Noise-robust representation for fault identification with limited data via data augmentation

Zahra Taghiyarrenani[1], Amirhossein Berenji[2]

[1] *Center for Applied Intelligence Systems Research, Halmstad University, Halmstad, Halland, 30118, Sweden*
zahra.taghiyarrenani@hh.se

[2] *Department of Mechanical and Energy Engineering, Shahid Beheshti University, Tehran, Tehran, 1983969411, Iran*
a.berenji@mail.sbu.ac.ir

ABSTRACT

Noise will be unavoidably present in the data collected from physical environments, regardless of how sophisticated the measurement equipment is. Furthermore, collecting enough faulty data is a challenge since operating industrial machines in faulty modes not only has severe consequences to the machine health, but also may affect collateral machinery critically, from health state point of view. In this paper, we propose a method of denoising with limited data for the purpose of fault identification. In addition, our method is capable of removing multiple levels of noise simultaneously. For this purpose, inspired by unsupervised contrastive learning, we first augment the data with multiple levels of noise. Later, we construct a new feature representation using Contrastive Loss. The last step is building a classifier on top of the learned representation; this classifier can detect various faults in noisy environments. The experiments on the SOUTHEAST UNIVERSITY (SEU) dataset of bearings confirm that our method can simultaneously remove multiple noise levels.

1. INTRODUCTION

Measurement noise is an integral part of instrumentation processes. It introduces noticeable amount of uncertainties, which complicates the decision making procedure. From the classification problem point of view, addition of noise results in severe reduction of separability between different classes, as it would scatter observations of different classes, which were fairly separable before addition of noise, all over the feature space. As it would result in poor classification performance, the employment of denoising techniques is an essential step in environments with significant level of noise presence.

Recently, Deep Learning has gained significant attention towards itself; however, coping with noise presence is still a challenge for Deep Learning-based methods (Liu, Zhou, Zhao, Shen, & Xiong, 2019). Despite of being highly admired due to their performance, deep learning methods are notorious for the requirement of huge amounts of information for training. Even with undeniable technological advancements during recent decades, it is still quite challenging to provide Deep Learning methods with sufficient training data; therefore, it is crucially important to employ strategies and techniques that make Deep Learning methods applicable in limited data scenarios. This matter comes to higher level of importance in the fields related to machinery health diagnosis, as running industrial pieces of equipment in faulty modes would bring up severe consequences (Wang et al., 2020).

Moreover, in a fault identification problem, different faults can be distinguished to some extent, but interference resulting from various factors, including measurement noise, inevitably weakens their separation. Therefore, achievement of acceptable separability of faults according to the set of features extracted, becomes a matter of great importance in the implementation of a fault identification model. Contrastive learning is a well established strategy to extract a feature space where different faults are properly discriminated in the constructed space(Le-Khac, Healy, & Smeaton, 2020). Being focused on the construction of a feature space where the distance between observations from similar classes (faults) is minimized, while clusters of different classes (faults) orient farthest from each other, approaches based on contrastive learning are discriminative feature extractors.

In this paper, we propose a new method for learning a new representation using contrastive learning and Siamese neural network for the denoising task. In addition to having noise-robustness and class discriminative properties, the proposed method addresses situations where enough labeled samples

are not available.

The rest of the paper is organized as follow: In section 2 the theoretical background about contrastive learning and Siamese neural networks are discussed. In section 3, we review the most recent and related works in denoising for fault identification. Section 4 defines the problem and the proposed method is presented in section 5. In Section 6, we describe experiments conducted to validate the effectiveness of the proposed method. Lastly, we will conclude our work in section 7.

2. BACKGROUND

2.1. Contrastive representation learning

In contrastive representation learning, an embedding space is constructed to close similar samples and separate dissimilar ones. In addition, the samples are paired, if two paired samples are similar we call the pair as a positive pair, otherwise a negative pair. To perform contrastive learning, it is, therefore, necessary to find similarities between samples. Moreover, contrastive learning can be accomplished either supervised or unsupervised. The similarity between samples is specified according to the labels of the samples in supervised contrastive learning. Therefore, in the constructed embedding by supervised contrastive learning, samples with the same label are placed near each other and those with different labels are placed far apart from each other. This approach is an example of few-shot learning methods (Jadon, 2020). Unsupervised contrastive learning incorporates data augmentation. In fact, it is assumed that the augmentation process will not change the class of a sample. Therefore, each sample and its augmented counterpart are considered to be similar and construct a positive pair, and consequently will be placed near to each other in the new constructed space. This approach is an example of a self-supervised learning methods (Le-Khac et al., 2020; Chen, Kornblith, Norouzi, & Hinton, 2020).

