LEVERAGING SELF-SUPERVISED LEARNING FOR VIBRATION DATA IN INDUSTRIAL SEPARATORS

Tim Heuwinkel[1], **Silke Merkelbach**[1], **Nils Janssen**[2], **Sebastian von Enzberg**[3], **and Roman Dumitrescu**[1]
[1]Fraunhofer Institute for Mechatronic Systems Design IEM, Paderborn, Germany
{firstname.lastname}@iem.fraunhofer.de
[2]University of Wuppertal, Wuppertal, Germany
njanssen@uni-wuppertal.de
[3]Hochschule Magdeburg-Stendal, Magdeburg, Germany
sebastian.von.enzberg@h2.de

ABSTRACT

Industrial separators play a pivotal role in production processes of various sectors such as chemical, pharmaceutical, biotechnology, oil extraction and food industries, with over 3000 distinct applications. Operating these separators involves managing several process parameters as well as discharge and cleaning cycles, which are hard to control mainly due to deficiencies of current physical sensor technology. Recent studies have shown that machine learning can be utilized to detect faults and particle presence in separators via vibration data. However, traditional machine learning methods require domain expertise or vast amounts of labeled data. We propose the use of self-supervised learning to resolve this issue by learning useful representations from unlabeled data, which is significantly easier and cheaper to obtain. An empirical validation on data from a disk stack separator shows that self-supervised learning can improve upon manual feature engineering and supervised approaches in terms of cost, accuracy and data efficiency.

Keywords machine learning · industrial separators · vibration data · self-supervised learning

1 Introduction

Industrial separators, crucial across diverse sectors such as chemical, pharmaceutical, biotechnology, and food, serve over 3000 distinct applications involving various fluids and solids [11]. Operating these separators involves managing multiple process parameters, including rotational speeds, temperatures, volumetric flow rates, discharge cycles, and cleaning cycles. The accumulation of solids during the separation process

necessitates discharge at specific times for optimal efficiency and product quality [28]. However, the absence of physical sensors capable of directly capturing the quality and quantities of the separation process makes accurate and intelligent control systems challenging [11]. Recent advancements demonstrate progress in using vibration data and machine learning (ML) approaches for intelligent control systems in separators [24, 11]. Traditionally, creating expressive representations for classification or regression tasks related to intelligent control involved manual feature engineering, requiring domain expertise and data science skills [26, 41]. Alternatively, fully data-driven representations in a supervised fashion required large quantities of labeled data, which can be expensive, especially in industrial settings [11].

We propose the use of self-supervised learning (SSL) to mend this trade-off. SSL utilizes raw, unlabeled data recorded from ongoing production with few sensors and a pretext task, which generates feedback signals from the data itself. This approach ensures that meaningful feature spaces are learned in the intermediate representations of the model, allowing for subsequent extraction. Simple models for downstream tasks, in this case related to intelligent control, can then be trained using these intermediate representations as input, combining the advantages of minimal domain expertise and a small number of labeled samples. To test the efficacy of SSL for complex systems like industrial separators systematically, a general and modular structure of SSL approaches is established and implemented for validation. The validation shows what prepeprocessing steps, pretext tasks and architectures are most suitable, how much data efficiency is gained in comparison to supervised baselines, and how well these SSL approaches can generalize concerning process parameters.

2 Prior Work

Traditionally, domain experts manually engineered algorithms for tasks using vibration data, a time-consuming process [26, 41]. Supervised learning enhances this by autonomously solving tasks with labeled data, but there is very limited literature on using supervised learning for industrial separators. The closest work is Merkelbach et al. [24] who employ manual feature engineering on short-time Fourier transform (STFT) representations, achieving 91.27% accuracy using supervised random forest for classifying yeast presence in industrial separators. Zamorano et al. [39] detect wear in disk stack separators using vibration data and an SVM classifier with wavelet transform preprocessing. Vekteris et al. [32] propose a correlation-based approach for predicting bearing failures in dairy industry separators. Although the assignment of manually engineered representations to labels can be done autonomously, experts of the domain and of machine learning are needed to create the representations, leading to high development cost and long innovation cycles [12, 26, 41], as well as labeling cost.

