

A Priori Domain Shift Index for Assessing Transferability in Federated Fault Diagnosis of Rotating Machinery

Javad Afshari

J.Afshari13610220@gmail.com

Islamic Azad University of Birjand <https://orcid.org/0009-0001-0199-9202>

Mahmoud Afshari

Birjand University of Technology <https://orcid.org/0000-0002-8556-4264>

Hossein Afshari

University of Birjand <https://orcid.org/0000-0003-4981-2504>

Research Article

Keywords: Federated learning, domain shift, Wasserstein distance, transferability, fault diagnosis, rotating machinery

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11501644/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Federated learning (FL) has emerged as a promising approach for fault diagnosis in rotating machinery, particularly where raw data cannot be shared across machines or factories. However, differences in machine structure, operating conditions, and sensing systems introduce inter-machine domain shifts that degrade a global model on previously unseen machines. Existing studies mainly mitigate these shifts after model training, and few quantify their severity beforehand. This study investigates whether such severity can be measured before training. A Domain Shift Index (DSI) is defined as the data-volume-weighted sliced Wasserstein distance between model-independent vibration-feature distributions; it is computed from quantile summaries of shared random projections, so no raw samples are exchanged, and it is linked to a Lipschitz-type error bound whose constants are calibrated by quantile regression. The index is evaluated against six baseline discrepancy measures on 40 controlled synthetic machines (960 source–target cases) and on 16 domains of the public CWRU bearing benchmark (432 cases), using machine-level cluster bootstrap, Holm-corrected paired tests, and leave-one-machine-out calibration. The DSI is numerically stable (median coefficient of variation of about 2–3% with 200 projections) and moderately rank-correlated with zero-shot transfer error (Spearman ρ between 0.35 and 0.49 on synthetic data and between 0.30 and 0.49 on CWRU), but it is not superior to simple baselines such as MMD, and the calibrated bound, although it attains its nominal coverage, is loose and improves only marginally on a no-shift predictor. The index is therefore best regarded as a privacy-friendly screening dissimilarity rather than a quantitative predictor of accuracy loss; validation on industrial field data remains future work.

Introduction

With the rapid development of the Industrial Internet of Things (IIoT) and smart manufacturing, early fault detection in rotating machinery such as bearings, gearboxes, and electric motors has become increasingly important for predictive maintenance. Unexpected equipment failures can disrupt production and cause significant economic losses [1]. Although deep learning has achieved strong results in automated fault diagnosis, these methods are often developed under centralized settings where training data are collected in one location. In practice, however, industrial organizations may be unable or unwilling to share raw sensor data because of privacy, security, intellectual-property, and competitive concerns [2].

Federated learning (FL) provides a suitable alternative by allowing multiple machines or factories to collaboratively train a shared model without exchanging their raw data [2, 3]. However, differences in data distributions across clients, commonly referred to as data heterogeneity or non-IIDness, can reduce the stability and performance of the global model [4]. This issue becomes more challenging when machines differ in their mechanical structures, operating conditions, or sensing systems, leading to significant inter-machine domain shifts [5, 6].

Existing studies have mainly addressed this problem through model architectures and training strategies, including contrastive prototype learning [5], corrective two-stage training [6], wavelet-convolutional transfer learning [7], and oversampling-based federated transfer [8]. These approaches generally evaluate domain-shift effects after model training, based on the resulting change in diagnostic performance. Moreover, recent studies indicate that many transfer-learning methods primarily address changes in operating conditions within individual machines rather than knowledge transfer between physically different machines [7], and that evaluation practice in this field is often not realistic enough to support claims of generalization [9, 10]. As a result, reliable assessment of cross-machine transferability remains an open challenge.

From an industrial perspective, this reactive approach can require substantial computational time and resources before the compatibility of a new machine with an existing federated model becomes clear. A practical a priori assessment of inter-machine domain shift is therefore desirable. This paper examines how far such an assessment can be supported by a distributional distance computed on unlabeled, model-independent features.

The contributions of this revised study are fourfold. (i) A Domain Shift Index (DSI) based on the sliced Wasserstein distance is defined together with a federated computation protocol that exchanges only quantile summaries of random projections, and a tighter mixture variant is introduced. (ii) The index is related to a Lipschitz-type error bound whose constants are calibrated on held-out machines. (iii) The index is evaluated on a large set of synthetic and real (CWRU) domain pairs against six baseline discrepancy measures, with explicit uncertainty quantification. (iv) A decision rule for deployment, fine-tuning, or local training is derived and tested. The results are reported as obtained, including the finding that the DSI is moderately informative but not superior to simpler measures.

RELATED WORK

2.1 Federated Learning for Inter-Machine Fault Diagnosis

Recent studies on federated learning for fault diagnosis have mainly focused on improving model architectures and training strategies. Fang et al. [5] introduced contrastive prototype learning to align representations across clients and reduce the effects of domain shifts. He and Shen [11] proposed a federated cross-machine diagnostic framework for machine-level motors under extreme label shortage. Yan et al. [6] developed a dual-correction training strategy (FTSDC) for knowledge transfer across different machines without requiring target-domain samples. Zhao et al. [7] used a multistep wavelet-convolutional network for cross-machine bearing fault diagnosis and highlighted the limited focus of existing transfer-learning methods on transfer across physically different machines. Xu et al. [8] addressed imbalanced client data through an oversampling-based federated transfer approach for rotating-machinery clusters.

Despite these efforts, the severity of domain shift is generally assessed only after model training, based on the resulting change in diagnostic performance. Thus, these studies do not provide an independent measure of inter-machine domain shift before training.

2.2 Distributional Measures and Transferability Estimation

Statistical measures of distributional discrepancy are central to domain-adaptation theory and practice. The A-distance and its proxy, obtained from the error of a domain classifier, enter the classical generalization bound of Ben-David et al. [12, 13]. Kernel-based measures such as the maximum mean discrepancy (MMD) [14], second-order statistics alignment (CORAL) [15], and the Fréchet distance between Gaussian fits [16] are widely used as training losses or diagnostics. Optimal-transport distances provide bounds in which the Wasserstein distance between source and target marginals directly controls the target risk [17, 18], and have been used as training objectives in domain-adversarial networks [19]. Within bearing diagnosis, Wasserstein-type constraints and discriminators have been used to improve robustness to unknown operating conditions [20] and to align subdomains [21]. Outside this field, transferability estimators such as LEEP [22] score a source model on target data, but they require a trained model, which is exactly what is unavailable in the a priori setting considered here.

Overall, in the existing literature these distances are used mainly inside training or optimization, and their ability to act as an independent, model-free, a priori indicator of cross-machine transfer error in a federated setting has not been systematically benchmarked against one another on controlled and real bearing data. This is the question examined here.