Different loss functions have been proposed in the literature to perform contrastive learning, including the Contrastive Loss, Triplet Loss, Lifted Structured Loss, etc (Le-Khac et al., 2020; Oh Song, Xiang, Jegelka, & Savarese, 2016). In this paper, we use Contrastive Loss function for training process.

Let X and $\mathbf{X}$ represent input and embed spaces, respectively; $f : X \rightarrow \mathbf{X}$ is a function that maps the original space to the embed space. In addition, c is a contrastive label associated with a pair of samples. It is equal to zero if the pair of samples is similar, and to one if they are dissimilar. Equation 1 shows the Contrastive Loss.

$$ContrastiveLoss = (1-c)D_w^2 + (c)(max(0, m - D_w))^2 \quad (1)$$

where D_w is a similarity index, such as Euclidean Distance and m is a parameter known as margin. Margin is supposed to be the distance between different classes, in the constructed feature space. According to the contrastive loss equation, the first term is supposed to represent similar observations as closely as possible, while the second term is supposed to increase the distance between dissimilar observations. (Jadon, 2020).

2.2. Siamese Neural Network

A siamese neural network consists of a dual and symmetric architecture, in which a pair of identical models are used to extract embedding corresponding to the given pairs of observations (Chicco, 2021). Given observations are either positive pairs (belonging to the same classes) or negative pairs (belonging to the different classes). During the training process, by comparing the observations available in training pairs, the network mines the input data. An architecture of a siamese neural network is shown in the figure 1c.

3. RELATED WORKS

Being capable of reconstructing a noise-free version of given noisy corrupted observations, Denoising Autoencoders (DAE) (Vincent, Larochelle, Bengio, & Manzagol, 2008) are highly used for fault diagnosis in noisy environments. For example, in (Zhao, Lu, Ma, & Wang, 2015), a hybrid classification approach consisting of an unsupervised feature learning using a Stacked Denoising Autoencoder and a consecutive fine-tuning using softmax regression is used for fault detection in bearing. Noise presence in this study is modeled by setting a randomly chosen fraction of the units in the input of the network to zero. Moreover, dropout is used during the training process of Denoising Autoencoders to not only prevent the network from overfitting, but also improve the robustness of the network towards noise presence. Similarly, in (Liu et al., 2019) the performances of a 1-D Convolutional Denosing Autoencoder to reconstruct noise free versions of given noisy observations and a conventional neural network to use the reconstructed version to identify the health state, is evaluated. This study uses a dual approach to take into account for noise presence. The presence of noise for cases involving training the denoising autoencder is done by adding Gaussian noise with various signal to noise ratio, from -2 to 12 dB, while noise presence during the training of the conventional neural network is achieved by randomly setting units of input to zero and the rate of chosen units varies from 0.2 to 0.8. The proposed method is evaluated on test sets with varying Signal-to-noise Ratio (SNR) from -2 to 12 dB, while the SNR level is kept constant in each test set. In (Vincent et al., 2008) also a stacked denoising autoencoder is used to learn features from unlabeled information and consecutively limited labeled data is used to post train the encoder, making it suitable for classification purposes. In this study, in addition to the variation of levels of noise which is modeled by the addition of

Gaussian Noise from 0 to 30 dB, the amount of labeled information available for the post-training process is also taken into account and it is shown that acceptable performances from classification point of view are achievable in even extremely limited labeled information scenarios, using the proposed method. Distance learning methods, due to their intrinsic ability in dedication of regions of space to specific classes, regardless of the presence of noise in the environment, have gained significant attraction to cope with noise presence. For example, in (Zhang et al., 2019) a Siamese network employing a deep convolutional neural network as its feature extractor is used to provide reliable performance in rolling bearing fault diagnosis. In this study, presence of noise is modeled by the addition of Gaussian Noise with varying SNRs, from – 4 to 10 dB. Moreover, this study investigated the effect of data availability on the goodness of classification, by varying the amount of available information and monitoring its effect on the classification accuracy.

