Few-shot Learning for Plastic Bearing Fault Diagnosis – An Integrated Image Processing and NLP Approach

David He[1] and Miao He[2]

[1] *The University of Illinois-Chicago, Chicago, IL, 60607, USA*
davidhe@uic.edu

[2] *Siemens Technology, Princeton, NJ, 08540, USA*
miao.he@siemens.com

ABSTRACT

Plastic bearings have a wide range of industrial applications due to their many desirable properties such as lightweight, low friction coefficient, chemical resistance, and ability to operate without lubrication. Timely bearing fault diagnosis can prevent equipment failure and costly downtime. In recent years, developing machine learning based bearing fault diagnosis with few labelled data has attracted a lot of attention as datasets with fault labels are rare in many industrial applications. One effective approach to meet the challenge is few-shot learning. Among many approaches, utilizing a good pre-trained deep learning model to achieve few-shot learning is an effective and efficient alternative. In this paper, a pre-trained deep learning model called CLIP that combines image processing and natural language processing (NLP) is adopted to few-shot learning for plastic bearing fault diagnosis. We explore the feasibility of leveraging CLIP model in the realm of bearing fault diagnosis via few-shot learning. Specifically, we tackle the challenges posed by CLIP's creation of requisite text prompt embeddings for the diagnosis of mechanical faults, within a few-shot learning framework. Our investigation illuminates the remarkable capability of CLIP to adapt to new tasks with minimal examples, a feature we exploit to devise a solution for plastic bearing fault diagnosis. The effectiveness of the few-shot learning method with CLIP is demonstrated using vibration data collected from plastic bearing seeded fault tests in the laboratory.

1. INTRODUCTION

Bearings play a crucial role in the performance and reliability of various types of machinery. Plastic bearings have a wide range of industrial applications due to their many desirable properties, including their lightweight, low friction coefficient, chemical resistance, and ability to operate without lubrication. Here are some industrial applications where plastic bearings are commonly used: (1) Plastic bearings are commonly used in food processing equipment due to their ability to resist corrosion and chemicals. (2) Plastic bearings are used in various medical equipment such as X-ray machines, MRI scanners, and dental chairs. They offer excellent corrosion resistance and are non-magnetic, making them ideal for medical applications. (3) Plastic bearings are used in packaging machinery such as bottling and filling machines. They offer low friction and are capable of withstanding high-speed rotations, making them suitable for high-speed packaging lines. (4) Plastic bearings are used in various automotive applications such as power steering systems, suspension systems, and door hinges. They offer excellent wear resistance and reduce noise and vibration. Unlike their steel counterparts, limited research has been conducted in developing effective fault diagnosis methods for plastic bearings.

Recent developments in machine learning and deep learning provide a great opportunity to develop effective fault diagnosis methods for plastic bearing fault diagnosis. However, machine learning and deep learning algorithms require a large amount of labeled data for training, which is time-consuming and expensive to collect. Few-shot learning, which aims to learn from a few labeled examples, has recently emerged as a promising solution to this problem. Among many approaches for implementing few-shot learning, transfer learning is becoming an attractive one for many industrial applications. Transfer learning is a technique that repurposes and fine-tunes a model pre-trained on a large dataset (often on a different but related task) with a smaller dataset for a specific task. The idea of transfer learning is to allow the model "transfer" knowledge learned from the larger dataset to the new task.

In recent years, a surge in the development and success of deep learning models, pre-trained on expansive datasets, has been observed across a multitude of applications. Notable examples include GPT-3, excelling in natural language processing (NLP) tasks, and VGG16 and ResNet50, which have become staples in the field of image processing. The emergence of these sophisticated pre-trained models has propelled transfer learning to the forefront as an immensely promising approach for tackling few-shot learning problems. This enables the leveraging of insights gleaned from comprehensive training sets to effectively handle tasks where the availability of labeled data is scarce. The recent success of OpenAI's CLIP model serves to underline the efficacy of this approach. CLIP represents an innovative blend of computer vision and natural language processing, trained on a vast corpus of internet text and images. This extensive training enables it to understand and link images and their textual descriptions in a versatile and flexible manner. When applied in a transfer learning context, the CLIP model, with its pre-existing knowledge of general visual and textual patterns, can be adapted to more specific tasks with limited data - the essence of few-shot learning. The model's ability to generate meaningful representations (embeddings) for both images and their corresponding textual prompts is key to its power in these applications. Thus, the use of models like CLIP further underscores the potential of transfer learning, particularly in scenarios where data availability is a challenge.

In this paper, we propose to exploit the transfer learning capability of CLIP with few-shot learning for plastic bearing fault diagnosis. We will investigate the issues and propose our approach to creating few-shot learning solutions for plastic bearing fault diagnosis. To the best of our knowledge, this is the first attempt to apply the pre-trained CLIP deep learning structures to few-shot learning for plastic bearing fault diagnosis.

