Research on Engineering Structures & Materials
www.jresm.org

Research Article

Cross-load generalization of bearing fault diagnosis via prototypical few-shot learning on vibration benchmarks

Kai Wang [*,a]

School of Automotive Engineering, Henan College of Industry & Information Technology, China

Article Info

Article History:

Received 16 July 2026

Accepted 01 Sep 2026

Keywords:

Few-Shot learning;
Prototypical networks;
EV gearbox fault diagnosis;
Cross-Load generalization;
Deep metric learning;
Vibration analysis;
1D-CNN

Abstract

Limited labelled bearing-fault data are a practical challenge for EV gearbox condition monitoring, motivating diagnostic methods that can learn effectively from only a few examples. EV gearboxes also operate under varying speed and load conditions and may experience torque transients, which can alter vibration signatures and further challenge the generalization of diagnostic models trained with limited data. To investigate this problem in a controlled and reproducible setting, this study evaluates a Prototypical Network for few-shot bearing fault diagnosis using the publicly available CWRU vibration benchmark. A 4-way K-shot protocol is used with K = 1, 3, 5, and 10, and a lightweight 1D-CNN is employed to extract robust embeddings while preventing overfitting on small support sets. In the pooled-load benchmark evaluation, ProtoNet achieves 98.87% accuracy in the 1-shot case, 99.80% in the 3-shot case, 99.58% in the 5-shot case, and 99.78% in the 10-shot case, outperforming SVM, KNN, and a supervised 1D-CNN under the same comparison protocol. Additional leave-one-load-out tests across all three CWRU load permutations and K = 1, 3, 5, and 10 maintain mean accuracies above 99.80% over 1,000 independent test episodes per condition. These results demonstrate strong few-shot performance and robustness to the limited load shifts represented by the CWRU benchmark, providing encouraging methodological evidence for the intended EV gearbox monitoring context. Because CWRU does not reproduce the full speed, transient-torque, noise, and installation variability of an EV gearbox, hardware-specific EV validation remains an important next step.

1. Introduction

1.1. Background and Motivation

Drive train safety for electric vehicles (EVs) is becoming increasingly important as bearing failures inside gearboxes could start off as a minor vibration or noise and develop into catastrophic failure of the whole transmission. Recent reviews of EV fault diagnosis and automotive powertrain monitoring also identify vibration-based condition monitoring as a central route for drivetrain health assessment [1-4]. Current industry practice for drivetrain diagnostics is based on using accelerometer data and processing it through segmentation and feature extraction methods in order to identify impact signatures within the vibration signal that can indicate the presence of inner-ring (IR), ball (B) or outer-ring (OR) faults or healthy bearings. Much of the state-of-the-art in this area has been achieved through the use of deep learning, where models have obtained high classification accuracy on bearing fault diagnosis datasets. However, there is a significant problem of obtaining sufficient representative labelled vibration data to train such models for EV transmission gearboxes. Bearing faults in practice are random events that do not occur very often, which makes labelling sufficient data samples very time consuming and, in practice, it is unlikely

*Corresponding author: kaiwang892@outlook.com
[a]orcid.org/0009-0009-0726-6198
DOI: http://dx.doi.org/10.17515/resm2026-1889vk0716rs

that more than a few samples per fault type will be available for training. Further complicating the problem are the variable operating conditions under which EVs operate including load, speed, temperature and lubrication, which all affect how a bearing fault signal is propagated to the accelerometer attachment point. In addition, EV traction systems can operate over wide and high rotational-speed ranges, which broaden the frequency range of vibration components and increase the bandwidth and sampling-rate requirements for reliable vibration acquisition. The regenerative braking process also results in abrupt torque transients, leading to non-stationary vibration characteristics and possible distribution shifts between operating conditions. Such variations can reduce the generalization ability of diagnostic models trained under a limited set of conditions, particularly when only a few labelled samples are available for adaptation. To the best of the authors' knowledge, there is no publicly available, comprehensive vibration dataset specifically recorded from an EV gearbox for fault diagnosis. Even if a few samples are available for each fault class at one particular operating load, it is unlikely that enough samples would be available to train a model to generalize across loads. This paper therefore investigates few-shot bearing fault diagnosis and controlled cross-load generalization using an open vibration benchmark, motivated by the data-scarcity and operating-condition challenges encountered in EV gearbox monitoring.

1.2. Research Gaps and Challenges

Few-shot learning (FSL) typically trains a model on a small dataset and bypasses the need for a large labelled training set by learning a metric in some embedding space or network structure to recast recognition as a similarity comparison problem. Hence, FSL has been shown to be effective for bearing and gearbox diagnosis in recent literature using Prototypical and Siamese network architectures. By further employing transfer learning and pseudo-labelling, FSL can even be extended to cross-domain scenarios. Nevertheless, three methodological gaps remain relevant to data-efficient bearing diagnosis for potential EV gearbox monitoring.

First, existing few-shot learning studies have predominantly focused on generic rolling-element bearings or other industrial applications, with limited focus on EV-relevant operational profiles, particularly variations in load and speed. In the present study, CWRU is used as a controlled and reproducible vibration benchmark for evaluating the proposed few-shot learning framework under limited load shifts, rather than as a direct proxy for an EV-specific gearbox. Secondly, although standard evaluation protocols for the CWRU dataset exist, a standardized evaluation protocol specifically tailored for few-shot learning in bearing diagnostics-defining fixed episode construction and data splitting-is still lacking. Even for those papers that do utilize a fixed dataset and split, key hyperparameters-such as segment length, overlap, and K-are typically to be determined by the individual researcher, producing unreliable cross-study comparisons. Finally, models trained at single operating points may exhibit degraded generalization under changes in loading conditions, which vary continuously in real-world EV driving. In addition to linear running between high and low speed, EV gearbox bearings are subjected to start–stop cycling and part-load operation, meaning that single-load evaluation provides little guarantee of in-vehicle performance. Quantitative cross-load generalization performance therefore remains a relatively under researched area in few-shot learning for diagnostic applications. The framework outlined subsequently addresses these methodological gaps through a fixed open-benchmark protocol and explicit cross-load evaluation, while EV gearbox monitoring remains the motivating application for subsequent hardware-specific validation.

