Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.084403

ARTICLE

Few-Shot Bearing Fault Diagnosis under Joint Fault-Severity and Load Shift: A Leak-Free Cross-Domain Benchmark

Safa Alsafari[1] **and Ayman Yafoz**[2,*]

[1]Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

[2]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

*Corresponding Author: Ayman Yafoz. Email: ayafoz@kau.edu.sa

Received: 22 April 2026; Accepted: 22 June 2026; Published: 27 July 2026

ABSTRACT: Bearing fault diagnosis in industrial deployment must contend with two simultaneous distributional shifts: fault severity increases as damage progresses, and motors operate at loads unseen during training. We define this compound setting as the *double domain shift* and present a rigorous few-shot benchmark on the Case Western Reserve University (CWRU) and Paderborn University (PU) bearing datasets. Six architectures spanning distinct learning paradigms—a multilayer perceptron (MLP), a capsule network (CapsNet), a residual capsule network (ResCaps), a prototypical network (ProtoNet), a modified residual convolutional network (MRCN), and Deep Correlation Alignment (Deep CORAL)—are evaluated under a strict three-way split (support/validation/held-out test) that prevents the data-leakage patterns prevalent in prior CWRU protocols. Models are adapted using $K \in \{5, 10, 20\}$ labelled target samples and assessed on three complementary metrics: accuracy, macro F1-score, and Cohen's κ. A lightweight 1-D convolutional neural network (CNN) with a capsule routing head (CapsNet) leads on 15 of 18 CWRU conditions and on all PU conditions at $K \geq 10$ (with MRCN leading at PU Target-A $K = 5$), achieving macro F1 of 0.860 and $\kappa = 0.818$ at $K = 5$ on the harder CWRU target—with 18,624 parameters (roughly one-third of the MLP baseline) and without any distribution-alignment objective. The multi-metric evaluation reveals findings invisible to accuracy alone: several baselines fall below moderate agreement ($\kappa < 0.60$) at low K, and MRCN's accuracy–F1 gap of 5.3 percentage points (pp) at $K = 10$ exposes class-selective failure that accuracy conceals. On PU, a macro F1 of 0.443 at $K = 5$ on the real inner-race target quantifies the artificial-to-real fatigue transfer gap, and all models produce $\kappa < 0.06$ on the outer-race target at $K = 5$, establishing a realistic lower bound for future work. Wilcoxon signed-rank tests confirm the CapsNet advantage is statistically significant against the weaker baselines in nearly all conditions. Capsule output norms provide interpretable, per-class confidence-like activation scores without requiring post-hoc attribution methods.

KEYWORDS: Bearing fault diagnosis; capsule network; cross-domain adaptation; few-shot learning; double domain shift

1 Introduction

Rolling-element bearing faults account for approximately 40% of rotating machine failures in industrial settings [1]. Data-driven deep learning methods have demonstrated strong diagnostic accuracy on standard benchmarks, but their performance degrades substantially when the operating conditions at test time differ from those seen during training [2,3]. In practice, deployment introduces two simultaneous distributional shifts: fault severity increases as bearing damage progresses, and motors operate across a range of loads and

speeds that may not have been present during training. We refer to this compound setting as the double domain shift. The joint shift is harder to handle than either factor alone, because representations that survive a load change do not necessarily survive a simultaneous change in fault geometry.

Existing methods largely address these two shifts in isolation. Transfer learning and domain adaptation approaches target load or speed variation while holding fault severity fixed [4,5], whereas few-shot meta-learning methods address severity variation under fixed operating conditions [6,7]. To our knowledge, neither family has been systematically evaluated under simultaneous severity and load shift with a held-out, leak-free protocol. A related concern is evaluation integrity: several CWRU benchmark studies have been noted to use validation procedures that allow indirect access to held-out statistics, inflating reported figures [1]. A further concern is metric selection—the overwhelming majority of bearing fault diagnosis papers report only accuracy, which can overstate performance when per-class recall is unequal or when small support sets cause low-shot instability. Cohen's κ and macro F1 jointly expose these failure modes and are reported throughout this work alongside accuracy.

The double domain shift also connects to the emerging digital twin (DT) paradigm, which models the coupling between physical degradation and operational context to maintain diagnostic fidelity as both co-evolve [8,9]. The benchmark proposed here probes a core DT capability—virtual diagnostic generalisation under coupled physical–operational drift—using a data-only approach that requires neither a physics model nor adversarial alignment, providing a controlled reference point for future DT-integrated methods.

This paper investigates three hypotheses. First, capsule dynamic routing, by aggregating votes across temporal sub-regions, should be more robust to the amplitude and shape changes introduced by double domain shift than flat classifiers or prototype-distance heads (H1). Second, models that generalise under single-factor shift will degrade more severely when both factors change simultaneously, including meta-learning approaches designed for rapid adaptation (H2). Third, accuracy alone is insufficient to reveal class-selective failure and low-shot instability; Cohen's κ and macro F1 will expose weaknesses that accuracy conceals (H3).

The main contributions of this work are: a leak-free three-way benchmark protocol for few-shot cross-domain fault diagnosis under double domain shift on CWRU and PU (i); an empirical comparison of six architectures including Deep CORAL as an explicit domain-adaptation baseline, showing that a lightweight capsule network matches or exceeds all baselines with no alignment objective and a parameter count roughly one-third that of the MLP baseline (ii); a multi-metric analysis demonstrating that κ and macro F1 reveal model weaknesses invisible to accuracy (iii); ablation evidence that dynamic routing, not encoder capacity, drives the CapsNet advantage (iv); a negative result showing that MAML-based meta-learning provides no benefit under joint severity and load shift (v); and cross-dataset validation on PU real fatigue damage, where a macro F1 of 0.443 at $K = 5$ quantifies the artificial-to-real transfer gap (vi).

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the datasets, benchmark protocol, and model architectures. Section 4 reports experimental results. Section 5 presents ablation and interpretability analyses. Section 6 discusses the findings and limitations, and Section 7 concludes.

2 Related Work

2.1 Deep Learning for Bearing Fault Diagnosis

Convolutional neural networks (CNNs) have become a central paradigm for vibration-based fault diagnosis, as demonstrated by a succession of high-performing methods on standard benchmarks [2,3]. Real-time 1-D CNNs applied directly to raw current/vibration signals were shown to be effective for

motor fault detection by Ince et al. [10]. Wen et al. [3] showed that 1-D CNNs applied directly to raw vibration signals match or exceed approaches that rely on hand-crafted spectral features, establishing the case for end-to-end training from raw waveforms. Architectural advances originally developed for image recognition—most notably residual connections [11]—have since been adapted to fault diagnosis; residual-depth networks [2], multi-scale spatial-temporal encoders [12], and transformer-based architectures [5] have all been applied to the CWRU dataset and report strong diagnostic accuracy under matched operating conditions. However, performance degrades substantially when the operating condition at test time differs from training; cross-domain evaluation on the same dataset typically produces a noticeable drop in accuracy relative to matched-condition baselines [4].

2.2 Capsule Networks for Fault Diagnosis

Capsule networks [13] replace scalar activations with vector-valued capsules and use dynamic routing to model part-whole relationships in the input signal. A recent survey covering literature from 2018 to 2025 [14] found that capsule networks appear in 75.2% of mechanical fault diagnosis papers that employ them, with 35.6% focused specifically on bearing faults. Most existing capsule-based diagnostic methods use 2-D time-frequency representations as input [15,16] and are evaluated under matched operating conditions; Wang and Chen [12] propose a multi-scale spatial-temporal capsule architecture that operates on sequential waveform encodings. The CICN model [17] introduces causal gating into capsule routing to improve domain generalisation but does not address the few-shot regime. Our work applies standard dynamic routing to raw 1-D waveforms in a carefully controlled few-shot cross-domain benchmark, allowing clean comparison across baselines.

2.3 Few-Shot and Cross-Domain Fault Diagnosis

Few-shot learning for fault diagnosis has been explored through prototypical networks [18], model-agnostic meta-learning (MAML) [19], and their variants [6,7]. Zhang et al. [6] showed that MAML achieves competitive accuracy on CWRU with as few as 5 samples per class under cross-condition transfer. Lin et al. [7] extended MAML to handle heterogeneous signal types across domains. Joint transfer and fine-grained metric learning approaches [20] further improve cross-condition diagnosis in the few-shot regime. Vu et al. [21] combined Transformer-based encoders with Mahalanobis-distance prototypes for multi-scale few-shot diagnosis. More recently, Rezazadeh et al. [22] proposed a prototype-attention domain adaptation framework that aligns class centroids and weights target samples by prototype similarity and confidence, achieving strong cross-condition accuracy on both CWRU and PU datasets with limited labelled data and built-in interpretability via prototype-to-instance similarity maps—an approach that shares the few-shot, prototype-based motivation of our benchmark but addresses single-factor operating-condition shift rather than the joint severity-and-load shift studied here. In contrast to all of the above, our benchmark simultaneously shifts fault severity and operating load, uses raw 1-D waveforms without manual feature extraction, enforces a strict held-out test split, and reports accuracy, macro F1, and Cohen's κ jointly. Table 1 summarises the key protocol differences between our benchmark and the three most closely related prior studies.

Table 1: Comparison of evaluation protocols across the four most closely related CWRU few-shot studies.

