Application of the LeJEPA self-supervised learning technique for the diagnosis of rotating looseness faults using vibration data

João Paulo Vieira[1], Victor Afonso Bauler[1], Thiago Carvalho Bittencourt[1], Bruno Forte de Lunardi Pinto[1], Karoline da Rocha[2], Diego Marques do Carmo[2], Danilo Silva[1]

[1]Federal University of Santa Catarina (UFSC)

[2]Dynamox

joaopaulovieirajpv@gmail.com, victor.bauler@gmail.com

Abstract. *Rotating looseness is an important industrial fault, yet the limited availability of public datasets has restricted the investigation of machine learning (ML) approaches for its diagnosis. This work investigates self-supervised learning (SSL) for rotating looseness diagnosis in bearing housings using an industrial vibration dataset. A LeJEPA-based framework is evaluated under a machine-wise protocol, where models are tested on unseen machines. Results show that SSL consistently improves deep learning (DL) performance in low-label settings, with gains of up to 13% to the DL baseline and a mean AUROC of 89.12% ± 4.78%. As more labeled data becomes available, a strong classical ML baseline prevails, reaching a mean AUROC of 90.73% ± 3.98%. These findings highlight the potential of SSL for industrial fault diagnosis under realistic deployment conditions, specially when few data is available.*

1. Introduction

Condition-based monitoring has become an essential component of modern industrial maintenance, supported by the widespread deployment of sensors and the increasing use of ML techniques for automated fault diagnosis [Randall 2021]. By continuously analyzing measurements such as vibration, temperature and electric current, ML-based systems can assist maintenance teams in detecting abnormal operating conditions, preventing catastrophic failures, and reducing operational costs [Zaidi et al. 2026].

Among the various fault modes encountered in rotating machinery, **rotating looseness** [Suechoey et al. 2021] in bearing housings is particularly relevant because it can introduce mechanical instability, increasing vibration levels and accelerating the development of other faults, including bearing damage [Jung et al. 2016]. This fault arises from excessive clearance in the bearing housing assembly, leading to impacts and modulation effects that generate harmonics of the shaft rotational frequency, often observable in the envelope spectrum [Jung et al. 2016]. Despite its practical importance, the limited availability of public datasets containing rotating looseness faults restricts the development and evaluation of ML diagnostic approaches. To address this gap, this work investigates the diagnosis of rotating looseness faults in bearing housings using a private industrial vibration dataset.

Recent studies have also shown that many fault diagnosis models are evaluated using protocols that introduce data leakage, leading to overly optimistic estimates of generalization performance [Wheat et al. 2024, Varejão et al. 2025, Vieira et al. 2025]. Consequently, realistic data partitioning strategies are essential when assessing diagnostic systems intended for deployment in industrial environments.

At the same time, deep learning (DL) methods have become increasingly attractive for fault diagnosis because they reduce the need for manual feature engineering. However, their performance often depends on the availability of labeled data, which can be costly to obtain in industrial settings. Since large quantities of unlabeled vibration data are typically available, SSL has emerged as a promising alternative for improving representation quality under limited-label regimes. Among recent SSL methods, LeJEPA [Balestriero and LeCun 2025] has demonstrated strong performance while maintaining relatively modest computational requirements.

In this work, we investigate the application of LeJEPA for rotating looseness diagnosis in conveyor-belt bearing housings. Unlike many previous studies that rely on randomized train-test splits, we employ a *machine-wise* evaluation protocol in which models are tested on machines that were not observed during training. This setting provides a more realistic assessment of generalization performance and avoids the risk of data leakage [Kapoor and Narayanan 2023]. We further compare SSL-based approaches against supervised classical and DL baselines under different label availability regimes.

In summary, the main contributions of this paper are:

- The investigation of a recent SSL technique, LeJEPA, to enhance the diagnosis of rotating looseness faults in bearing housings. To the best of our knowledge, this is the first study to investigate the application of SSL techniques for rotating looseness diagnosis.
- A comprehensive comparison of multiple training paradigms under different label availability regimes, including classical feature-based and DL models trained in a fully supervised setting, a lower-bound configuration based on a frozen randomized encoder, and SSL-based linear probing and fine-tuning approaches.
- A realistic evaluation protocol based on machine-wise data partitioning, where models are tested on machines unseen during training.