4. Problem Definition

Given X and Y as input and output spaces, respectively, we are provided with n labeled samples, $D_{train} = \{(x_i, y_i)\}_{i=1}^{n}$ where $y_i \in Y, x_i \in X$. We aim to construct a function f that maps input space X to the new space $\mathbf{X}$, $f : X \rightarrow \mathbf{X}$. To this end, we design a method that ensures that $\mathbf{X}$:

1. is capable of multi-level denoising.
2. can be constructed using limited available labeled samples.
3. is class discriminating.

Therefore a conventional classifier on top of the new space $\mathbf{X}$ classifies original and noisy samples effectively.

5. The proposed Method

We construct a new feature space for denoising inspired by both supervised and unsupervised contrastive learning. From one hand, taking advantage of the availability of labeled samples, we employ supervised contrastive learning. As a Few-shot learning technique, supervised contrastive learning can construct a class discriminated space based on a few labeled samples. On the other hand, inspired by unsupervised contrastive learning, we augment the samples with different levels of noise. It is noteworthy that, in the case of unsupervised contrastive learning, augmentation is required due to the lack of labels for the samples. However, in this paper, we perform data augmentation for denoising purposes. Figure 1 summarizes the steps of the proposed method.

First, we augment the samples, D_{train}, with the any arbitrary levels of noise. This step is shown in the figure 1a. Therefore, considering L levels of noise, we construct $D_{train}^{noisy} = \cup_{l=1}^{L} D_{train}^{l}$ where D_{train}^{l} is the augmented samples with noise level l. To be consistent, let call D_{train} as $D_{train}^{original}$.

After that we construct pairs using the original and noisy samples. This step is shown in figure 1b. The possible group of pairs are:

1. (x_i, x_j) where $x_i \in D_{train}^{original}$ and $x_j \in D_{train}^{noisy}$. This group of pairs by aggregating original samples with noisy ones helps in denoising. In addition, because of pairing original samples with augmented samples with different levels noise, this group helps in multiple level denoising.
2. (x_i, x_j) where $x_i, x_j \in D_{train}^{noisy}$. This group of pairs enhances the aggregation of multiple levels of noise. Consequently, this group of pairs, when combined with the previous group of pairs, contributes to multilevel denoising. In contrast with cases where original data is only corrupted with a single level noise, known as single level denoising, multilevel denoising involves reduction of noise where data is corrupted using various levels of noise.
3. (x_i, x_j) where $x_i, x_j \in D_{train}^{original}$. By using limited labeled samples, this group of pairs is able to construct a class discriminated feature space. In fact, this group of pairs is similar to those used in supervised contrastive learning in few-shot learning(Jadon, 2020).

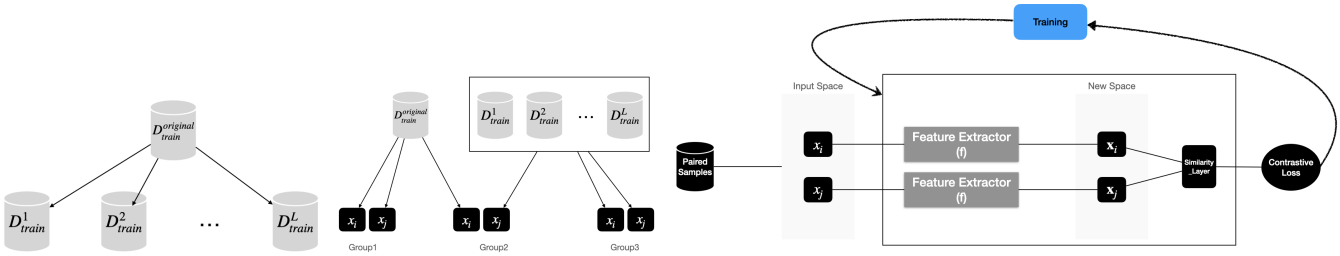
We calculate the similarity between paired samples (in all three groups) based on their labels. Although we augment the data, they are already labeled, which allows us to calculate the similarity between paired samples directly from their labels. Consequently, any training example takes the form of $\{(x_i, x_j), c_{i,j}\}$ where $c_{i,j}$ is the contrastive label corresponding to the pair (x_i, x_j).

Following the preparation of the training examples (the pairs with respective contrastive labels), they are fed into a Siamese neural network shown in figure1c; This network is constructed with two copy of feature extractor f. the network is trained using the Contrastive Loss described in equation 1. After training the network, the function f is used to map samples from input space X to the new feature space $\mathbf{X}$.

