m$, the goal of BN is to normalize each sample as follows,

$$\mu^{(j)} = \frac{1}{m}\sum_{i=1}^{m} x_i^{(j)} \tag{2}$$

$$\delta^2(j) = \frac{1}{m}\sum_{i=1}^{m}(x_i^{(j)} - \mu^{(j)})^2 \tag{3}$$

$$\hat{x}^{(j)} = \frac{x^{(j)} - \mu^{(j)}}{\sqrt{\sigma^2(j) + \epsilon}} \tag{4}$$

$$y^{(j)} = \gamma^{(j)}\hat{x}^{(j)} + \beta^{(j)} \tag{5}$$

To make BN more suitable for transfer learning tasks, Li et al. (2018) proposed the concept of adaptive batch normalization (AdaBN), which extends BN to domain adaptation problems. Inspired by the efficiency of AdaBN in the field of computer vision, this paper introduces AdaBN into the proposed model to deal with the fault diagnosis task of rolling bearings under different working conditions. The core of AdaBN is to

adopt domain-specific normalization for different domains, that is, first, to operate on the data of source domain *Ds* with BN, and then use the BN layer to update the corresponding statistics when processing the data of target domain *Dt*, as shown in Algorithm 1.

Algorithm 1. The pseudocode of AdaBN. Adapted from literature (Li et al. 2018).

For neuron *j* in model:
Collect the neuron responses $\{x_j(m)\}$ on all original vibration signal of D_t
Compute the mean and variance of D_t: μ_j^t and δ_j^t respectively by Eq. (2) Eq. (3)
end for
for neuron *j* in model, testing vibration signal *m* in D_t:
Compute BN output $y_j(m)$ by Eq. (5)
end for

2.2.2 *The Noise Reduction Attention Mechanism*

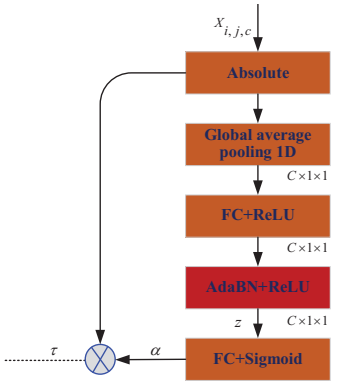


Figure 1. The noise reduction attention mechanism.

To overcome the problem that the model does not perform well under noisy environment, this paper designs a noise reduction attention mechanism. As shown in Fig. 1, the noise reduction attention mechanism provides a weight threshold τ for each channel of the feature map to filter the features. The input feature x will be reduced to a 1-D vector through an absolute operation and global average pooling layer, next, input to the fully connected layer, like (Hu, Shen, and Sun 2018). The number of neurons in the fully connected layer is set to be equal to the number of channels in the input feature map. The features that output of the fully connected layer will be scaled to (0, 1) by the following Eq. (6),

$$\alpha_c = \frac{1}{1+e^{-z_c}} \tag{6}$$

where α_c represents the c-th scaling parameter, and z_c represents the feature at the c-th neuron. The output threshold for each channel can be calculated as Eq. (7),

$$\tau_c = \alpha_c \cdot \underset{i,j}{average}\left|x_{i,j,c}\right| \tag{7}$$

where *i, j,* and *c* are the indexes of the width, the height, and the channel of the feature map x, respectively, and τ_c is the threshold for the *c*-th channel of the feature map. It has been verified by (Zhao et al. 2020) that the output threshold will always remain a positive value throughout the training process, and the value range will also be within a reasonable range, which ensures that the output features keep from all being zeros. The model can automatically learn the τ through backpropagation.

After obtaining τ, the model can use soft threshold segmentation (Isogawa et al. 2018) to filter the feature map to achieve the purpose of removing noise-related features, as shown in Eq. (8),

$$y = \begin{cases} x-\tau & x > \tau \\ 0 & -\tau \le x \le \tau \\ x+\tau & x < -\tau \end{cases} \tag{8}$$

here x and y represent the input and output features, respectively.