2. Related works

Driven by the success of SSL in domains such as Computer Vision and Natural Language Processing, recent studies have begun applying these techniques to rotating machinery fault diagnosis, proposing methods that improve classification performance, particularly in scenarios where labeled data is scarce [Wei et al. 2021, Yan et al. 2023, Wang et al. 2024, Tang et al. 2024, Zhou and Barati Farimani 2024, Liang et al. 2025, Lu et al. 2026]. Most of these works evaluate multiple label availability regimes, comparing SSL approaches against fully supervised baselines. A wide range of techniques have been explored, including contrastive learning [Liang et al. 2025, Lu et al. 2026], novel pretext tasks [Wang et al. 2024, Tang et al. 2024], reconstruction-based methods using Masked Autoencoders (MAE) [Zhou and Barati Farimani 2024], self-distillation techniques [Yan et al. 2023] and methods from the Canonical Correlation Analysis family [Wei et al. 2021].

Most of these studies rely on public bearing fault datasets, particularly those from Case Western Reserve University (CWRU) and Paderborn University (PU), although datasets from the University of Connecticut (UOC), University of Ottawa (UORED-VAFCLS), Hanoi University of Science and Technology (HUST), and Xi'an Jiaotong University (XJTU) are also frequently used. A common characteristic among these studies is the adoption of randomized train-test splits, which previous works have identified as a major source of data leakage, often leading to overly optimistic performance estimates [Wheat et al. 2024, Vieira et al. 2025]. These studies further demonstrate that even splits based solely on operating conditions may not provide a realistic assessment of model generalization.

In contrast to the existing literature, this work investigates the application of SSL techniques to the diagnosis of rotating looseness faults in bearing housings. Furthermore, our study is conducted on a large industrial dataset containing data collected from a diverse set of machines (see Section 3.1), contrary to the majority of public datasets, which are typically acquired from a single or few machines.

In this scenario, we are able to employ a more realistic evaluation protocol. As discussed in [Vieira et al. 2025] in the context of bearing fault diagnosis, to avoid data leakage when using data collected from a single machine, it is strictly necessary to adopt a bearing-wise splitting methodology, which evaluates models on data from bearings unseen during training. The present dataset allows us to further extend this idea through a *machine-wise* split, where models are evaluated on entirely unseen machines.

3. Methodology

3.1. Datasets

The datasets used in this work were collected from multiple acquisition points installed on low-speed conveyor belt bearings. Wireless sensors were configured to acquire vibration signals along three axes: horizontal, vertical, and axial. Both the pretraining and training datasets share similar rotational speed ranges (40 to 200 RPM) and acquisition durations, with most recordings lasting between 4 s and 5 s. As this work uses 3 s segments for training and evaluation, all shorter signals were discarded. Furthermore, while most recordings were acquired at a sampling frequency of 2048 Hz, signals collected at different sampling rates were resampled to 2048 Hz. For signals requiring downsampling, an anti-aliasing low-pass filter was applied prior to resampling to suppress frequency components above the target Nyquist frequency (1024 Hz).

The **pretraining** dataset consists of 25654 unlabeled signals, with a mean duration of 4.48 s, collected from 60 conveyor belts. The **training (downstream)** dataset consists of 2171 labeled signals collected from 47 conveyor belts (distinct from the pretraining ones), a mean duration of 4.07 s and a fault rate of 20.50%. Figure 1 summarizes the main characteristics of both datasets, including the distribution of rotational speeds, the number of signals per machine, and the acquisition durations. Additionally, Figure 2 exhibits an example comparing healthy and faulty signals on the time- and frequency-envelopes.

3.2. Data splitting and evaluation methodology

Our data splitting methodology is based on evaluating models on data from machines that were unseen during training, as this better reflects real industrial deployment scenarios.

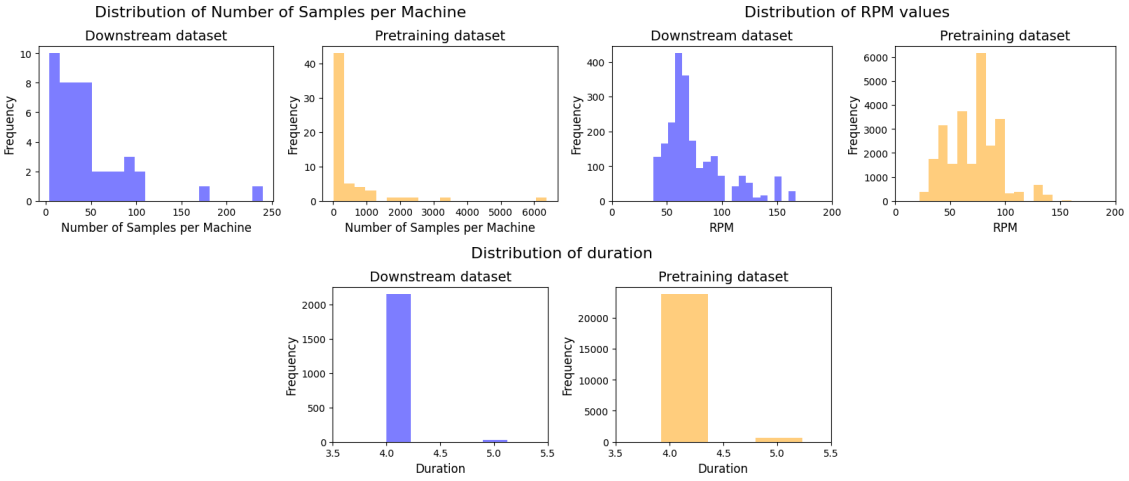


Figure 1. Comparison on downstream and pretraining datasets: RPM and duration range and number of samples per machine (conveyor-belt).