2. RELATED WORK

In this section, we first introduce some related work on few-shot learning in the context of transfer learning. We then explain the learning and classification mechanism of CLIP using a simple example. At the end, a brief review of previous research on plastic bearing fault diagnosis is provided.

2.1. Few-shot Learning for Fault Diagnosis

The goal of few-shot learning for fault diagnosis is to diagnose faults in a system or machine with high accuracy using only a small amount of labelled data. Typically, few-shot learning for fault diagnosis involves using transfer learning, domain adaptation, and meta-learning. Transfer learning can be used to transfer knowledge learned from one domain to another, while domain adaptation can be used to adapt a model trained on one domain to another. Meta-learning can be used to learn how to learn, enabling a model to quickly adapt to new domains with few samples.

One of the popular few-shot learning methods is *K*-way *N*-shot learning, which is used to classify new samples based on a small set of labeled samples (*N*) taken from *K* different categories.

Liu *et al*. (2021) conducted a survey on meta-learning for few-shot cross-domain fault diagnosis. Their paper provides an overview of various algorithms, including model-agnostic meta-learning (MAML), Reptile, and Prototypical Networks, which have been shown to be effective in adapting to new fault diagnosis tasks with limited labeled data. It suggests that meta-learning has great potential for few-shot cross-domain fault diagnosis. Feng *et al*. (2022) explored the use of meta-learning for fault diagnosis. The paper reviews various meta-learning algorithms and their applications in fault diagnosis across different domains, such as manufacturing, aerospace, and automotive industries. The authors also discussed the prospects of meta-learning for fault diagnosis and highlighted some of the challenges that need to be addressed to make this approach more practical and effective. For example, they pointed out that the lack of labeled data and the high cost of obtaining it are significant barriers to the adoption of meta-learning in industrial settings. Yan *et al*. (2021) proposed a few-shot learning framework for fault diagnosis in industrial machine. The proposed framework is based on a transformer architecture with attention mechanisms and uses contrastive learning to learn a feature representation that can discriminate between normal and faulty conditions. It is trained on a small labeled dataset and can quickly adapt to new machines with few labeled samples. Domain shift caused by changes in machine speed can be handled by the proposed framework. Hu *et al*. (2019) explored the use of external data and fine-tuning for improving the performance of few-shot learning pipelines. They found that fine-tuning the image encoder on the target task can improve performance of few-shot learning.

Even though few-shot learning doesn't not always involve transfer learning, transfer learning often forms the basis for few-shot learning approaches. By leveraging a model that has been pre-trained on a large and diverse dataset, we can extract useful features or representations that can be used to quickly adapt to new tasks, even when only a few examples are available - thus enabling few-shot learning.

2.2. The CLIP Architecture and Its Learning and Classification Mechanism

CLIP (Contrastive Language–Image Pretraining) was first introduced by OpenAI in a paper entitled "Learning transferable visual models from natural language supervision," in June 2021 (Radford et al, 2021). It is a cutting-edge artificial intelligence model that integrates image processing and NLP in a meaningful way.

CLIP uses a learning mechanism known as contrastive learning. The idea is to train an image encoder and a text encoder to generate embeddings for a given image-text pair that are similar to each other. More formally, for a given pair (image, text) that are a match, the model is trained to make the image embedding (representation) and the text embedding (representation) close in the embedding space, while pushing apart the representations for pairs that do not match.

Here we use a simple example to explain how CLIP works. Suppose we want to build a CLIP model to classify 3 types of objects: bottles, book, and cups. It would start with generating training pairs (image, text prompt) as the examples shown in Figure 1.

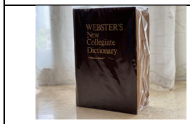
Image	Text prompt
	“This is a bottle.”
	“This is a book.”
	“This is a cup.”

Figure 1. Image and text prompt pair examples

The training process of a CLIP model is shown in Figure 2. As shown in Figure 2, for each pair in the training dataset, the image is passed through the image encoder to generate image embeddings: I_1, I_2, and I_3 and the text prompt is passed through the text encoder to generate text embeddings: T_1, T_2, and T_3. Various image processing and NLP models can be used as the image encoder and text encoder, respectively. Currently, CLIP uses a transformer model as its text encoder and a vision transformer (ViT) or ResNet50 as its image encoder.