The last step is training a classifier, utilizing the feature space derived by the network, $\mathbf{X}$. In this study, K-nearest neighbor (KNN) is used to carry out the classification step. We consider KNN as the best choice of classification model in this study, mainly due to its pure distance-based mechanism, which makes it a great metric to evaluate the effectiveness of a feature space, in providing sufficient separablity of classes.

6. Experiments

In this study, the effectiveness of the proposed method is evaluated using the dataset provided by the Southeast University (Shao, McAleer, Yan, & Baldi, 2018). The referenced dataset consists of both bearing and gearbox signals, however, in this study we only utilized the bearing dataset. Five different health classes are taken into account for bearings,



(a) Data Augmentation (b) paring the samples (c) Training the network

Figure 1. The three steps of the proposed method

including healthy operation, inner ring fault, outer ring fault, rotating ball fault and combination of inner and outer ring fault. Moreover, two rotational speeds, as loading conditions, are included in this dataset (20 and 30 Hz). Various channels of accelerations coming from various locations of the test bench are provided in this dataset and we used the signals provided by the second channel in this study. The original time series identical to each load-fault combinations are split to 1024 point long signals in time domain. Consecutively, Fast Fourier Transform is used to derive the frequency domain observations from time domain observations, as bearing faults are significantly easier to diagnose in frequency domain. Employment of FFT on the time domain signals would provide us with 512 point long frequency domain signal.

For the experiments, we augment the data by adding Gaussian noise with two different Signal-to-noise ratios (SNR), -2 and -4 dB. Mathematical definition of SNR can be seen in 2, where P_{signal} and P_{noise} are powers of original signal and noise, respectively. It is worth noting that any arbitrary noise can be added to data. In addition, all of the provided results are the average of five runs.

$$SNR_{db} = 10log_{10}(\frac{P_{signal}}{P_{noise}}) \tag{2}$$

The Siamese network used in these experiments utilizes a three-layered multi-layered perceptron as the base feature extractor. Hyperbolic tangent is used as the activation function for all the neurons in the network; moreover, number of neurons per layer are as 512-256-128. ADAM optimizer with the learning rate equal to 0.0001 was used to train the network and 1000 iterations provided satisfactory training process in all of the experiments.

As a demonstration that our proposed method can construct a noise-robust feature space, we show that there is no distinction between original and noisy samples (regardless of the noise level) in the new space.

To conduct experiments that demonstrate our method's ability to work with limited data, first of all, we separate 40 percent of all samples for use as a test dataset, in order to have the same test data for all experiments. We construct the training datasets using the remaining 60 percent of samples. By varying the number of training examples, we conduct three experiments. For the first one, we use all 60% of the remaining samples as the training dataset. The second, third training datasets consist of 20%, and 1% of the 60% remaining samples, respectively.

In addition, we aim to construct a model that is robust to the different levels of noise (strong noise levels) that may occur in an environment where the model will be used; to simulate such a condition, we add noise to the test data and evaluate the performance of our method and provide the results for the four following situations.

1. original test samples. These results are shown in the first set of bars in each sub-figure named by *Result on No noise*.
2. The corrupted test samples with -2 dB noise. These results are shown in the second set of bars in each sub-figure named by *Result on -2 dB*.
3. The corrupted test samples with -4 dB noise. These results are shown in the third set of bars in each sub-figure named by *Result on -4 dB*.
4. The combination of original and corrupted samples with -2 and -4 dB noise. These results are shown in the last set of bars in each sub-figure named by *Result on overall*.

Furthermore, it is worth mentioning that our method, regardless of how many levels of noise are taken into consideration, works by unifying the original and noisy data (augmented data by noise); so that in the constructed space, they cannot be distinguished (In this way the effect of noise will be removed). To demonstrate that the original and noisy data are adequately unified by our method, for all experiments, we first extract a new feature representation, then on the top of the constructed space, we train KNN (with $k = 5$) using

1. original samples (The red bars in the plots that is named with *train KNN on No-Noise*),
2. samples corrupted with -2 dB noise (The blue bars in the plots that is named with *train KNN on -2 dB*),

3. samples corrupted with -4 dB noise (The purple bars in the plots that is named with *train KNN on -4 dB*),
4. all original and noisy samples (The gray bars in the plots that is named with *train KNN on Overall*).

