Advances in Reliability, Safety and Security, Part 5 – Kołowrocki, Kosmowski (eds.)

Advances in Reliability, Safety and Security

ESREL 2024
Monograph Book Series

Human Guided Fault Diagnosis With Contrastive Sensor Transformer

Zaharah Bukhsh[a], Irina Stipanovic[b,c]

[a]*School of Industrial Engineering, Eindhoven University of Technology, Eindhoven, Netherlands*
[b]*Department of Construction Management and Engineering, Univeristy of Twente, Enschede, Netherlands*
[c]*Infra Plan Consulting, Zagreb, Crotia*

Abstract

Reliable fault detection in industrial assets requires predictive models that can generalize with limited labeled data, especially as new fault types emerge. Traditional supervised techniques perform poorly in these low-resource settings. We propose a human-in-the-loop approach combining Contrastive Sensor Transformer (CST) self-supervised pre-training and active learning for data-efficient fault diagnosis. CST learns representations from unlabeled condition monitoring data using a contrastive learning approach. This pre-trained model captures general fault patterns without direct supervision. To introduce supervision with minimal labeling, we iteratively query samples for an expert to annotate based on the model's uncertainty. The encoder provides the learned embeddings and samples are incrementally labeled through active learning. This focuses the expert's labeling effort on the most informative instances to refine the classification boundaries. We evaluate our approach on two bearing fault datasets, namely the Case Western Reserve University (CWRU) dataset and KAT dataset provided by Paderborn University. Our approach archives over 90% accuracy on CWRU and 65% on KAT using less than 1% of labels. Compared to autoencoder and fully supervised baselines, our method reduces the labeling demands while maintaining high performance, especially valuable as new fault types emerge in real-world settings.

Keywords: fault diagnosis, contrastive learning, active learning, human in the loop, transformer

1. Introduction

Condition-based maintenance of industrial assets relies on machine learning models trained to detect and diagnose faults from sensor data (Fink et al., 2020). However, collecting comprehensive labeled examples of all possible fault scenarios under varying machine conditions is challenging and resource-intensive. As a result, models typically have access to limited annotated training data collected from targeted fault injection experiments and scheduled maintenance periods. This poses challenges for applying data-driven approaches at scale, as most machine learning algorithms require large labeled datasets to learn robust representations.

Traditional supervised learning approaches also struggle to generalize to novel scenarios not present in the training data. Recent works have explored self-supervised learning techniques that leverage large unlabeled datasets to address these limitations (Bukhsh, 2022; Ding et al., 2022; Golyadkin et al., 2023; Hu et al., 2022; Senanayaka et al., 2020; Wang et al., 2020; Zhang et al., 2022). The key idea is to leverage vast unlabeled datasets to learn transferable representations of machine health. While self-supervised methods make better use of unlabeled data than supervised approaches, they are still constrained by the initial set of labeled examples used for fine-tuning. If this initial set does not adequately represent the full data distribution, important fault patterns may be missed. Human-guided learning, also known as active learning, aims to iteratively select the most informative instances for annotation, maximizing model performance with minimal labeling effort (Settles, 2009). Previous works have applied active learning to prognostic and health management applications. Some notable examples include Jian et al., (2021), who proposed the use of active learning with ensemble classifiers, Jin et al., (2021) proposed a residual attention network combined with active learning and (Chen et al., 2019) introduced an empirical model singular value decomposition, active learning and random forest for gearbox fault

detection. These approaches relay on supervised and semi-supervised learning to address the limited availability of labeled data.

In the paper, we propose a novel method called Human Guided Contrastive Sensor Transformer (HG-CST), which combines powerful self-supervised learning, contrastive learning, and transformer architecture with an active learning paradigm for robust and data-efficient fault diagnosis. Our paper offers the following key contributions: (1) HG-CST leverages Contrastive Sensor Transformer (Bukhsh, 2022) to extract representations from unlabeled bearing sensor dataset and use learned embeddings as input to active learning loop, (2) Two sampling strategies are evaluated - class-agnostic versus class-aware targeting one sample per class, (3) Experiments on CWRU and KAT bearing datasets demonstrate HG-CST achieves over 90% accuracy on CWRU and 65% on KAT using less than 1% of labels, and (4) HG-CST outperforms supervised autoencoder baselines with limited labeled data. Both class-aware sampling and prioritizing uncertain examples improve active learning performance. Similarly, sampling based on uncertainty yields superior results compared to random sampling. Our results validate that combining self-supervised learning, contrastive representations, transformers and human guidance maximizes data value for predictive maintenance under real constraints.