Study	Held-Out Test Split	Both Shifts Simultaneous	Metrics Reported	PU Val.
Zhang et al. [6]	No	No	Acc only	No
Lin et al. [7]	No	No	Acc only	No

(Continued)

Table 1 (continued)

Study	Held-Out Test Split	Both Shifts Simultaneous	Metrics Reported	PU Val.
Vu et al. [21]	Yes	No	Acc, F1	No
This work	Yes	Yes	Acc, F1, κ	Yes

2.4 Digital Twins for Bearing Fault Diagnosis

Digital twin frameworks have recently been applied to bearing diagnostics to bridge the gap between physics-based modelling and data-driven classification. By constructing a virtual counterpart that tracks both fault progression and operating-regime variation, DT approaches can generate labelled training data for conditions that are rare or unsafe to induce in practice. Qiao et al. [8] showed that a DT-guided denoising stage, calibrated to severity-dependent vibration models, improves early fault detection accuracy before any classifier is applied. Ning et al. [9] proposed a DT-driven adversarial domain adaptation method that jointly handles cross-load and cross-severity transfer—the same compound shift studied in our benchmark—using explicit covariance alignment guided by the virtual model. Our work is complementary: rather than relying on a physics model or adversarial alignment objective, we establish a data-only, leak-free benchmark that quantifies how far a lightweight capsule network can generalise under the same coupled shift, providing a controlled reference point for future DT-integrated approaches.

3 Dataset, Protocol, and Architectures

3.1 CWRU Double Domain-Shift Benchmark

Drive-end vibration signals from the Case Western Reserve University (CWRU) Bearing Dataset [1] are sampled at 12 kHz and segmented into four fault classes: Normal (N), Inner Race (IR), Ball Fault (BF), and Outer Race (OR).

3.1.1 Domain Split

Table 2 defines the three non-overlapping domains. The source domain uses bearings with 0.007″ electro-discharge-machined (EDM) drilled faults recorded at motor loads of 0 and 1 HP (1797 and 1772 rpm, respectively). Target-A uses the same fault types at the next severity level (0.014″) and at higher loads of 2 and 3 HP (1750 and 1730 rpm). Target-B uses the largest fault size (0.021″) at the same 2–3 HP loads as Target-A.

Table 2: CWRU domain split. Source and target domains differ in both fault diameter and motor load, realising the double domain shift.

Domain	Fault Diam.	Load	N	Role
Source	0.007″	0–1 HP	948	Training
Target-A	0.014″	2–3 HP	944	Test domain 1
Target-B	0.021″	2–3 HP	948	Test domain 2

Both fault severity and motor load therefore shift simultaneously between source and each target—a challenging and industrially motivated cross-domain setting, since bearing damage in service naturally progresses in severity while machinery operates across a range of loads. Assuming circular EDM pit

geometry, fault cross-sectional area scales approximately as the square of fault diameter, implying that Target-A presents roughly 4× and Target-B roughly 9× the fault area of the source domain, though this scaling is an approximation since actual EDM pit geometry is not perfectly circular [1].

3.1.2 Preprocessing

Raw signals are segmented into non-overlapping windows of $N = 1024$ samples (≈85 ms) and arranged as $(N, 1, 1024)$ tensors with a channel dimension. Per-sample z-score normalisation is applied as $\hat{x} = (x - \mu_x)/(\sigma_x + \varepsilon)$, $\varepsilon = 10^{-8}$, keeping each window independent of global statistics. Classes are balanced per domain by capping to the minimum class count, yielding 237 samples per class for Source and Target-B, and 236 for Target-A.

3.1.3 Leak-Free Evaluation Protocol

A stratified 30% holdout is drawn once from each target domain using non-overlapping window indices and fixed for all experiments; support, validation, and held-out test windows are drawn from disjoint index ranges within each recording, ensuring no adjacent or temporally overlapping segments appear across the three splits. This held-out test set is never seen during training or model selection. The remaining 70% forms the adaptation pool. For each of $N_{\text{runs}} = 10$ independent runs, K support samples per class are drawn from the pool for training, and 15 validation samples per class are used for checkpoint selection; seeds are fixed globally to ensure reproducibility.

3.2 Paderborn University (PU) Cross-Dataset Benchmark

To verify that results generalise beyond CWRU, we evaluate all six models on the Paderborn University (PU) Bearing Dataset [23]. PU differs from CWRU in three important respects: it contains both artificially drilled (EDM) and real accelerated-fatigue damage; signals are sampled at 64 kHz (downsampled to 12 kHz here using polyphase resampling); and real fatigue vibration patterns are far less structured than EDM damage, making the domain shift substantially harder.

PU domain split

The source domain uses bearings with artificially drilled EDM damage (specimens K001, K002, KI01, KA01, KB23), mirroring the CWRU source domain. Target-A uses bearings with real inner-race fatigue faults (KI03, KI05) and Target-B uses real outer-race fatigue faults (KA07, KA08), both recorded under operating condition $N09\_M07\_F10$ (900 rpm, 0.7 Nm). Four classes are used: Healthy, Inner Race, Outer Race, and Combined (concurrent inner and outer race damage), with 200 balanced segments per class per domain. Note that the PU class set differs from CWRU: the Combined class replaces CWRU's Ball Fault, so PU and CWRU evaluate related but non-identical fault taxonomies; cross-dataset comparisons should be interpreted accordingly. The same stratified 30% held-out split and 10-run protocol as CWRU are applied.

3.3 Model Architectures

To evaluate performance under the proposed double domain shift, we compare six architectures that span distinct learning paradigms—from a simple fully connected baseline to a meta-learning framework designed for rapid adaptation. All six share a **1-D CNN encoder** that maps raw 1024-sample waveforms to a d-dimensional embedding; they differ only in the classification head appended to this shared front-end (Table 3, Fig. 1).

Table 3: Summary of the six evaluated architectures. All models share the same 1-D CNN encoder; they differ only in the classification head.

Model	Head	Meta?	Params
MLP	2-layer MLP	No	55,060
CapsNet	Dynamic routing	No	18,624
ResCaps	Residual + capsule	No	71,224
ProtoNet	Prototypical (cosine)	No	50,900
MRCN	2 × Res + capsule	Yes (MAML)	87,928
CORAL	MLP head + CORAL loss	No	55,060

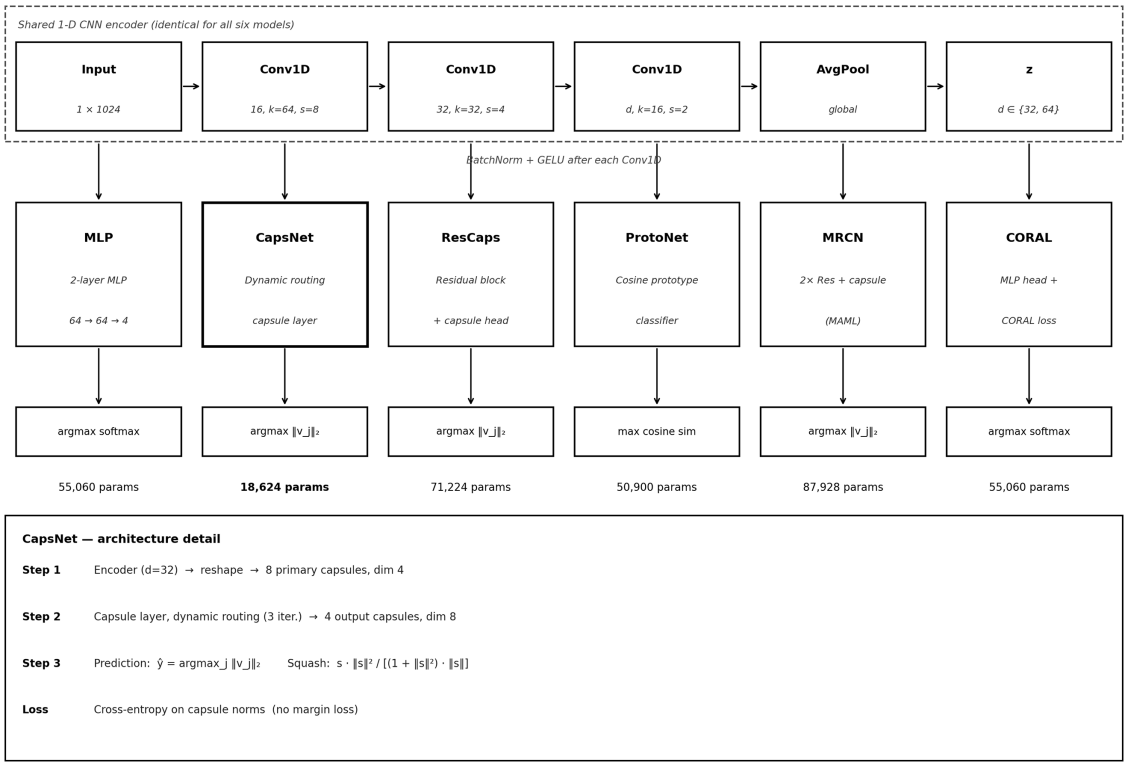

Figure 1: Shared 1-D CNN encoder pipeline for all six models. The encoder maps a raw 1024-sample vibration window to a d-dimensional embedding $\mathbf{z}$; classification heads differ per architecture. CapsNet uses dynamic routing [13] to produce one output capsule per fault class; the predicted class is the capsule with the largest ℓ_2 norm.

3.3.1 Shared 1-D CNN Encoder

The shared front-end f_θ consists of three convolutional blocks, each followed by Batch Normalisation and a Gaussian Error Linear Unit (GELU) activation, and a final global average pooling step:

$$f_\theta : \mathbb{R}^{1\times 1024} \longrightarrow \mathbb{R}^{d}, \qquad d \in \{32,\ 64\}. \tag{1}$$

The layer sequence is

$$\text{Conv1D}(1 \to 16,\ k = 64,\ s = 8) \ \to\ \text{BN} \ \to\ \text{GELU}$$
$$\to\ \text{Conv1D}(16 \to d/2,\ k = 32,\ s = 4) \ \to\ \text{BN} \ \to\ \text{GELU}$$

$$\rightarrow \text{Conv1D}(d/2 \rightarrow d,\ k = 16,\ s = 2) \rightarrow \text{BN} \rightarrow \text{GELU}$$
$$\rightarrow \text{AdaptiveAvgPool1D}(1) \rightarrow \mathbf{z} \in \mathbb{R}^d$$

where k denotes kernel size and s denotes stride. All convolutional weights are initialised with Kaiming uniform initialisation.