1.3. Objectives and Contributions

We seek a reproducible few-shot learning framework for bearing defect detection and explore its pooled-load and cross-load performance using data from CWRU. Specifically, we present a 1D-CNN Prototypical Network (1D-CNN ProtoNet) designed for the aforementioned segmented vibration time series data. For robust model evaluation, we perform 4-way K-shot learning using the CWRU drive-end (DE) bearing vibration data under 1, 2, and 3 HP load conditions, corresponding to approximately 1772, 1750, and 1730 RPM, respectively. We choose a 1024-point window with 50% overlap. We consider Normal and three levels of defect (IR, Ball, and OR) for classification. For the pooled-load evaluation, episodes are sampled from the available CWRU load conditions, with support and query samples kept separate within each episode. For the cross-load evaluation, the

model is trained using two load conditions and tested on the remaining unseen load without target-load fine-tuning. All three leave-one-load-out load permutations are evaluated. We use three supervised learning baselines: SVM, KNN and a supervised 1D-CNN. All experiments in the present study use CWRU as a controlled vibration benchmark rather than data recorded from an EV gearbox. Accordingly, the present work is positioned as a pilot benchmark study of the proposed few-shot learning framework, while its generalization to the wider speed, load, and transient operating conditions of real EV gearboxes requires further hardware-specific validation.

1.4. Paper Organization

Section 2 reviews the relevant state-of-the-art; Section 3 details the proposed 1D-CNN ProtoNet and few-shot learning framework; Section 4 describes the experimental setup; Section 5 presents and discusses the results, including cross-load robustness; and Section 6 concludes the study and outlines future work.

2. Related Work

2.1. Vibration-Based Fault Diagnosis

Impacts of bearing faults in rotating machineries are manifested in vibration signals. The frequency of such impacts is related to the location of faults, i.e., ball pass frequency inner (BPFI), ball pass frequency outer (BPFO) and ball spin frequency (BSF). The fault related frequency components are embedded or masked within the broadbands of vibration signals. Thus, time-frequency analysis and some statistical features are usually employed to improve the fault diagnosis performance. The short-time Fourier Transform (STFT) and wavelet transform are commonly used time-frequency transforms. Several features including Root Mean Square (RMS), kurtosis, spectral centroid, spectral bandwidth, spectral excess kurtosis, spectral kurtosis, Crest factor, and Shape factor are extracted from time-frequency transformed signals. Two of the most commonly used features, namely, RMS and kurtosis have been exploited widely. The energy of a signal band (in spectral domain) is proportional to its RMS value, and it correlates well correlates with wear and overall loading on a machine. The kurtosis feature of a signal captures the Transient Impulses within the signal and is commonly used for incipient fault detection. Although these fixed features have been widely and successfully applied for bearing fault diagnosis, they suffer significant limitations. Beyond conventional feature-based methods, recent studies have also explored advanced signal-processing, simulation-driven learning, and digital-twin-based approaches for bearing and gear diagnostics. For example, variable-speed bearing diagnosis has been investigated through envelope-spectrum characteristic-frequency analysis [5], while simulation-driven machine learning [6] and digital-twin frameworks [7] have been applied to bearing fault classification and gear degradation prognosis. Large variations in operating conditions or in machine dynamics result in variations that are large enough to significantly degrade the performance of classifiers that are trained under a specific operating condition. Due to large variability of images of the same object, there is an increasing interest in data-driven approach capable of learning to recognize an object from very few examples. Deep learning methods drastically changed the rule of feature extraction in time domain. More recently, 1D-CNN, LSTM, TCN and other deep learning architectures have been applied to time-series fault diagnosis in rotating machinery and EV transmission bearings [8,9]. Features can be extracted directly from the amplitude of the raw signal and, unlike traditional methods, feature extraction is no longer a manually designed process. The entire sensing to diagnosis process of automotive powertrains can be reviewed in the light of order analysis techniques and the processing of time varying signals using time-frequency methods to tackle the non-stationary fault characteristics. In particular for EVs, data-driven approaches are becoming increasingly prominent for motor, battery and gearboxes etc. diagnostics. A crucial assumption of the above methods is that a large labelled dataset is available. While this can be achieved through real world EV driving, the EV test rigs can be utilized in the meantime to generate the required data. This may not be true on real EV gearboxes where faults are extremely rare and annotating such faults is very time consuming and expensive.

2.2. Few-Shot Learning for Fault Diagnosis

Foundational few-shot learning formulations include Siamese one-shot learning [10], model-agnostic meta-learning (MAML) [11], and Prototypical Networks [12]. In FSL, many classification methods can be viewed as searching in the embedding space, either fixed prototypes or Siamese networks learning embeddings of small support sets and then classifying a new query sample based on distance or pairwise scores. Recently, such methods have been applied to vibration-based diagnosis. The wavelet-prototypical method in [13] combines wavelet transforms with a prototypical network for few-shot gearbox detection and reports strong small-sample performance. Weighted prototype and anti-noise Siamese variants in [14] and [15] further show that bearing diagnosis can be handled under limited-data conditions. The hydrodynamic-transmission study in [16] extends few-shot diagnosis through transfer learning, while the pseudo-labeling framework in [17] uses time-frequency preprocessing for aero-engine bearings. Cross-domain few-shot diagnosis has also been studied through meta-learning and domain-adversarial graph modeling [18]. The review in [19] shows that metric-learning methods remain competitive at very small data sizes, whereas the industrial-machinery study in [20] compares preprocessing choices across MIMII, CWRU, and MFPT. Meta-transfer and MAML-style approaches further address operating-condition shifts [21,22]. Related motor and EV-powertrain studies include wide-kernel CNN/Siamese models for PMSM faults [23] and residual-current few-shot learning for EV inverter faults [24]. Data augmentation and multimodal fusion have been explored through GAN/CBAM rolling-bearing diagnosis [25] and MMFSL for industrial bearings [26]. Wavelet-prototypical fusion also confirms the usefulness of embedding-based few-shot learning on vibration signals [27]. Compared with wavelet-prototypical and anti-noise Siamese approaches, the present method adopts a lightweight 1D-CNN Prototypical Network operating directly on segmented vibration signals, without additional time-frequency preprocessing or dedicated anti-noise modules. However, systematic few-shot evaluation using a fixed open-benchmark protocol, particularly under cross-load conditions, remains limited in bearing fault diagnosis.

2.3. Open-Access Vibration Benchmarks

The experiments in this study use the publicly available CWRU bearing dataset, including drive-end vibration measurements, EDM-induced bearing faults, a 12 kHz sampling rate, and three motor load conditions. CWRU is retained because it provides a public and widely used vibration benchmark with controlled fault classes and operating conditions, allowing the proposed few-shot protocol to be evaluated reproducibly against established bearing-diagnosis studies. Nevertheless, recognized limitations and anomalies in the CWRU dataset have been documented in previous benchmark analyses [28]. Accordingly, the present study treats CWRU as a controlled methodological benchmark rather than as representative of the broader operating variability of an EV gearbox. Differences in dataset partitioning, preprocessing, and episodic construction make direct comparison across few-shot diagnostic studies difficult. However, due to the lack of a standardized evaluation protocol, it is generally not possible to perform a direct comparison between different approaches in the field, Liang et al. [19]. Using time–frequency prototypical network by Wang et al. on CWRU bearing data, wavelet-based frequency information improves the classification performance when only a small number of training samples are available, which makes this network a good starting point for a few-shot learning protocol. We utilize CWRU at three different motor loads (1, 2 and 3 HP), each having different shaft speeds (approximately 1772–1730 RPM). These operating speeds are determined by the CWRU benchmark and are used here as a controlled starting point for methodological evaluation; they do not represent the wider speed range encountered in real EV gearboxes. We test whether a model trained at one load can generalize to other loads, in addition to the existing fault severity levels and bearing fault types included in the CWRU dataset. We use a single experimental protocol throughout the work to enable reproducibility. This includes fixed window length, overlap, and number of ways for a fixed number of training cases (K-shot learning), as well as the pooled-load and leave-one-load-out cross-load evaluation settings.