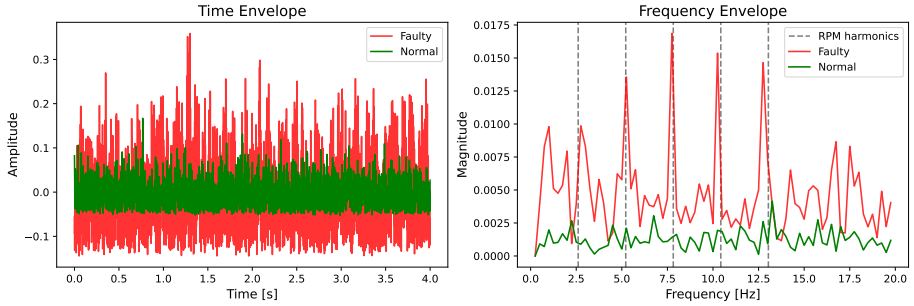


Figure 2. Example of healthy and faulty signals on time- and frequency-envelope representations. The rotating looseness fault is mostly diagnosed through the presence of one or multiple RPM harmonics along the frequency envelope.

As discussed in a previous work [Vieira et al. 2025], this evaluation strategy is necessary to avoid data leakage, since models (particularly high-capacity DL models) may learn machine- or bearing-specific characteristics instead of fault-related patterns, which could overestimate results if models were evaluated on data from the same training machines.

Since the training dataset contains data from 47 machines, we are able to explore multiple train-test partitioning scenarios using the CVM-CV strategy [Tsamardinos et al. 2015]. In this work, the CVM phase (model development) was performed using three train-test splits, while the CV phase (model evaluation) employed ten train-test splits. Splits are generated using machines as grouping units, with approximately 80% of the machines allocated to the training set and the remaining allocated to the test set. Additionally, the fault rate in both validation and test sets is constrained to remain between 15% and 30%, preventing the generation of splits with insufficient faulty samples.

Model performance is evaluated using the Receiver Operating Characteristic (ROC) curve, as it enables the analysis of classifier performance across different operating points. This is particularly useful in industrial applications, where specific performance metrics, such as the true positive rate (TPR), may need to be prioritized according to op-

erational requirements. Another advantage of the ROC curve is its independence from class prevalence, which is especially important given the class imbalance present in our dataset. To facilitate model comparison, we summarize each ROC curve using the Area Under the ROC Curve (AUROC). Since multiple splits are evaluated during the CV phase, the reported results correspond to the mean AUROC computed across all test splits.

3.3. Pretraining, linear probing and fine-tuning pipelines

Previous studies have shown that the Deep Convolutional Neural Network with Wide First-layer Kernels (WDCNN) [Zhang et al. 2017] architecture provides a strong baseline for fault diagnosis tasks while maintaining a relatively small number of trainable parameters. In this work, we use the encoder portion of the WDCNN architecture as the feature extractor. Following common practice in the SSL literature, the classification head is replaced by a single linear layer, enabling a more direct evaluation of representation quality during the linear probing stage.

During the pretraining phase, the encoder is trained using the LeJEPA framework. Multiple augmented views are generated for each input signal, and the model is optimized to learn representations that remain consistent across different views of the same sample. Following the formulation proposed in [Balestriero and LeCun 2025], the encoder f_θ produces a latent representation z, which is subsequently processed by the projector g_ϕ to obtain the projected representation p. The projector is implemented as a multilayer perceptron (MLP) composed of three fully connected layers. The first two layers expand the encoder representation to twice its original dimensionality and are each followed by batch normalization and ReLU activation, while the final layer projects the resulting features to a lower-dimensional latent space whose dimensionality is one quarter of the encoder embedding dimension. The SSL loss is computed from these projected representations and used to update the encoder and projector parameters. The data augmentation procedures used to generate the different views are described in Section 3.4.