The rest of the paper is structured as follows: Section 2 provides the background on key machine learning techniques that are core to our proposed method. Section 3 introduces our method, a human-guided contrastive sensor transformer (HG-CST). Section 4 presents the dataset description, experimental setup, and model evaluation results. Section 5 provides the concluding remarks.

2. Background

This section provides background about key machine learning techniques used in our approach, namely self-supervised contrastive learning and transformers.

2.1. Self-supervised contrastive learning

Machine learning models require large amounts of labeled data to achieve high performance. However, collecting comprehensive labels is prohibitively expensive, especially for industrial equipment condition monitoring. Self-supervised learning has emerged as a promising approach to leverage the abundant unlabeled data readily available from sensors. Rather than relying entirely on human-provided labels, self-supervised learning techniques design pretext tasks on unlabeled data to learn general representations (Saeed et al., 2019). Contrastive learning is a widely used self-supervised technique that leverages data transformations to construct contrastive examples (Saeed et al., 2021). The key idea of contrastive learning is to apply different transformations (e.g., wrapping, shifting) to the same sample, creating two different views of the same data point. These augmented views are deemed as similar, while views from other samples act as contrasts that should be dissimilar.

Triplet contrastive loss is one of the loss function that is used in self-supervised learning frameworks to train models in differentiating between similar and dissimilar instances within an embedding space. The learning process revolves around the use of triplets, which consist of an anchor sample A, a positive sample P, and a negative sample N. The positive sample shares some form of similarity with the anchor, while the negative sample does not.

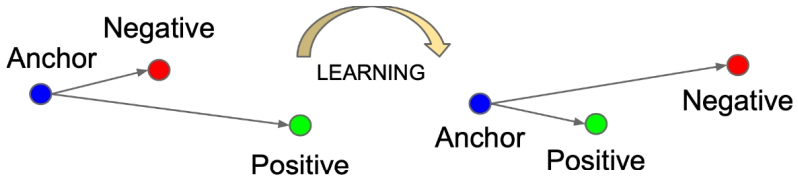

Fig. 1. Contrastive (triplet) loss function with one query (anchor), one positive key and one negative key (Figure taken from Schroff et al., 2015).

The aim of the triplet loss is to learn embeddings such that the distance $d(A, P)$ between the anchor and the positive is less than the distance between the anchor and the negative by a predefined margin. Figure 1 provides illustration of learning/updating mechanism contrastive loss function.

By pre-training on such self-supervised tasks, general-purpose features can be learned from unlabeled data that transfer well to downstream tasks with limited labeled data. Contrastive learning has achieved state-of-the-art results across diverse domains like computer vision, NLP and time series modeling. In this work, we explore its potential for representation learning from industrial monitoring data for predictive maintenance applications.

2.2. Transformer architecture

Transformers were introduced by Vaswani et al. as a novel architecture for machine translation tasks (Vaswani et al., 2017). Unlike previous recurrent networks that process sequences recursively (Hochreiter and Schmidhuber, 1997), Transformers rely on self-attention mechanisms. Self-attention allows Transformers to directly model relationships between different positions in the input sequence. Through self-attention, each position in the input sequence is able to attend to all other positions to draw connections between them. Figure 2 provides overview of self-attention mechanism. Concretely, for an input sequence of length n, self-attention first projects the sequence into three vectors - queries ($q_i \in Q$), keys ($k_i \in K$) and values ($v_i \in V$) where i is patch from a single sequence. The model then computes the dot products of the query vector of interest q_i with the keys of all input vectors, resulting in a vector of weights a_i for all the input tokens, which produces an attention matrix of shape [$n \times n$]. These dot products represent the similarity between each query-key pair. To normalize these raw attention scores, the model applies a softmax operation across each row to obtain the final attention weights. These weights essentially indicate the importance of each input position with respect to the query position. Finally, the attention matrix a_i is multiplied by the values vector to produce the attended outputs. By aggregating information from the entire sequence using attention weights, each output receives contextualized representations.