2.4. Summary of Research Gaps

Despite substantial progress in vibration-based fault diagnosis and few-shot learning, several methodological issues remain in reproducible few-shot bearing diagnosis. First, existing studies employ different preprocessing procedures, episodic settings, and operating-condition partitions, which makes direct comparison across studies difficult. Existing few-shot learning studies have predominantly focused on generic bearing benchmarks, while systematic evaluation under changes in operating load remains limited. Secondly, open benchmarks have not been consistently utilized in the literature. For example, different window lengths, data split ratios and/or different episode construction methodologies have been used at different times for supposedly the same dataset (e.g. [19]). Thirdly, and most importantly, cross-load generalization has rarely been quantified. As mentioned earlier, EV gearboxes are subject to dynamic changes in load and speed, therefore explicit evaluations of train-on-one-load and test-on-another methodologies (e.g. [16, 20]) are required. The present study evaluates the proposed few-shot learning framework using a fixed protocol on the CWRU bearing benchmark, with EV gearbox monitoring retained as the motivating application context. Results were evaluated under both pooled-load and leave-one-load-out cross-load settings.

3. Methodology

3.1. Overall Framework

As shown in Figure 1, the architecture consists of standard few-shot learning stages to process raw CWRU vibration data: windowing and normalization of the time series. Following this, an episodic few-shot learning loop is implemented with N classes and K support samples per class. A single 1D-CNN is employed as a shared-weight encoder for both support and query samples, such that the same network parameters are used to embed all windowed and normalized time-series segments into the same embedding space. For each class, the mean embedding of the support samples is computed as the class prototype, and each query sample is classified according to its Euclidean distance to these prototypes. The use of an episodic framework as opposed to training on a large fixed dataset is due to the small number of labelled samples per fault type.

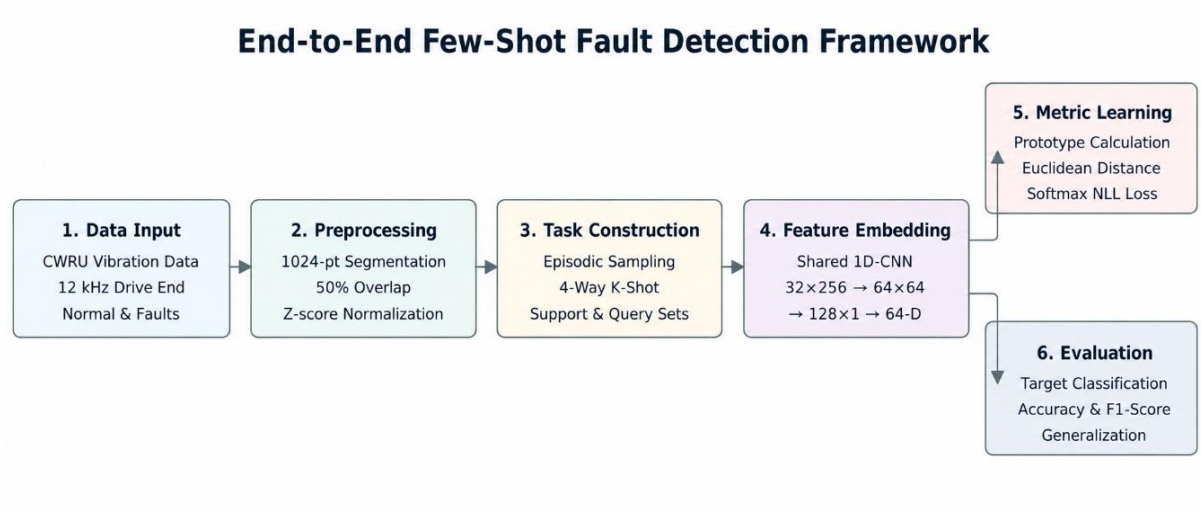

Fig. 1. End-to-end few-shot fault detection pipeline

In practice the CWRU segments are of a fixed window size with 50% overlap between successive segments. Hence all individual segments are first separately Z-score normalized, and then collectively treated as a single set of samples for training a given network. For training a given network, the same four classes are always used, K supports and query samples are drawn for each class, and then passed through the same 1D-CNN, after which prototypes for each class are computed and the negative log-likelihood loss is evaluated. After that Adam is used for back propagation to update the CNN's weights. The test accuracy and F1 score are then computed using the trained CNN according to the evaluation protocol specified in Section 3.3. The codebase for this

work consists entirely of Python, and only uses the following libraries: PyTorch for all of the neural network layers/functions, and scikit-learn for an SVM baseline as well as functions to calculate class weights for K. The functions to calculate the mean and standard deviation of time series samples for Z-score normalization are implemented using SciPy. All results were generated with a fixed random seed for reproducibility.

3.2. Signal Acquisition and Preprocessing

The CWRU Bearing Data Center provides drive-end bearing data for a test rig that includes a variable speed drive motor with a torque transducer, a dynamometer, and coupling to the test bearing trial. The bearing is an SKF 6205-2RS JEM. Three different defects were machined on the IR and OR bands and on the balls using EDM at 4 different diameters (0.007–0.028 in.). Three motor load conditions were considered: 1 HP at approximately 1772 RPM, 2 HP at approximately 1750 RPM, and 3 HP at approximately 1730 RPM. The DE accelerometer was sampled at 12 kHz. The bearing data are provided in the form of MATLAB files with the ".mat" extension. The relevant parameters are listed in Table 1.