3.3.2 MLP Baseline

The multilayer perceptron (MLP) serves as the architectural lower bound. Its head appends a LayerNorm and two fully connected layers to the encoder output:

$$\mathbf{z} \xrightarrow{\text{LN}} \xrightarrow{\text{Linear}(64\rightarrow 64)} \xrightarrow{\text{GELU}} \xrightarrow{\text{Linear}(64\rightarrow 4)} \hat{\mathbf{y}}.$$

Using encoder depth $d = 64$, this baseline measures the performance gain that specialised feature extraction (capsule routing, residual blocks) provides over a plain nonlinear classifier under double domain shift.

3.3.3 CapsNet

Unlike standard CNNs, which use pooling layers that can discard spatial relationships, CapsNet [13] replaces scalar activations with vector-valued *capsules* and uses dynamic routing to preserve part-whole relationships in the signal. We hypothesise that this routing mechanism is more robust to the amplitude and shape changes introduced by shifting fault severity. The encoder output ($d = 32$) is reshaped into $N_{\text{prim}} = 8$ primary capsules of dimension 4:

$$\mathbf{u}_i \in \mathbb{R}^4, \quad i = 1, \ldots, 8. \tag{2}$$

A dynamic-routing capsule layer [13] with 3 routing iterations maps these to $N_{\text{cls}} = 4$ output capsules of dimension 8:

$$\mathbf{v}_j \in \mathbb{R}^8, \quad j = 1, \ldots, 4, \tag{3}$$

where the squash non-linearity is applied after each routing step:

$$\text{squash}(\mathbf{s}) = \frac{\|\mathbf{s}\|^2}{1 + \|\mathbf{s}\|^2} \cdot \frac{\mathbf{s}}{\|\mathbf{s}\|}. \tag{4}$$

The predicted fault class is the capsule with the largest ℓ_2 norm, since a larger norm indicates higher routing confidence:

$$\hat{y} = \arg\max_j \|\mathbf{v}_j\|_2. \tag{5}$$

Training minimises cross-entropy applied to the vector of capsule norms.

3.3.4 ResCaps

ResCaps integrates residual skip connections with capsule routing. The residual backbone facilitates deeper feature extraction while mitigating vanishing gradients, providing richer representations of high-frequency vibration patterns before they are passed to the capsule routing module.

Starting from the encoder ($d = 64$), a residual block is applied:

$$\mathbf{h} = \mathbf{z} + \mathbf{W}_2\,\mathrm{GELU}(\mathbf{W}_1\,\mathrm{LN}(\mathbf{z}))\,, \tag{6}$$

where $\mathbf{W}_1 \in \mathbb{R}^{128\times64}$, $\mathbf{W}_2 \in \mathbb{R}^{64\times128}$, and $\mathbf{W}_2$ is zero-initialised so the block begins as an identity map, stabilising few-shot training. The output $\mathbf{h}$ is then projected via LN → Linear(64→40) → GELU and reshaped into 8 primary capsules of dimension 5, which are processed by the same dynamic routing head as CapsNet.

3.3.5 ProtoNet

ProtoNet [18] treats fault diagnosis as a metric-learning problem and is specifically suited to the low-shot instability of the $K \in \{5, 10, 20\}$ regime. The encoder ($d = 64$) is trained with cross-entropy on the support set. At inference, a class prototype is computed as the mean of the ℓ_2-normalised support embeddings:

$$\mathbf{c}_k = \frac{1}{|S_k|}\sum_{(\mathbf{x},\,y)\in S_k}\phi(\mathbf{x}), \quad \phi(\mathbf{x}) = \frac{f_\theta(\mathbf{x})}{\|f_\theta(\mathbf{x})\|_2}, \tag{7}$$

where S_k is the support set for class k. Each query sample is assigned to the class whose prototype has the highest cosine similarity:

$$\hat{y} = \arg\max_k \frac{\phi(\mathbf{x}_q)^\top \mathbf{c}_k}{\|\phi(\mathbf{x}_q)\|\,\|\mathbf{c}_k\|}. \tag{8}$$

3.3.6 MRCN (MAML Baseline)

MRCN pairs a Modified Residual Convolutional Network backbone with first-order MAML [19,24]. The strength of MAML lies in its bi-level optimisation: the *inner loop* adapts to a specific fault condition using K labeled samples, while the *outer loop* updates the model initialisation so that adaptation requires as few gradient steps as possible.

The backbone appends two ResidualBlocks (6) to the shared encoder ($d = 64$), followed by the same LN → Linear(64→40) → GELU projection and capsule routing head as ResCaps. Meta-training uses 500 episodes on the source domain with a stochastic gradient descent (SGD) inner loop ($\alpha_{\text{inner}} = 0.05$, 5 steps) and an Adam outer loop. At test time the encoder is frozen and only the head parameters are adapted on the K support samples (40–50 Adam steps, $\alpha = 5 \times 10^{-3}$, cosine-annealed):

$$\theta^*_{\text{head}} = \theta_{\text{head}} - \alpha\,\nabla_{\theta_{\text{head}}}\mathcal{L}(f_\theta(X_s),\,y_s)\,. \tag{9}$$

3.3.7 Deep CORAL (Explicit Domain-Adaptation Baseline)

To position the benchmark against methods that *explicitly* target distribution shift, we include Deep CORAL (Correlation Alignment) [25]. CORAL minimises the Frobenius distance between the second-order statistics of the source and target feature distributions, jointly with the supervised loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\text{source}) + \lambda\,\mathcal{L}_{\text{CORAL}}, \quad \mathcal{L}_{\text{CORAL}} = \frac{1}{4d^2}\left\|C_S - C_T\right\|_F^2, \tag{10}$$

where C_S and C_T are the $d \times d$ covariance matrices of the source and target embeddings and $\lambda = 1$. For a fair comparison, CORAL reuses the *identical* 1-D CNN encoder and the same 2-layer MLP head as the MLP baseline (hence the same 55,060 parameters); the *only* difference is the added alignment term. The target embeddings entering $\mathcal{L}_{\text{CORAL}}$ are drawn from the unlabelled portion of the adaptation pool, with the support

and validation indices removed, so the held-out test set is never accessed during training. CORAL therefore isolates the contribution of an explicit alignment objective under our leak-free protocol, with all other factors held constant.

3.4 Benchmark Procedure

Algorithm 1 summarises the complete leak-free evaluation procedure applied identically to both CWRU and PU.

Algorithm 1: Leak-free few-shot benchmark procedure

Input: Domain data $(X_{\text{src}}, y_{\text{src}})$, $(X_{\text{tgt}}, y_{\text{tgt}})$; shot counts $\mathcal{K} = \{5, 10, 20\}$; runs $N_{\text{runs}} = 10$
Output: Accuracy, macro F1, Cohen's κ for each model $\times$ K pair
// **Stage 1: Fixed held-out split (done once, never changed)**
Draw stratified 30% holdout $\mathcal{H}$ from X_{tgt} using disjoint index ranges
Remaining 70% forms adaptation pool $\mathcal{P}$
// **Stage 2: Source training**
Train each model on $(X_{\text{src}}, y_{\text{src}})$ for 150 epochs with Adam
// **Stage 3: Few-shot adaptation and evaluation (repeated)**
for each $K \in \mathcal{K}$ **do**
  for $r = 1$ **to** N_{runs} **do**
    Sample K support examples per class from $\mathcal{P} \rightarrow \mathcal{S}$
    Sample 15 validation examples per class from $\mathcal{P} \setminus \mathcal{S} \rightarrow \mathcal{V}$
    Fine-tune model head on $\mathcal{S}$; select checkpoint via $\mathcal{V}$
    Evaluate on $\mathcal{H}$; record accuracy, F1, κ
  end for
end for
Aggregate over N_{runs}; report mean $\pm$ std and Wilcoxon tests

3.5 Training Details

MLP, CapsNet, ResCaps, and ProtoNet use Adam [26] (lr = 3×10^{-3}, weight decay 10^{-4}, cosine annealing to lr/50, 150 epochs, batch 32, gradient clip 1.0) with best-validation-accuracy checkpointing. MRCN uses SGD for the inner loop and Adam (lr = 5×10^{-4}) for the outer loop. All experiments use $N_{\text{runs}} = 10$ independent support draws; global seed fixed to 42. All experiments were implemented in Python 3.12 with PyTorch 2.1.0 (CUDA 12.8) and executed on an NVIDIA Tesla T4 GPU (15.6 GB VRAM) via Google Colab; key dependencies: NumPy 2.0.2, scikit-learn 1.6.1, SciPy 1.16.3, and Pandas 2.2.2.

4 Experimental Results

4.1 CWRU Results

Tables 4–6 report mean accuracy, macro F1, and Cohen's κ across 10 independent runs on the held-out test set. CapsNet leads on 15 of 18 conditions (three metrics $\times$ two targets $\times$ three K values); in the remaining three conditions (accuracy and F1 at Target-B $K = 20$, and κ at Target-A $K = 10$) CapsNet and the next-best model tie within one standard deviation, while CapsNet retains the highest mean on Cohen's κ across all six K/target combinations. For most models and conditions the accuracy and F1 results are closely aligned, as expected for a balanced four-class problem; any divergence reflects unequal per-class performance rather than distributional artefacts.