2.3.3 *Overall Architecture of the Proposed Model*

The overall architecture of the proposed method in this paper is shown in Fig. 2. The input of the model is the original vibration signal, and the artificial design for features is not required, which can accomplish end-to-end rolling bearing fault diagnosis. Embedding the AdaBN layer into the proposed model allows it to deal with transfer learning tasks that can realize cross-domain fault diagnosis of rolling bearings. The proposed model

does not have additional branches, and still retains the single-stream structure. Furthermore, the noise reduction attention mechanism is introduced into the proposed model to help it suppress noise-related features under noisy environments. By adding shortcut connections (He et al. 2016), the proposed method will enable the input signals to propagate from any bottom layer to higher layers, which is beneficial for the model to learn crucial local information and global information about the failure of rolling bearing. The output after model training is the fault label, and the total loss function is the cross-entropy loss, as shown in Eq. (9),

$$\mathcal{L} = -\sum_{i=1}^{N} p_i \log(q_i) \tag{9}$$

where p represents the targeted output, and p_i is the actual probability of an observation belonging to the i-th class, q_i denote estimated distribution.

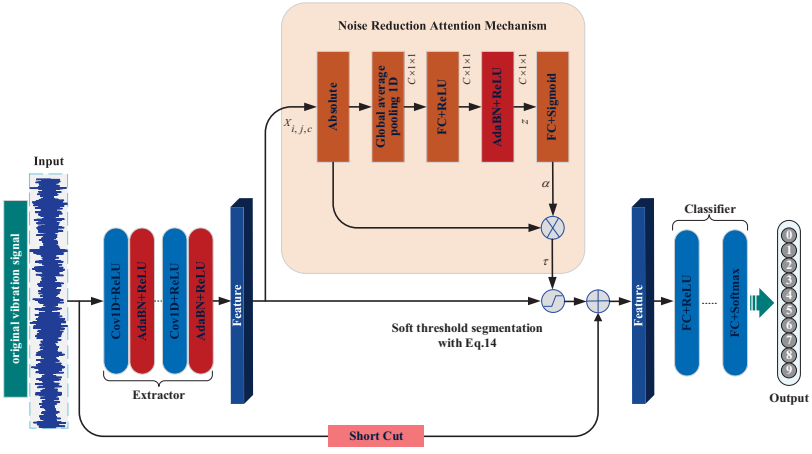


Figure 2. Overall architecture of the proposed model.

3. Experimental Design and Analysis

Two sets of experiments are conducted to evaluate the performance of the proposed model with the CWRU-bearing dataset (Smith and Randall 2015). Bearings in the dataset contain a total of 10 states, including 1 healthy state and 9 fault states. The CWRU bearing data were collected under four loads, which can be considered as four different working conditions {0, 1, 2, and 3hp}. The cross-domain fault diagnosis from one working condition to another is defined as a set of transfer learning tasks. For example, Task 0-1 is denoted as the following: the source domain is work condition 0, and the target domain is work condition 1. The first 12,000 points of the original vibration signal are selected as the total number of samples, and the overlapping sampling is performed to form 4,000 training samples of length 4,096 and 800 testing samples of the same length in the source and target domains, respectively.

The hyperparameters and structure of the model during the experiments are shown in Table 1. Except for the experimental results cited in the literature, the models participating in the comparison all use the same hyperparameter settings. Referring to the conclusions in the literature (Zhang et al. 2018), a wider convolution kernel 64*1 is set in the first layer of the model so that a larger receptive field can be obtained when processing the original time domain information. The model is optimized by the stochastic gradient descent algorithm (SGD), the initial value of the learning rate is set to be 0.01, the decay coefficient is set to be 1e-6, and the momentum is set to be 0.9. In the experiment, the models were trained for 1000 epoch and with a batch size of 128.

Table 1. Architecture and hyperparameters of the model in the experiments.

Layer	Kernel size/stride/ hyperparameters	Output
Input	/	[(None, 4096, 1)]
Conv1D	64*1 / 1, relu	(None, 4096, 16)
Batch Normalization	/	(None, 4096, 16)
Conv1D	5*1 / 1, relu	(None, 4096, 32)
Batch Normalization	/	(None, 4096, 32)
Conv1D	5*1 / 1, relu	(None, 4096, 64)
Batch Normalization	/	(None, 4096, 64)
the noise reduction attention	/	(None, 4096, 64)
MaxPooling1D	2*1 / 1	(None, 4096, 64)
Flatten	/	(None, 262144)
Fully-connected	128, relu	(None, 128)
Dropout	0.5	(None, 128)
Batch Normalization	/	(None, 128)
Fully-connected	10, softmax	(None, 10)