As we can see in the bottom sub-figures in figures 2 to 4, in all cases, the results of KNN with different training sets are the same; this means the training data for KNN (No-noise, -2 dB, -4 dB, and overall) are aligned to each other in the constructed feature space. Thus, the same result from, for example, KNN on 'No-Noise' (red) and train with 'all noises' (grey) is proof that our proposed method is functioning and the constructed space is robust to noise with varying levels. In addition, the fact that such results have been obtained even after decreasing the number of training examples indicates that our method is capable of denoising data even when the number of training data is insufficient, although the performance in terms of the detection accuracy is affected by this factor.

In order to emphasize the effect of the augmentation in our method, we remove this step and redo the experiments. The first sub-figures (sub-figures on the top) in figures 2 to 4 show the results. In fact, we can interpret the first sub-figures as the results of applying contrastive learning to data; we just pair the available labeled samples and train the Siamese neural network with contrastive loss. Comparing every two sub-figures demonstrates how data augmentation part of the proposed method is crucial for denoising; As we can see in the first sub-figure of figures 2 to 4, when we do not perform the augmentation process, training KNN on different levels of noise provides different results regardless of the test case. In fact, these differences indicates that the constructed space which is obtained without data augmentation is not robust to the noise.

Moreover, we conduct another experiment and compare the results of our method with KNN and denoising autoencoder. The results are shown in the figure 5. In fact, we compare the results of our method with the results of KNN applied to the original samples. The reason we perform this comparison is because we applied KNN to the constructed feature space as well. Therefore, this comparison illustrates the capabilities of the constructed feature space.

In addition, we compare the results of the proposed method with those of the conventional denoising autoencoder. We designed this comparison to demonstrate the superiority of our methods in removing multiple noise levels with insufficient available samples. Our approach is to train a denoising autoencoder, using the -2 dB, -4 dB and original training sets as the input and corresponding noise-free version of training sets as the output. On top of the constructed space with the denoising autoencoder, we train a KNN classifier with $k = 5$. As with other experiments, we use the same 40% of the samples as a test dataset. We consider 5% of the remaining samples as labeled datasets.

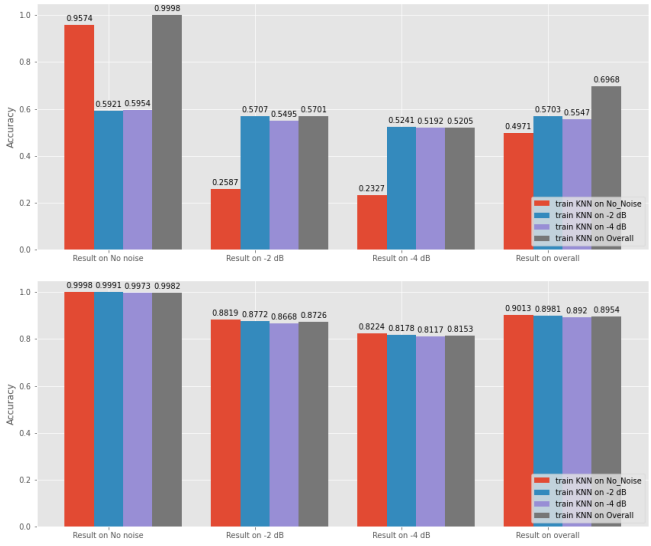


Figure 2. All training samples are used in this experiment. The first sub-figure is the results of eliminating data-augmentation. The second sub-figure shows the results of our proposed method.

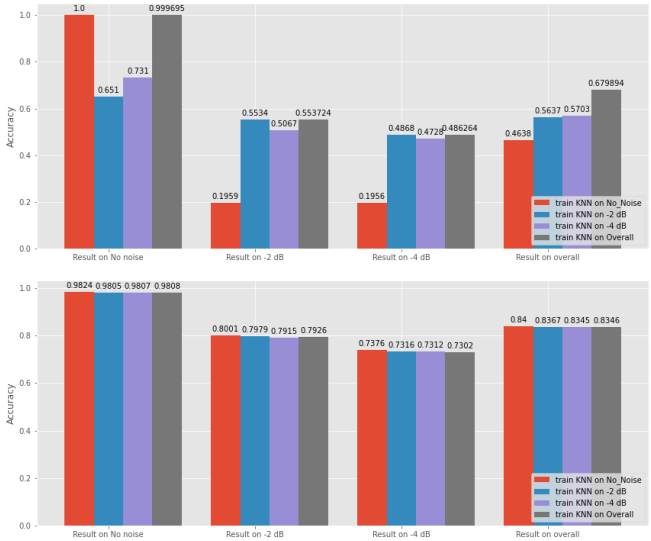


Figure 3. 20 Percent of training samples are used in this experiment. The first sub-figure is the results of eliminating data-augmentation. The second sub-figure shows the results of our proposed method.