Table 4: CWRU results: mean ± std **accuracy** over 10 independent runs on the held-out test set. Bold indicates best per column.

Model	Target-A (0.014″, 2–3 HP)			Target-B (0.021″, 2–3 HP)		
	$K = 5$	$K = 10$	$K = 20$	$K = 5$	$K = 10$	$K = 20$
MLP	0.789 ± 0.052	0.922 ± 0.042	0.938 ± 0.040	0.881 ± 0.040	0.977 ± 0.013	0.981 ± 0.010
CapsNet	**0.864** ± 0.057	**0.948** ± 0.041	**0.985** ± 0.011	**0.970** ± 0.025	**0.985** ± 0.012	**0.986** ± 0.014
ResCaps	0.593 ± 0.070	0.875 ± 0.040	0.916 ± 0.032	0.780 ± 0.041	0.964 ± 0.013	0.982 ± 0.014
ProtoNet	0.798 ± 0.057	0.933 ± 0.040	0.941 ± 0.017	0.892 ± 0.051	0.977 ± 0.020	**0.986** ± 0.008
MRCN	0.717 ± 0.056	0.716 ± 0.071	0.775 ± 0.038	0.906 ± 0.034	0.940 ± 0.064	0.954 ± 0.020
CORAL	0.845 ± 0.048	0.944 ± 0.029	0.964 ± 0.017	0.917 ± 0.054	0.978 ± 0.012	0.984 ± 0.012

Note: Source: 0.007″, 0–1 HP. CORAL = Deep CORAL domain-adaptation baseline.

Table 5: CWRU results: mean ± std **macro F1-score** over 10 independent runs on the held-out test set. Bold indicates best per column.

Model	Target-A (0.014″, 2–3 HP)			Target-B (0.021″, 2–3 HP)		
	$K = 5$	$K = 10$	$K = 20$	$K = 5$	$K = 10$	$K = 20$
MLP	0.781 ± 0.054	0.921 ± 0.044	0.936 ± 0.042	0.878 ± 0.041	0.977 ± 0.013	0.981 ± 0.009
CapsNet	**0.860** ± 0.058	**0.948** ± 0.041	**0.985** ± 0.011	**0.970** ± 0.024	**0.985** ± 0.012	**0.986** ± 0.014
ResCaps	0.570 ± 0.065	0.871 ± 0.043	0.913 ± 0.033	0.773 ± 0.043	0.964 ± 0.013	0.982 ± 0.014
ProtoNet	0.792 ± 0.058	0.930 ± 0.044	0.940 ± 0.018	0.889 ± 0.054	0.977 ± 0.021	**0.986** ± 0.008
MRCN	0.676 ± 0.085	0.663 ± 0.094	0.775 ± 0.039	0.904 ± 0.036	0.933 ± 0.082	0.953 ± 0.020
CORAL	0.842 ± 0.048	0.944 ± 0.029	0.963 ± 0.017	0.915 ± 0.056	0.977 ± 0.012	0.984 ± 0.012

Note: Source: 0.007″, 0–1 HP. CORAL = Deep CORAL domain-adaptation baseline.

Table 6: CWRU results: mean ± std **Cohen's κ** over 10 independent runs on the held-out test set. Bold indicates best per column.

Model	Target-A (0.014″, 2–3 HP)			Target-B (0.021″, 2–3 HP)		
	$K = 5$	$K = 10$	$K = 20$	$K = 5$	$K = 10$	$K = 20$
MLP	0.719 ± 0.069	0.896 ± 0.056	0.917 ± 0.053	0.842 ± 0.053	0.969 ± 0.018	0.975 ± 0.013
CapsNet	**0.818** ± 0.076	**0.930** ± 0.055	**0.980** ± 0.014	**0.960** ± 0.033	**0.980** ± 0.016	**0.982** ± 0.019
ResCaps	0.457 ± 0.093	0.833 ± 0.054	0.888 ± 0.042	0.707 ± 0.054	0.953 ± 0.017	0.976 ± 0.019
ProtoNet	0.731 ± 0.076	0.910 ± 0.054	0.922 ± 0.023	0.855 ± 0.069	0.969 ± 0.027	0.981 ± 0.010
MRCN	0.622 ± 0.075	0.622 ± 0.095	0.700 ± 0.050	0.875 ± 0.046	0.920 ± 0.085	0.938 ± 0.026
CORAL	0.793 ± 0.065	0.926 ± 0.038	0.951 ± 0.023	0.889 ± 0.072	0.970 ± 0.016	0.978 ± 0.016

Note: Source: 0.007″, 0–1 HP. CORAL = Deep CORAL domain-adaptation baseline.

The most informative contrasts emerge at low K. At $K = 5$ on Target-A, CapsNet achieves accuracy 0.864 and macro F1 0.860, compared with 0.789 and 0.781 for MLP—gains of 7.5 pp in accuracy and 7.9 pp in macro F1, obtained despite having 18,624 parameters (compared with 55,060 for MLP) and relying on no alignment objective. ResCaps, despite having 3.8× more parameters than CapsNet, achieves only 0.593 accuracy at the same condition; its $\kappa = 0.457$ places it below moderate agreement on the Landis–Koch scale and suggests that accuracy alone does not fully describe its behaviour. MRCN performs weakly on Target-A, particularly at $K = 10$ and $K = 20$ (peak accuracy 0.775), and its accuracy–F1 gap of 5.3 pp at $K = 10$ reveals that it concentrates correct predictions on a subset of fault types while failing on others. Deep CORAL, the

explicit alignment baseline, is the second-strongest method overall: it consistently outperforms MLP and matches CapsNet closely at $K \in \{5, 10\}$, though CapsNet retains the higher mean in every condition.

4.2 Statistical Significance

Wilcoxon signed-rank tests (Table 7) confirm that the CapsNet advantage is statistically significant ($p < 0.05$) against ResCaps and MRCN at every K on Target-A, and against MLP and ProtoNet at $K \in \{5, 20\}$. At $K = 10$, MLP ($p = 0.10$) is not significantly worse, as the three strongest models fall within one standard deviation of each other. Against Deep CORAL the gap is significant only at $K = 20$ ($p = 0.035$) on Target-A; on Target-B, the gap is significant at $K \in \{5, 10\}$ ($p = 0.027$, 0.031; separate pairwise Wilcoxon tests, results not tabulated). The Friedman test confirms overall model differences at all K on both targets (df = 5; Target-A: $\chi^2 = 32.29$ at $K = 5$, $\chi^2 = 34.71$ at $K = 10$, $\chi^2 = 38.07$ at $K = 20$; Target-B: $\chi^2 = 34.34$ at $K = 5$, $\chi^2 = 15.19$ at $K = 10$, $\chi^2 = 17.44$ at $K = 20$; all $p < 0.01$). The pairwise Wilcoxon tests in Table 7 are reported without a multiple-comparison correction; given the 15 comparisons presented (5 pairs × 3 K values), borderline p-values (e.g., $p = 0.035$, $p = 0.037$) should be interpreted with caution.

Table 7: Wilcoxon signed-rank test: CapsNet vs. each baseline on Target-A accuracy ($\alpha = 0.05$, two-sided, 10 paired raw per-run scores).

Comparison	$K = 5$	$K = 10$	$K = 20$
CapsNet vs. MLP	$p = 0.0098$✓	$p = 0.0996$ n.s.	$p = 0.0039$✓
CapsNet vs. ResCaps	$p = 0.0020$✓	$p = 0.0039$✓	$p = 0.0020$✓
CapsNet vs. ProtoNet	$p = 0.0371$✓	$p = 0.1523$ n.s.	$p = 0.0020$✓
CapsNet vs. MRCN	$p = 0.0039$✓	$p = 0.0020$✓	$p = 0.0020$✓
CapsNet vs. CORAL	$p = 0.3047$ n.s.	$p = 0.4199$ n.s.	$p = 0.0352$✓

Note: ✓ $p < 0.05$; n.s. not significant.

4.3 Cross-Dataset Validation: PU Benchmark

The PU benchmark probes a different axis of transfer from CWRU: models are trained on artificially drilled (EDM) damage and tested on real accelerated-fatigue faults at a fixed operating condition, rather than the simultaneous severity-and-load shift of the CWRU setting. The two datasets therefore measure complementary aspects of generalisation, and the PU results should be read as an artificial-to-real transfer test.

Tables 8–10 report the PU results. CapsNet leads on all three metrics in both PU domains at $K \geq 10$. At $K = 5$ on Target-A, MRCN records the highest mean accuracy (0.493), macro F1 (0.454), and κ (0.324) across all six models, suggesting that source-domain meta-training provides a useful initialisation at the lowest data budget; however, this advantage disappears by $K = 10$, where CapsNet surpasses MRCN on all three metrics and continues to improve monotonically through $K = 20$. Deep CORAL is the second-strongest method at $K \geq 10$ on Target-A but never exceeds CapsNet's mean in any condition. The substantially lower PU numbers relative to CWRU—CapsNet F1 0.443 vs. 0.860 at $K = 5$ on their respective Target-A—reflect the qualitatively harder nature of real fatigue damage compared with controlled EDM pits. On Target-B, all models produce $\kappa < 0.06$ at $K = 5$, confirming that five labelled samples per class are insufficient for the real outer-race fatigue transfer problem regardless of architecture. Table 11 places these results in context by comparing CapsNet and MLP directly across all four domains at $K = 5$; CapsNet leads on every domain and metric, with F1 gains ranging from +2.0 pp (PU Target-B) to +9.2 pp (CWRU Target-B).