By stacking multiple self-attention layers, Transformers can draw on information from the entire input sequence to derive contextualized representations of each individual element. This ability to reason over global relationships sets Transformers apart from sequential or convolutional networks and helps them learn more general representations from variable-length input/output sequences. Their parallelizable nature also makes Transformers highly efficient for time series modeling tasks.

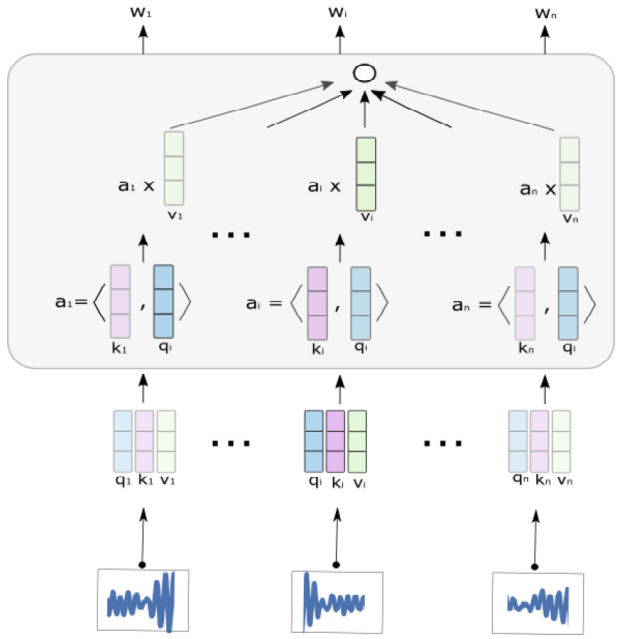

Fig. 2. Self-attention mechanism of transformer encoder
(figure adapted from Bukhsh et al., 2021)

3. Method

We propose to use a self-supervised pre-trained model with an active learning strategy to address the problem of limited labeled data for fault diagnosis. We refer to our approach as a human-guided contrastive sensor transformer (HG-CST). This methodology builds upon our previous work (Bukhsh, 2022) utilizing self-supervised learning, contrastive learning, and transformer architecture. Figure 3 provides an overview of our proposed methodology, which consists of two main components: pre-training a Contrastive Sensor Transformer (CST) model on unlabeled data and iteratively procuring the labels from oracle (or human annotator) for the most informative instances.

The CST model learns general-purpose representations through a combination of the transformer architecture and contrastive learning. Unlike conventional RNN/LSTM models, we utilize a transformer encoder $f(.)$ to encode condition monitoring signals for fault diagnosis. Due to its self-attention mechanism, the transformer architecture can better capture the complex temporal dynamics within input sensor data compared to traditional sequential models. The contrastive learning framework aims to maximize the agreement between embeddings of an original input segment and its augmented versions. During pre-training using only unlabeled data, we generate different views of each input segment through various transformations. The original segment acts as the query q, while its perturbed versions become positive keys $k+$. All other samples in a batch are considered negative keys $\mathbb{K}$. Through this framework, the CST encoder learns representations that are invariant to various transformations by maximizing similarity between the query and its matching positive keys, while reducing similarity to negatives. No direct supervision is required during pre-training. Further implementation details of the CST modeling approach can be found in (Bukhsh, 2022).