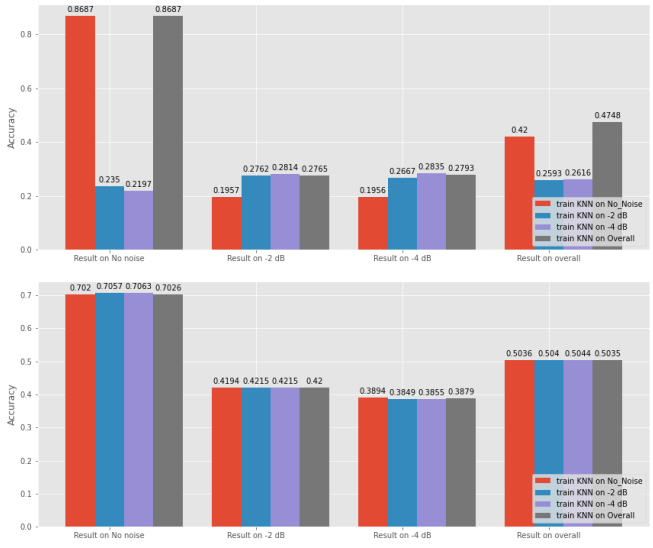


Figure 4. 1 Percent of training samples are used in this experiment. The first sub-figure is the results of eliminating data-augmentation. The second sub-figure shows the results of our proposed method.

Similar to the previous experiment, we evaluate the methods on the original test samples, corrupted test samples with -2 and -4 dB noises and the combination of original and corrupted ones. As we can see in figure 5, in the noisy environments, our method outperforms other.

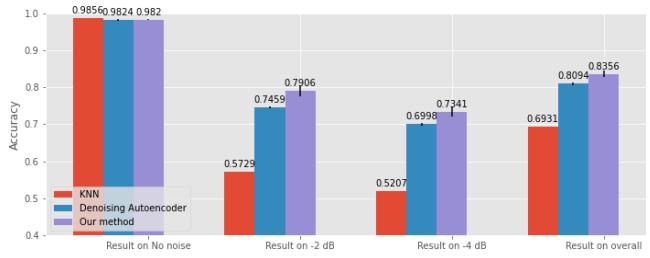
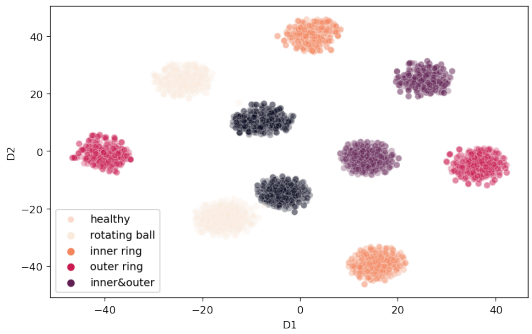
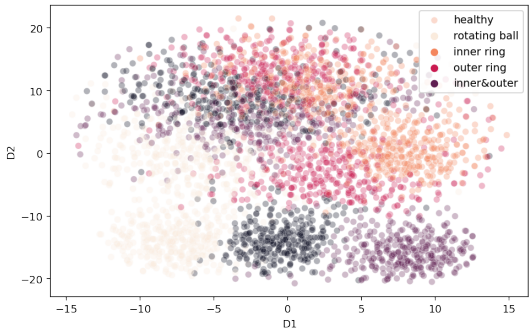


Figure 5. Evaluation of our method in comparison with others; The results are obtained using 5 percent of the samples

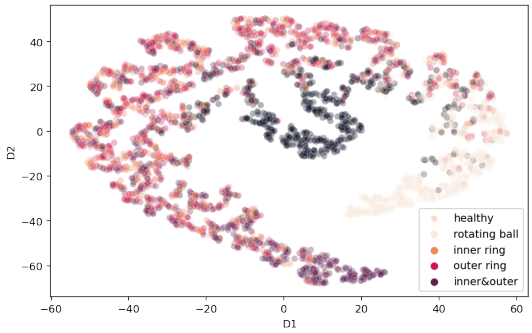
Moreover, we use t-SNE to visualize the effects of our proposed method. Figure 6a illustrates the training samples. As can be seen, this dataset is not noisy. However, due to the fact that we use data from two different loads, there are two distinct clusters per health class. In Figure 6b, we see the test samples that have been corrupted by -2dB noise. Considering these two figures, we are able to conclude that a model trained with clean training samples will not be robust to noise and therefore will experience performance degradation. To address this problem, we aim to reduce the effect of noise, in a new representation space. As a critical part of our proposed method is data augmentation, we show the constructed