Table 8: PU benchmark results: mean ± std **accuracy** over 10 independent runs. Source domain uses artificial EDM damage; target domains use real accelerated-fatigue faults. Bold indicates best per column.

Model	PU Target-A (Real Inner-Race Fatigue)			PU Target-B (Real Outer-Race Fatigue)		
	$K = 5$	$K = 10$	$K = 20$	$K = 5$	$K = 10$	$K = 20$
MLP	0.401 ± 0.073	0.575 ± 0.041	0.577 ± 0.038	0.265 ± 0.021	0.331 ± 0.033	0.395 ± 0.033
CapsNet	0.465 ± 0.032	**0.601** ± 0.032	**0.632** ± 0.043	**0.290** ± 0.037	**0.362** ± 0.044	**0.470** ± 0.045
ResCaps	0.389 ± 0.039	0.517 ± 0.032	0.561 ± 0.042	0.262 ± 0.024	0.311 ± 0.040	0.326 ± 0.033
ProtoNet	0.421 ± 0.064	0.568 ± 0.038	0.584 ± 0.029	0.272 ± 0.029	0.325 ± 0.039	0.405 ± 0.045
MRCN	**0.493** ± 0.046	0.495 ± 0.037	0.515 ± 0.019	0.284 ± 0.022	0.268 ± 0.025	0.283 ± 0.037
CORAL	0.446 ± 0.044	0.565 ± 0.030	0.590 ± 0.039	0.266 ± 0.033	0.343 ± 0.037	0.429 ± 0.037

Note: Random baseline = 0.250 (four balanced classes). MRCN's lead at PU Target-A $K = 5$ collapses at $K \geq 10$; CapsNet leads all six conditions at $K \geq 10$ on both targets.

Table 9: PU benchmark results: mean ± std **macro F1-score** over 10 independent runs. Bold indicates best per column.

Model	PU Target-A (Real Inner-Race Fatigue)			PU Target-B (Real Outer-Race Fatigue)		
	$K = 5$	$K = 10$	$K = 20$	$K = 5$	$K = 10$	$K = 20$
MLP	0.377 ± 0.086	0.553 ± 0.045	0.554 ± 0.034	0.238 ± 0.025	0.320 ± 0.039	0.385 ± 0.041
CapsNet	0.443 ± 0.040	**0.581** ± 0.032	**0.622** ± 0.037	0.258 ± 0.055	**0.355** ± 0.047	**0.465** ± 0.044
ResCaps	0.368 ± 0.043	0.504 ± 0.032	0.546 ± 0.048	0.223 ± 0.050	0.303 ± 0.037	0.322 ± 0.032
ProtoNet	0.409 ± 0.063	0.553 ± 0.040	0.569 ± 0.023	**0.262** ± 0.027	0.317 ± 0.038	0.398 ± 0.045
MRCN	**0.454** ± 0.050	0.477 ± 0.045	0.496 ± 0.022	0.236 ± 0.053	0.253 ± 0.033	0.272 ± 0.036
CORAL	0.434 ± 0.042	0.555 ± 0.029	0.582 ± 0.039	0.253 ± 0.040	0.330 ± 0.041	0.421 ± 0.037

Note: CapsNet leads all six conditions at $K \geq 10$ on both PU targets.

Table 10: PU benchmark results: mean ± std **Cohen's κ** over 10 independent runs. Bold indicates best per column.

Model	PU Target-A (Real Inner-Race Fatigue)			PU Target-B (Real Outer-Race Fatigue)		
	$K = 5$	$K = 10$	$K = 20$	$K = 5$	$K = 10$	$K = 20$
MLP	0.202 ± 0.097	0.433 ± 0.055	0.436 ± 0.051	0.021 ± 0.028	0.108 ± 0.044	0.194 ± 0.044
CapsNet	0.287 ± 0.043	**0.468** ± 0.042	**0.509** ± 0.057	**0.053** ± 0.050	**0.150** ± 0.059	**0.293** ± 0.060
ResCaps	0.186 ± 0.051	0.356 ± 0.043	0.414 ± 0.057	0.016 ± 0.032	0.081 ± 0.053	0.102 ± 0.044
ProtoNet	0.228 ± 0.085	0.424 ± 0.051	0.445 ± 0.038	0.029 ± 0.039	0.099 ± 0.052	0.207 ± 0.059
MRCN	**0.324** ± 0.061	0.327 ± 0.050	0.354 ± 0.026	0.045 ± 0.029	0.023 ± 0.033	0.044 ± 0.049
CORAL	0.261 ± 0.059	0.420 ± 0.040	0.454 ± 0.052	0.021 ± 0.044	0.123 ± 0.050	0.238 ± 0.049

Note: At PU Target-B $K = 5$ all six models have $\kappa < 0.06$, near chance agreement; this benchmark is intractable at $K = 5$ for every method tested.

Table 11: CapsNet vs. MLP at $K = 5$: mean accuracy and macro F1 across all four evaluation domains (10 runs).

Domain	Accuracy		Macro F1	
	CapsNet	MLP	CapsNet	MLP
CWRU Tgt-A (0.014″)	0.864	0.789	0.860	0.781
CWRU Tgt-B (0.021″)	0.970	0.881	0.970	0.878
PU Tgt-A (real IR)	0.465	0.401	0.443	0.377
PU Tgt-B (real OR)	0.290	0.265	0.258	0.238

5 Analysis

5.1 Sample Efficiency and Ablation

Table 12 reports the minimum K at which each model first exceeds three accuracy thresholds on CWRU. CapsNet and CORAL are the only two models to reach ≥95% accuracy on Target-A within $K \leq 20$—both require $K = 20$—and CapsNet additionally exceeds ≥90% on Target-B with just five labelled samples per class, a budget at which only CapsNet, MRCN, and CORAL clear that threshold on Target-B. MRCN never exceeds ≥80% on Target-A at any tested K, which aligns with its poor absolute performance in Table 4 and confirms that meta-training on the source domain does not improve sample efficiency under joint severity and load shift.

Table 12: Minimum K at which each model first exceeds three accuracy thresholds (mean over 10 runs).

	Target-A			Target-B		
	≥0.80	**≥0.90**	**≥0.95**	**≥0.80**	**≥0.90**	**≥0.95**
MLP	10	10	>20	5	10	10
CapsNet	**5**	**10**	**20**	**5**	**5**	**5**
ResCaps	10	20	>20	10	10	10
ProtoNet	10	10	>20	5	10	10
MRCN	>20	>20	>20	5	5	20
CORAL	5	10	20	5	5	10

Note: Bold indicates the lowest (best) K per column.

Table 13 isolates the contribution of the classification head by holding the encoder fixed and varying only the head. CapsNet (B) outperforms MLP (A) by 2.0 pp in accuracy, 2.0 pp in macro F1, and 2.6 pp in Cohen's κ, suggesting that dynamic routing contributes beyond a nonlinear classifier on the same encoder. To distinguish the routing mechanism from parameter efficiency, Variant E constrains a multilayer perceptron head to 18,624 parameters—matching CapsNet's count exactly. MLP-18K (E) achieves 0.889 accuracy and $\kappa = 0.852$, trailing CapsNet (B) by 5.6 and 7.5 pp, respectively, establishing that the gap is not an artefact of model size and that routing appears to account for a substantial part of the improvement. ResCaps (C), despite having 3.8× more parameters than CapsNet, is the weakest variant by 9.7 pp in accuracy, confirming that adding residual depth before the capsule head induces support-set memorisation rather than improved generalisation at low K. ProtoNet (D) trails CapsNet by 3.0 pp, indicating that cosine prototype distance does not fully capture what iterative routing achieves. All three metrics rank the variants in the same order, ruling out class-imbalance artefacts as a confound. Fig. 2 visualises the accuracy comparison.

Table 13: Ablation study on Target-A (0.014″), $K = 10$, 10 runs. The shared encoder is fixed; only the classification head varies. Variant E isolates routing from parameter efficiency by matching CapsNet's parameter count.

Variant	Head	Acc	F1	κ
A: MLP	2-layer MLP (55K params)	0.926 ± 0.021	0.924 ± 0.022	0.901 ± 0.028
B: CapsNet	Dynamic routing (18K params)	**0.945** ± 0.029	**0.944** ± 0.031	**0.927** ± 0.039
C: ResCaps	Residual + capsule (71K params)	0.848 ± 0.062	0.842 ± 0.067	0.797 ± 0.082
D: ProtoNet	Prototypical cosine (51K params)	0.916 ± 0.050	0.911 ± 0.056	0.888 ± 0.067
E: MLP-18K	1-layer MLP ≈ 18K params	0.889 ± 0.039	0.885 ± 0.043	0.852 ± 0.052

Note: Bold indicates best per column.