While there is limited related literature on industrial separators, the analysis of ball bearings is described in much more detail. Due to the similarity of the used measurement technology and the basic mechanical principle, insights from this domain can also be included in the investigation of separators. In this domain, Transformer models gain traction due to improved capabilities and larger datasets [9, 1, 30, 3]. Some focus on sequential signals with vanilla Transformers, while others use vision Transformers for time-frequency representations [9]. Spectral transforms, mainly STFT, are common in Vision Transformers for bearing fault analysis [1, 3, 14]. The general architecture involves creating time-frequency representations, splitting them into patches, linearly transforming, and feeding them into a vanilla Transformer encoder [1, 30, 3, 20].

A common weakness in all supervised approaches is their reliance on a large quantity of labeled data, which can be expensive [21]. As stated in the introduction, SSL promises to learn expressive and useful representations without any labeled data, by creating supervision signals from the data itself [21, 25, 18]. Similar to the supervised setting, there is no literature utilizing SSL for complex assemblies like separators and therefore no evidence of their efficacy as well as no information on optimal selection of SSL methodologies. Different pretext tasks have been proposed for SSL in bearing fault analysis though [35, 16, 43, 42]. Most SSL pretext tasks can be classified as "contrastive learning" and "generative learning" [21, 25].

Contrastive SSL aims to create a representation space where similar data points are close, achieved by contrasting positive and negative samples. The SimCLR framework, a prominent example, involves an augmentation module, an encoder, a projection head, and a contrastive loss function [6]. Positive and negative pairs are created through augmentation, with the loss function maximizing agreement between positive pairs while minimizing agreement with other views in the batch [6]. Various works in bearing fault analysis leverage the SimCLR framework, demonstrating its effectiveness in tasks such as cutting tooth fault diagnosis and bearing fault diagnosis [35]. Modifications, like introducing multiple heads and uncertainty-based dynamic weighting, can enhance the framework's performance [16].

Generative SSL aims to learn meaningful representations by encoding input into a latent space and reconstructing it using a decoder [25]. The information bottleneck is crucial for meaningful representation learning, as without it, the task becomes trivial [25]. One classic generative method is the autoencoder (AE), where the encoder maps input to a condensed internal representation, and the decoder reconstructs the input from it [15, 43, 42]. The dimensionality of the internal representation serves as the information bottleneck. In Transformer models, creating a bottleneck in the latent space is challenging due to equal input and output dimensions in encoder and decoder blocks. The bottleneck can also be applied at the input, masking portions of the sequence before passing it to the encoder [5]. This method, tested in bearing fault diagnosis, can improve slightly upon state-of-the-art supervised models [5].

3 Self-Supervised Learning Approach

To test related approaches from the literature in a well-structured manner on their efficacy for industrial separators, general SSL approaches for bearing data are differentiated by the different choices of algorithms at the following stages of the training process: preprocessing, pretext task and architectures. Viable combinations of algorithm choices at different stages of the training process are added to the list of models to test (see Figure 1). The subsection on preprocessing motivates viable choices for transforming the input data, while the second subsection explores considerations related to pretext tasks. Viable deep learning architectures are proposed in the last subsection.

3.1 Preprocessing

Spectral transformations like wavelet- and Fourier transform have shown to benefit self-supervised learning (SSL) and supervised models in bearing fault analysis. While SSL works often use no transform, state-of-the-art supervised models commonly employ continuous wavelet transform (CWT) and short-time Fourier transform (STFT) [16, 10, 1, 14, 3]. CWT reveals non-stationary signals and accentuates singularities, offering advantages for separator-related tasks [30, 9]. The Morlet wavelet is popular for CWT in bearing fault analysis [40, 30] and is therefore chosen as the mother wavelet. Similar

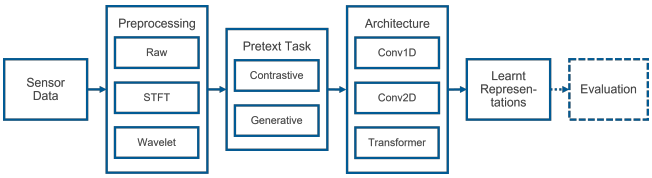

Figure 1: Stages of the learning process with viable algorithm choices in the domain of industrial separators.

to the CWT, STFT can capture time-varying characteristics in signals and highlights frequency components, which makes it successful in supervised bearing fault analysis, albeit less prevalent in SSL literature [1, 3, 14, 20]. To compare the effect of the wavelet transform and STFT to no transform on SSL with data from separators, all three are chosen for a comparison in the preprocessing stage.