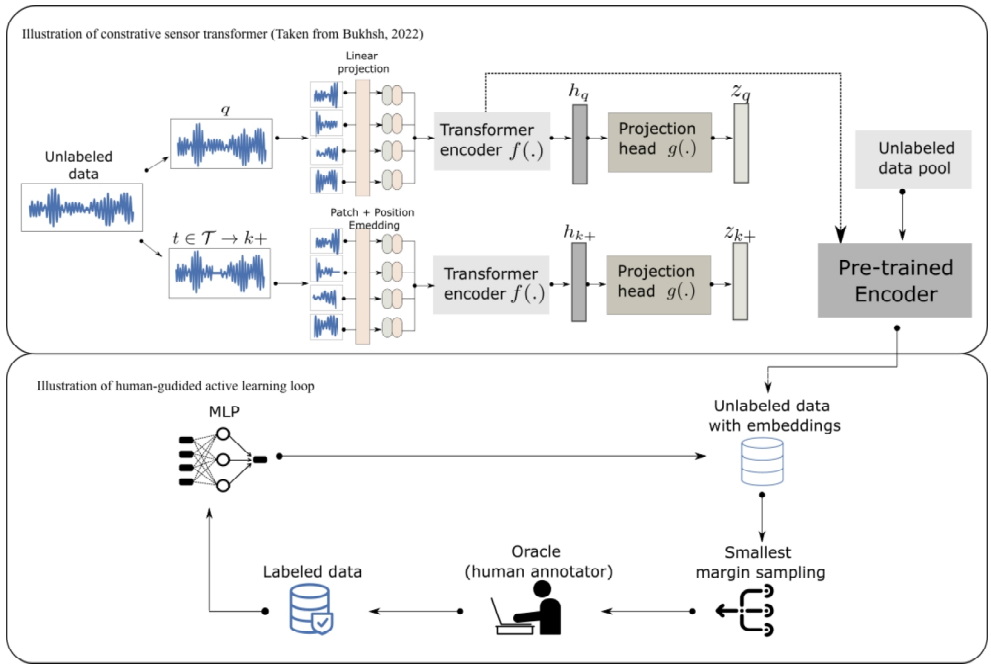

Fig. 3. Human guided contrastive sensor transformer (HG-CST).

Instead of using raw input data, we employ the pre-trained CST encoder to obtain generalized feature encodings of input signals, which capture patterns relevant to fault diagnosis. These embeddings provide a more robust representation of the data compared to the original high-dimensional sensor measurements. We begin with an initial small set of labeled samples (D_l) and a larger pool of unlabeled instances (D_u). The goal is to incrementally label more samples from D_u to maximize model performance while minimizing the labeling effort required. Our active learning process first trains a simple multi-layer perceptron (MLP) classifier on the features embeddings from CST of the original labeled D_l samples. This MLP is then used to predict the classes of instances in D_u. To select the most informative samples from D_u for labeling, we employ an uncertainty

sampling method called Smallest Margin (SM). SM is defined as the difference between the probabilities of the most likely and second most likely predicted classes for each sample (Lewis and Catlett, 1994). Samples with lower SM scores, where the classifier is less certain of the predicted class, are considered more uncertain. The k most uncertain samples selected from D_u using SM are then queried from an oracle (human annotator) to acquire their true labels. These newly labeled instances are merged back into the labeled set D_l. The MLP is retrained on the augmented D_l and the process repeats - selecting uncertain samples from D_u for labeling until the predefined active learning budget (e.g. maximum iterations) is reached.

4. Experiments

4.1. Datasets

We evaluate the learning capabilities of our approach on open benchmark bearing datasets, i.e., Case Western Reserve University (CWRU) (Smith and Randall, 2015) and the KAT dataset provided by Paderborn University (Lessmeier et al., 2016). Both of these datasets have been used widely for different fault diagnostic and detection tasks (Neupane and Seok, 2020). Table 1 provides an overview of classes in datasets for specific fault severity and fault type. The accelerometer sensor collected the CWRU vibration data at locations near and far off the motor bearing. The data is recorded with a sampling frequency of 48kHz and divided into sequence lengths of 512 points, with two channels based on the measurement location. The KAT dataset provides high-resolution data consisting of six healthy and 26 damaged bearings. The data is collected at a sampling rate of 64kHz, which is further segmented into fixed-length windows of 1200 points.

Table 1. Classes of statistics in the considered datasets with respect to fault severity & type
(fault severity refers to the damage level applied on the bearings, except for healthy state H,
the fault types present the specific damage location such as ball fault B, inner race fault IR & outer race fault OR).