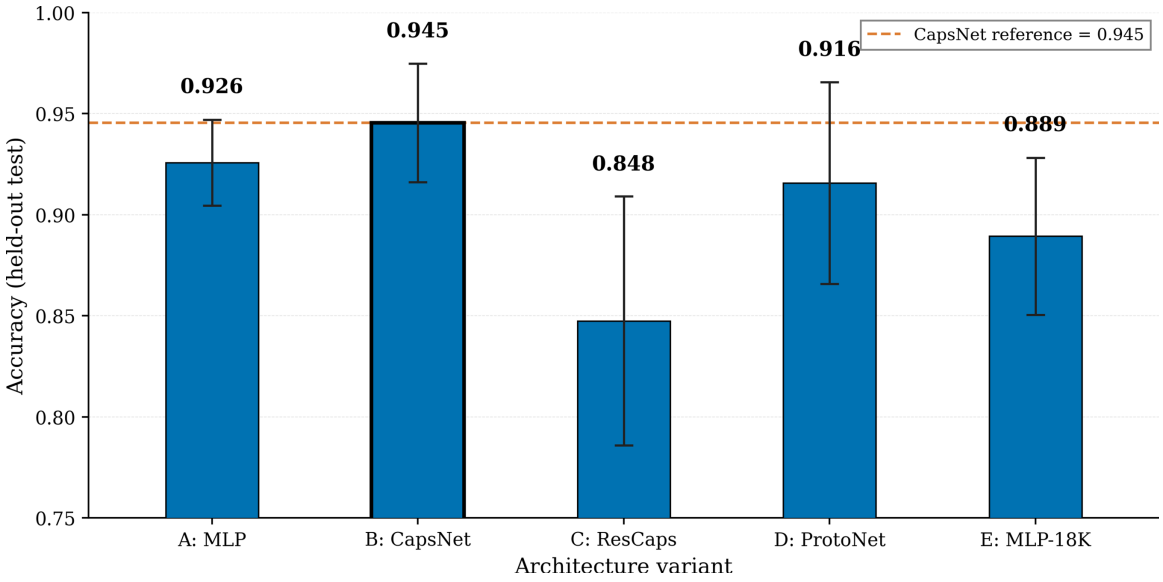

Figure 2: Ablation results (accuracy, Target-A, $K = 10$, 10 runs). CapsNet (B) leads; ResCaps (C) underperforms MLP (A) despite 3.8× more parameters. Dashed line = CapsNet reference at 0.945.

5.2 Routing Behaviour and Interpretability

Fig. 3 shows that the accuracy gain from CapsNet is not evenly distributed across fault classes. MLP misclassifies 10 Inner Race samples as Outer Race and 4 as Ball Fault; CapsNet reduces Inner Race → Outer Race errors to 4 and eliminates Inner Race → Ball Fault errors entirely. The Inner Race–Outer Race confusion is physically motivated: both fault types produce impulsive vibrations and differ mainly in repetition frequency, which is a subtle distinction under limited support. The pattern confirms that CapsNet's improvement is concentrated on the class pairs most susceptible to low-shot confounding.

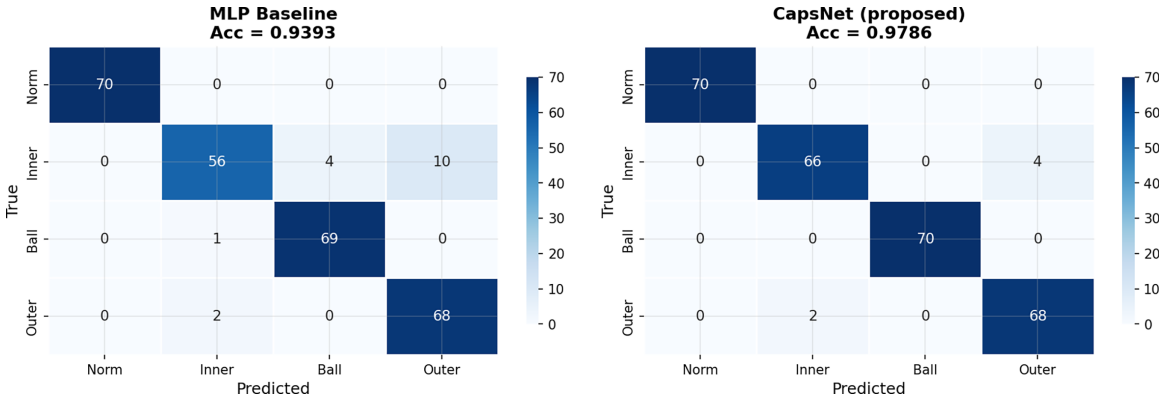

Figure 3: Confusion matrices: MLP (acc =0.939, **left**) vs. CapsNet (acc =0.979, **right**), Target-A, $K = 10$, seed = 42.

Fig. 4 shows mean capsule output norms per true class on the held-out test set. The diagonal values—Normal 0.977, Inner Race 0.804, Ball Fault 0.861, Outer Race 0.814—exhibit strong dominance, indicating that routing concentrates activation in the correct capsule for the vast majority of test samples. The largest off-diagonal entry (0.206, Outer Race activating the Inner Race capsule) is consistent with the physical proximity of these two fault types and matches the confusion pattern in Fig. 3. The correct-class norm distributions (right panel) concentrate near 1.0 for all four classes, consistent with routing producing class-specific

confidence-like activation scores. These patterns constitute behavioural evidence of class-discriminative routing; formal mechanistic attribution is left for future work.

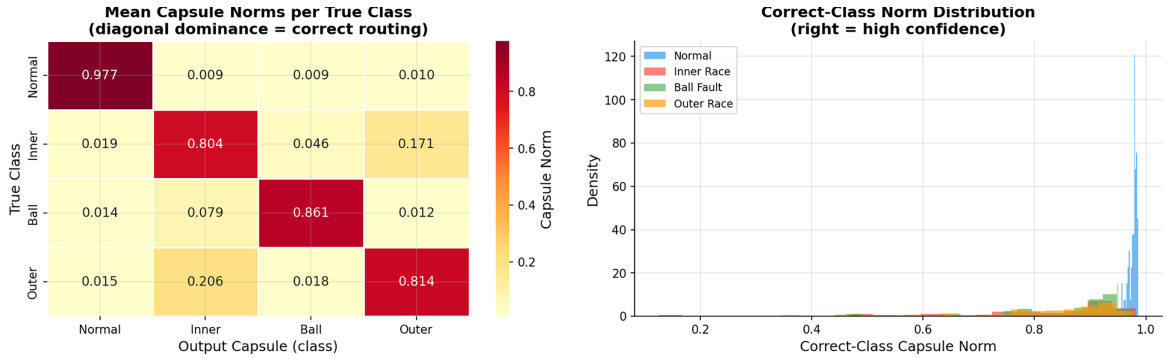

Figure 4: Left: mean capsule norm per true class. Strong diagonal dominance is consistent with class-discriminative routing without post-hoc attribution. Right: correct-class norm distributions concentrate near 1.0, indicating high routing confidence. Target-A, $K = 10$.

The t-distributed stochastic neighbor embedding (t-SNE) projections in Fig. 5 provide complementary evidence about the encoder's role. The left panel shows that fault class clusters are not well separated in encoder space. This is consistent with the shared encoder being trained to produce a general embedding under cross-entropy loss, while the capsule routing head is responsible for final class discrimination; however, we note that this pattern is an empirical observation rather than an architectural prediction, as the encoder is also trained with a discriminative objective. The right panel shows substantial overlap between source and Target-A embeddings in the projected space. This is consistent with the observation that convolutional fault encoders can exhibit partial domain invariance through shared temporal features [4], which may contribute to the model's ability to adapt from as few as $K = 5$ support samples. We treat this as a qualitative observation; t-SNE projections do not preserve global distances and cannot confirm domain alignment in the original feature space.

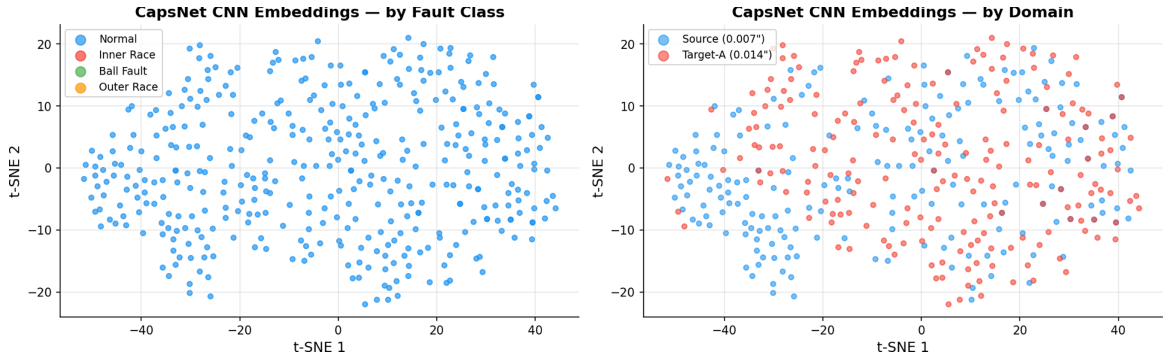

Figure 5: t-SNE of CapsNet encoder embeddings (200 samples per domain). (**Left**): by fault class—limited cluster separation suggests that class discrimination is not fully captured by the encoder alone. (**Right**): by domain—source (blue) and Target-A (red) overlap substantially, suggesting implicit domain consistency without explicit alignment.

Finally, Fig. 6 examines which waveform regions CapsNet relies on by measuring loss sensitivity to zero-masking each of 16 equal segments (≈5.3 ms each). Two peaks at 27–32 ms and 43–48 ms are separated by ≈16 ms, corresponding to approximately 1–2 impulse periods at the ball-pass frequency outer-race (BPFO ≈ 107 Hz, $T \approx 9.3$ ms). CapsNet appears more sensitive to waveform regions consistent with recurring

fault impulses rather than a single transient, which is consistent with the routing mechanism aggregating agreement votes from primary capsules at different temporal positions.