2.3 Research Gap and Contribution

Studies on federated fault diagnosis [5–8, 11] address domain heterogeneity mainly through model architectures and training strategies, while domain-adaptation studies use distributional distances inside the optimization. The gap addressed here is an empirical and theoretical assessment of a model-independent index that can be computed before training without sharing raw samples. The index is linked to a risk bound, and its predictive value is tested with calibration, baseline comparison, and uncertainty quantification rather than with a single correlation coefficient.

PROPOSED DOMAIN-SHIFT ASSESSMENT FRAMEWORK

3.1 Problem Definition and Formulation

Let K denote the number of clients (machines) participating in the federation, each maintaining its own labeled local dataset. To ensure that the index remains computable before any fault-detection model exists, a fixed, predefined feature extraction function $\varphi: X \rightarrow \mathbb{R}^d$ is used. Rather than relying on learned representations, it uses established vibration-signal characteristics: time-domain statistics, envelope-spectrum energies at the characteristic fault frequencies, and the maximum spectral kurtosis [23, 24]. These hand-crafted features are chosen so that the index is independent of any model that has yet to be

trained. Let μ_k and μ_m denote the feature distributions of client k and of the newly introduced machine m , respectively, and let $\pi_k = n_k / \sum_j n_j$ be the data-volume weight of client k , as in FedAvg [3]. Features are compressed by a fixed monotone transform ($\log_{10}$ for scale-type features, $\log(1 + x)$ for spectral kurtosis) and standardized with federation-wide moments aggregated from the client moments only; no statistic of the new machine enters the standardization.

3.2 Domain Shift Index

The distance between feature distributions is quantified with the first-order Wasserstein distance $W_1(\mu_k, \mu_m)$ [25] in its sliced form [26], which averages one-dimensional Wasserstein distances over random projection directions θ on the unit sphere. The Domain Shift Index of a new machine m is defined as

$$S_m = \sum_{k=1}^K \pi_k \cdot SW_1(\mu_k, \mu_m), \quad SW_1(\mu, \nu) = E_{\theta} [W_1(\theta \# \mu, \theta \# \nu)] \quad (1)$$

where $\theta \# \mu$ is the distribution of the projection $\theta \cdot \phi(x)$. Because $|\theta \cdot a - \theta \cdot b| \leq \|a - b\|$, $SW_1 \leq W_1$ always holds; the reverse inequality holds only up to a dimension-dependent power, so S_m is an index motivated by the bound below rather than a strict upper bound of the Wasserstein term. By convexity of W_1 in its first argument, the distance to the data-volume mixture is never larger than the weighted sum, which motivates the tighter mixture variant

$$S_m^{\text{mix}} = SW_1(\sum_k \pi_k \mu_k, \mu_m) \leq S_m \quad (2)$$

The mixture variant was introduced after the weakness of the weighted sum for federated models had been observed in a pilot run; it is therefore reported as an exploratory variant, and its comparison with S_m should be read with that selection in mind.

Federated computation protocol. Computing the distance directly requires access to raw data from both parties, which conflicts with the privacy principles of FL. Privacy-preserving estimators of the Wasserstein distance exist [27, 28]. In this study, a simpler protocol exploiting the sliced form is used: the server broadcasts a random seed from which every party draws the same projection directions; each client returns, for every direction, $Q = 100$ quantiles of its projected features; the server forms the one-dimensional distances from these summaries (and the mixture quantiles for S_m^{mix} by weighting the summary points with π_k/Q). No raw sample leaves a client, and the quantities exchanged are exactly those used in the experiments. The quantile-summary computation reproduces the exact sliced distance with a median relative error of $\{\text{SENS}[\text{Synthetic}][\text{protocol_rel_err}][\text{median}] * 100 : .2f\}\%$ (maximum $\{\text{SENS}[\text{Synthetic}][\text{protocol_rel_err}][\text{max}] * 100 : .1f\}\%$) on the synthetic data and $\{\text{SENS}[\text{CWRU}][\text{protocol_rel_err}][\text{median}] * 100 : .2f\}\%$ (maximum $\{\text{SENS}[\text{CWRU}][\text{protocol_rel_err}][\text{max}] * 100 : .1f\}\%$) on CWRU. Quantile summaries still leak distributional information, and a formal privacy analysis (for example with differential-privacy noise [28]) is outside the scope of this paper.

3.3 Theoretical Bound for Performance Degradation

Let ℓ be a bounded loss, h a hypothesis, and $R_k(h) = E_{P_{\{k\}}}[\ell(h(x), y)]$ the risk of h on client k , with P_m the distribution of the new machine. Define $g_h^k(x) = E_{P_{\{k\}}}[\ell(h(x), y) | x]$. **Proposition 1.** If g_h^m is L -Lipschitz in x for every h in the hypothesis class, then for every h

$$R_m(h) \leq \sum_k \pi_k R_k(h) + L \cdot \sum_k \pi_k W_1(\mu_k, \mu_m) + \lambda^* \quad (3)$$

with $\lambda^* = \sum_k \pi_k \lambda_k$ and $\lambda_k = \sup_h \int |g_h^m - g_h^k| d\mu_k$. *Proof sketch.* For one client, $R_m(h) - R_k(h) = [\int g_h^m d\mu_m - \int g_h^m d\mu_k] + \int (g_h^m - g_h^k) d\mu_k$. The bracket is at most $L \cdot W_1(\mu_k, \mu_m)$ by the Kantorovich–Rubinstein duality, and the last term is at most λ_k . Averaging with the weights π_k gives Eq. (3). This is the weighted form of the optimal-transport bounds of Redko et al. [17] and Shen et al. [19]; finite-sample terms are omitted. Three remarks are essential. First, $R_k(h)$ denotes the risk of the *global* model on client k (the first version of this paper wrote $R_k(h_k)$, which is not what the argument yields). Second, the Lipschitz argument applies to a surrogate loss, not to the 0–1 error; the measured accuracy loss is related to the bound only empirically. Third, λ^* is the label-conditional discrepancy, which marginal-distribution distances cannot capture; it is estimated, together with L , from data.

Calibrated bound. In the experiments, S_m replaces W_1 , and (L, λ^*) are estimated by τ -quantile regression [29] of the excess error $e_m - \sum_k \pi_k R_k$ on S_m with $L \geq 0$ and $\lambda^* \geq 0$ ($\tau = 0.90$), solved as a linear program. The resulting quantity $B_m = \sum_k \pi_k R_k + \hat{L} S_m + \hat{\lambda}^*$ is a calibrated 90% bound on the observed error rather than a theoretical guarantee. An ordinary least-squares fit of the same form provides the point predictor.

3.4 Practical Procedure

Data collection: the new machine provides a small batch of unlabeled feature vectors. **Estimation:** the server collects the projection quantile summaries of every client and of the new machine. **Index calculation:** S_m (and S_m^{mix}) are computed. **Calibration:** $\hat{L}$ and $\hat{\lambda}^*$ are fitted on a reference bank of previously observed (federation, machine) cases; calibration cases must not include the new machine. **Decision logic:** the predicted error is compared with two thresholds $\tau_1 \leq \tau_2$: direct deployment below τ_1 , fine-tuning with a few labeled samples between τ_1 and τ_2 , and independent local training above τ_2 . The thresholds are tuned on the reference bank (Section 4.5), not fixed a priori.