CWRU dataset										
Class	0	1	2	3	4	5	6	7	8	9
Fault severity (am)	-	7	14	21	7	14	21	7	14	21
Fault type	H	B	IR	OR	B	IR	OR	B	IR	OR
KAT dataset										
Class	0	1	2	3	4					
Fault severity (mm)	-	≤ 2	≤ 2	>2	>2					
Fault type	H	IR	OR	IR	OR					

4.2. Experimental setup

We train the CST model to serve as an encoder for feature extraction in the active learning loop. The CST takes non-overlapping signal segments of size N as inputs (see Section 3). The transformer encoder consists of 4 multi-headed attention blocks with 64-unit MLPs. We apply global max pooling followed by layer normalization and tanh activation to obtain embeddings for pre-training. The embeddings are passed to a contrastive model with bilinear similarity to compute the loss. The CST is trained for 100 epochs using the Adam optimizer with a learning rate of 0.0001 to minimize the negative log likelihood loss. For baselines, we consider a Convolutional Autoencoder (Conv-AE) and fully supervised Convolutional network (Conv-FS). The Conv-AE encoder has 4 1D conv layers with increasing filters from 16-128 and kernel size 7, stride 2, with Reu activations and 0.1 dropout. For classification, we keep the Conv-AE encoder, apply global max pooling and add a dense layer. Conv-FS uses the Conv-AE encoder structure directly on labeled data.

We implement active learning using a pool-based sampling approach with a budget of 200 queries to the human oracle. At each iteration of the active learning loop, the simple MLP classifier is first trained for 100 epochs on the current labeled set D_l. The trained MLP is then used to predict the labels of all unlabeled instances in the pool D_u and calculate an uncertainty score (e.g. smallest margin) for each sample. Next, the most uncertain k samples are selected from D_u using either class-aware or class-agnostic sampling, where k is the batch size. ***Class-aware sampling*** selects 1 sample per class, while ***class-agnostic sampling*** picks samples randomly from the entire unlabeled pool D_u, providing fewer labels per iteration compared to class-aware. The selected k samples are queried to the oracle for labeling, after which they are removed from D_u and added to the labeled set

D_l. This process is repeated at each iteration, querying the oracle to label the most uncertain samples, until the total number of queries reaches the defined budget of 200. By iteratively focusing on the most informative samples according to model uncertainty, our goal is to maximize classification performance within the limited annotation budget.

4.3. Results

Figures 4 and 5 compare class-agnostic and class-aware sampling across datasets and methods. It shows how the validation accuracy improves as more labeled examples are added after each iteration. It is important to note that the CWRU has ten classes, and the KAT dataset has four classes (see Table 1); therefore, CWRU had more labeled examples to begin with and would have more total samples to train under a class-aware setting. Figures 4a and 4b show the performance of CST with HG learning on the CWRU and KAT datasets under a class-agnostic setting. For both, uncertainty sampling yields higher validation accuracy than random sampling when limited to a small percentage of labeled examples. Figures 4c and 4d show the results using Conv-AE as an encoder for both datasets. Here, the uncertainty sampling also performs better on CWRU, but on KAT, random sampling eventually surpasses it despite initially lagging. This could be attributed to KAT containing fewer classes (4 vs 10 for CWRU), limiting the effectiveness of class-agnostic uncertainty sampling with fewer within-class examples to refine the model early on. Figure 5 reports the validation accuracy under a class-aware setting. Here again, CST with HG learning obtained better results for both datasets, and uncertainty sampling consistently performed better.

This analysis is further confirmed with the evaluation of the test set. Table 2 reports the test accuracy and provides several key insights into the performance of our proposed CST with HG learning and baselines under limited labeled data settings. The top portion shows that the fully supervised Conv-FS model achieves near-perfect performance close to 99% for CWRU and 96% for the KAT dataset when trained on 100% of the labeled training data. This establishes an upper bound on accuracy that emphasizes the challenge of data scarcity. Our proposed CST with HG learning approach significantly outperforms autoencoder-based baseline when labeled data is limited to less than 1%. On both datasets, our approach with class-aware uncertainty sampling achieves the best results, exceeding 90% accuracy on CWRU and 65% on KAT. In class-aware sampling, using only the uncertainty approach provides a slight boost over random sampling. Class-aware sampling consistently outperforms class-agnostic sampling across methods and datasets. This indicates that actively querying the most informative samples according to the model uncertainty is valuable for training with minimal labels.

Class-aware sampling enables a more balanced exploration of the decision boundary by targeting one sample per class at each iteration. While still lagging HG-CST, Conv-AE combined with HG labeling improves over a purely random baseline. This demonstrates the value of (self-supervised) pre-training and active learning working together to enhance a model with limited labeled data.

Table 2. Comparison of test accuracy (%) with our approach (CST with HG) to other baselines (our approach obtains good performance while using less than 1% of the labeled training data).