After pretraining, two downstream evaluation strategies are considered: linear probing and fine-tuning. In the linear probing setting, the encoder f_ξ remains frozen and only a linear classification head h_ψ is trained, providing a direct assessment of the quality of the learned representations. In the fine-tuning setting, both the encoder f_θ and the classification head h_ψ are jointly optimized using labeled data. In both cases, a single augmented view is generated for each training sample using the same augmentation pipeline adopted during pretraining. Since the task is formulated as a binary classification problem, models are trained using the Binary Cross-Entropy loss. Figure 3 illustrates the complete pipeline, including the pretraining, linear probing, and fine-tuning stages.

To reduce evaluation bias when employing the CVM-CV strategy, each train-test split seed is also used to control the randomization of several components of the experimental pipeline, including encoder initialization, data augmentation procedures, sample ordering in the dataloaders, and linear head initialization. Consequently, each split is evaluated under an independent combination of model parameters and training conditions, reducing the influence of a particular random configuration on the reported results.

3.4. Input representation and data augmentation techniques

In this work, preliminary experiments using the development (CVM) splits comparing raw time-domain, frequency-domain, frequency-envelope, and time-envelope representa-

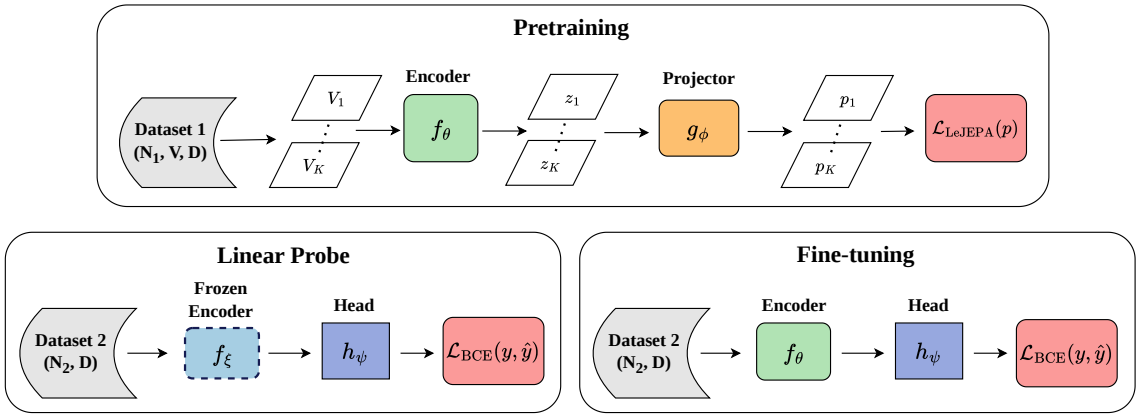


Figure 3. Pretraining, linear probing and fine-tuning pipeline descriptions. During pretraining, the unlabeled dataset with $N_1 = 188705$ samples is used; during the linear probing and fine-tuning phase, the labeled dataset with $N_2 = 2171$ samples is used.

tions showed that the time-envelope input consistently achieved the best performance. Consequently, it was adopted throughout the SSL experiments. Envelope analysis is a well-established technique that demodulates high-frequency resonance bands, facilitating the identification of fault-related frequency components [Smith and Randall 2015, Randall 2021]. To extract the temporal envelope, all signals were first band-pass filtered between 50 and 800 Hz using a 12th-order Butterworth filter, implemented as a cascade of second-order sections. This frequency range was found to contain relevant fault information under the low-speed operating conditions considered in this work. Then, the Hilbert transform was then applied to obtain the envelope.

Based on previous studies [Yan and Liu 2022, Liang et al. 2025, Vieira et al. 2025], two data augmentation techniques were employed to generate multiple views of each signal. The first, Random Crop, extracts a 3 s segment from a randomly selected position within the signal. The second, Random Gain, introduces stochastic amplitude variations by multiplying the signal by a scalar sampled from a normal distribution $\mathcal{N}(\mu = 1, \sigma^2)$. Consequently, multiple views generated from the same acquisition correspond to different temporal segments with independently randomized amplitudes, as illustrated in Figure 4. Since each acquisition contains three channels corresponding to the horizontal, vertical, and axial directions, augmentations are applied independently to each channel, increasing the variability of the generated views and strengthening the regularization effect.

3.4.1. Handcrafted features for classical models

To provide a strong classical ML baseline, all DL models are directly compared against a random forest (RF) classifier (implemented by Yggdrasil Decision Forests [Guillame-Bert et al. 2023]) trained on specialized handcrafted features. The selected feature set combines domain knowledge from vibration analysis and was designed to capture characteristics commonly associated with rotating looseness faults.