Table 1. CWRU bearing test-rig technical parameters

Parameter	Specification
Sensor Type	Drive End (DE) accelerometer
Sampling Rate	12,000 Hz
Defect Introduction	Electrical discharge machining (EDM)
Fault Types	Inner Race (IR), Ball (B), Outer Race (OR)
Defect Diameters (inch)	0.007, 0.014, 0.021, 0.028
Motor Load (HP)/ Rotational Speed (RPM)	1/1772, 2/1750, 3/1730

For this work, 27 fault files and 3 normal time history files have been recorded. The normal files were recorded at three different loads, and the IR, Ball, and OR defects are recorded at three different loads each. All files are recorded on the DE transducer channel. All of the time history files are recorded at full length without any sampling down. The length of the time history files therefore varies, with the fault files being approximately 121k-123k samples (or ~10 s) long, and the normal files approximately 484k-486k samples (or ~40 s) long. In this work, all four defect diameters and all three loads are included, rather than studying diameter effects or load effects in this paper. Window lengths of 512, 1024, and 2048 points, corresponding to approximately 42.7, 85.3, and 170.7ms at the 12kHz sampling rate, were evaluated. A 1024-point window was selected as a compromise between capturing sufficient fault-related temporal information and retaining an adequate number of training segments. At a representative shaft speed of 1750 RPM, one revolution lasts approximately 34.3ms, so a 1024-point window spans approximately 2.5 shaft revolutions. For the CWRU drive-end bearing, the BPFI and BPFO are 5.4152 and 3.5848 times the shaft rotational frequency, corresponding to approximately 157.9 Hz and 104.6 Hz at 1750RPM. Thus, one 1024-point window contains approximately 13.5 BPFI periods and 8.9 BPFO periods, providing multiple fault-related impact opportunities within each input segment. The window steps by 512 points, which is roughly twice the number of segments that would be obtained with 0% overlap. After segments were Z-score normalized (mean subtracted and then divided by standard deviation) within files, the following segment counts per file were used (Table 2).

Table 2. Dataset segmentation summary

Class	Fault Type	Sample Segments	Percentage
0	Normal	2835	30.73%
1	Inner Race (IR) Fault	2841	30.79%
2	Rolling Element (Ball) Fault	2129	23.07%
3	Outer Race (OR) Fault	1422	15.41%

A total of 9,227 segments were produced and labeled for the total data set (2,835 Normal, 2,841 IR, 2,129 Ball, 1,422 OR) and numbered accordingly for classification (0-3). Figure 2 provides some

waveforms from the different fault classes. In the IR and Ball faults, there are strong periodic effects observed within the waveform, whereas the OR faults have effects that are more evenly spaced within the waveform. For segment class definition, this project followed CWRU's guidelines for segmenting roller bearing data so the research can be comparable with past studies.

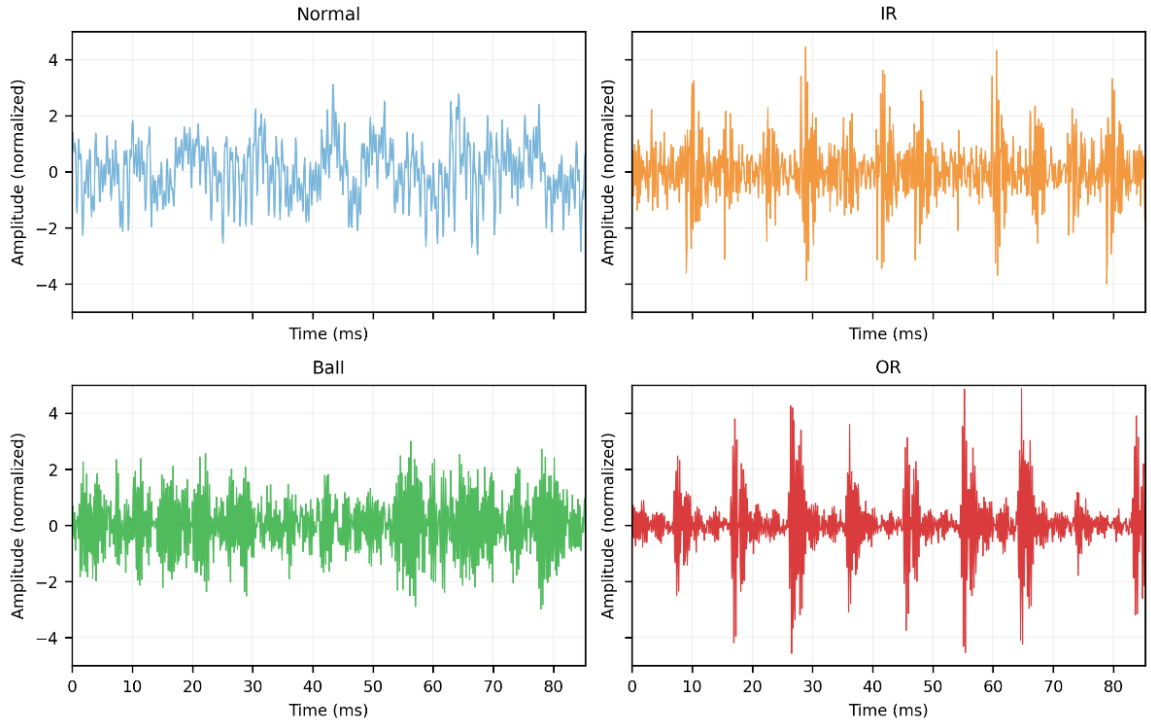

Fig. 2. Time-domain waveforms of Z-score-normalized vibration segments for the four classes

3.3. Few-Shot Task Construction

We fix N=4 (i.e., all four classes are in every episode) and vary K in {1, 3, 5, 10}. Increasing K from 1 to 10 significantly changes both the accuracy and stability of the network as it sees a support set of increasing size. K=1 is the most relevant scenario, given the data scarcity assumption of this work. For each class, we use an average of 15 queries per episode (resulting in 60 queries per episode). For each K-shot setting, the ProtoNet is trained using 200 episodically generated training tasks. For the pooled-load evaluation, performance is evaluated over 100 independently sampled test episodes. For the leave-one-load-out cross-load evaluation, each K-shot setting and load permutation is evaluated over 1000 independently sampled test episodes. Accuracy and macro-F1 are reported as the mean and standard deviation across the test episodes.

- N = 4: Normal, IR, Ball, OR appear in every episode.
- K ∈ {1, 3, 5, 10}: number of support samples per class.
- Query: 15 samples per class per episode.
- Training: 200 episodes; pooled-load testing: 100 episodes; added cross-load testing: 1,000 episodes per condition; results reported as mean ± std.