The goal of the training process is to make the similarity high for matching image-text pairs, e.g., (I_1, T_1), (I_2, T_2), (I_3, T_3) and low for non-matching pairs, e.g., (I_1, T_2, (I_1, T_3), etc. Therefore, the model uses a contrastive loss function, which encourages the model to make the embeddings of matching pairs similar and the embeddings of non-matching pairs different. Various metrics can be used to measure the similarity between the embeddings of an image-text pair. For example, cosine similarity can be used. Let **I** be the embedding of an image and **T** be the embedding of a text prompt. Then the cosine similarity between the image and the text prompt can be computed as:

$$\frac{\mathbf{I}\mathbf{T}^{T}}{\|\mathbf{I}\|_2\|\mathbf{T}\|_2} \qquad (1)$$

As the result of the training, the cosine similarity matrix in Figure 2 should have its highest values on the diagonal elements that are highlighted with blue color.

After the training, the text embedding obtained will be used to classify the new unlabeled images. The classification process is shown in Figure 3.

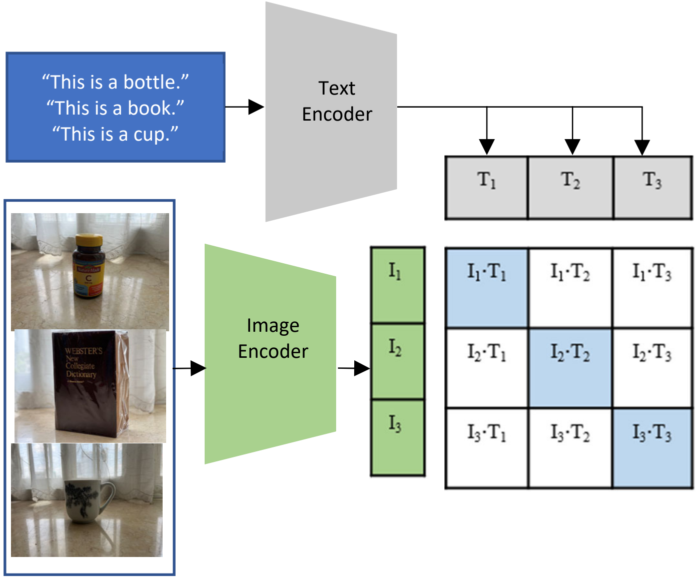


Figure 2. The training process of CLIP

For the simple example with the 3 objects, after the training, we obtained the following cosine similarity matrix:

0.2791	0.2190	0.2125
0.1891	0.2493	0.2137
0.1892	0.2151	0.2603

As shown in the similarity matrix, the diagonal elements in the matrix have the highest values. This indicates that the training was successful, the text prompts of the corresponding images were appropriate, and the correct text embeddings were generated.

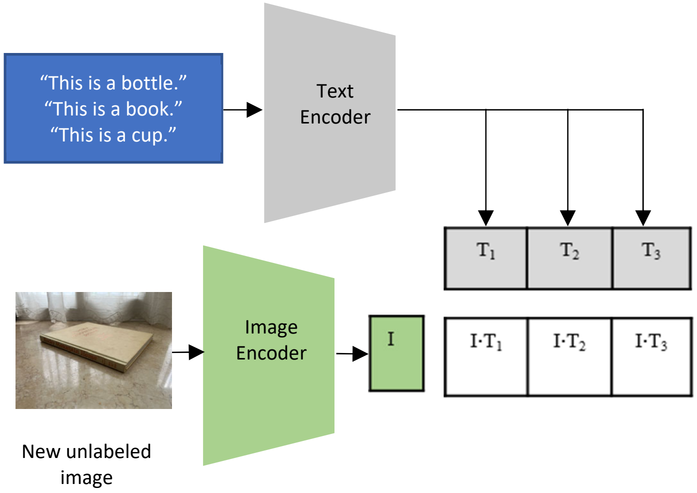


Figure 3. The classification process of CLIP

For the classification showcase, we chosen a book image as the new unlabeled image. After the image was passed through the image encoder, the image embedding of the new image was generated and the cosine similarity was computed as following:

0.1879	0.2437	0.2133

From the above computed cosine similarity scores, we can see that since the new image has the highest similarity score with book, the new image was classified as a book.

OpenAI's CLIP model was trained on a large dataset of 400 million image-text pairs collected from the internet. This extensive dataset is essential for the model's ability to generalize from text to images and perform learning tasks, as it has been exposed to a wide variety of concepts and associations between texts and images during training. The CLIP model pre-trained with huge dataset represents a great opportunity for transfer learning in the field of PHM. Many successful applications of CLIP have been reported since its introduction in June 2021. It has been leveraged for tasks such as zero-shot image classification, where it has demonstrated impressive results by using only textual descriptions of classes without requiring explicit examples of each class. Furthermore, when used with text generation models like GPT-3, CLIP can even produce image descriptions or generate images from textual descriptions.