Dataset	Model name	100% training data			
CRWU	Conv- FS	98.97			
KAT	Conv- FS	96.68			
Active learning results					
Less than 1% of labeled data		Class agnostic		Class aware	
		Random	Uncertainty	Random	Uncertainty
CWRU	HG-CST	80.25	**82.88**	91.07	**92.82**
KAT	HG-CST	51.58	**53.90**	63.33	**65.68**
CWRU	Conv-AE with HG	50.49	55.55	71.95	75.33
KAT	Conv-AE with HG	39.99	36.66	40.95	43.07

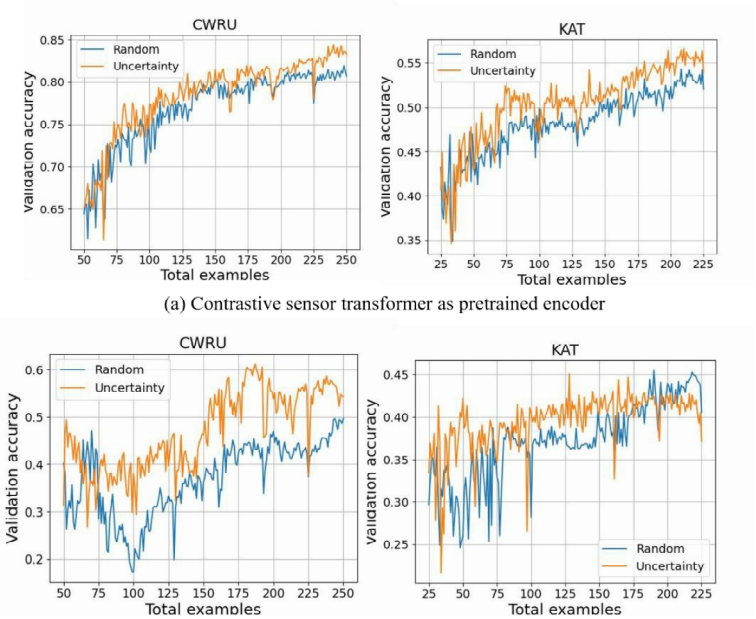

Fig. 4. Validation accuracy of class-agnostic sampling on both CWRU and KAT datasets using CST and AE as pretrained encoders.

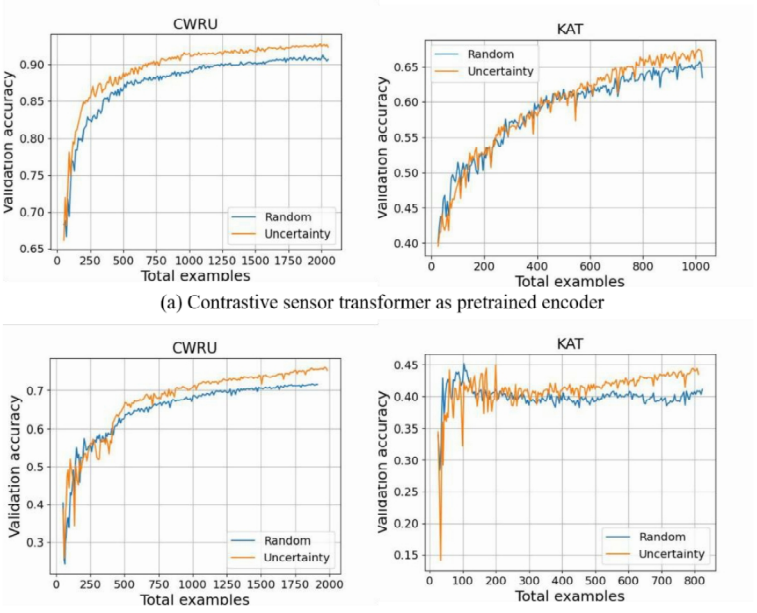

Fig. 5. Validation accuracy of class-aware sampling on both CWRU and KAT datasets using CST and AE as pretrained encoders.