3.4. Model Architecture and Metric Learning

To learn segment embeddings, the network uses three sequential 1D convolutional blocks, followed by a fully connected projection layer to obtain the final embedding representation. Large feature maps from early layers capture high level local structure while later layers capture finer details. The first block utilizes a kernel size of 7 to incorporate a slightly longer local region of the signal, corresponding to approximately 0.58 ms at the 12 kHz sampling rate. The following two blocks have kernel sizes of 5 and 3 respectively. The first two convolutional blocks are followed by stride-4 max-pooling, reducing the temporal dimension from 1024 to 256 and then to 64, respectively. The third block is followed by adaptive average pooling, reducing the resulting feature map to 128×1 before the final 128-to-64 linear projection. The 1D-CNN therefore learns an

embedding of a 1024-point segment into a 64-dimensional vector. Step by step the capacity of the network is increased by powers of two, by increasing the number of filters from 32 to 64 to 128 dimensional filters, within three corresponding blocks of convolutional layers. After the third such block, adaptive average pooling is utilized in order to step down the temporal dimension to a single index, the corresponding channel. Then a linear projection from 128 to 64 dimensions follows, without any activation function. We then L2 normalize the 64-dimensional vector, which means that we keep the distances between the vectors scale consistent, which is important for prototype matching.

- Conv Block 1: 32 filters, kernel 7 (padding 3), BN, ReLU, max pool stride 4 → (32, 256).
- Conv Block 2: 64 filters, kernel 5 (padding 2), BN, ReLU, max pool stride 4 → (64, 64).
- Conv Block 3: 128 filters, kernel 3 (padding 1), BN, ReLU, adaptive avg pool → (128, 1).
- FC: 128 → 64, output L2-normalized.

Following the standard Prototypical Network formulation [12], the prototype of class n is computed as the mean of its K support embeddings c_n:

$$c_n = \frac{1}{K}\sum_{i=1}^{K} f_\theta(x_{n,i}) \tag{1}$$

A query embedding $f_\theta(\cdot)$ is assigned to the class whose prototype is nearest under Euclidean distance:

$$d(q, c_n) = \|q - c_n\|_2 \tag{2}$$

Euclidean distance is retained (rather than a learned metric or cosine) because, at K = 1, the prototype is simply the single support embedding and the assignment rule is unambiguous; for small K the mean already provides a low-variance estimate, and preliminary trials showed no benefit from more elaborate metrics. SoftMax over negative distances yields probabilities, and training minimizes NLL:

$$\mathcal{L} = -log\frac{exp[-d(q, c_y)]}{\sum_{n=1}^{N} exp[-d(q, c_n)]} \tag{3}$$

with y the true class of the query.

3.5. Training Strategy and Implementation

The model is optimized with Adam (lr 0.001, β_1=0.9, β_2=0.999) with a StepLR scheduler that halves the learning rate every 50 training steps. For each model setting, training is performed over 200 episodically generated tasks. In each episode, support and query samples are independently sampled according to the predefined 4-way K-shot protocol, and previously generated episodes are not replayed. To ensure reproducibility, fixed random seeds are used for NumPy (42) and PyTorch (123), including the episodic sampling procedure. The code runs without errors with Python 3.10, PyTorch 2.2.2, scikit-learn 1.7.2, SciPy 1.15.3, NumPy 1.26.4. No proprietary software is required.

4. Experimental Setup

4.1. Benchmark Dataset and Partitioning Scheme

The training and evaluation procedures use the 9,227 segments described in Section 3.2. The four classes contain 2,835 Normal, 2,841 IR, 2,129 Ball, and 1,422 OR segments. Although these raw totals are imbalanced, each episode is constructed with equal numbers of samples from all four classes. For the pooled-load benchmark comparison, episodes are sampled from the three motor load levels (1, 2, and 3 HP), with support and query samples kept non-overlapping within each episode rather than using a fixed segment-level train/test ratio. For cross-load testing, a leave-one-load-out scheme is used: Loads 1+2 → 3, Loads 1+3 → 2, and Loads 2+3 → 1. The held-out load is not used for fine-tuning.

4.2. Baseline Methods

Three reference baselines are included to compare metric-based few-shot classification with conventional small-sample classifiers and a supervised neural-network baseline.

- SVM (RBF): C=10, γ= `scale. Trained per episode on the support set and evaluated on the query set.
- KNN: $K = min(K_{shot}, 5)$the support set is used as the reference, Euclidean nearest-neighbor classification. No training step is required.
- Supervised 1D-CNN: shares the same 1D-CNN backbone as the embedding network, with an added linear classifier; fine-tuned on the support set for 50 epochs per episode before evaluation on the query set [8]. This is used as a conventional small-sample supervised reference rather than as an episodically trained FSL method. No early stopping was used for this baseline.

SVM and KNN use the same 1024-point segments as ProtoNet. Fourteen fixed handcrafted features are extracted from each segment: mean, standard deviation, maximum, minimum, RMS, mean absolute value, crest factor, kurtosis, shape factor, skewness, spectral centroid, spectral energy, peak frequency, and spectral standard deviation. The feature set is kept fixed for all comparisons.

4.3. Evaluation Metrics

- Accuracy: mean ± std over independent test episodes (100 episodes for the pooled-load comparison and 1,000 for each added cross-load condition).
- Macro F1: average of the four per-class F1 scores. Because every episode contains equal query counts from all four classes, Macro F1 and Accuracy can be numerically very close at high classification accuracy.
- Confusion matrix: normalized, ProtoNet 5-shot, to reveal which fault pairs are most often confused.

4.4. Ablation and Sensitivity Analysis

All methods in the pooled-load comparison are evaluated at K = 1, 3, 5, and 10. The cross-load analysis is also evaluated at K = 1, 3, 5, and 10 for each of the three leave-one-load-out permutations. No additional architecture or dataset ablation is introduced in this study.

5. Results and Discussion

See Tables 3 and 4 for the Accuracy and macro F1 of all methods under the 4-way K-shot setting (K = 1, 3, 5, 10).

Table 3. 4-way K-shot classification accuracy comparison (mean ± std, 100 test episodes)

Method	1-shot	3-shot	5-shot	10-shot
ProtoNet	98.87 ± 2.04	99.80 ± 0.59	99.58 ± 0.79	99.78 ± 0.56
Supervised 1D-CNN	63.07 ± 7.75	77.42 ± 7.81	84.63 ± 6.71	91.48 ± 5.03
SVM	57.80 ± 7.93	67.03 ± 7.16	71.80 ± 6.40	78.17 ± 5.61
KNN	57.75 ± 7.97	62.13 ± 5.99	63.68 ± 7.66	73.45 ± 6.28

Table 4. 4-way K-shot macro-averaged F1 score comparison

Method	1-shot	3-shot	5-shot	10-shot
ProtoNet	98.87	99.80	99.58	99.78
Supervised 1D-CNN	63.12	77.44	84.64	91.48
SVM	57.83	67.05	71.80	78.16
KNN	57.77	62.10	63.61	73.45

ProtoNet achieves 98.87 ± 2.04% accuracy in the 1-shot setting, compared with 63.07 ± 7.75% for the supervised 1D-CNN, 57.80 ± 7.93% for SVM, and 57.75 ± 7.97% for KNN. At 3-shot, ProtoNet

reaches 99.80 ± 0.59%. A small decrease to 99.58 ± 0.79% is observed at 5-shot, followed by 99.78 ± 0.56% at 10-shot; this small non-monotonic variation is treated as sampling variability rather than as a systematic trend. The strong low-shot performance is consistent with the metric-learning formulation: instead of fitting a new decision boundary from a very small support set, ProtoNet compares queries directly with class prototypes in a learned embedding space. The baselines improve as K increases, but remain below ProtoNet under the reported protocol. Figures 3 and 4 visualize the corresponding accuracy comparisons.