Independent of the transform, standardization and data augmentation are employed to improve stability and generalization. Data augmentations help generalization by learning useful invariances [37, 17]. We chose jitter and scaling, since they have proven their effectiveness for vibration data [38, 44, 27].

3.2 Pretext Tasks

One of the most important aspects of an SSL framework is its pretext task. We incorporate generative and contrastive approaches as viable choices for the pretext task, since both types of tasks have been prevalent in SSL, not only in bearing fault diagnosis but also in state-of-the-art natural language processing (NLP) and computer vision (CV) benchmarks [21, 18].

For **contrastive learning**, a SimCLR-style pretext task is adopted, particularly suited for bearing fault analysis SSL [6, 35, 40, 33]. In this task, a recording is divided into short sequences, allowing the learning of representations invariant to variance within a recording. Beyond the typical two augmented views, a third view is created by randomly selecting and augmenting a different sequence from the same recording, enhancing performance over the vanilla SimCLR formulation [16]. The normalized temperature-scaled cross-entropy loss function (NT-Xent) is extended to handle three positive views, by adding all pair-wise losses of the views.

In the realm of **generative learning**, autoencoder models are chosen, given their prevalent use in SSL for bearing fault analysis [43, 42, 5]. For Convolutional Neural Network (CNN) models, the latent representation dimension is restricted to ensure effectiveness. The reconstruction loss utilized is the mean squared error, a standard for autoencoders. Transformer models, with the challenge of equal in- and output dimensions, adopt a modified masked autoencoder inspired by masked language modeling in NLP [8]. This involves masking a portion of the input data, training the encoder on unmasked patches, and padding the encoded patches to restore original dimensions. This generative pretext task offers advantages in reduced processing time and meaningful task creation [13].

3.3 Architectures

Convolutional Neural Networks (CNNs) dominate the self-supervised learning (SSL) landscape for bearing fault analysis, with ResNet being the preferred architecture due to its established performance on diverse datasets [10, 36, 38, 40] and is therefore

investigated for the architecture choice. For non-transformed data, the conventional choice is 1D CNNs, while 2D CNNs are used for three-dimensional time-frequency representations resulting from spectral transforms [16, 33, 35]. The 1D and 2D ResNet architecture can be employed for contrastive learning without modification, but a decoder is necessary for generative pretext tasks to map back to the input space. The decoder replicates the ResNet encoder but uses transposed convolutions for upsampling instead of regular convolutions for downsampling.

Transformer models have also demonstrated efficacy in supervised bearing fault analysis [9, 1, 30, 3] and are therefore chosen as an architecture to be tested. In the proposed approach, a single convolutional layer is added before the Transformer encoder to learn local features and adjust dimensions for three-dimensional spectral-transformed data [8, 13, 4]. To retain information about position, positional encoding based on geometric functions is applied, followed by the use of vanilla Transformer encoder layers. For generative pretext tasks, a vanilla Transformer decoder is added, while for contrastive pretext tasks, a small projector with average pooling and a linear layer is introduced. The average pooling operation is also utilized for fine-tuning on downstream tasks, reducing the output dimension of the decoder [31].

4 Validation

In the validation, all viable combinations of choices for each stage of the learning process are tested on their efficacy for solving downstream tasks related to industrial separators and compared to supervised baselines using vibration data from a separator in a laboratory environment.