The extracted features include the magnitudes of rotational harmonics in

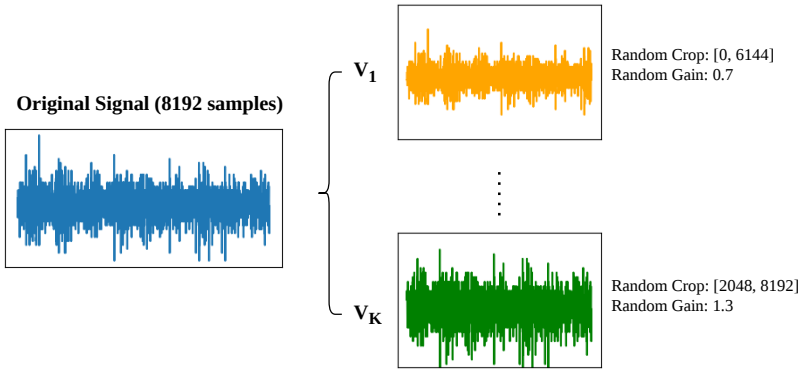


Figure 4. Generation of multiple views of a signal by applying the Random Crop and Random Gain data augmentation techniques.

the frequency-envelope, as well as the Peak-to-Floor Ratios (PFRs) proposed in [Vieira et al. 2024], computed from these harmonics. Additionally, several statistical descriptors are extracted from the time-envelope, including peak-to-peak amplitude, RMS, carpet RMS, and skewness. The feature selection was based on previous studies [Abburi et al. 2023, González et al. 2023]. All features are computed independently for each vibration axis and for two frequency bands: 5 Hz–200 Hz and 50 Hz–800 Hz. Table 1 summarizes the complete set of handcrafted features used in this work.

Domain	Features
Time-envelope	Peak-to-peak, rms, carpet RMS and skewness values.
Frequency-envelope	Magnitude of rotation harmonics (1x - 3x); PFRs calculated over the same harmonics.

Table 1. Handcrafted envelope features used to train the Random Forest model. All features were calculated for the three axes and on envelopes extract on two distinct bands: 5 Hz–200 Hz and 50 Hz–800 Hz.

3.5. Model development

During model development, the quality of the learned representations is periodically assessed throughout the pretraining process. For each optimization split, the encoder is evaluated on the downstream task by freezing its parameters, reinitializing the linear classification head, and training the head from scratch. By repeating this procedure at different stages of pretraining and across all optimization splits, we are able to identify the pretraining duration that maximizes the mean AUROC.

The hyperparameter optimization process includes LeJEPA-specific parameters, namely the coefficient λ that weights the sum of the invariance and SIGReg loss terms and the number of views V; pretraining parameters, including the learning rate lr_{pt}, batch size bs_{pt}, and the number of warmup samples n_w (using a linear warmup schedule with a start factor of 0.1); and downstream training parameters, namely the learning rate lr_d and batch size bs_d. The complete hyperparameter search space is presented in Table 2.

To monitor downstream performance throughout pretraining, the linear probe is evaluated every 20000 pretraining samples. For each evaluation, the classification head is

Hyperparameter	Search space
Deep Learning	
lr_{pt}	[0.00001, 0.0001, 0.0005]
bs_{pt}	[64, 128, 256]
n_w	[25000, 50000, 100000]
V	[2, 3, 5]
λ	[0.05, 0.1, 0.15, 0.2, 0.25]
lr_d	[0.001, 0.0001]
bs_d	[32, 64]
Random Forest (Yggdrasil)	
max_depth	[3, 9, 15, 21, 27, 33, 39, 45, 51, 60]
min_examples	[13, 24, 35, 57, 68, 79, 90, 101, 115, 126, 137]

Table 2. Hyperparameter search space for model development.

trained for up to $n_t = 20000$ samples, and its performance is recorded every 2048 training samples, producing a complete learning curve. The maximum pretraining budget is set to $n_{pt} = 200000$ samples.

Throughout the experiments, the selected DL hyperparameters were $lr_{pt} = 5 \times 10^{-4}$, $bs_{pt} = 128$, $n_w = 50000$, $V = 2$, $\lambda = 0.25$, $lr_d = 1 \times 10^{-3}$, $bs_d = 64$. For the RandomGain augmentation, a standard deviation of 0.3 yielded the best performance.

3.6. Implementation details

All of the DL experiments presented in this work were implemented in Python using the PyTorch framework and executed on a workstation equipped with an NVIDIA RTX 3090 GPU and 64 GB of RAM. Both pretraining and downstream training employed the AdamW optimizer with a fixed weight decay of 5×10^{-2}.

To increase sample diversity within each batch, we implemented a custom batch sampler that leverages the dataset metadata. Since multiple acquisitions are available for each sensor location, batch construction is performed in two stages. First, a sensor location is randomly selected with replacement using a uniform distribution. Subsequently, one acquisition associated with that location is randomly sampled. This strategy promotes a more balanced representation of sensor locations during training and reduces the influence of locations with a larger number of acquisitions.