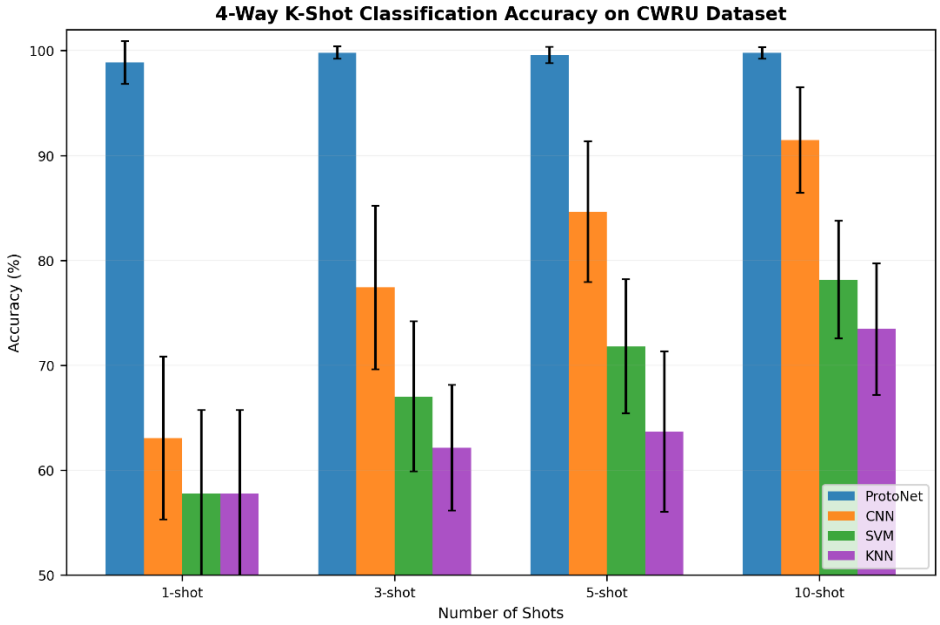

Fig. 3. Classification accuracy comparison bar chart for the four methods under 4-way K-shot

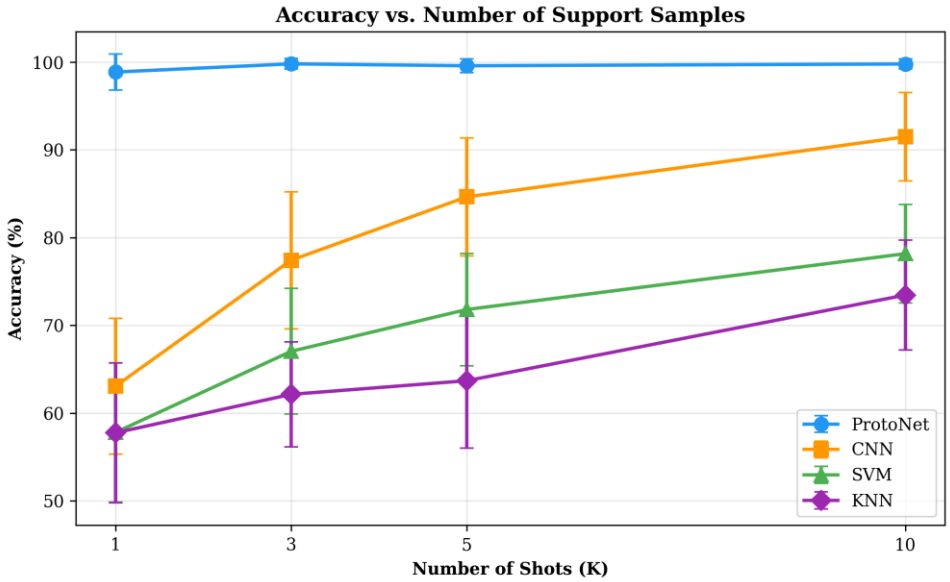

Fig. 4. Accuracy vs. number of support samples for each method

The pooled-load results are obtained on the CWRU benchmark and should therefore be interpreted as methodological evidence for the proposed few-shot bearing-classification framework. In the context of this study, they support the feasibility of a data-efficient diagnostic approach relevant to EV gearbox monitoring, while direct EV gearbox performance remains to be established using hardware-specific data.

5.2. Cross-Load Generalization Performance

Table 5 reports the added leave-one-load-out cross-load evaluation. The model is trained using two CWRU load levels and evaluated on the third held-out load, and all three permutations are tested at K = 1, 3, 5, and 10. Across all three held-out-load configurations, mean accuracy remains above 99.80% for every tested K. For Loads 1+2→3, the accuracy ranges from 99.81% to 99.97%; Loads 1+3→2 yields 100.00% in the reported runs; and Loads 2+3→1 ranges from 99.96% to 99.98%. Macro-F1 values differ from Accuracy by less than 0.01 percentage points because each episode uses balanced query sets. These additional permutations show that the original cross-load observation is not unique to one particular held-out CWRU load.

Table 5. Cross-load generalization under leave-one-load-out evaluation (accuracy %, mean ± std over 1,000 test episodes)

Training loads	Unseen load	1-shot	3-shot	5-shot	10-shot
1 + 2	3	99.94 ± 0.32	99.81 ± 0.56	99.90 ± 0.40	99.97 ± 0.23
1 + 3	2	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
2 + 3	1	99.96 ± 0.24	99.97 ± 0.24	99.98 ± 0.20	99.98 ± 0.20

The three operating points nevertheless have relatively close shaft speeds and represent controlled laboratory loading. The results therefore support robustness to modest operating-condition shifts within CWRU and provide encouraging evidence for further evaluation under the wider speed, transient-torque, noise, temperature, and installation variability encountered in EV gearboxes. Quantitative extrapolation to those broader conditions would require additional hardware-specific validation.

5.3. Error Analysis

Figure 5 shows the normalized confusion matrix for ProtoNet at 5-shot in the pooled-load evaluation. The diagonal entries dominate, with only a small number of off-diagonal errors. The few Ball–IR errors may be associated with similarities in their transient vibration signatures; however, the present experiment does not establish a specific physical mechanism for these errors. Figure 6 provides a qualitative t-SNE visualization of the 64-dimensional embedding space. The visualization is used only to illustrate the learned representation and is not interpreted as quantitative evidence of K-dependent cluster evolution.