4.1 Dataset and Preparation

The dataset for the proposed approach is generated using an experimental separator from the Chair of Fluid Mechanics at the University of Wuppertal. Three-axis accelerometers are attached at five positions to measure vibrations at different parts of the separator, namely at the bowl (2), top, bearing and base (see [24]). With a sampling frequency of 51,200 Hz, sequences of 4 seconds are recorded at regular intervals during the separation process, resulting in 204,800 values per recording, sensor and axis. A total of 1,337 recordings are used. To ensure unbiased results, the data is split into pre-training (935 recordings) and fine-tuning (402 recordings) sets. Each subset is further divided into training (841 and 253 recordings respectively) and validation sets (94 and 28 recordings respectively) for hyperparameter testing and early stopping. Additionally, the fine-tuning data requires a test set for final unbiased metrics calculation (121 recordings). Sequences for training are obtained by applying an overlapping sliding window, creating 200 sequences for Raw/Wavelet transformed data and 6 sequences for STFT transformed data per recording. Additionally, only the magnitude of the spherical coordinates per sensor was used to reduce input dimensionality [24]. No data selection or balancing strategies are employed, reflecting real-world scenarios where certain labels or conditions may be scarce and true distributions unknown.

4.2 Downstream Tasks

Adapting a pre-trained SSL model to downstream tasks can be achieved through fine-tuning on a small quantity of labeled data. In partial fine-tuning, only the classification or regression head and a few of the last layers are finetuned, while in full fine-tuning, every parameter of the model is adjusted [2]. We evaluate the SSL models and the baselines on

three tasks: (1) Determining if yeast particles are present in the separator. In our dataset 10.7% of recordings contain only water and 89.3% of recordings have varying yeast concentrations. (2) Estimating the yeast particle concentration of the input solution. The input yeast particle concentration in our dataset ranges from 0 to 2.5 g/l. (3) Assessing the amount of deposits on the disk stack. For this, the structural similarity index measure (SSIM) of a reference image before separation can be compared to an image during or after separation, to gauge the level of fouling [34]. The SSIM of the images ranges between 0.58 and 1.0 in our dataset.

4.3 Models and Baselines

Since almost all possible combinations are tested, it is easier to justify why certain combinations are not included. Data that was transformed by a spectral transform is two-dimensional in multiple channels, which limits the possibility to be processed by 1D convolutions and therefore only raw signals are processed with 1D CNNs. Similarly, processing one-dimensional data in multiple channels with 2D CNNs has no meaningful interpretation and is therefore excluded.

In addition to the SSL models, several baselines are implemented to represent general approaches in the literature and to isolate the effects of pre-training. The first baseline is the exact implementation of Merkelbach et al. [24], which was validated on the same data set as this paper and represents manual feature engineering with supervised learning. In the following, this approach will be called **STFT+BaselineManual**. The other baselines are based on supervised representation learning. One baseline model is trained for each type of architecture (**BaselineConv1D**, **BaselineConv2D** and **BaselineTransformer**) with the preprocessing strategy that led to the best results in the SSL model comparison.

4.4 Implementation Details

The most important hyperparameters for our approach are the latent dimension for contrastive and generative pretext tasks as well as masking rate, embedding dimension and number of decoder layers for (generative) Transformers. Since the chosen approach is heavily inspired by the SimCLR framework, we use a 128-dimensional latent space for contrastive tasks, like in the original implementation [6]. The latent dimension was set depending on the chosen transform for the generative learning tasks, since vastly more data needs to be compressed if a spectral transform was applied. For spectral transformed data a latent dimension of 1024 was used, while for combinations with no transform, a latent dimension of 256 was chosen. The choice of masking rate is the main adjustment for the difficulty of the generative pretext task for the Transformer models. Similar approaches for time series data report best results with a masking rate of 75%, hence a masking rate of 75% is also employed for our approach [29, 19]. Additionally the embedding dimension for the Transformer architectures was set at 64 (see [14, 19]) and a lightweight decoder with only one Transformer layer was implemented [7].

The training schedule for pre-training included a maximum of 100 epochs, and for fine-tuning 50 epochs, with early stopping based on validation loss. The binary cross-entropy loss was applied for the classification downstream task and a mean-squared error loss for the regression tasks. The base learning rate was set as $1e-4$ and the batch size was set as high as possible for the used GPU, since contrastive learning benefits from larger batch sizes to have more negative pairs [6]. In addition, a cosine decay learning rate scheduler was employed to improve convergence [22]. The AdamW optimizer was utilized to update the weights [23].