3.1 *Results and Discussion*

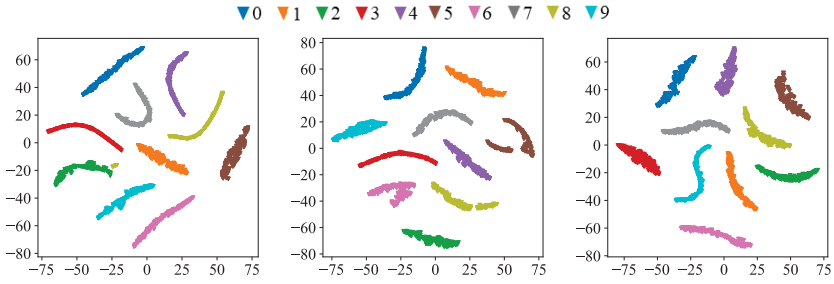
3.1.1 *The Experiment on the CWRU Bearing Dataset*

Existing methods proposed in the literature were used for comparison with the models proposed in this paper, including Generative two-stage (Li, Zhang, and Ding 2019), DANN (Li et al. 2018, Wang, Michau, and Fink 2019), DAUA (Wang, Michau, and Fink 2021). Taking the classification accuracy on the test set as the evaluation index of performance, the comparison results of each method are shown in Table 2, such as 0-1, 1-0, and 1-2. The proposed method also achieved 100% transfer accuracy on tasks 1-2 and 2-1. The results indicate that the proposed method achieved the highest accuracy on multiple transfer tasks among the five methods. Overall, the transfer accuracy of the proposed method on 12 tasks both all exceeded 95%, which shows that the proposed method has good applicability and stability.

Table 2. Comparison of transfer accuracy on original CWRU datasets.

	Generative-two-stage(Li, Zhang, and	DANN(Ganin et al. 2016)	DAUA(Wang, Michau, and Fink 2021)	Our Method
0-1	97.81	97.27±0.76	98.08±0.16	**99.12**
0-2	96.02	97.86±1.72	**99.56±0.18**	99.50
0-3	94.24	96.97±2.91	98.22±0.65	**99.38**
1-0	97.27	97.82±0.16	98.08±0.32	**99.62**
1-2	96.32	99.95±0.06	**100.00±0.00**	**100.00**
1-3	94.59	98.93±0.36	99.20±0.19	**99.56**
2-0	95.44	92.99±1.25	96.43±0.43	**96.75**
2-1	96.55	97.55±0.40	97.48±0.40	**100.00**
2-3	96.13	**99.62±0.21**	98.97±0.21	99.37
3-0	92.82	87.89±0.53	94.85±2.16	**96.49**
3-1	93.04	92.53±2.45	96.18±0.50	**97.37**
3-2	95.63	**99.90±0.07**	99.78±0.09	99.74

Taking tasks 0-1, 0-2, and 0-3 as examples, T-SNE was used to visualize output features of the final layer of our model proposed in this paper, as shown in Fig. 3. The proposed model can well distinguish the features between different failure modes in the target domain. Combining the analysis of the results in Table 2, further proved that the method proposed in this paper is feasible and effective in dealing with the task of cross-domain fault diagnosis for rolling bearings.



(a) task 0-1 (b) task 0-2 (c) task 0-3

Figure 3. Visualization of the output features at the final layer with T-SNE in tasks 0-1, 0-2, and 0-3 on original data.

3.1.2 *The Experimental on the CWRU Bearing Dataset with Noise*

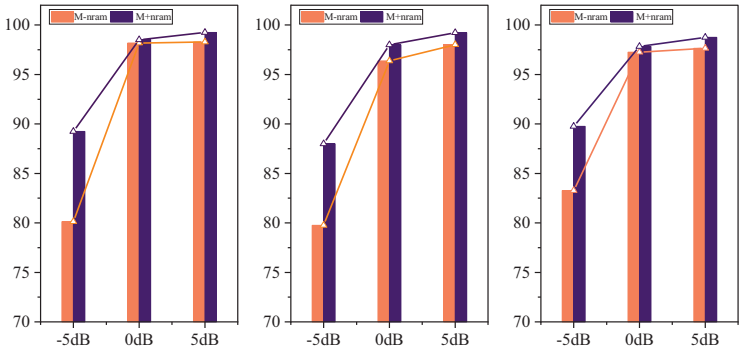
The purpose of this experiment is to demonstrate the performance of the proposed method in noisy environments and to validate the effectiveness of the noise reduction attention mechanism. By adding SNR -5dB, SNR 0dB, and SNR 5dB Gaussian noise to the original CWRU bearing dataset, three new noisy datasets were constructed. The model with the noise reduction attention mechanism was compared with the model without the noise reduction attention mechanism, and the hyperparameter settings were the same as in Experiment 1, see Table 1. For the convenience of representation, the model with the noise reduction attention mechanism is designated as M+nram, and the model without the noise reduction attention mechanism is designated as M-nram. The experiment results are shown in Table 3. It can be seen that after adding the noise reduction attention mechanism, the anti-noise ability of the model has been greatly improved. The experimental results show that the noise reduction attention mechanism in the model allows the model to have better performance even in noisy environments.