3.5 Computational Considerations

The communication cost per client is $Q \times n_{\text{dir}}$ numbers (here 100×200), independent of the number of samples, and the index is computed once when a machine is introduced. Federations of tens of clients and the cost of secure aggregation have not been tested.

EXPERIMENTAL DESIGN, RESULTS, AND IMPLEMENTATION

4.1 Implementation Environment and Code Structure

All experiments ran on a single-core CPU with 4 GB RAM (Ubuntu, Python 3.12) using NumPy, SciPy, scikit-learn, pandas, and matplotlib [30, 31]; the complete run takes about 33 minutes. The complete code is provided in Appendices A–G, and it is written to run unchanged on Google Colab (a CPU runtime suffices). To reproduce the results, open a new Colab notebook, paste Appendices A–G into consecutive cells in order (or save them as one file and execute it with `%run`), and execute the final line of Appendix G, which calls `run_all()`. Appendix A downloads the required CWRU files with a sparse Git clone of the public mirror <https://github.com/s-whynot/CWRU-dataset> into `/content/CWRU` (about 175 MB); alternatively, the official .mat files of the CWRU Bearing Data Center [32] can be copied to Google Drive, mounted, and the folder passed as `cwr_root`, because the loader finds files by their numeric ID in any sub-folder. Tables and figures are written to `/content/dsi_outputs` (results.json, case-level CSV files, and PNG figures). Random seeds are fixed in CFG; the earlier use of Python’s `hash()` for seeding, which is not reproducible across runs, was removed.

4.2 Datasets

Synthetic machines. Forty machines were generated with the standard bearing-fault signal model, in which a point defect excites a decaying structural resonance that recurs at BPFO, BPFI, or BSF and is modulated by the shaft frequency [23]. Six machines are the named configurations of the first version of this paper (Table 1); thirty-four further machines were drawn by a random sampler whose scale parameter $s \sim U(0,1)$ widens all parameter ranges (Table 2), so that the shift between machines forms a continuum rather than a handful of levels. Each machine has four classes (Normal, IR, OR, Ball), 40–60 signals per class of 0.5 s, and within-machine variability in amplitude and resonance frequency. The generator is in Appendix B.

Table 1
Named synthetic machines (reference A and test machines)

ID	f_s (Hz)	f_{sh} (Hz)	N_b	α (°)	f_{res} (Hz)	σ_{noise}
A	12000	29.95	9	0	3000	0.30
B1	12000	32.0	9	0	3000	0.30
B2	12000	39.9	9	0	3000	0.30
C1	12000	45.0	8	5	3400	0.35
C2	25600	24.4	8	9	4200	0.45
D	48000	58.0	7	12	7000	0.60

Here f_s is the sampling frequency, f_{sh} the shaft frequency, N_b the number of balls, α the contact angle, f_{res} the structural resonance, and σ_{noise} the noise level. For C2 and D the ball diameter and damping also differ (Appendix B). The first version of this paper described Table 1 twice under two section headings and its parameter table was incomplete; the full definitions are now in the code.

Table 2
Random machine sampler ($s \sim U(0,1)$ scales every spread)

Parameter	Distribution
f_s	12 kHz ($s \leq 0.35$); otherwise {12, 12, 25.6, 48} kHz
f_{sh} , f_{res} , damping, noise, amplitude	reference $\times 2^{N(0, c \cdot s)}$, $c = 0.8, 0.8, 0.7, 0.9, 0.6$
N_b	round($9 + N(0, 2.5 s)$), clipped to [6, 13]
α , ball diameter	$ N(0, 8 s) ^\circ$; $\times(1 + U(-0.3, 0.3) s)$
modulation depth	$U(0.3, 0.8)$

CWRU benchmark. Real data come from the Case Western Reserve University Bearing Data Center [32], using the 12 kHz records. A domain (“machine”) is defined by the accelerometer position (drive end, DE, bearing SKF 6205; or fan end, FE, bearing SKF 6203), the fault diameter (0.007, 0.014, 0.021 in for DE; 0.007 in for FE), and the motor load (0–3 hp), which gives 16 domains (Table 3). Each domain has four classes: Normal (48 kHz baseline records decimated to 12 kHz), inner-race, ball, and outer-race (6 o’clock position) faults. Per class and domain, 50 windows of 4096 samples were taken at evenly spaced positions of the record. Fault frequencies were computed from the recorded shaft speed and the bearing multiples published with the dataset. Domains therefore differ in load, sensor location and bearing, and fault severity, but all come from a single test rig; this limits the independence of domains and is discussed in Section 5.5. Paderborn data [33] and field data from an industrial plant were not used in this revision.

Table 3
CWRU domains (16 machines)

Sensor	Fault diameter (in)	Loads (hp)	Domains
DE	0.007, 0.014, 0.021	0, 1, 2, 3	12
FE	0.007	0, 1, 2, 3	4

4.3 Feature Extraction

An eight-dimensional feature vector $\phi(x)$ is extracted from each signal: RMS, kurtosis, skewness, crest factor, envelope-spectrum energy within ± 15 Hz of BPFO, BPFI, and BSF, and the maximum STFT spectral kurtosis [24]. The envelope spectrum uses the Hilbert transform and Welch estimation with a fixed 6 Hz resolution for every sampling frequency (the first version used a fixed segment length, which changed the resolution with f_s). The earlier “spectral kurtosis” was in fact the kurtosis of the normalized power spectrum and has been replaced by the estimator of Antoni (Appendix C).

4.4 Federated Training, Cases, and Baselines

A case is a federation of K clients plus one target machine, with $K = 1$ (a source–target pair) or $K = 3$. For each of the 40 synthetic targets, 12 pairs and 12 federations were sampled from the other machines (960 cases); for each of the 16 CWRU targets, all 15 pairs and 12 federations were used (432 cases). The global model is a softmax-regression classifier trained by FedAvg [3] (30 rounds, 10 local steps); a Random Forest (100 trees) trained on pooled client data is a classifier-agnostic check. Each client keeps 30% of its data (stratified at random for synthetic data, last time block for CWRU) to measure R_k . The primary outcome is the zero-shot error of the global model on all target data. Fine-tuning uses 5 labeled target samples per class (3 random draws). The DSI is compared with six baselines computed on the same standardized features: MMD with a Gaussian kernel and median bandwidth [14], CORAL distance [15], Fréchet distance [16], Euclidean distance of feature means, the average one-dimensional marginal W_1 , and the proxy A-distance [12] from a cross-validated logistic domain classifier. Baselines are aggregated in the same weighted-sum form as the DSI and, for $K = 3$, also on the pooled client data (suffix “pooled”, not privacy-preserving).