4.5 Empirical Results

Table 1 shows results for all selected SSL combinations as well as baselines, downstream tasks and metrics in the partial fine-tuning setting to highlight which SSL combinations work best. To show the effect of pre-training on labeled data efficiency and the learnt representations, the supervised baselines serve as a comparison. Since the best results for the Conv1D and Transformer architecture were achieved with no transform, the corresponding baselines are chosen as Raw+BaselineConv1D and Raw+BaselineTransformer (see Section 4.3). Applying the same criterion to Conv2D, the baseline STFT+BaselineConv2D is chosen. The hyperparameters of each baseline architecture is exactly the same as their self-supervised counterparts. The best overall combination according to the water/yeast classification and concentration regression task is STFT+Generative+Conv2D, with an F1-Score of 0.997 and an RMSE of 0.317. When partially fine-tuning, STFT+Generative+Conv2D outperforms the corresponding baseline, STFT+BaselineConv2D, by 4.7 percentage points in terms of F1 score for classifying water/yeast runs and 12.9% in terms of RMSE for predicting yeast concentration. However, two other combinations with no transform, a contrastive pretext task, Raw+Contrast+Conv1D and Raw+Contrast+TF perform similarly and could only be worse due to fluctuations in the test data. In general, contrastive pretext tasks seem to perform better than generative pretext tasks and convolutional architectures seem to perform better than transformer architectures.

Table 1: Results for partial fine-tuning of all selected SSL combinations as well as baselines and full training data sorted by Water/Yeast Accuracy/F1 score. Best results for each metric and overall model in **bold**. Wavelet transform is abbreviated with WT, Transformer with TF, (vanilla) neural network with NN and contrastive with Contrast.

Combination			Water/Yeast		Concentration		SSIM	
Pre	Pretext	Arch	Acc	F1	RMSE	MAE	RMSE	MAE
STFT	Generative	Conv2D	**0.995**	**0.997**	**0.317**	**0.243**	0.124	0.108
Raw	Contrast	Conv1D	0.985	0.990	0.348	0.278	0.105	0.082
Raw	Contrast	TF	0.980	0.987	0.377	0.308	0.121	0.105
Raw	Baseline	TF	0.972	0.981	0.393	0.332	0.106	0.090
STFT	Contrast	Conv2D	0.967	0.981	0.331	0.259	0.128	0.097
Raw	Baseline	Conv1D	0.954	0.972	0.421	0.348	0.116	0.100
WT	Contrast	Conv2D	0.928	0.957	0.380	0.298	0.106	0.088
STFT	Baseline	Conv2D	0.918	0.950	0.363	0.297	0.103	0.083
WT	Contrast	TF	0.893	0.935	0.393	0.314	0.119	0.104
Raw	Generative	TF	0.860	0.919	0.447	0.383	0.097	0.080
Raw	Generative	Conv1D	0.859	0.921	0.456	0.390	0.120	0.089
WT	Generative	TF	0.851	0.917	0.453	0.392	0.096	0.075
STFT	Contrast	TF	0.851	0.917	0.334	0.264	0.115	0.098
STFT	Generative	TF	0.851	0.916	0.468	0.393	0.105	0.092
Manual	Baseline	NN	0.851	0.915	0.427	0.359	0.116	0.089
WT	Generative	Conv2D	0.851	0.914	0.462	0.391	**0.093**	**0.074**

In essence, labeled data efficiency can be understood as maximizing accuracy for a given number of labeled data points. Partially fine-tuning the SSL models as well as the supervised baselines isolates the effect of pre-training on labeled data efficiency. This is because SSL models with pre-trained feature extractors are compared to randomly initialized feature extractors in the supervised models. While the partial fine-tuning might be optimal for SSL models from an data efficiency standpoint [13, 2], it is more realistic to finetune the full supervised model to harness the full representation learning