Table 3. Comparison of transfer accuracy on noisy datasets.

	M-nram			M+nram		
	-5dB	0dB	5dB	-5dB	0dB	5dB
0-1	83.26	97.25	97.65	89.75	97.84	98.75
0-2	80.12	98.17	98.29	89.24	98.51	99.25
0-3	79.75	96.38	98.02	88.00	98.00	99.23
1-0	82.93	97.11	98.12	90.87	98.73	99.62
1-2	82.77	99.37	99.37	90.77	99.72	99.85
1-3	84.00	93.77	94.50	89.63	98.59	99.56
2-0	76.75	91.39	94.10	90.61	93.00	95.38
2-1	76.25	97.33	97.61	91.11	98.31	99.87
2-3	86.11	99.12	99.12	92.37	98.24	99.07
3-0	75.13	87.63	89.25	85.12	95.39	96.13
3-1	75.82	88.63	89.38	90.49	96.25	97.25
3-2	86.62	98.66	99.13	92.62	96.38	99.33

Taking tasks 0-1, 0-1, and 0-3 as examples, the transfer performance of the two model under different SNR noises was analyzed. The results are shown in Fig. 4. It can be seen that after introducing the noise reduction attention mechanism, the performance of the model under noise environment significantly improved. Additionally, with the increase in SNR, the

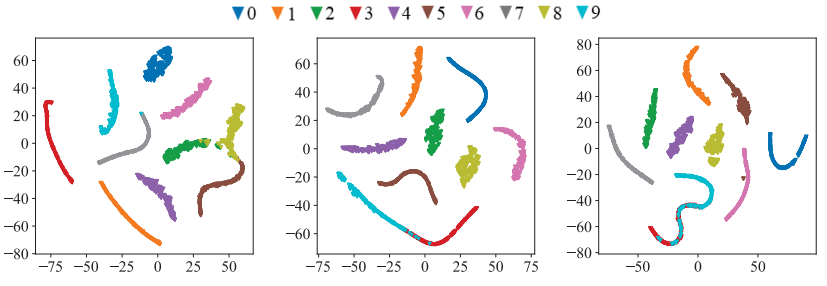
performance of the model gradually returned to the normal level. This phenomenon shows that strong noise can seriously degrade the performance of the model, so it is essential to introduce a noise reduction attention mechanism.



(a) task 0-1 (b) task 0-2 (c) 0-3 task

Figure 4. Comparison of transfer accuracy between AdaBN and our method on noise datasets.

Also taking 0-1, 0-2, and 0-3 as examples, T-SNE is used to visualize the final output of the model on the SNR -5 noisy dataset, as shown in Fig. 5. With the help of the noise reduction attention mechanism, the proposed model was capable to distinguish the failure modes of the rolling bearing under different working conditions even in a noisy environment. A comprehensive analysis of Table 3, Fig. 4, and Fig. 5 shows that the performance of the model improved in noisy environments after introducing the noise reduction attention mechanism.



(a) task 0-1 (b) task 0-2 (c) task 0-3

Figure 5. Visualization of the output features at the final layer with T-SNE in tasks 0-1, 0-2, and 0-3 on SNR -5dB noise datasets.

4. Conclusion

This paper presents a new approach that addresses the aforementioned issues by introducing the AdaBN layer and the noise reduction attention mechanism into the deep network. The effectiveness and feasibility of the proposed method are verified on the CWRU-bearing dataset and the generated noisy dataset. In addition, the method proposed in this paper requires only a single branch and is easier to implement and train. And the input of the model can be the original vibration signal which enables end-to-end fault diagnosis task processing.