Despite its impressive capabilities and potential for transfer learning, CLIP is not without limitations, particularly in specialized domains such as mechanical fault diagnosis. The challenge arises from the difficulty in finding suitable image-text pairs on the internet that adequately describe specific mechanical faults. The scarcity of these relevant pairs could potentially limit the model's ability to accurately understand and categorize images related to mechanical faults. Hence, while CLIP holds immense potential, its application in such specialized fields may require additional strategies to effectively capture domain-specific knowledge.

In this paper, we explore the feasibility of leveraging OpenAI's CLIP model in the realm of mechanical fault diagnosis via transfer learning. Specifically, we tackle the challenges posed by CLIP's creation of requisite text prompt embeddings for the diagnosis of mechanical faults, within a few-shot learning framework. Our investigation illuminates the remarkable capability of CLIP to adapt to new tasks with minimal examples, a feature we exploit to devise a solution for plastic bearing fault diagnosis. Our work thereby underscores the potential of this innovative approach within the context of mechanical fault diagnosis.

2.3. Previous Research on Plastic Bearing Fault Diagnosis

Previous research on plastic bearing fault diagnosis using data mining method was conducted by He, Li, and Zhu (2013). In their two-step data mining approach, frequency domain features were first extracted from vibration signals to classify the bearing faults into either outer race faults or non-outer race faults. Then the empirical mode decomposition (EMD) approach was applied to extract time domain features to further classify non-outer race faults into inner race fault, ball fault, and cage fault with *K*-nearest neighbor (KNN) approach. Their method provided good classification accuracy as high as to 97.1%. In their study, the fault diagnosis was done at each input speed separately and the training was not done with mixed speed data. Therefore, their study doesn't involve transfer learning.

3. THE METHODOLOGY

As discussed in Section 2.2, even though the CLIP model was pre-trained with 400 million image-text pairs collected from the internet, it is still difficult to associate effective text prompts for bearing fault images with those in the CLIP's pre-training dataset. Therefore, it will not be useful if we use any text prompt for our bearing fault diagnosis. For example, for the plastic bearing fault diagnosis, we could have the following text prompts for each of the 5 bearing conditions:

"It is a ball fault." ← text prompt for ball fault

"It is a cage fault." ← text prompt for cage fault

"It is a normal bearing." ← text prompt for normal bearing

"It is an inner race fault." ← text prompt for inner race fault

"It is an outer race fault." ← text prompt for outer race fault

By using these text prompts, the following cosine similarity matrix is generated using CLIP model:

0.2277	0.2323	0.2384	0.2266	0.2266
0.2261	0.2319	0.2357	0.2252	0.2352
0.2257	0.2318	0.2367	0.2251	0.2352
0.2266	0.2335	0.2354	0.2255	0.2351
0.2224	0.2274	0.2334	0.2208	0.2318

As we can see from the above cosine similarity matrix that the highest similarity score doesn't show on the diagonal elements of the matrix. Instead, the 3rd column of the matrix contains the highest similarity scores. This means that all the images of the bearing conditions are most likely associated with a normal bearing. The reason for this could be that there are much more images in the 400 million pre-training dataset used for CLIP model associated with words "normal" and "bearing" than words like "inner race", "outer race", "cage fault" and "ball fault".

To overcome the limitation of using CLIP model for plastic bearing fault diagnosis with few-shot learning in creating propriate text prompts, our approach is to use the image embeddings as the text embeddings and adding wavelet coefficient statistics into the text embeddings. The rationale behind this is that the image embeddings are the good representation of the faults in the embedding space. In addition, by adding wavelet coefficient statistics into the text embedding, we enhance the representation of the faults in the embedding space.

The proposed few-shot learning methodology with pre-trained CLIP model is presented in Figure 4.

As we can see from Figure 4, the vibration data is first processed using continuous wavelet transform (CWT) to extract CWT coefficients and convert the vibration data into corresponding scalogram images. The images are then passed through CLIP to generate image embeddings. The following statistics: mean, variance, skewness, kurtosis, and entropy, are then computed from CWT coefficients and added into the image embeddings to form the text embeddings. The image embeddings and text embeddings are then used to compute the cosine similarity matrix. Once the cosine similarity matrix is computed, any incoming unlabeled images will be classified by comparing their image embeddings with the cosine similarity matrix.