4.5 Statistical Protocol

Rank (Spearman) correlation between each index and the target error is the primary association measure; Pearson correlations are in the supplementary output. Cases that share a target machine are not independent, so 95% confidence intervals come from a cluster bootstrap that resamples target machines (1000 replicates) [34]. Differences between the correlation of the DSI and of each baseline use the same bootstrap draws, with two-sided p-values adjusted by the Holm procedure [35]. A target-machine-level analysis (mean index versus mean error per target, permutation test) removes the pair structure. Calibration and decisions are evaluated by leave-one-machine-out validation: for every held-out machine, (L, λ^*) and the thresholds are fitted only on cases in which that machine appears neither as target nor as client. The oracle action for a case is “deploy” if the zero-shot error is at most $\epsilon = 0.10$, “fine-tune” if the fine-tuned error is at most ϵ , and “local training” otherwise. Reported decision metrics are balanced accuracy and the unsafe-deployment rate, i.e. the fraction of “deploy” decisions for which deployment was not acceptable.

RESULTS AND DISCUSSION

5.1 Association Between the Index and Transfer Error

Table 4 reports the rank correlation for source–target pairs. The DSI reaches $\rho = 0.49$ [0.37, 0.60] on the synthetic machines and 0.49 [0.41, 0.60] on CWRU. The result of the first version ($r = 0.900$, $p = 0.037$, $n = 5$ hand-designed pairs) is therefore not reproduced at scale and is superseded. On synthetic data MMD is more strongly associated with the error than the DSI (0.69 [0.60, 0.76]); on CWRU the DSI is statistically indistinguishable from MMD, Fréchet, CORAL, and mean shift, and the marginal Wasserstein distance is slightly higher (0.52 [0.42, 0.65]).

Table 4
Spearman ρ [95% cluster-bootstrap CI] with the zero-shot error, $K = 1$ pairs

Index	Synthetic (480 pairs)	CWRU (240 pairs)
DSI	0.49 [0.37, 0.60]	0.49 [0.41, 0.60]
MMD	0.69 [0.60, 0.76]	0.48 [0.41, 0.57]
CORAL	0.12 [-0.04, 0.31]	0.46 [0.36, 0.58]
Frechet	0.47 [0.35, 0.59]	0.47 [0.39, 0.57]
MeanShift	0.52 [0.41, 0.62]	0.45 [0.35, 0.57]
MargW1	0.47 [0.35, 0.60]	0.52 [0.42, 0.65]
PAD	0.55 [0.46, 0.62]	0.26 [0.11, 0.39]

After Holm adjustment, on the synthetic data the DSI is significantly better than CORAL ($\Delta\rho=+0.37$, $p_{\text{Holm}}=0.006$), Frechet ($\Delta\rho=+0.02$, $p_{\text{Holm}}=0.048$) and significantly worse than MMD ($\Delta\rho=-0.20$, $p_{\text{Holm}}=0.006$); on CWRU it is better than PAD ($\Delta\rho=+0.23$, $p_{\text{Holm}}=0.006$) and worse than MargW1 ($\Delta\rho=-0.03$, $p_{\text{Holm}}=0.006$). The machine-level analysis shows that the pair-level association is not a property of the target alone: the target-level correlation of the DSI is $\rho = 0.01$ ($n = 40$, $p = 0.95$) and 0.02 ($n = 16$, $p = 0.95$) for pairs.

Table 5 reports the federated case with $K = 3$. The associations are weaker than for pairs. On the synthetic machines the DSI gives $\rho = 0.35$ [0.16, 0.54] and the mixture variant 0.40 [0.22, 0.57], against 0.60 [0.45, 0.72] for MMD; on CWRU, the DSI gives 0.30 [0.19, 0.43], the mixture variant 0.25 [0.13, 0.44], and the marginal W_1 0.39 [0.28, 0.51]. Replacing the weighted sum by the mixture distance improved the association on synthetic data ($\Delta\rho = +0.05$, $p_{\text{Holm}} = 0.013$) but not on CWRU ($\Delta\rho = -0.05$, $p_{\text{Holm}} = 1.00$). With the Random Forest as the downstream model the ordering of the indices is similar (Spearman ρ of the DSI: 0.25 on synthetic and 0.44 on CWRU), so the weak association is not specific to the softmax classifier.

Table 5
Spearman ρ [95% CI] with the zero-shot error, K = 3 federations

Index	Synthetic (480)	CWRU (192)
DSI	0.35 [0.16, 0.54]	0.30 [0.19, 0.43]
DSI-mix	0.40 [0.22, 0.57]	0.25 [0.13, 0.44]
MMD	0.60 [0.45, 0.72]	0.20 [0.00, 0.45]
CORAL	0.10 [-0.10, 0.31]	0.05 [-0.12, 0.24]
Frechet	0.33 [0.13, 0.53]	0.21 [0.09, 0.37]
MeanShift	0.41 [0.24, 0.57]	0.29 [0.17, 0.46]
MargW1	0.34 [0.14, 0.53]	0.39 [0.28, 0.51]
PAD	0.49 [0.33, 0.62]	-0.01 [-0.29, 0.29]
MMD (pooled)	0.51 [0.36, 0.65]	0.25 [0.09, 0.48]
CORAL (pooled)	0.15 [-0.04, 0.35]	0.04 [-0.15, 0.25]
Frechet (pooled)	0.37 [0.18, 0.55]	0.10 [-0.07, 0.37]
MeanShift (pooled)	0.41 [0.25, 0.57]	0.27 [0.12, 0.46]
MargW1 (pooled)	0.39 [0.20, 0.56]	0.35 [0.23, 0.51]
PAD (pooled)	0.42 [0.26, 0.56]	0.27 [0.09, 0.46]

5.2 Calibration of L and λ^* and Bound Quality

Table 6 gives the leave-one-machine-out calibration for K = 3. The 90% quantile bound attains its nominal coverage (synthetic 0.90, CWRU 0.90), but this is of limited value because the bound is loose: its mean is 0.25 on synthetic data and 0.53 on CWRU, against mean observed errors of 0.08 and 0.25, and the intercept λ^* is estimated at 0.00 and 0.00, respectively. The fitted Lipschitz-type slope is $\hat{\mathcal{L}} = 0.27 \pm 0.01$ (synthetic) and 0.94 ± 0.04 (CWRU), values being means and standard deviations over the held-out folds. As a point predictor, the calibrated DSI improves only slightly on a null model that ignores the shift (mean absolute error 0.093 versus 0.103 on synthetic data and 0.156 versus 0.166 on CWRU), and the coefficient of determination stays close to zero (0.04 and -0.00). On synthetic data MMD gives the best calibrated predictor (MAE 0.074, $R^2 = 0.36$). The quantitative use of Eq. (3) as an error predictor is therefore not supported by these data.