As a consequence of this sampling procedure, training is controlled by the total number of processed samples rather than by the number of epochs. This design choice enables direct optimization of both the pretraining budget and the downstream training budget through the hyperparameters n_{pt} and n_t, respectively.

4. Experimental results

In the following experiments, we compare five training scenarios under different levels of label availability:

- **Classical ML:** a Random Forest classifier with 100 estimators is trained using the handcrafted features described in Section 3.4.
- **Fully Supervised Baseline (FS):** the customized WDCNN model is trained from scratch using only the labeled training dataset.

- **Lower Bound (Random Features):** the encoder is randomly initialized and kept frozen throughout training, while only the classification head is optimized. This scenario serves as a reference point for evaluating whether the learned representations obtained through SSL pretraining provide information beyond that of a random feature extractor.
- **Linear Probing:** the encoder is first pretrained using LeJEPA and subsequently frozen, while only the classification head is trained using labeled data.
- **Fine-Tuning:** the encoder is first pretrained using LeJEPA, after which both the encoder and classification head are jointly optimized using labeled data.

Using the hyperparameters selected during the model development phase with 100% of the available training data, we evaluate all five scenarios under progressively smaller training sets corresponding to 100%, 80%, 60%, 40%, 20%, 10%, 5%, and 2.5% of the original training dataset. For each training-set size, samples are drawn in a nested manner, while preserving the class proportions described in Section 3.2. This sampling procedure is repeated independently for each CV split. This experimental design enables the analysis of model performance across a wide range of label availability regimes, including extremely low-label scenarios. As a result, we are able to identify which learning paradigm provides the best performance under each data availability condition. The main results are summarized in Figure 5 and detailed in Table 3.

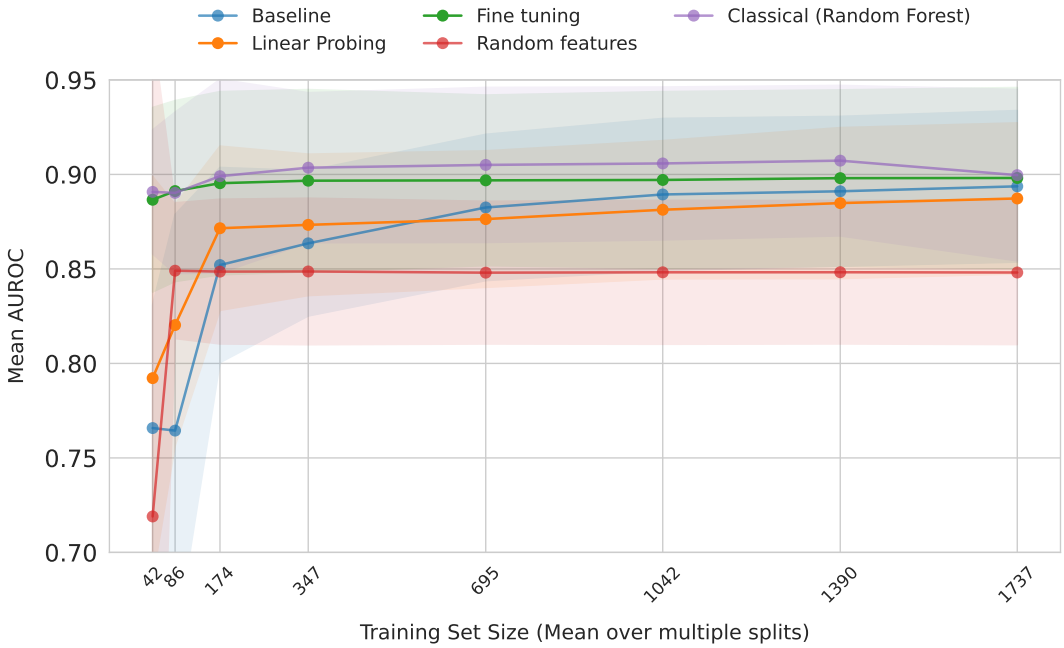


Figure 5. Performance curves for the proposed experimental scenarios.