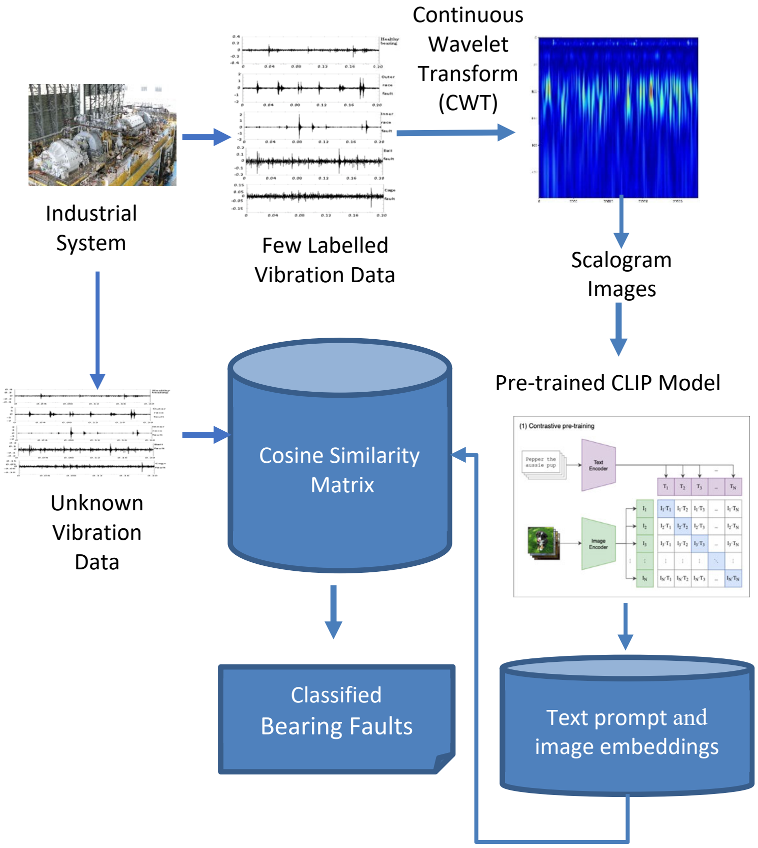

Industrial System
Few Labelled Vibration Data
Continuous Wavelet Transform (CWT)
Scalogram Images
Pre-trained CLIP Model
Unknown Vibration Data
Cosine Similarity Matrix
Classified Bearing Faults
Text prompt and image embeddings
(1) Contrastive pre-training
Text Encoder
Image Encoder

Figure 4. The flowchart of the methodology

Let $x(t)$ be the vibration signal value at time $t, t = 1, \dots, N$. Then CWT coefficient $\omega(t, s)$ at time t and scale s can be computed as:

$$\omega(t,s) = \frac{1}{\sqrt{s}} \int x(u)\psi[(u-t)/s]du \qquad (2)$$

where: $\psi(u-t)$ is the scaled and translated wavelet function evaluated at $u-t$.

The CWT coefficients then can be used to compute at scale s the following time-frequency features: mean, variance, skewness, kurtosis, and entropy, respectively:

$$\omega_{mean}(s) = \frac{1}{N}\sum_{t=1}^{N} \omega(t,s) \qquad (3)$$

$$\omega_{var}(s) = \frac{1}{N}\sum_{t=1}^{N}[\omega(t,s) - \omega_{mean}(s)]^2 \qquad (4)$$

$$\omega_{skew}(s) = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{\omega(t,s)-\omega_{mean}(s)}{\omega_{var}(s)}\right]^3 \qquad (5)$$

$$\omega_{kurtosis}(s) = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{\omega(t,s)-\omega_{mean}(s)}{\omega_{var}(s)}\right]^4 - 3 \qquad (6)$$

$$\omega_{entropy}(s) = -\sum_{t=1}^{N} p(t,s) * \log_2 p(t,s) \qquad (7)$$

Note that in Eq. (7), $p(t,s)$ is defined as normalized probability and can be computed as follow:

$$p(t,s) = \frac{|\omega(t,s)|^2}{\sum_{u=1}^{u=N} |\omega(u,s)|^2} \tag{8}$$

The normalization by Eq. (8) basically is to ensure that the sum of probabilities across all time points at a given scale s is equal to 1.

The CWT coefficients are also used to convert the vibration signals into scalogram images as follow:

$$scalogram\ \ image(t,s) = normalize(|\omega(t,s)|^2) \tag{9}$$

Eq. (9) computes values to create a scalogram image where the x-axis represents the time or position and the y-axis represents the scale. Function normalize() is to scale the coefficients to a suitable range for visualization purposes. Eq. (9) basically assigns the normalized coefficients to the corresponding positions in the image, where the intensity or color value represents the magnitude or power of the coefficient.

Since a CWT scalogram image of a vibration signal only captures the magnitude of the CWT coefficients, to maximize the similarity among the matched pairs or dissimilarity among unmatched pairs, additional CWT coefficient features as expressed in Eq. (3) – Eq. (7) will be added to the image embedding to form the text embeddings of the (image, text prompt) pairs. Finally, the text embedding is computed as:

$$text_{embedding} = image_{embedding} + \omega_{mean} + \omega_{var} + \omega_{skew} + \omega_{kurtosis} + \omega_{entropy} \tag{10}$$

4. THE DATASET AND ANALYSIS RESULTS

4.1. The Plastic Bearing Dataset

To evaluate the performance of the proposed methodology for plastic bearing fault diagnosis, vibration signals collected from plastic bearing seeded fault tests performed on a bearing test rig in the laboratory are used. For vibration data acquisition, two 603C01 wide range accelerometers and a data acquisition card NI PCI-4472B were used. The accelerometers were mounted on the surface of the bearing housing. Figure 5 shows the bearing test rig and the vibration sensors on the bearing housing.