Table 6
Leave-one-machine-out calibration, $K = 3$ (MAE and R^2 of the point predictor; coverage of the 90% bound)

Data	Index	MAE [95% CI]	R^2	Cover.
Synth.	DSI	0.093 [0.071, 0.118]	0.04	0.90
Synth.	DSI-mix	0.091 [0.069, 0.116]	0.06	0.90
Synth.	MMD	0.074 [0.057, 0.095]	0.36	0.92
Synth.	MargW1	0.091 [0.070, 0.116]	0.06	0.91
Synth.	No shift	0.103 [0.082, 0.128]	-0.07	0.90
CWRU	DSI	0.156 [0.134, 0.181]	0.00	0.90
CWRU	DSI-mix	0.158 [0.135, 0.185]	-0.02	0.91
CWRU	MMD	0.162 [0.138, 0.190]	-0.07	0.90
CWRU	MargW1	0.151 [0.130, 0.175]	0.06	0.91
CWRU	No shift	0.166 [0.147, 0.189]	-0.12	0.89

5.3 Decision Logic

With $\varepsilon = 0.10$ the oracle actions are deploy / fine-tune / local training in 368 / 87 / 25 synthetic and 56 / 125 / 11 CWRU federations. Using thresholds tuned on held-out calibration cases, the DSI-based rule reaches a balanced accuracy of 0.51 [0.33, 0.70] on synthetic data and 0.22 [0.15, 0.42] on CWRU, where a random three-way guess gives 0.33 (Table 7). The unsafe-deployment rate is 7% and 42%, respectively. The decision logic of Section 3.4 therefore does not currently perform well enough to support operational use, and no baseline measure is clearly better.

Table 7
Decision rule, $K = 3$ (balanced accuracy [95% CI];
unsafe-deployment rate)

Data	Index	Bal. acc. [CI]	Unsafe dep.
Synth.	DSI	0.51 [0.33, 0.70]	7%
Synth.	DSI-mix	0.53 [0.36, 0.71]	7%
Synth.	MMD	0.61 [0.46, 0.76]	6%
Synth.	PAD	0.51 [0.24, 0.59]	16%
CWRU	DSI	0.22 [0.15, 0.42]	42%
CWRU	DSI-mix	0.19 [0.13, 0.35]	38%
CWRU	MMD	0.27 [0.19, 0.50]	51%
CWRU	PAD	0.29 [0.19, 0.52]	57%

5.4 Sensitivity and Stability of the Index

The index itself is numerically stable (Fig. 3 and Table 8). With 200 projections the median coefficient of variation of S_m over repeated draws of the directions is 2.6% (synthetic) and 2.1% (CWRU); with 50 unlabeled target samples it is 3.7% and 5.8%, and the association with the error is essentially unchanged down to 25–50 samples on synthetic data, whereas on CWRU it drops for 25 samples or fewer. By contrast, the association depends on the modelling choices. Standardizing with statistics that include the target data (not a priori) changes p from 0.36 to 0.48 on synthetic data and from 0.30 to 0.23 on CWRU, and omitting standardization lowers it to 0.24 and 0.03. The best feature subset differs between the two data sets (statistical features on synthetic data, envelope and spectral-kurtosis features on CWRU), so the feature function ϕ , which the first version kept fixed without examination, is a main source of variation. These sensitivity analyses vary only the index; the classifier errors are those of the full feature set.

Table 8
Sensitivity of the DSI (Spearman ρ with federated error, $K = 3$)

Variant	Synthetic	CWRU
Default (200 directions, federation moments)	0.36	0.30
10 directions	0.34	0.27
50 directions	0.36	0.29
1000 directions	0.36	0.31
10 unlabeled target samples	0.35	0.21
25 unlabeled target samples	0.36	0.21
50 unlabeled target samples	0.36	0.28
100 unlabeled target samples	0.36	0.28
Standardization with target data included	0.48	0.23
No standardization	0.24	0.03
Features: statistical (4)	0.44	0.26
Features: envelope + SK (4)	0.13	0.35
Features: no scale (RMS, env)	0.26	-0.09

5.5 Analysis, Limitations, and Future Directions

The evidence supports a modest conclusion. A model-independent, privacy-friendly distance between feature distributions is moderately associated with the cross-machine error of a federated model, and the association survives clustering-aware statistics and a classifier change; it does not, however, outperform simpler discrepancy measures, and it does not yield a calibrated quantitative predictor or a reliable decision rule with the present feature function. Several reasons are plausible but were not all tested. The index compares marginal feature distributions, whereas much of the error on CWRU arises from changes in the class-conditional distribution (different fault diameters, bearings, and sensor positions), which marginal distances cannot see (the λ^* term of Eq. (3)). The federated model's error depends on how close the target is to the nearest client rather than to the average client, which the mixture variant addresses only on synthetic data. The eight hand-crafted features mix machine-specific scale and geometry with fault information. The pair-level association is weak at the level of target machines.

The main limitations are as follows. First, only two data sources were used: synthetic machines generated by the same model family that defines the features, and 16 domains from a single test rig, which are not independent machines; Paderborn and industrial field data were not analysed, so claims about industrial deployment are not made. Second, the downstream model is a linear softmax classifier

trained with FedAvg, and Lipschitz assumptions for deep networks were not examined. Third, the oracle thresholds ($\epsilon = 0.10$, five labeled samples per class) are arbitrary, and the mixture variant was introduced after exploratory analysis. Fourth, privacy of quantile summaries was not analysed formally, and communication with tens of clients was not measured.

CONCLUSION AND FUTURE WORK

6.1 Summary

This paper examined whether the severity of inter-machine domain shift can be measured before training a federated fault-diagnosis model. A sliced-Wasserstein Domain Shift Index computed from shared quantile summaries was defined, related to a Lipschitz-type risk bound, and evaluated on 960 synthetic and 432 CWRU cases against six baseline measures. The index is stable and moderately correlated with transfer error ($\rho \approx 0.3-0.5$), but it is not better than simple baselines, the calibrated bound is valid in coverage yet loose, and the decision rule is not reliable enough for deployment. The earlier result based on five simulated pairs is not supported at scale.

6.2 Future Research Directions

Directions suggested by these findings are: (i) label-aware or class-conditional shift measures, for example with pseudo-labels or prototype-wise Wasserstein distances, to address the λ^* term; (ii) learned, model-independent encoders in place of the manual feature function, with a re-examination of the a priori property; (iii) validation on further public benchmarks such as Paderborn and on industrial field data, with a calibration bank of many independent machines; (iv) extension of the bound to personalized and clustered FL; (v) a formal privacy analysis of the quantile-summary protocol; and (vi) an operational decision-support tool only after the decision rule reaches acceptable accuracy.