5. Conclusion

We present a human-in-the-loop approach called HG-CST that leverages self-supervised pre-training and active learning for efficient fault classification. The CST model captures general representations from unlabeled data, which are then refined through iterative feedback from domain experts. By querying only the most informative samples according to the model's uncertainty, our active learning procedure aims to maximize diagnostic performance using a minimal labeling budget. This balancing of automatic modeling and human input allows HG-CST to address real-world challenges of limited labeled data as new fault types emerge over time. Our experimental results demonstrate that HG-CST can achieve over 90% and 65% accuracy on two bearing benchmark datasets using less than 1% of labels, significantly outperforming supervised baselines under limited data conditions. In particular, class-aware sampling that balances queries across classes and prioritizes uncertain examples is shown to improve active learning effectiveness. This confirms that jointly leveraging self-supervised feature extraction, contrastive representations, transformers, and human guidance maximizes value from scarce labeled data. The proposed HG-CST framework provides a data-efficient solution for developing reliable predictive maintenance capabilities with effective human involvement.

The future work plan is to evaluate the proposed approach on more datasets and expand the methodology to regression problems. While evaluated on bearing datasets, the underlying principles of self-supervised pre-training, contrastive learning, and active sampling are broadly applicable to various machine monitoring domains.

References

Bukhsh, Z. 2022. Contrastive Sensor Transformer for Predictive Maintenance of Industrial Assets, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3558–3562.

Chen, J., Zhou, D., Guo, Z., Lin, J., Lyu, C., Lu, C. 2019. An active learning method based on uncertainty and complexity for gearbox fault diagnosis. IEEE Access 7, 9022–9031.

Ding, Y., Zhuang, J., Ding, P., Jia, M. 2022. Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings. Reliab. Eng. Syst. Saf. 218, 108126.

Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.-J., Ducoffe, M. 2020. Potential, challenges and future directions for deep learning in prognostics and health management applications. Eng. Appl. Artif. Intell. 92, 103678.

Golyadkin, M., Pozdnyakov, V., Zhukov, L., Makarov, I., 2023. SensorSCAN: Self-supervised learning and deep clustering for fault diagnosis in chemical processes. Artif. Intell. 324, 104012.

Hochreiter, S., Schmidhuber, J. 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Hu, C., Wu, J., Sun, C., Yan, R., Chen, X. 2022. Inter-Instance and Intra-Temporal Self-Supervised Learning with Few Labeled Data for Fault Diagnosis. IEEE Trans. Ind. Inform.

Jian, C., Yang, K., Ao, Y. 2021. Industrial fault diagnosis based on active learning and semi-supervised learning using small training set. Eng. Appl. Artif. Intell. 104, 104365.

Jin, Y., Qin, C., Huang, Y., Liu, C. 2021. Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network. Measurement 173, 108500.

Lessmeier, C., Kimotho, J.K., Zimmer, D., Sextro, W. 2016. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification, in: PHM Society European Conference.

Lewis, D.D., Catlett, J. 1994. Heterogeneous uncertainty sampling for supervised learning, in: Machine Learning Proceedings 1994. Elsevier, pp. 148–156.

Neupane, D., Seok, J. 2020. Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: A review. IEEE Access 8, 93155–93178.

Saeed, A., Grangier, D., Zeghidour, N. 2021. Contrastive learning of general-purpose audio representations, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3875–3879.

Saeed, A., Ozcelebi, T., Lukkien, J. 2019. Multi-task self-supervised learning for human activity detection. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3, 1–30.

Senanayaka, J.S.L., Van Khang, H., Robbersmyr, K.G. 2020. Toward self-supervised feature learning for online diagnosis of multiple faults in electric powertrains. IEEE Trans. Ind. Inform. 17, 3772–3781.

Settles, B. 2009. Active Learning Literature Survey (Computer Sciences Technical Report No. 1648). University of Wisconsin–Madison.

Smith, W.A., Randall, R.B. 2015. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. Mech. Syst. Signal Process. 64, 100–131.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, \Lukasz, Polosukhin, I. 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, T., Qiao, M., Zhang, M., Yang, Y., Snoussi, H. 2020. Data-driven prognostic method based on self-supervised learning approaches for fault detection. J. Intell. Manuf. 31, 1611–1619.

Zhang, T., Chen, J., He, S., Zhou, Z. 2022. Prior knowledge-augmented self-supervised feature learning for few-shot intelligent fault diagnosis of machines. IEEE Trans. Ind. Electron. 69, 10573–10584.