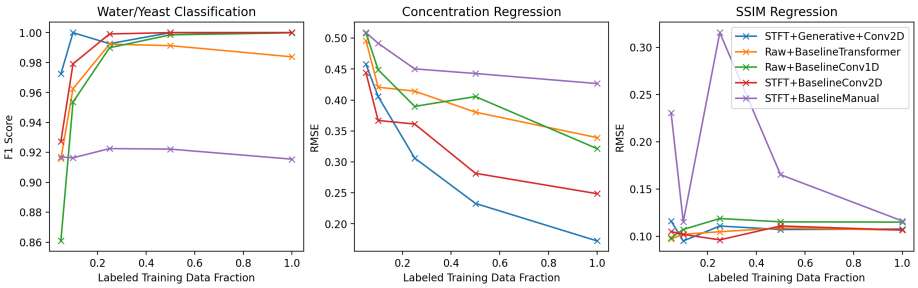

Figure 2: Results for the best SSL combination and baselines in all tasks over increasing fractions of training data with full fine-tuning. A labeled data fraction of 1.0 equates to 402 recordings.

power of deep models. In the full fine-tuning case, pre-training can be seen more as a sophisticated weight initialization scheme, since all pre-trained weights are adjusted.

Figure 2 hence visualizes the labeled data efficiency in the case of full fine-tuning. Similar to the partial fine-tuning case, one might expect that the supervised models would catch up to the pre-trained model, since the pre-trained weights are only used as a starting point for initialization. Because this is not the case, one might conclude that the initial weights set by the pre-training not only create immediately useful representations, but also weights that help convergence for settings with more training data. Apart from a dip in performance at 25% of labeled data for the first task, the SSL model seems to perform very well in comparison to the baselines. With only 5% of labeled data it already reaches an F1-score of 0.97 and an F1-score of 1.0 with 10% of labeled data (40 recordings) on the first task. While the supervised counterpart STFT+BaselineConv2D slightly outperforms the SSL model for very low labeled data settings on the second task, the positive margin between them steadily increases after 25% of labeled data. RMSE and MAE for the SSIM task are quite high for all models (compare range of labels for particle concentration and SSIM), suggesting a poor fit. Moreover, performance on this task does, not seem to be correlated with labeled data fraction at all, for any model, possibly indicating poor label quality.

Table 2: Results for partial fine-tuning for the best selected SSL combination and different out-of-sample (OOS) test sets.

		Water/Yeast		Concentration		SSIM	
Model	OOS	Acc	F1	RMSE	MAE	RMSE	MAE
STFT+Gen+Conv2D	-	0.995	0.997	0.317	0.243	0.124	0.108
STFT+Gen+Conv2D	RPM	1.000	1.000	0.417	0.327	0.179	0.140
STFT+Gen+Conv2D	VF	1.000	1.000	0.442	0.339	0.207	0.161

To actually be useful in practice, the SSL model should generalize reasonably well over possible process parameter ranges, since adjustments to these parameters happens regularly in production. Generalizability can be tested by performing out-of-sample (OOS) validations for different process parameters to simulate inference on unseen production conditions. The two process parameters available in the given data set are the revolutions per minute (RPM) and the volumetric flow rate (VF). The model was trained on recordings which had a nominal range of RPM and VF respectively and then tested only on recordings which were outside the nominal range. Table 2 outlines the changes

in accuracy when testing on out-of-sample data in terms of process parameters. In this setting the first downstream task can still be solved accurately, while the performance for the second downstream task degrades by 0.1 on non-nominal RPM ranges and by 0.125 for non-nominal VF ranges in terms of RMSE. For the third downstream task, testing on non-nominal RPM ranges degrades performance by 0.055 and 0.083 for RPM and VF respectively. While this represents an overall large degradation in performance for the second and third downstream task, a complete shift in process parameters without any similar training data in terms of the tested process parameters should be considered an extreme test of generalization, almost akin to adaptation.