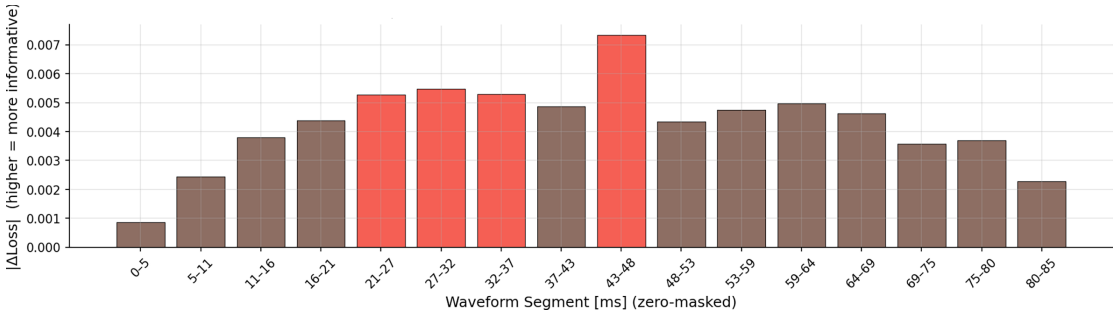

Figure 6: Loss sensitivity to zero-masking 16 waveform segments. Highlighted bars (top 25% sensitivity) peak at 27–32 ms and 43–48 ms, consistent with 1–2 bearing fault impulse repetition periods at the target operating speed (BPFO ≈ 107 Hz).

To examine whether CapsNet's sensitive waveform regions correspond to physically meaningful signal content, Fig. 7 compares envelope spectra of the high-sensitivity segment (27–32 ms, top-25% sensitivity from Fig. 6) and the low-sensitivity segment (0–5 ms, bottom-25% sensitivity) for a representative Inner Race fault sample from Target-A. Each spectrum is computed by bandpass filtering (1–5 kHz), applying the Hilbert transform to extract the envelope, and taking the fast Fourier transform (FFT). Because each segment is only 64 samples (≈5.3 ms), the frequency resolution is coarse (≈188 Hz/bin), so individual harmonics cannot be resolved precisely. The dashed reference line marks the ball-pass frequency inner-race (BPFI = 157.9 Hz), which is the expected diagnostic frequency for Inner Race faults at 1750 rpm. The high-sensitivity segment shows a clear spectral elevation near BPFI that is absent from the low-sensitivity segment, whose spectrum rises monotonically with no comparable structure. This contrast suggests that the waveform region CapsNet attends to carries fault-related energy, though the coarse resolution prevents a definitive harmonic match. Taken together, the four analyses in this section—confusion matrices, capsule norms, t-SNE embeddings, and this sensitivity–envelope comparison—consistently point to the same interpretation: CapsNet selectively attends to waveform regions that carry periodic fault-impulse structure, though we treat this as supporting evidence rather than formal proof.

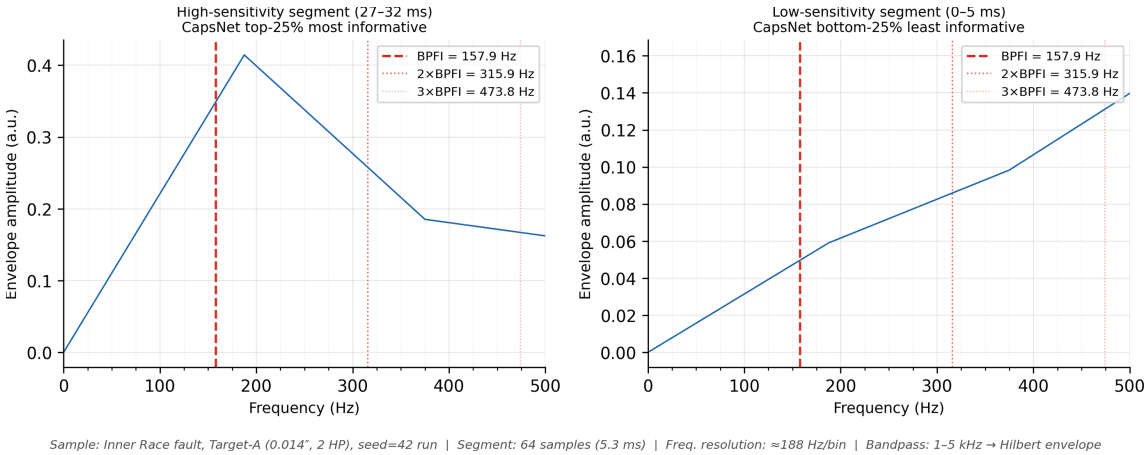

Figure 7: Envelope spectra of the high-sensitivity segment (27–32 ms, **left**) and low-sensitivity segment (0–5 ms, **right**) for a representative Inner Race fault sample, Target-A, K = 10, seed = 42. Spectra are computed via bandpass filtering

(Continued)

Figure 7: (continued) (1–5 kHz) → Hilbert envelope → FFT on 64-sample (≈5.3 ms) segments (frequency resolution ≈188 Hz/bin). Dashed line: BPFI = 157.9 Hz (expected Inner Race diagnostic frequency at 1750 rpm); dotted lines: harmonics. A spectral elevation near BPFI is present in the high-sensitivity segment and absent from the low-sensitivity segment, consistent with fault-related energy in the waveform region CapsNet attends to.

6 Discussion

6.1 CapsNet Performance under Double Domain Shift

The strong performance of CapsNet across all three metrics can be attributed to three complementary factors. The temporal sensitivity analysis shows that CapsNet attends to two impulse-period-separated waveform regions rather than a single transient, which is consistent with the routing mechanism aggregating agreement votes from primary capsules at different temporal positions. MLP heads collapse the full embedding into a single logit vector, and ProtoNet uses cosine prototype distance that does not weight temporal position; neither mechanism inherently favours periodically structured fault signals. Beyond temporal selectivity, capsule routing concentrates activation mass into the correct output capsule, so the output norm serves as a class confidence-like score that does not require a separate linear probe or softmax head—though, as noted in Section 6.4, formal probability calibration is not claimed and remains future work. At $K = 5$, where the MLP softmax is poorly conditioned, this difference is reflected in the κ gap (0.818 vs. 0.719 on Target-A), indicating meaningfully stronger agreement with ground truth. Finally, CapsNet's compact size (18,624 parameters—one-third as many as MLP, 4.7× fewer than MRCN) limits support-set memorisation. The ablation confirms this directly: adding residual depth before the capsule head (ResCaps) consistently degrades all three metrics at $K \leq 10$ despite a 3.8× increase in parameter count, indicating overfitting rather than richer representation.

6.2 Failure of Meta-Learning under Joint Shift

MRCN, the only meta-learning baseline, performs weakly on Target-A, particularly at $K = 10$ and $K = 20$, despite source-domain meta-training. It is worth noting that Target-B results are generally higher than Target-A across most models, which is consistent with the larger 0.021″ fault producing more pronounced vibration signatures that are easier to discriminate despite the simultaneous load shift; this suggests that the relationship between fault severity and classification difficulty is non-monotonic and depends on the signal-to-noise ratio of the fault impulse relative to background vibration. MRCN's relatively stronger performance on Target-B ($\kappa = 0.875$ at $K = 5$) compared with Target-A ($\kappa = 0.622$) is likely attributable to this same effect: the larger fault amplitude on Target-B provides a stronger initialisation signal for the inner-loop adaptation steps, partially compensating for the distributional gap that undermines MRCN on the harder Target-A setting. Raghu et al. [24] showed that MAML's practical benefit derives primarily from feature reuse rather than rapid adaptation. When both fault amplitude and operating speed shift simultaneously, the source encoder provides a poor initialisation for the target domain, and five inner-loop gradient steps are insufficient to recover. This failure pattern is consistent with Lin et al. [7], who observed similar degradation under large cross-domain gaps. The result suggests that source-domain meta-training is not a reliable strategy for compound distributional shift, a finding of practical relevance to prognostics and health management (PHM) system design.

6.3 Metric Selection and Diagnostic Implications

The multi-metric evaluation reveals failure modes that accuracy alone cannot expose. MRCN's accuracy–F1 gap of 5.3 pp at $K = 10$ on Target-A shows that it concentrates correct predictions on a

subset of fault classes while failing on others—a pathology that accuracy masks but macro F1 makes explicit. ResCaps at $K = 5$ achieves 0.593 accuracy yet $\kappa = 0.457$, below moderate agreement on the Landis–Koch scale, demonstrating that a model can appear functional on an accuracy scale while providing little genuine discriminative value. Cohen's κ is particularly informative in the few-shot regime not because class distributions are skewed (they are explicitly balanced) but because with $K = 5$ support samples a classifier can accumulate non-trivial accuracy through low-shot instability alone; κ adjusts for this by measuring agreement beyond what random guessing would predict. CapsNet is the only model to exceed $\kappa = 0.80$ at $K = 5$ on Target-A, placing it in the strong-agreement range of the Landis–Koch scale [27]. In the present context this means CapsNet maintains consistent class-level discrimination even with fewer than ten training examples per fault type—a regime in which all other evaluated models fall short of moderate agreement on at least one target condition.

6.4 Limitations and Future Work

Both CWRU and PU are laboratory benchmarks with controlled damage and balanced class distributions. Although PU introduces real accelerated-fatigue damage, neither dataset captures the noise levels, variable speeds, or class imbalance typical of continuous industrial operation. Evaluation on run-to-failure benchmarks such as FEMTO and PRONOSTIA [28], variable-speed recordings, and datasets with additive sensor noise would strengthen the external validity of these findings.

The baseline set is intentionally constrained to isolate the effect of architecture under a controlled, leak-free protocol. Deep CORAL is included as a representative covariance-alignment domain-adaptation method, but adversarial domain adaptation, maximum mean discrepancy approaches, and domain-generalisation frameworks [7] address distributional shift [4] through distinct mechanisms and may offer complementary strengths under double domain shift. Extending the comparison to these methods is a natural next step. We view the present study as establishing a controlled, leak-free benchmark protocol rather than an exhaustive leaderboard across all modern domain-adaptation, domain-generalisation, self-supervised, and Transformer paradigms; the goal is to isolate the effect of routing-based adaptation under tightly matched training conditions, upon which broader benchmark suites can be systematically layered in future work.