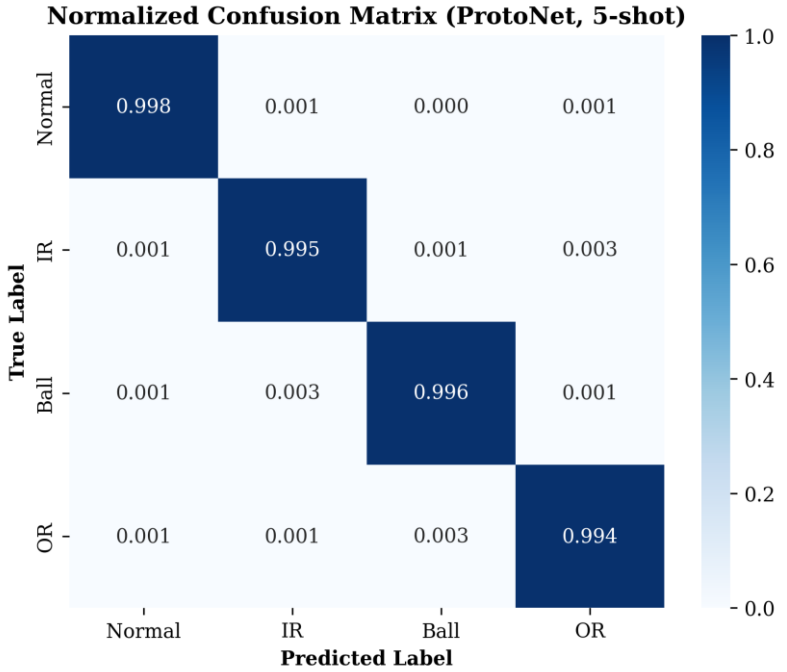

Fig. 5. ProtoNet 5-shot normalized confusion matrix

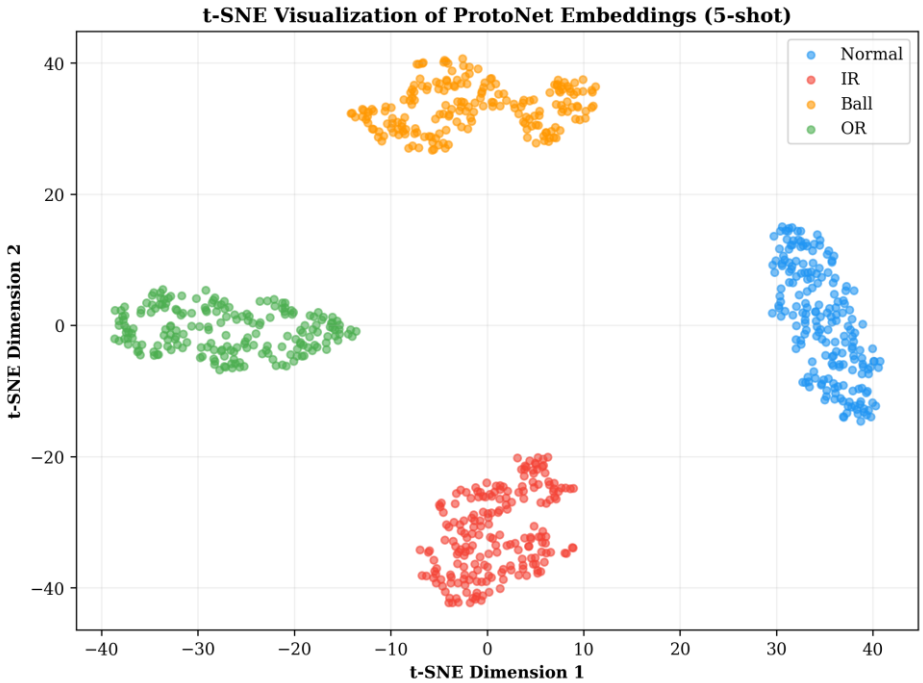

Fig. 6. ProtoNet 5-shot embedding-space t-SNE visualization

5.4. Engineering Implications

The results indicate that, within the controlled CWRU benchmark, the ProtoNet can achieve high classification accuracy with only a few labelled support segments per class. Performance changes little beyond approximately 3-5 shots in the pooled-load results, suggesting diminishing returns from additional support examples for this particular benchmark. This data-efficient behaviour is relevant to the limited-labelled-data motivation of EV gearbox monitoring, but it should not be interpreted as a complete field-deployment requirement. Real gearbox measurements include additional sources of variability, including road and structural noise, temperature, sensor placement, wider speed ranges, and transient loading. These factors define the next stage of hardware-specific validation and, where necessary, domain adaptation.

6. Conclusions and Future Work

6.1. Principal Findings

This study evaluates a 1D-CNN Prototypical Network for 4-way few-shot bearing fault diagnosis on the CWRU benchmark. In the pooled-load comparison, ProtoNet achieves 98.87 ± 2.04% at 1-shot, 99.80 ± 0.59% at 3-shot, 99.58 ± 0.79% at 5-shot, and 99.78 ± 0.56% at 10-shot, outperforming the reported SVM, KNN, and supervised 1D-CNN baselines. The additional leave-one-load-out evaluation covers all three CWRU load permutations and K = 1, 3, 5, and 10, with mean accuracy above 99.80% in every tested condition. Together, these results establish a reproducible benchmark-level demonstration of strong few-shot classification and robustness to the limited load shifts represented by CWRU, and provide encouraging methodological evidence for the intended EV gearbox monitoring context. Direct performance under the broader operating conditions of real EV gearboxes remains to be established through hardware-specific validation.

6.2. Practical Significance

The practical significance of the present result is its data efficiency under a controlled laboratory benchmark. The method requires only a small support set at inference, while the metric-learning representation avoids fitting a separate high-capacity classifier from those few examples. This characteristic is relevant to EV gearbox monitoring scenarios where labelled fault examples may be scarce. Before field application, however, the approach should be validated on hardware-specific data that include realistic speed variation, transient torque, noise, temperature, mounting, and

sensor-location effects. The present results therefore provide a methodological basis for subsequent EV gearbox validation rather than a claim of completed field deployment.

6.3. Limitations and Future Work

Three limitations should be emphasized. First, the experiments use only the CWRU bearing benchmark, whose load and speed variations are limited and whose laboratory fault generation does not represent the full diversity of field faults. Second, no vibration measurements from an EV gearbox are used in the present evaluation, so the influence of road noise, temperature, sensor placement, wide speed ranges, and torque transients remains to be quantified. Third, the present study evaluates only the Normal, IR, Ball, and OR bearing classes and a lightweight 1D-CNN embedding; it does not compare alternative datasets or more complex architectures. Future work will therefore extend the benchmark findings through additional bearing datasets and, importantly, dynamometer and in-service EV gearbox measurements.