Apart from predictive performance, no domain knowledge was required at any point of the process, greatly reducing development costs and making it possible to adapt the learnt representations autonomously through new unlabeled data. As of current pricing, pre-training the best SSL model (STFT+Generative+Conv2D) can be achieved with under 5 USD using an AWS ml.p3.2xlarge instance and 1000 four second recordings. Furthermore The approach is real-time capable with a processing time of 66ms for one sequence of 600ms sensor data on an NVIDIA RTX 3090 GPU and a 51,200 Hz sensor.

5 Conclusion

We proposed the use of self-supervised learning to improve labeled data efficiency for downstream tasks in industrial separators by learning useful representations from unlabeled vibration data, which is significantly easier and cheaper to obtain. Different combinations of preprocessing, pretext tasks and architectures have been tested and compared to supervised baselines inspired by the related literature. The best SSL model (STFT+Generative+Conv2D) requires little domain expertise and money to develop, while being label efficient, well generalized and flexible as well as real-time capable on the given data set. The empirical validation data is constrained to a single machine in a laboratory, limiting generalizability to diverse production environments. Moreover, real-life production data may be less diverse and contain fewer instances from non-nominal process parameter ranges than the artificially diverse datasets commonly used in experiments. Less diverse training data could inhibit inference for rare occurences in practice.

Since effectiveness of self-supervised learning for numerous domains and now industrial separators has been shown, it should be more closely examined by researchers and practitioners for real-life production processes, potentially paving the way for more efficient and environmentally friendly industrial separation processes.

Acknowledgments

The work in this paper was supported by the German Federal Ministry for Economic Affairs and Climate Action under grant 03EN4004B.

References

[1] C. T. Alexakos, Y. L. Karnavas, M. Drakaki, and I. A. Tziafettas. A combined short time fourier transform and image classification transformer model for rolling element bearings fault diagnosis in electric motors. *Machine Learning and Knowledge Extraction*, 3(1):228–242, 2021.

[2] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[3] Z. Bao, J. Du, W. Zhang, J. Wang, T. Qiu, and Y. Cao. A transformer model-based approach to bearing fault diagnosis. In *Data Science: 7th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2021, Taiyuan, China, September 17–20, 2021, Proceedings, Part I 7*, pages 65–79. Springer, 2021.

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[5] J. Cen, Z. Yang, Y. Wu, X. Hu, L. Jiang, H. Chen, and W. Si. A mask self-supervised learning-based transformer for bearing fault diagnosis with limited labeled samples. *IEEE Sensors Journal*, 2023.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Y. Ding, M. Jia, Q. Miao, and Y. Cao. A novel time–frequency transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings. *Mechanical Systems and Signal Processing*, 168:108616, 2022.

[10] Y. Ding, J. Zhuang, P. Ding, and M. Jia. Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings. *Reliability Engineering & System Safety*, 218:108126, 2022.

[11] R. Dumitrescu and M. Fleuter. *Intelligenter Separator: Optimale Veredelung von Lebensmitteln*. Springer-Verlag, 2019.

[12] D. Goyal, A. Saini, S. Dhami, B. Pabla, et al. Intelligent predictive maintenance of dynamic systems using condition monitoring and signal processing techniques—a review. In *2016 international conference on advances in computing, communication, & automation (ICACCA)(Spring)*, pages 1–6. IEEE, 2016.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[14] Q. He, S. Li, Q. Bai, A. Zhang, J. Yang, and M. Shen. A siamese vision transformer for bearings fault diagnosis. *Micromachines*, 13(10):1656, 2022.

[15] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[16] C. Hu, J. Wu, C. Sun, R. Yan, and X. Chen. Inter-instance and intra-temporal self-supervised learning with few labeled data for fault diagnosis. *IEEE Transactions on Industrial Informatics*, 2022.

[17] B. K. Iwana and S. Uchida. An empirical survey of data augmentation for time series classification with neural networks. *Plos one*, 16(7):e0254841, 2021.

[18] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[19] Z. Li, Z. Rao, L. Pan, P. Wang, and Z. Xu. Ti-mae: Self-supervised masked time series autoencoders. *arXiv preprint arXiv:2301.08871*, 2023.