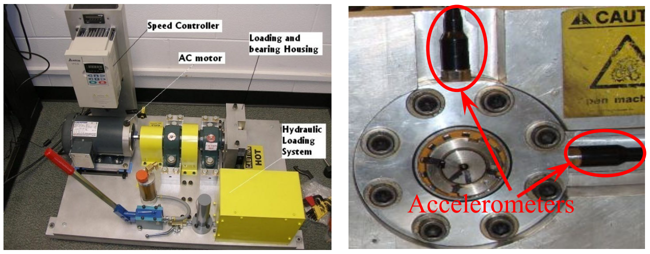


Figure 5. Bearing test rig (left) and accelerometers (right)

To simulate the localized faults on a plastic bearing, four different bearing fault types were generated: inner and outer race contact surface faults, rolling element fault, and cage fault (see Fig. 6). The damages of the contact surface on the bearing inner race and outer race were generated by scratching the race surface using an electric solder iron with a heated tip. The diameter of the damaged surface area was about one third of the ball diameter. The rolling element damage was created by scratching one of bearing balls with a grinding wheel. Roughly 40% of the ball volume was ground off. The broken cage was created by cutting the Teflon retainer using a pair of sharp scissors.

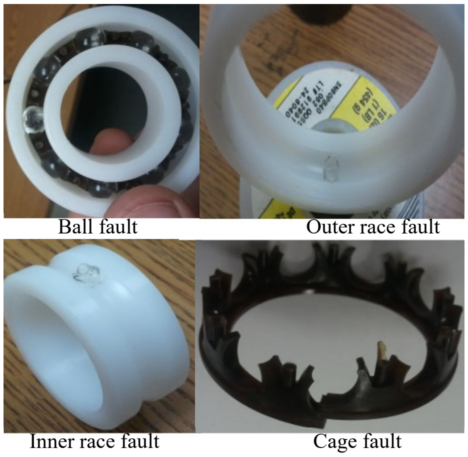

Ball fault
Outer race fault
Inner race fault
Cage fault

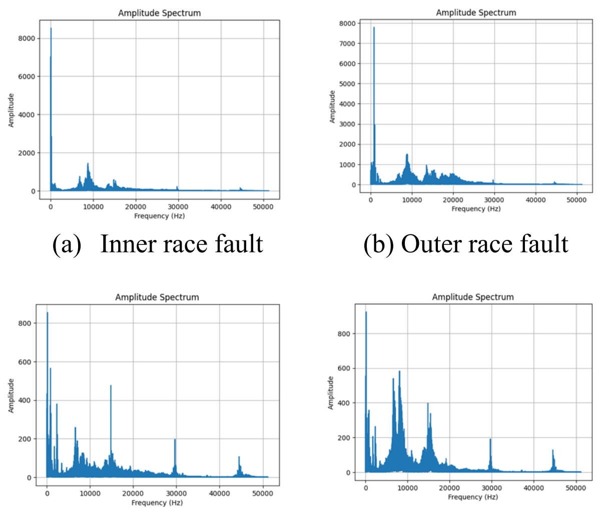
Figure 6. The 4 plastic bearing seeded faults

Note that in addition to the bearings with the faults shown in Figure 6, a healthy bearing without any fault was used in collecting the vibration data. During the testing, vibration signals were collected with a sampling rate of 102.4 kHz. Totally, four input shaft speeds 10 Hz, 20 Hz, 40 Hz, and 60 Hz were used during the testing.

4.2. Signal Processing with CWT

In order to generate the appropriate scalogram images and compute the time-frequency features using CWT, the right CWT decomposition scales have to be used. To determine the scales of the CWT, the frequency spectrums of the vibration signals were first obtained and analyzed. Figure 7 provides the frequency spectrum of the vibration signals for the 5 bearings at 60Hz input shaft speed obtained using fast Fourier transform (FFT).

From Figure 7, we can see that most of the frequency activities are concentrated in the frequency range between 5kHz and 25kHz. To extract the scalogram images and time-frequency features in this frequency range, scales spaced evenly on a logarithmic scale between 5kHz and 25kHz will be computed.