5. Discussion

Our results reveal three main findings. First, SSL-based approaches consistently improve upon the fully supervised DL baseline when only a limited amount of labeled data is available. The largest gains are observed in the most extreme low-label settings (42 to 86 training samples), where fine-tuning substantially outperforms training from scratch,

	Run Type				
Training Percentage	**Baseline**	**Lower bound**	**Linear probing**	**Fine-tuning**	**Classical (RF)**
2.5% (42)	76.58% ± 6.51%	71.90% ± 25.68%	79.22% ± 10.67%	88.66% ± 4.88%	89.08% ± 3.28%
5.0% (86)	76.45% ± 11.43%	84.91% ± 3.59%	82.03% ± 6.39%	**89.12% ± 4.78%**	89.02% ± 4.27%
10.0% (174)	85.20% ± 5.16%	84.86% ± 3.83%	87.15% ± 4.34%	89.54% ± 4.84%	89.91% ± 5.12%
20.0% (347)	86.35% ± 3.84%	84.87% ± 3.87%	87.33% ± 3.74%	89.67% ± 4.82%	**90.36% ± 3.96%**
40.0% (695)	88.25% ± 3.87%	84.80% ± 3.77%	87.64% ± 3.61%	89.69% ± 4.51%	**90.50% ± 4.10%**
60.0% (1042)	88.94% ± 4.02%	84.82% ± 3.81%	88.13% ± 3.65%	89.71% ± 4.67%	**90.58% ± 4.04%**
80.0% (1390)	89.11% ± 3.96%	84.82% ± 3.79%	88.48% ± 3.99%	89.80% ± 4.67%	**90.73% ± 3.98%**
100.0% (1737)	89.37% ± 4.01%	84.81% ± 3.81%	88.73% ± 3.99%	89.81% ± 4.78%	**89.96% ± 4.53%**

Table 3. Experimental results comparing five training scenarios on multiple train sizes: Baseline (Supervised DL), Lower Bound (linear probing on random features), Linear probing, Fine-tuning and Classical (RF). Train percentage values are reported along with the corresponding estimated number of samples (mean along test splits). The performances are reported through the mean $\pm$ std AUROC values calculated over 10 distinct train-test splits.

achieving up to a 13% AUROC gain in the most extreme scenario, while linear probing provides smaller but still significant improvements. Notably, fine-tuning reaches performance close to that of the Random Forest in these low-label settings. These results indicate that the representations learned during pretraining capture information that remains useful even when only a small number of labeled samples is available for downstream training.

Second, the lower-bound configuration achieves remarkably stable performance across all training sizes, maintaining AUROC values close to 85%. In the most challenging low-label scenarios, it even surpasses the fully supervised baseline, suggesting that optimizing the entire network from scratch becomes difficult when only a few labeled samples are available. These results will be investigated in future works and may be partly attributed to the large-kernel convolutional filters in WDCNN, which could provide a useful inductive bias for automatically extracting relevant signal characteristics.

Finally, we observe that the handcrafted-feature Random Forest achieves the best performance at most training percentages, except at 5.0%, indicating that a high-quality handcrafted representation can surpass the WDCNN in the evaluated scenarios. However, the performance gap between the Random Forest and the SL baseline progressively decreases as more labeled data becomes available, suggesting that, beyond a certain point, DL-based approaches may eventually outperform the classical model.

Overall, our results indicate that SSL is an effective strategy for fault diagnosis under limited-label conditions, providing a favorable trade-off between predictive performance and the development effort required for feature engineering.

6. Conclusions

In this paper, we investigated the use of the LeJEPA self-supervised learning framework for diagnosing rotating looseness faults in bearing housings. To the best of our knowledge, this is one of the first studies evaluating SSL techniques for this fault mode. By employing a machine-wise data-splitting methodology, we assessed model performance under realistic industrial conditions while avoiding data leakage.

The results showed that SSL-based approaches consistently outperform a fully supervised DL baseline when only a limited amount of labeled data is available. As the

number of labeled samples increases, fine-tuning achieves performance comparable to a strong Random Forest baseline trained on handcrafted features. These findings suggest that SSL can narrow the performance gap between end-to-end DL models and carefully engineered classical ML pipelines while retaining the advantage of automatic feature extraction. Future work will investigate additional data augmentation strategies, alternative encoder architectures such as Transformers, and the extension of the evaluation to additional fault modes and operating conditions.

References

Abburi, H., Chaudhary, T., Ilyas, S. H. W., Manne, L., Mittal, D., Williams, D., Snaidauf, D., Bowen, E., and Veeramani, B. (2023). A closer look at bearing fault classification approaches. In *Annual Conference of the PHM Society*, volume 15.

Balestriero, R. and LeCun, Y. (2025). LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics. Version Number: 3.

González, M., Díaz, V. G., Pérez, B. L., G-Bustelo, B. C. P., and Anzola, J. P. (2023). Bearing fault diagnosis with envelope analysis and machine learning approaches using cwru dataset. *IEEE Access*, 11:57796–57805.

Guillame-Bert, M., Bruch, S., Stotz, R., and Pfeifer, J. (2023). Yggdrasil decision forests: A fast and extensible decision forests library. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4068–4077. ACM.