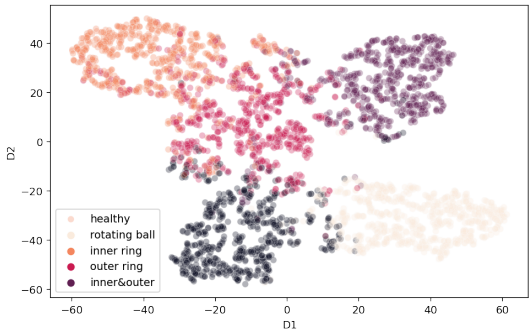
(a) Training samples



(b) Corrupted test sample. The samples are corrupted by -2 dB noise.



(c) Test samples in the new constructed feature representation when the augmentation step is ignored.



(d) Test sample in the new constructed feature representation when the data augmentation step is performed.

Figure 6. t-SNE is used to visualize training and test samples in the original and constructed space, and to compare the constructed space with and without the data augmentation step.

space with and without this step. Figure 6c shows the test samples in the constructed space when data augmentation is ignored. We can see that samples from different health states are overlapped. Figure 6d demonstrates the test samples in the constructed new representation when dataaugmentation is employed. As can be seen, our method can reduce noise as well as gather samples from different loads, resulting in a higher degree of accuracy. This visualization is related to the experiments for which all training samples are used to construct a new space, whose results can be found in Figure 2.

7. Conclusion

In this paper, we propose a feature representation learning method for the purpose of denoising. It is possible to remove multiple levels of noise through this technique, even when enough labeled samples are not available. To achieve this goal, we augment the samples with different levels of noise, inspired by unsupervised contrastive learning techniques. Using the original samples and the corrupted ones, we pair the samples. Then, using the prepared paired samples, we train a Siamese neural network with Contrastive Loss function. Training the network results in a feature extractor that maps the samples to a new space. In this space, the corrupted samples are aggregated with the original samples, resulting in denoising. Moreover, since the new space is constructed using contrastive learning, not only are the classes separated but also the new space can be achieved by using a small number of labeled samples. With the SEU dataset, we conduct several experiments with different amounts of labeled samples. The effects of denoising can be observed in each case. We also compare our method to the results of the denoising autoencoder in the absence of sufficient labeled data.

8. Acknowledgements

This research has been funded in part by the Knowledge Foundation and by Vinnova, Strategic Vehicle Research and Innovation programme.

References

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Chicco, D. (2021). Siamese neural networks: An overview. *Artificial Neural Networks*, 73–94.

Jadon, S. (2020). An overview of deep learning architectures in few-shot learning domain. *arXiv preprint arXiv:2008.06365*.

Le-Khac, P. H., Healy, G., & Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. *IEEE Access*, *8*, 193907–193934.

Liu, X., Zhou, Q., Zhao, J., Shen, H., & Xiong, X. (2019). Fault diagnosis of rotating machinery under noisy environment conditions based on a 1-d convolutional autoencoder and 1-d convolutional neural network. *Sensors*, *19*(4), 972.

Oh Song, H., Xiang, Y., Jegelka, S., & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4004–4012).

Shao, S., McAleer, S., Yan, R., & Baldi, P. (2018). Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Transactions on Industrial Informatics*, *15*(4), 2446–2455.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on machine learning* (pp. 1096–1103).

Wang, S., Wang, D., Kong, D., Wang, J., Li, W., & Zhou, S. (2020). Few-shot rolling bearing fault diagnosis with metric-based meta learning. *Sensors*, *20*(22), 6437.

Zhang, A., Li, S., Cui, Y., Yang, W., Dong, R., & Hu, J. (2019). Limited data rolling bearing fault diagnosis with few-shot learning. *IEEE Access*, *7*, 110895–110904.

Zhao, W., Lu, C., Ma, J., & Wang, Z. (2015). A deep learning method using sda combined with dropout for bearing fault diagnosis. *Vibroengineering Procedia*, *5*, 151–156.