References

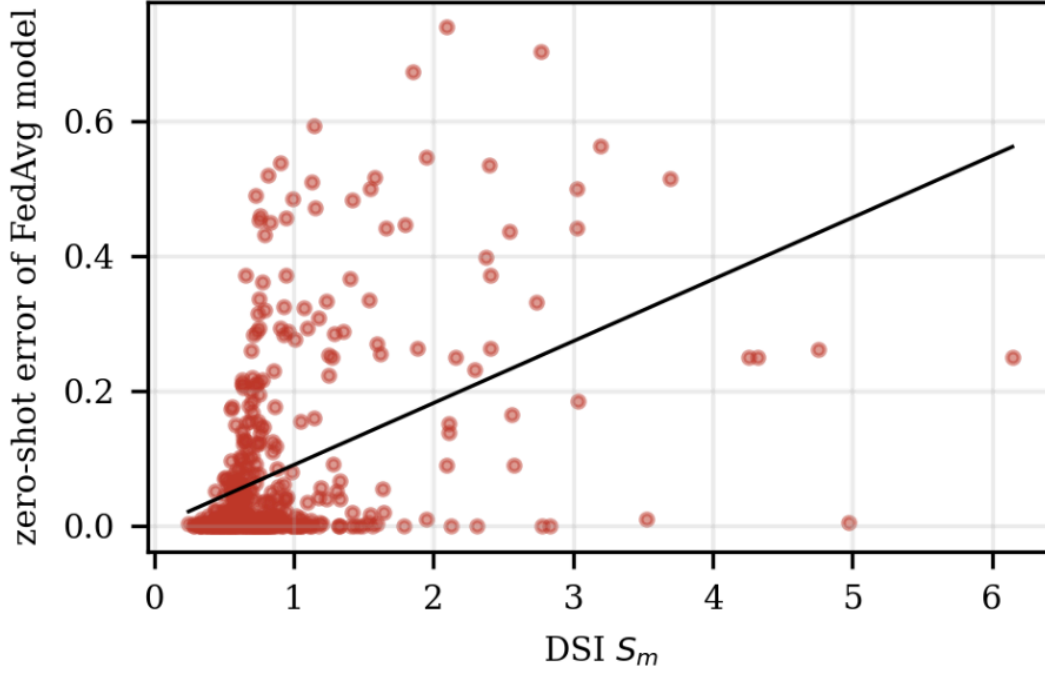
1. Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi AK (2020) Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech Syst Signal Process* 138:106587
2. Kairouz P, McMahan HB et al (2021) Advances and open problems in federated learning. *Found Trends Mach Learn* 14(1–2):1–210
3. McMahan HB, Moore E, Ramage D, Hampson S, Agüera B (2017) Communication-efficient learning of deep networks from decentralized data. In *Proceedings of AISTATS*, PMLR 54, 1273–1282
4. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V (2020) Federated optimization in heterogeneous networks. In *Proceedings of MLSys*, 2, 429–450
5. Fang L, Shi J, Qu H, Safran M, Tan J, Wan J (2025) Contrastive prototype guided federated learning for rotating machinery fault diagnosis under spatio-temporal domain shift. *Mech Syst Signal Process* 232:112707. <https://doi.org/10.1016/j.ymssp.2025.112707>

6. Yan Z, Xu Z, Zhang Y et al (2024) FTSDC: A novel federated transfer learning strategy for bearing cross-machine fault diagnosis based on dual-correction training. *Adv Eng Inform* 61:102499. <https://doi.org/10.1016/j.aei.2024.102499>
7. Zhao L, He Y, Zheng H, Dai D (2025) A novel multistep wavelet convolutional transfer diagnostic framework for cross-machine bearing fault diagnosis. *Sensors* 25(10):3141. <https://doi.org/10.3390/s25103141>
8. Xu Z, Yu L, Li S, Li C, Feng Y (2025) An optimized oversampling-based federated transfer learning approach for rotating machinery cluster fault diagnosis. *J Comput Des Eng* 12(8):154–172. <https://doi.org/10.1093/jcde/qwaf068>
9. Vieira JP, Bauler VA, Rosa RK, Silva D (2025) Towards a more realistic evaluation of machine learning models for bearing fault diagnosis. *arXiv preprint, arXiv:2509.22267*
10. Hendriks J, Dumond P, Knox DA (2022) Towards better benchmarking using the CWRU bearing fault dataset. *Mech Syst Signal Process* 169:108732
11. He Y, Shen W (2024) A federated cross-machine diagnostic framework for machine-level motors with extreme label shortage. *Adv Eng Inform*. <https://doi.org/10.1016/j.aei.2024.102511>
12. Ben-David S, Blitzer J, Crammer K, Pereira F (2007) Analysis of representations for domain adaptation. *Adv Neural Inf Process Syst* 19:137–144
13. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2010) A theory of learning from different domains. *Mach Learn* 79(1–2):151–175
14. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A (2012) A kernel two-sample test. *J Mach Learn Res* 13:723–773
15. Sun B, Saenko K (2016) Deep CORAL: Correlation alignment for deep domain adaptation. *ECCV Workshops LNCS* 9915:443–450
16. Dowson DC, Landau BV (1982) The Fréchet distance between multivariate normal distributions. *J Multivar Anal* 12(3):450–455
17. Redko I, Habrard A, Sebban M (2017) Theoretical analysis of domain adaptation with optimal transport. In *ECML PKDD 2017, LNCS* 10535, 737–753
18. Courty N, Flamary R, Tuia D, Rakotomamonjy A (2017) Optimal transport for domain adaptation. *IEEE Trans Pattern Anal Mach Intell* 39(9):1853–1865
19. Shen J, Qu Y, Zhang W, Yu Y (2018) Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1)
20. Ma Z et al (2025) Bearing fault diagnosis under unknown operating conditions using distribution robust optimization theory with distribution exploration strategy and Wasserstein constraints. *Meas Sci Technol* 36(6):066115. <https://doi.org/10.1088/1361-6501/add040>
21. Cai H, Yang B, Xue Y, Xu Y (2024) Residual adversarial subdomain adaptation network based on Wasserstein metrics for intelligent fault diagnosis of bearings. *Appl Sci* 14(19):9057. <https://doi.org/10.3390/app14199057>