Jung, J., Lee, S. B., Lim, C., Cho, C.-H., and Kim, K. (2016). Electrical Monitoring of Mechanical Looseness for Induction Motors With Sleeve Bearings. *IEEE Transactions on Energy Conversion*, 31(4):1377–1386.

Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804.

Liang, Q., Liang, X., Zhang, M., Liu, C., and Zhang, T. (2025). A Novel Comprehensive Semi-Supervised Learning Method for Fault Diagnosis Under Extremely Low Label Rate. *IEEE Transactions on Instrumentation and Measurement*, 74:1–14.

Lu, W., Wang, Y., Wang, W., Zhang, S., and Cai, Z. (2026). A Multi-Task Self-Supervised Fault Diagnosis Method for Rotating Machinery Under Limited Labeled Data. *Quality and Reliability Engineering International*, page qre.70272.

Randall, R. B. (2021). *Vibration-based condition monitoring: industrial, automotive and aerospace applications*. John Wiley & Sons.

Smith, W. A. and Randall, R. B. (2015). Rolling element bearing diagnostics using the case western reserve university data: A benchmark study. *Mechanical Systems and Signal Processing*, 64-65:100–131.

Suechoey, B., Siriporananon, S., Chupun, P., Boonkhun, C., and Chompooinwai, C. (2021). Performance analysis and fault classification in a large electric motor using vibration assessment technique. *International Journal of Intelligent Engineering and Systems*, 14:124–133.

Tang, J., Xiao, J., Chen, W., Li, X., Wei, C., Ding, X., and Huang, W. (2024). A prior knowledge-enhanced self-supervised learning framework using time-frequency invariance for machinery intelligent fault diagnosis with small samples. *Engineering Applications of Artificial Intelligence*, 133:108503.

Tsamardinos, I., Rakhshani, A., and Lagani, V. (2015). Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. *International Journal on Artificial Intelligence Tools*, 24(05):1540023.

Varejão, I. M. D. S., Costa, L. G. D. O., Silva, L. H. P. D., Rodrigues, A., Ribeiro, M. P., Varejão, F. M., and Oliveira-Santos, T. (2025). The similarity bias problem: What it is and how it impacts vibration based intelligent fault diagnosis. *Mechanical Systems and Signal Processing*, 235:112822. Publisher: Elsevier BV.

Vieira, J. P., Bauler, V. A., Rosa, R. K., and Silva, D. (2025). Towards a more realistic evaluation of machine learning models for bearing fault diagnosis. Version Number: 4.

Vieira, J. P., Rosa, R. K., Bauler, V. A., Braga, D., and Silva, D. (2024). Bearing fault diagnosis using machine learning and a novel set of fault-related spectral features. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2024)*, pages 120–131, Brasil. Sociedade Brasileira de Computação - SBC.

Wang, H., Wang, X., Yang, Y., Gryllias, K., and Liu, Z. (2024). A Few-Shot Machinery Fault Diagnosis Framework Based on Self-Supervised Signal Representation Learning. *IEEE Transactions on Instrumentation and Measurement*, 73:1–14.

Wei, M., Liu, Y., Zhang, T., Wang, Z., and Zhu, J. (2021). Fault Diagnosis of Rotating Machinery Based on Improved Self-Supervised Learning Method and Very Few Labeled Samples. *Sensors*, 22(1):192.

Wheat, L., Mohrenschildt, M. V., Habibi, S., and Al-Ani, D. (2024). Impact of Data Leakage in Vibration Signals Used for Bearing Fault Diagnosis. *IEEE Access*, 12:169879–169895. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Yan, Z. and Liu, H. (2022). 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. Publisher: MDPI AG.

Yan, Z., Liu, H., Tao, L., Ma, J., and Cheng, Y. (2023). A Universal Feature Extractor Based on Self-Supervised Pre-Training for Fault Diagnosis of Rotating Machinery under Limited Data. *Aerospace*, 10(8):681.

Zaidi, S. H. H., Shenfield, A., Zhang, H., and Ikpehai, A. (2026). A Systematic Review of Anomaly and Fault Detection Using Machine Learning for Industrial Machinery. *Algorithms*, 19(2):108.

Zhang, W., Peng, G., Li, C., Chen, Y., and Zhang, Z. (2017). A New Deep Learning Model for Fault Diagnosis with Good Anti-Noise and Domain Adaptation Ability on Raw Vibration Signals. *Sensors*, 17(2):425. Publisher: MDPI AG.

Zhou, A. Y. and Barati Farimani, A. (2024). FaultFormer: Pretraining Transformers for Adaptable Bearing Fault Classification. *IEEE Access*, 12:70719–70728.