(a) Inner race fault (b) Outer race fault

(c) Cage fault (d) Ball fault

Figure 7. The frequency spectrums of the bearing faults at input shaft speed of 60Hz

Define:

n = number of scales

$f\_sampling$ = sampling frequency

f_lower = lower frequency limit of the frequency range

f_upper = upper frequency limit of the frequency range

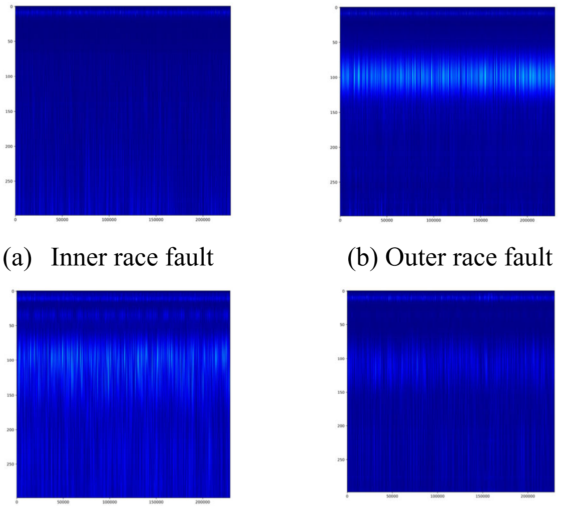
i = index of the scale, $i = 0, \dots, n-1$

Then the i^{th} CWT logarithmic scale s_i can be computed as:

$$s_i = 10^{\left\{\log_{10}\left(\frac{f\_sampling}{f\_upper}\right)+i*\left[\frac{\frac{f\_sampling}{f\_lower}-\frac{f\_sampling}{f\_upper}}{n-1}\right]\right\}} \tag{11}$$

Note that CLIP uses an embedding size of 512. Therefore, in our paper, the number of scales n was set as 512.

Using "Morlet" as the wavelet function and the scales generated between 5 kHz and 25 kHz frequency range with Eq. (11), the scalogram images of the 4 types of the bearing faults generated by CWT are provided in Figure 8.



(a) Inner race fault (b) Outer race fault

(c) Cage fault (d) Ball fault

Figure 8. Scalogram images of the bearing faults at input shaft speed of 60Hz

4.3. The Analysis Results

Once the scalogram images were generated using CWT, these images were then passed through the pre-trained CLIP model to generate the image embeddings. CWT coefficients were obtained to compute the statistics and then added into image embeddings to form the text embeddings. With the generated image and text embeddings, cosine similarity matrices were computed for *K*-way *N*-shot learning classification. In this paper, *K* = 5 and version 1 of CLIP was used. Note that *K* was set as 5 because in addition to the 4 bearing faults, a healthy bearing without any fault was also used in the dataset. In this version of CLIP, two image processing models are used: vision transformer with base size (ViT-B) and ResNet50 (RN50). Therefore, the results provided next include both models.

The classification results of 1-shot learning for each of the 4 input shaft speeds are provided in Table 1.

Table 1. 5-Way 1-Shot results

Classification Accuracy		Input Shaft Speed
ViT-B	RN50	
60%	60%	10 Hz
60%	80%	20 Hz
80%	40%	40 Hz
100%	100%	60 Hz

As we can see from Table 1, for both image processing models ViB-T and RN50, as the input shaft speed increases, the classification tends to be more accurate. This is reasonable. As the speed increases, the impact generated by the bearing faults tend to be more significantly recorded by the vibration signals and therefore be more distinctly reflected by the scalogram images. However, there is an exception for RN50 at 40Hz. This indicates that the results of ViT-B are more stable than that of RN50.

The classification results of 2-shot learning for each of the 4 input shaft speeds are provided in Table 2.

Table 2. 5-Way 2-Shot results

Classification Accuracy		Input Shaft Speed
ViT-B	RN50	
100%	100%	10 Hz
100%	100%	20 Hz
100%	40%	40 Hz
100%	100%	60 Hz

From Table 2, we can see that as the number of shots increases to 2, a perfect classification will be obtained except for RN50 at 40Hz. Table 2 again shows that the results of ViT-B are more stable than that of RN50. One explanation for this is that ViT-B is a transformer based deep learning model and has a more sophisticated deep structure than that of RN50.

To compare with the results in He, Li, and Zhu (2013), a comparison summary is provided in Table 3. Note that in Table 3, only the best results for the two methods are used for the comparison purpose.

Table 3. Comparison with He, Li, and Zhu (2013)

Classification Accuracy			Input Shaft Speed
Method in this paper		He, Li, and Zhu (2013)	
1-shot	2-shot		
60%	100%	90.0%	10 Hz
80%	100%	95.7%	20 Hz
80%	100%	96.3%	40 Hz
100%	100%	97.1%	60 Hz

From Table 3, we can see that the CLIP based few-shot learning method has shown a significant improvement over the previous research results. The highest classification accuracy achieved by the previous research is 97.1% and it took more than two multiple samples. However, the CLIP based method needs only two samples to achieve a 100% classification accuracy.