In the future, the integration of AdaBN and the noise reduction attention mechanism into other models should be further investigated to verify the general application of the methods.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China (52275116).

References

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor %J The journal of machine learning research Lempitsky. (2016). "Domain-adversarial training of neural networks." *The journal of machine learning research* 17 (1):2096-2030.

He, K., X. Zhang, S. Ren, and J. Sun. (2016). "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 June 2016.

Hu, J., L. Shen, and G. Sun. (2018). "Squeeze-and-Excitation Networks." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18-23 June 2018.

Ioffe, Sergey, and Christian Szegedy. 2015. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning.

Isogawa, K., T. Ida, T. Shiodera, and T. Takeguchi. (2018). "Deep Shrinkage Convolutional Neural Network for Adaptive Noise Reduction." *IEEE Signal Processing Letters* 25 (2):224-228. doi: 10.1109/LSP.2017.2782270.

Jeong, Jongheon, and Jinwoo %J arXiv e-prints Shin. (2021). Training GANs with Stronger Augmentations via Contrastive Discriminator. arXiv:2103.09742. Accessed March 01, 2021.

Lei, Yaguo, Bin Yang, Xinwei Jiang, Feng Jia, Naipeng Li, and Asoke K. Nandi. (2020). "Applications of machine learning to machine fault diagnosis: A review and roadmap." *Mechanical Systems and Signal Processing* 138:106587. doi: https://doi.org/10.1016/j.ymssp.2019.106587.

Li, J. P., R. Y. Huang, G. L. He, Y. X. Liao, Z. Wang, and W. H. Li. (2021). "A Two-Stage Transfer Adversarial Network for Intelligent Fault Diagnosis of Rotating Machinery With Multiple New Faults." *IEEE-Asme Transactions on Mechatronics* 26 (3):1591-1601. doi: 10.1109/tmech.2020.3025615.

Li, X., W. Zhang, and Q. Ding. (2019). "Cross-Domain Fault Diagnosis of Rolling Element Bearings Using Deep Generative Neural Networks." *IEEE Transactions on Industrial Electronics* 66 (7):5525-5534. doi: 10.1109/TIE.2018.2868023.

Li, Yanghao, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. (2018). "Adaptive Batch Normalization for practical domain adaptation." *Pattern Recognition* 80:109-117. doi: https://doi.org/10.1016/j.patcog.2018.03.005.

Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. (2016). "Improved techniques for training GANs." Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain.

Smith, Wade A., and Robert B. Randall. (2015). "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study." *Mechanical Systems and Signal Processing* 64-65:100-131. doi: https://doi.org/10.1016/j.ymssp.2015.04.021.

Wang, Q., G. Michau, and O. Fink. (2019). "Domain Adaptive Transfer Learning for Fault Diagnosis." 2019 Prognostics and System Health Management Conference (PHM-Paris), 2-5 May 2019.

Wang, Q., G. Michau, and O. Fink. (2021). "Missing-Class-Robust Domain Adaptation by Unilateral Alignment." *IEEE Transactions on Industrial Electronics* 68 (1):663-671. doi: 10.1109/TIE.2019.2962438.

Wang, X., C. Q. Shen, M. Xia, D. Wang, J. Zhu, and Z. K. Zhu. (2020). "Multi-scale deep intra-class transfer learning for bearing fault diagnosis." *Reliability Engineering & System Safety* 202. doi: 10.1016/j.ress.2020.107050.

Wen, L., L. Gao, and X. Li. (2019). "A New Deep Transfer Learning Based on Sparse Auto-Encoder for Fault Diagnosis." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49 (1):136-144. doi: 10.1109/TSMC.2017.2754287.

Zhang, Wei, Chuanhao Li, Gaoliang Peng, Yuanhang Chen, and Zhujun Zhang. (2018). "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load." *Mechanical Systems and Signal Processing* 100:439-453. doi: https://doi.org/10.1016/j.ymssp.2017.06.022.

Zhao, M., S. Zhong, X. Fu, B. Tang, and M. Pecht. (2020). "Deep Residual Shrinkage Networks for Fault Diagnosis." *IEEE Transactions on Industrial Informatics* 16 (7):4681-4690. doi: 10.1109/TII.2019.2943898.