Acknowledgements

The publicly available CWRU Bearing Data Center was used for the vibration data in this study. The dataset can be accessed at https://engineering.case.edu/bearingdatacenter. Lab colleagues were consulted during the development of the work.

References

[1] Kumar MP, Velpula S, Saiprakash C, Sahoo B. Advancements in fault detection and diagnosis methods for electric vehicles: a review. Discov Appl Sci. 2025;7(11):1235. https://doi.org/10.1007/s42452-025-07758-9

[2] Tao Y, Wang X, Zhang L, Bao X, Xue H, Yue H, et al. Fault diagnosis of in-wheel motors used in electric vehicles: state of the art, challenges, and future directions. Machines. 2025;13(8):711. https://doi.org/10.3390/machines13080711

[3] Lang W. Artificial intelligence-based condition monitoring techniques for powertrains in electric vehicles [PhD thesis]. York: University of York; 2023. Available from: https://etheses.whiterose.ac.uk/id/eprint/33975/

[4] Shah R, Mittal V, Lotwin M. Recent advances in vibration analysis for predictive maintenance of modern automotive powertrains. Vibration. 2025;8(4):68. https://doi.org/10.3390/vibration8040068

[5] Pei D, Yue J, Jiao J. A novel method for bearing fault diagnosis under variable speed based on envelope spectrum fault characteristic frequency band identification. Sensors. 2023;23(9):4338. https://doi.org/10.3390/s23094338

[6] Sobie C, Freitas C, Nicolai M. Simulation-driven machine learning: bearing fault classification. Mech Syst Signal Process. 2018;99:403-419. https://doi.org/10.1016/j.ymssp.2017.06.025

[7] Matania O, Bechhoefer E, Bortman J. Digital twin of a gear root crack prognosis. Sensors. 2023;23(24):9883. https://doi.org/10.3390/s23249883

[8] Wang P, Liu Y, Liu Z. A fault diagnosis method for rotating machinery in nuclear power plants based on long short-term memory and temporal convolutional networks. Ann Nucl Energy. 2025;213:111092. https://doi.org/10.1016/j.anucene.2024.111092

[9] Chen Y, Li G, Li A, He B. End-to-end intelligent fault diagnosis of transmission bearings in electric vehicles based on CNN. Lubricants. 2024;12(11):364. https://doi.org/10.3390/lubricants12110364

[10] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. In: ICML 2015 Deep Learning Workshop; 2015. Available from: https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf

[11] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. PMLR. 2017;70:1126-1135. https://proceedings.mlr.press/v70/finn17a.html

[12] Snell J, Swersky K, Zemel RS. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. 2017;30:4077-4087. https://arxiv.org/abs/1703.05175

[13] Chen X, Tian Z, Chen Y. A novel method based on wavelet transform and prototypical network for gearbox detection in few-shot learning. Expert Syst Appl. 2025;292:128601. https://doi.org/10.1016/j.eswa.2025.128601

[14] Deng F, Huang Z, Hao R, Gu X. An inception transformer-based weighted prototype network for few-shot defect recognition of wheelset bearing. J Comput Des Eng. 2025;12(3):36-50. https://doi.org/10.1093/jcde/qwaf019

[15] Fang Q, Wu D. ANS-net: anti-noise Siamese network for bearing fault diagnosis with a few data. Nonlinear Dyn. 2021;104(3):2497-2514. https://doi.org/10.1007/s11071-021-06393-4
[16] Sun D, Yang X, Yang H. A bearing fault diagnosis method for hydrodynamic transmissions integrating few-shot learning and transfer learning. Sci Rep. 2025;15:19011. https://doi.org/10.1038/s41598-025-04543-x
[17] Wu S, Yang L, Tao L. Synergistic WSET-CNN and confidence-driven pseudo-labeling for few-shot aero-engine bearing fault diagnosis. Processes. 2025;13(7):1970. https://doi.org/10.3390/pr13071970
[18] Hu J, Li W, Zhang Y, Tian Z. Cross-domain few-shot fault diagnosis based on meta-learning and domain adversarial graph convolutional network. Eng Appl Artif Intell. 2024;136:108970. https://doi.org/10.1016/j.engappai.2024.108970
[19] Liang X, Zhang M, Feng G, Wang D, Xu Y, Gu F. Few-shot learning approaches for fault diagnosis using vibration data: a comprehensive review. Sustainability. 2023;15(20):14975. https://doi.org/10.3390/su152014975
[20] Siraj FM, Ayon STK, Uddin J. A few-shot learning based fault diagnosis model using sensors data from industrial machineries. Vibration. 2023;6(4):1004-1029. https://doi.org/10.3390/vibration6040059
[21] Lin C, Kong Y, Han Q, Wang T, Dong M, Liu H, et al. An information fusion-based meta transfer learning method for few-shot fault diagnosis under varying operating conditions. Mech Syst Signal Process. 2024;220:111652. https://doi.org/10.1016/j.ymssp.2024.111652
[22] Zhao H, Liu C, Dang X, Xu J, Deng W. Few-shot cross-domain fault diagnosis of transportation motor bearings using MAML-GA. IEEE Trans Transp Electrification. 2026;12(1):1165-1174. https://doi.org/10.1109/TTE.2025.3625779
[23] Zhang C, Wang F, Li X, Dong Z, Zhang Y. Fault diagnosis method of permanent magnet synchronous motor based on WCNN and few-shot learning. Actuators. 2024;13(9):373. https://doi.org/10.3390/act13090373
[24] Lang W, Hu Y, Zhang Z, Gan C, Si J, Wen H. Few-shot learning with residual current for EV inverter fault diagnosis of EV powertrain. IEEE Trans Transp Electrification. 2024;10(4):9316-9327. https://doi.org/10.1109/TTE.2024.3349619
[25] Chen Y, Pu X, Li G, Bai Y, Hao L. Few-shot fault diagnosis of rolling bearings using generative adversarial networks and convolutional block attention mechanisms. Lubricants. 2025;13(12):515. https://doi.org/10.3390/lubricants13120515
[26] Cheng L, An Z, Guo Y, Ren M, Yang Z, McLoone S. MMFSL: a novel multi-modal few-shot learning framework for fault diagnosis of industrial bearings. IEEE Trans Instrum Meas. 2023;72:1-13. https://doi.org/10.1109/TIM.2023.3289549
[27] Wang Y, Chen L, Liu Y, Gao L. Wavelet-prototypical network based on fusion of time and frequency domain for fault diagnosis. Sensors. 2021;21(4):1483. https://doi.org/10.3390/s21041483
[28] Smith WA, Randall RB. Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study. Mech Syst Signal Process. 2015;64-65:100-131. https://doi.org/10.1016/j.ymssp.2015.04.021