Since the few-shot learning approach developed in this paper is based on the concept of transfer learning, we also did test the transfer learning capability of the developed method. To do that, we mixed the bearing faults with different input shaft speeds together. With *K*-way *N*-shot learning, *N* number of images for each speed will be used for training. For example, for 5-way 2-shot learning, two samples of each speed were used in the training set. The transfer learning results for mixed input shaft speeds are provided in Table 4.

Table 4. Transfer learning results for mixed input shaft speed conditions

Classification Accuracy		*K*-way *N*-shot
ViT-B	RN50	
85%	70%	5-way 1-shot
100%	80%	5-way 2-shot
100%	100%	5-way 3-shot

As we can see from Table 4, for the best results, the CLIP based few-shot learning can get up to 85% accuracy with only one sample. As the number of samples increases to two, it can obtain a 100% accuracy. Again, the CLIP based approach has shown its power for transfer learning with only few samples.

5. CONCLUSION

In this paper, a pre-trained deep learning model called CLIP that combines image processing and NLP is adopted to few-shot learning for plastic bearing fault diagnosis. We explored the feasibility of leveraging CLIP model in the realm of bearing fault diagnosis via few-shot learning. Specifically, we addressed the challenges posed by CLIP's creation of requisite text prompt embeddings for the diagnosis of mechanical faults, within a few-shot learning framework. Our investigation illuminated the remarkable capability of CLIP to adapt to new tasks with minimal examples, a feature we exploited to devise a solution for plastic bearing fault diagnosis. The effectiveness of the few-shot learning method with CLIP was demonstrated using vibration data collected from plastic bearing seeded fault tests in the laboratory and was compared with that of the previous results. The comparison has shown that the CLIP based approach presented in this paper provided significant performance improvement over the previous data mining method for plastic bearing fault diagnosis. Even though the presented methodology was demonstrated with a plastic bearing fault diagnosis case study, it should be applied to other types of fault diagnosis applications where both numerical and text data can be used.

REFERENCES

Liu, X., Zhang, Y., Liu, Y., Li, H., & Li, Z. (2021). Meta-learning as a promising approach for few-shot cross-domain fault diagnosis: Algorithms, applications, and prospects. *Neurocomputing*, vol. 449, pp. 93-103. https://doi.org/10.1016/j.neucom.2021.05.086.

Yan, Y., Liu, Y., Liu, X., Chen, Y., Peng, Y., & Li, X. (2021). Few-shot learning under domain shift: Attentional contrastive calibrated transformer of time series for fault diagnosis under sharp speed variation. *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2834-2844Chen, W., (1991). *Nonlinear Analysis of Electronic Prognostics.* Doctoral dissertation. The Technical University of Napoli, Napoli, Italy.

Feng, Y., Chen, J., Xie, J., Zhang, T., Lv, H., & Pan, T. (2022). Meta-learning as a promising approach for few-shot cross-domain fault diagnosis: Algorithms, applications, and prospects. Knowledge-Based Systems, 237, 107165. https://doi.org/10.1016/j.knosys.2021.107165.

Hu, S. X., Li, D., Stühmer, J., Kim, M., & Hospedales, T. M. (2019). Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 11803-11812). doi: 10.1109/CVPR.2019.01206.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. OpenAI. Retrieved from https://cdn.openai.com/papers/Learning_Transferable_Visual_Models_From_Natural_Language_Supervision.pdf.

He, D., Li, R., & Zhu, J. (2013). Plastic bearing fault diagnosis based on a two-step data mining approach. *IEEE Transactions on Industrial Electronics*, vol. 60, pp. 3429 – 3440.

BIOGRAPHIES

David He received a Ph.D. in Industrial Engineering from The University of Iowa. Dr. He is a Professor and Director of the Intelligent Systems Modeling & Development Laboratory in the Department of Mechanical and Industrial Engineering at The University of Illinois-Chicago. Dr. He is also a Fellow of the Prognostics and Health Management (PHM) Society. Dr. He's research areas include: PHM, Industrial AI, smart manufacturing systems modeling and analysis, quality and reliability engineering.

Miao He received a Ph.D. in Industrial Engineering and Operations Research from University of Illinois at Chicago, a B.E. degree in Vehicle Engineering from University of Science and Technology Beijing, Beijing, China and a M.S. degree in Mechanical Engineering from Purdue University, Hammond, IN. She is currently a senior research scientist at Siemens Corporation. Her research interests include digital signal processing, machinery health monitoring and fault diagnosis, prognostics, artificial neural network applications, and edge computing on factory automation.