[20] W. Liu, Z. Zhang, J. Zhang, H. Huang, G. Zhang, and M. Peng. A novel fault diagnosis method of rolling bearings combining convolutional neural network and transformer. *Electronics*, 12(8):1838, 2023.

[21] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[22] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[23] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[24] S. Merkelbach, L. Afroze, N. Janssen, S. von Enzberg, A. Kühn, and R. Dumitrescu. Using vibration data to classify conditions in disk stack separators. *Vibroengineering Procedia*, 46:21–26, 2022.

[25] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, et al. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[26] M. H. Mohd Ghazali and W. Rahiman. Vibration analysis for machine monitoring and diagnosis: a systematic review. *Shock and Vibration*, 2021:1–25, 2021.

[27] G. Nie, Z. Zhang, M. Shao, Z. Jiao, Y. Li, and L. Li. A novel study on a generalized model based on self-supervised learning and sparse filtering for intelligent bearing fault diagnosis. *Sensors*, 23(4):1858, 2023.

[28] W. H. Stahl. *Fest-Flüssig-Trennung. 2, Industrie-Zentrifugen. Maschinen-& Verfahrenstechnik*. DrM Press, 2004.

[29] P. Tang and X. Zhang. Mtsmae: Masked autoencoders for multivariate time-series forecasting. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 982–989. IEEE, 2022.

[30] X. Tang, Z. Xu, and Z. Wang. A novel fault diagnosis method of rolling bearing based on integrated vision transformer model. *Sensors*, 22(10):3878, 2022.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[32] V. Vekteris, A. Trumpa, V. Turla, N. Šešok, I. Iljin, V. Mokšin, A. Kilikevičius, A. Jakštas, and J. Kleiza. An investigation into fault diagnosis in a rotor-bearing system with dampers used in centrifugal milk separators. *Transactions of FAMENA*, 41(2):77–86, 2017.

[33] W. Wan, J. Chen, Z. Zhou, and Z. Shi. Self-supervised simple siamese framework for fault diagnosis of rotating machinery with unlabeled samples. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[34] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao. Ssim-motivated rate-distortion optimization for video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(4):516–529, 2011.

[35] M. Wei, Y. Liu, T. Zhang, Z. Wang, and J. Zhu. Fault diagnosis of rotating machinery based on improved self-supervised learning method and very few labeled samples. *Sensors*, 22(1):192, 2021.

[36] Y. Wei, X. Cai, J. Long, Z. Yang, and C. Li. Self-supervised contrastive representation learning for machinery fault diagnosis. In *Neural Computing for Advanced Applications: Second International Conference, NCAA 2021, Guangzhou, China, August 27-30, 2021, Proceedings 2*, pages 347–359. Springer, 2021.

[37] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu. Time series data augmentation for deep learning: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2021.

[38] Z. Yan and H. Liu. Smoco: A powerful and efficient method based on self-supervised learning for fault diagnosis of aero-engine bearing under limited data. *Mathematics*, 10(15):2796, 2022.

[39] M. Zamorano, D. Avila, G. N. Marichal, and C. Castejon. Data preprocessing for vibration analysis: Application in indirect monitoring of 'ship centrifuge lube oil separation systems'. *Journal of Marine Science and Engineering*, 10(9):1199, 2022.

[40] B. Zhang, Y. Mao, X. Chen, Y. Chai, and Z. Yang. Self-supervised learning advance fault diagnosis of rotating machinery. In *International Conference on Neural Computing for Advanced Applications*, pages 319–332. Springer, 2021.

[41] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler. Deep learning algorithms for bearing fault diagnostics—a comprehensive review. *IEEE Access*, 8:29857–29881, 2020.

[42] T. Zhang, J. Chen, S. He, and Z. Zhou. Prior knowledge-augmented self-supervised feature learning for few-shot intelligent fault diagnosis of machines. *IEEE Transactions on Industrial Electronics*, 69(10):10573–10584, 2022.

[43] W. Zhang, D. Chen, and Y. Kong. Self-supervised joint learning fault diagnosis method based on three-channel vibration images. *Sensors*, 21(14):4774, 2021.

[44] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35:3988–4003, 2022.