22. Nguyen C, Hassner T, Seeger M, Archambeau C (2020) LEEP: A new measure to evaluate transferability of learned representations. In Proceedings of ICML, PMLR 119
23. Randall RB, Antoni J (2011) Rolling element bearing diagnostics—A tutorial. *Mech Syst Signal Process* 25(2):485–520. <https://doi.org/10.1016/j.ymssp.2010.07.017>
24. Antoni J (2006) The spectral kurtosis: A useful tool for characterising non-stationary signals. *Mech Syst Signal Process* 20(2):282–307
25. Villani C (2009) *Optimal Transport: Old and New*. Springer, Berlin
26. Kolouri S, Rohde GK, Hoffmann H (2018) Sliced Wasserstein distance for learning Gaussian mixture models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3427–3436
27. Rakotomamonjy A, Nadjahi K, Ralaivola L (2024) Federated Wasserstein distance. *International Conference on Learning Representations (ICLR)*
28. Li W, Pang Y (2024) Private Wasserstein distance. *arXiv:2404.06787*
29. Koenker R, Bassett G (1978) Regression quantiles. *Econometrica* 46(1):33–50
30. Virtanen P et al (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods* 17:261–272
31. Pedregosa F et al (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830
32. Smith WA, Randall RB (2015) Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech Syst Signal Process* 64–65:100–131. <https://doi.org/10.1016/j.ymssp.2015.04.021>
33. Lessmeier C, Kimotho JK, Zimmer D, Sextro W (2016) Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In Proceedings of the PHM Society European Conference (Vol. 3, No. 1)
34. Efron B, Tibshirani RJ (1993) *An Introduction to the Bootstrap*. Chapman & Hall, New York
35. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6(2):65–70

Figures

(a) Synthetic, K=3 federations: Spearman $\rho=0.35$ [0.16, 0.54]



(b) CWRU, K=3 federations: Spearman $\rho=0.30$ [0.19, 0.43]

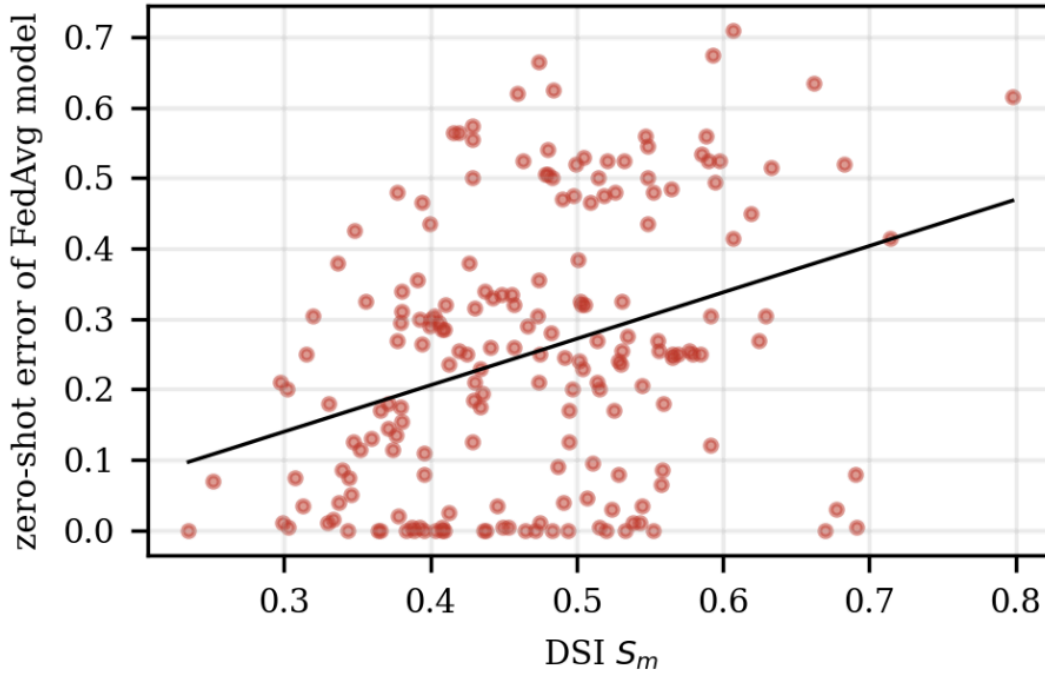


Figure 1

DSI versus zero-shot error of the FedAvg model for $K = 3$ federations, with a linear fit: (a) synthetic machines, (b) CWRU domains.

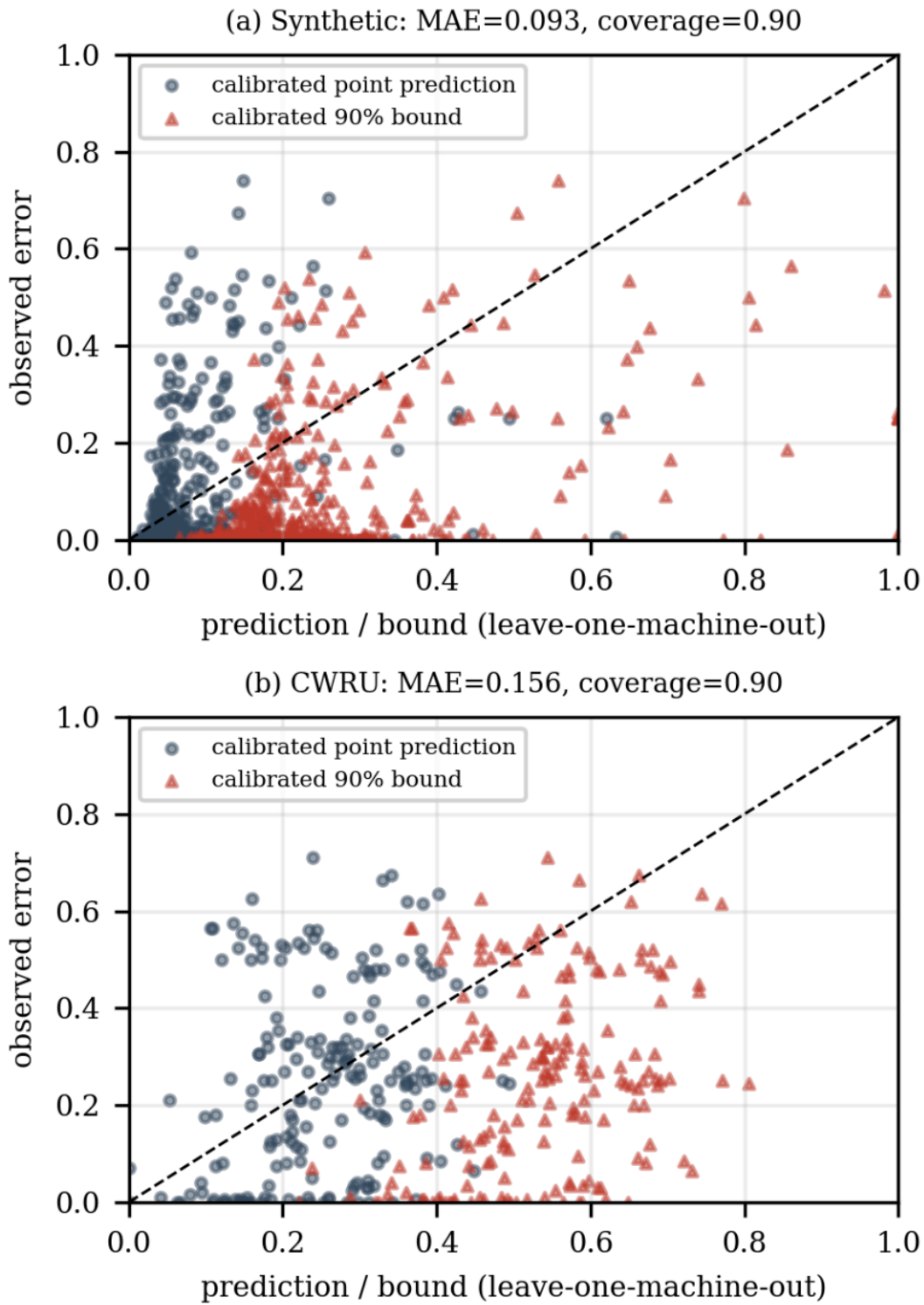


Figure 2

Calibrated prediction and 90% bound versus observed error (leave-one-machine-out, $K = 3$): (a) synthetic, (b) CWRU.