Standard dynamic routing [13] was chosen for reproducibility, but EM routing [29] and attention-based variants may yield further gains, particularly on hard transfer targets such as PU Target-B where all models produce $\kappa < 0.06$ at $K = 5$. The 70% unlabelled adaptation pool is also left unused under the inductive protocol adopted here; transductive inference and semi-supervised contrastive pre-training have shown consistent gains on harder cross-domain targets [7] and represent a direct extension of the present benchmark. Three fixed EDM fault diameters are used for controlled benchmarking; a continual learning extension with remaining useful life (RUL) regression on run-to-failure trajectories would address progressive fault evolution. Finally, capsule norms are treated as confidence proxies but are not formally calibrated; systematic evaluation via reliability diagrams or expected calibration error, alongside inference-speed benchmarking on resource-constrained hardware, are necessary before deployment in safety-critical embedded PHM systems.

7 Conclusion

This paper introduced a controlled few-shot benchmark for bearing fault diagnosis under simultaneous fault-severity and operating-load shift—the double domain shift—evaluated on both the CWRU and Paderborn University datasets using a leak-free three-way split and three complementary metrics: accuracy, macro F1, and Cohen's κ. A lightweight 1-D CNN with a dynamic capsule routing head outperforms all

five baselines in 15 of 18 CWRU conditions and in all PU conditions at $K \geq 10$, achieving $\kappa > 0.80$ at $K = 5$ on CWRU Target-A—a threshold reached by no other model in the benchmark—while using 18,624 parameters—roughly one-third of the MLP baseline—and no explicit alignment objective. The multi-metric evaluation surfaces findings invisible to accuracy alone: several baselines fall below moderate κ at low K, and MRCN's accuracy–F1 gap exposes class-selective failure that accuracy conceals.

Two important negative results emerge from the benchmark. Residual depth before the capsule head consistently degrades all three metrics at low K, confirming that parameter efficiency matters more than representational capacity in this regime. MAML-based meta-training provides no benefit under joint severity and load shift, extending the findings of prior work on single-factor cross-domain gaps. The consistently low PU Target-B scores ($\kappa < 0.06$ at $K = 5$ for all models) define a realistic lower bound for future work on the artificial-to-real fatigue transfer problem and suggest that transductive or semi-supervised approaches are a natural next step for this challenging target.

Acknowledgement: The authors acknowledge with thanks the KAU Endowment (WAQF) at King Abdulaziz University, Jeddah, Saudi Arabia, and the Deanship of Scientific Research (DSR) for technical and financial support. AI-assisted tools were used for language editing and formatting; all experimental design, data collection, analysis, and scientific conclusions are solely the work of the authors.

Funding Statement: The project was funded by KAU Endowment (WAQF) at King Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Safa Alsafari and Ayman Yafoz; data preparation and preprocessing: Safa Alsafari; implementation and experiments: Safa Alsafari; analysis and interpretation of results: Safa Alsafari and Ayman Yafoz; draft manuscript preparation: Safa Alsafari and Ayman Yafoz. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: This study uses two publicly available benchmark datasets: the Case Western Reserve University (CWRU) Bearing Data Set, available at https://engineering.case.edu/bearingdatacenter, and the Paderborn University (PU) Bearing Data Set, available at https://mb.uni-paderborn.de/kat/forschung/bearing-datacenter/data-sets-and-download. All experiment code, trained model weights, and result files generated in this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable. This study does not involve human participants or animals; it uses only publicly available, non-personal vibration benchmark datasets.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Smith WA, Randall RB. Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study. Mech Syst Signal Process. 2015;64–65:100–31. doi:10.1016/j.ymssp.2015.04.021.
2. Zhao M, Zhong S, Fu X, Tang B, Pecht M. Deep residual shrinkage networks for fault diagnosis. IEEE Trans Ind Inform. 2020;16(7):4681–90. doi:10.1109/tii.2019.2943898.
3. Wen L, Li X, Gao L, Zhang Y. A new convolutional neural network-based data-driven fault diagnosis method. IEEE Trans Ind Electron. 2018;65(7):5990–8. doi:10.1109/tie.2017.2774777.
4. Shao S, McAleer S, Yan R, Baldi P. Highly accurate machine fault diagnosis using deep transfer learning. IEEE Trans Ind Inform. 2019;15(4):2446–55. doi:10.1109/tii.2018.2864759.

5. Wu M, Zhang J, Xu P, Liang Y, Dai Y, Gao T, et al. Bearing fault diagnosis for cross-condition scenarios under data scarcity based on transformer transfer learning network. Electronics. 2025;14(3):515. doi:10.3390/electronics14030515.
6. Zhang S, Ye F, Wang B, Habetler T. Few-shot bearing fault diagnosis based on model-agnostic meta-learning. IEEE Trans Ind Appl. 2021;57(5):4754–64. doi:10.1109/tia.2021.3091958.
7. Lin J, Shao H, Zhou X, Cai B, Liu B. Generalized MAML for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals. Expert Syst Appl. 2023;230(3):120696. doi:10.1016/j.eswa.2023.120696.
8. Qiao Z, Ning S, Gai Y, Xie C. A digital twin guided physical-virtual denoising method for early fault detection of rolling element bearings. Mech Syst Signal Process. 2026;249(1):114108. doi:10.1016/j.ymssp.2026.114108.
9. Ning S, Qiao Z, Peng B, Zhu R, Zhang C. A digital twin-driven cross-domain adaptation method for bearing intelligent fault diagnosis. Nondestruct Test Eval. 2026:1–23. doi:10.1080/10589759.2025.2610726.
10. Ince T, Kiranyaz S, Eren L, Askar M, Gabbouj M. Real-time motor fault detection by 1-D convolutional neural networks. IEEE Trans Ind Electron. 2016;63(11):7067–75. doi:10.1109/tie.2016.2582729.
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
12. Wang Y, Chen L. A multi-scale spatial-temporal capsule network based on sequence encoding for bearing fault diagnosis. Complex Intell Syst. 2024;10(5):6189–212. doi:10.1007/s40747-024-01462-8.
13. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS); 2017 Dec 4–9; Long Beach, CA, USA. p. 3856–66.
14. Liu Y, Wang J, Li H, Chen X. Capsule networks for intelligent fault diagnosis: a roadmap of recent advancements and challenges. Expert Syst Appl. 2025;275:124327. doi:10.1016/j.eswa.2025.124327.
15. Jiang GJ, Zhao TH, Li HN, Sun CY, Li HY. A novel intelligent fault diagnosis method of rolling bearings based on capsule network with fast routing algorithm. Qual Reliab Eng Int. 2024;40(4):1944–58. doi:10.1002/qre.3496.
16. Lu W, Liu J, Lin F. The fault diagnosis of rolling bearings is conducted by employing a dual-branch convolutional capsule neural network. Sensors. 2024;24(11):3384. doi:10.3390/s24113384.
17. Xu H, Huang R, Li J, Liao W, Chen W. A causality-inspired capsule network for domain generalisation in cross-domain fault diagnosis of rolling bearing. IEEE Trans Instrum Meas. 2025;74:1–12. doi:10.1109/tim.2025.3569920.
18. Snell J, Swersky K, Zemel RS. Prototypical networks for few-shot learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS); 2017 Dec 4–9; Long Beach, CA, USA. p. 4077–87.
19. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the International Conference on Machine Learning (ICML); 2017 Aug 6–11; Sydney, Australia. p. 1126–35.
20. Hu X, Li W, Tian Z, Wu J. Novel joint transfer fine-grained metric network for cross-domain few-shot fault diagnosis. Knowl Based Syst. 2023;279(4):110958. doi:10.1016/j.knosys.2023.110958.
21. Vu MH, Nguyen VQ, Tran TT, Pham VT, Lo MT. Few-shot bearing fault diagnosis via ensembling transformer-based model with Mahalanobis distance metric learning from multiscale features. IEEE Trans Instrum Meas. 2024;73:1–18. doi:10.1109/tim.2024.3381270.
22. Rezazadeh N, Caputo F, Aversano A, Lamanna G, De Luca A, Perfetto D. Prototype-attention domain adaptation for explainable bearing fault diagnosis. Procedia Struct Integr. 2026;80:411–7. doi:10.1016/j.prostr.2026.02.039.
23. Lessmeier C, Kimotho JK, Zimmer D, Sextro W. 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: Proceedings of the European Conference of the PHM Society (PHMEurope); 2016 Jul 5–8; Bilbao, Spain.
24. Raghu A, Raghu M, Bengio S, Vinyals O. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. In: Proceedings of the International Conference on Learning Representations (ICLR); 2020 Apr 30; Addis Ababa, Ethiopia.

25. Sun B, Saenko K. Deep CORAL: correlation alignment for deep domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops; 2016 Oct 8–16; Amsterdam, The Netherlands. p. 443–50. doi:10.1007/978-3-319-49409-8_35.
26. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv:1412.6980v9. 2017. doi:10.48550/arXiv.1412.6980.
27. Landis JR. Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74. doi:10.2307/2529310.
28. Nectoux P, Gouriveau R, Medjaher K, Ramasso E, Chebel-Morello B, Zerhouni N, et al. PRONOSTIA: an experimental platform for bearings accelerated degradation tests. In: Proceedings of the IEEE International Conference on Prognostics and Health Management (PHM); 2012 Jun 18–21; Denver, CO, USA.
29. Hinton GE, Sabour S, Frosst N. Matrix capsules with EM routing. In: Proceedings of the International Conference on Learning Representations (ICLR); 2018 Apr 30–May 3; Vancouver, BC, Canada.