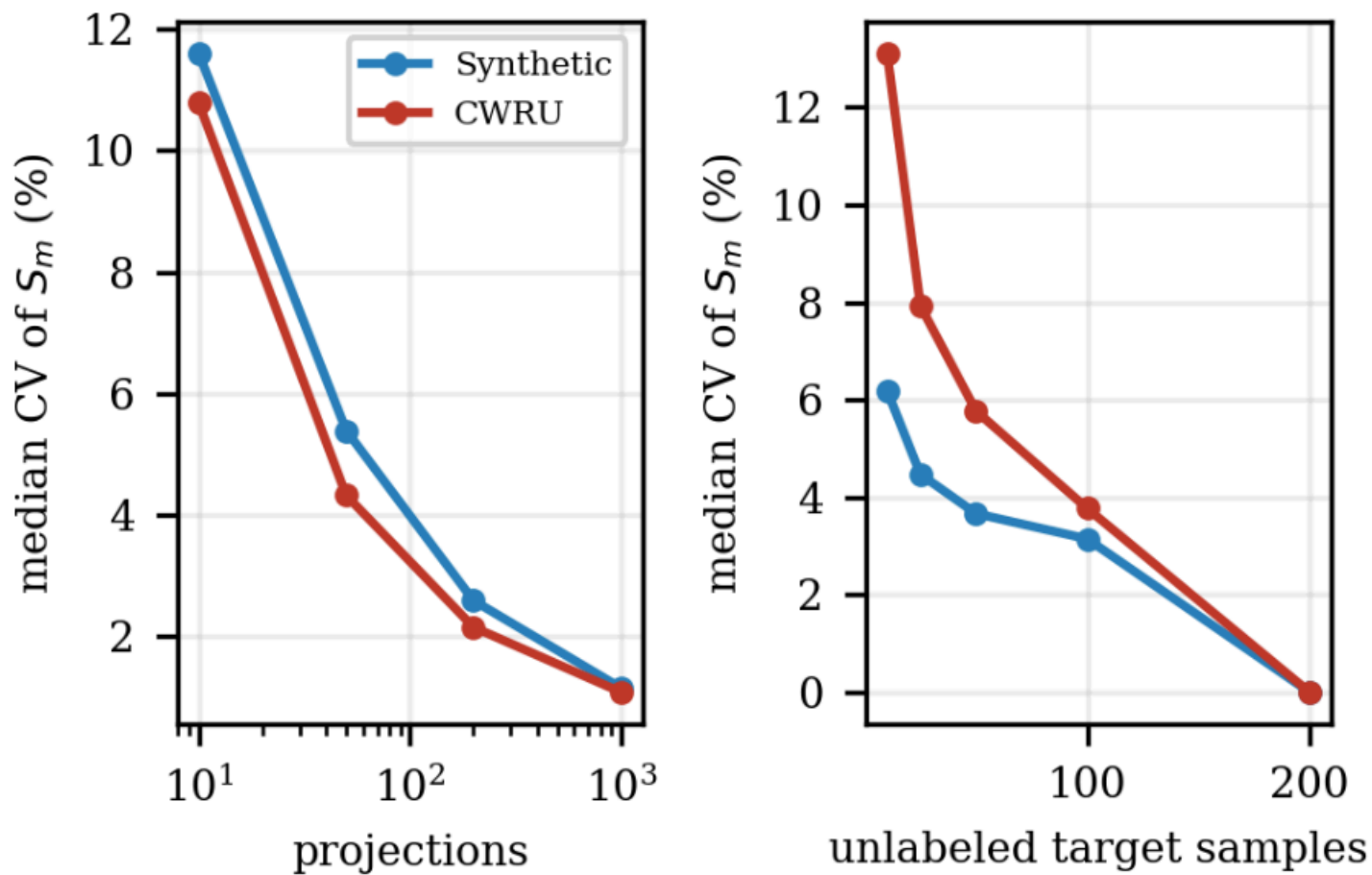


Figure 3

Stability of S_m : median coefficient of variation versus the number of projections (left) and the number of unlabeled target samples (right).

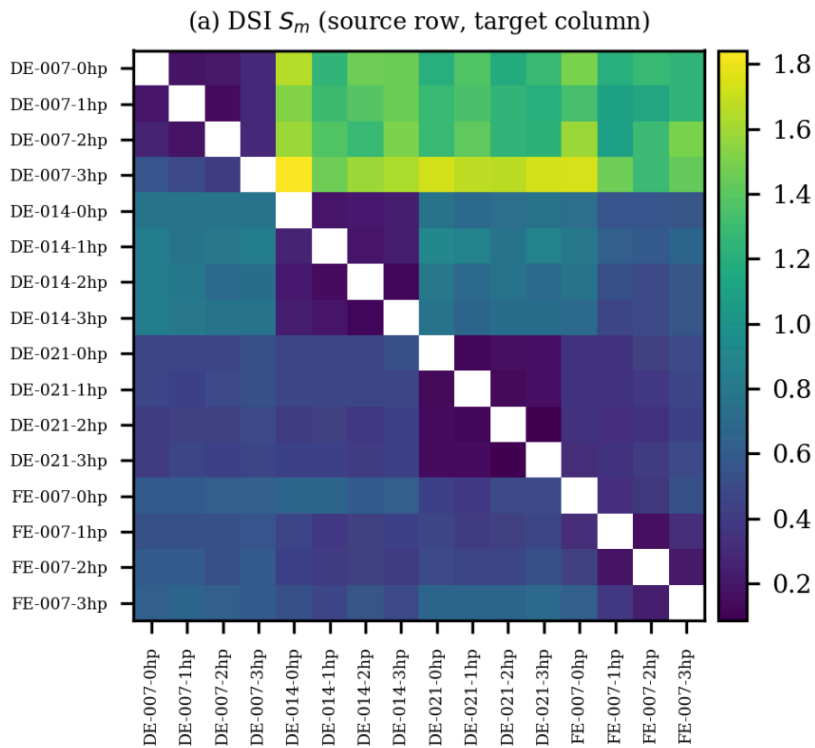
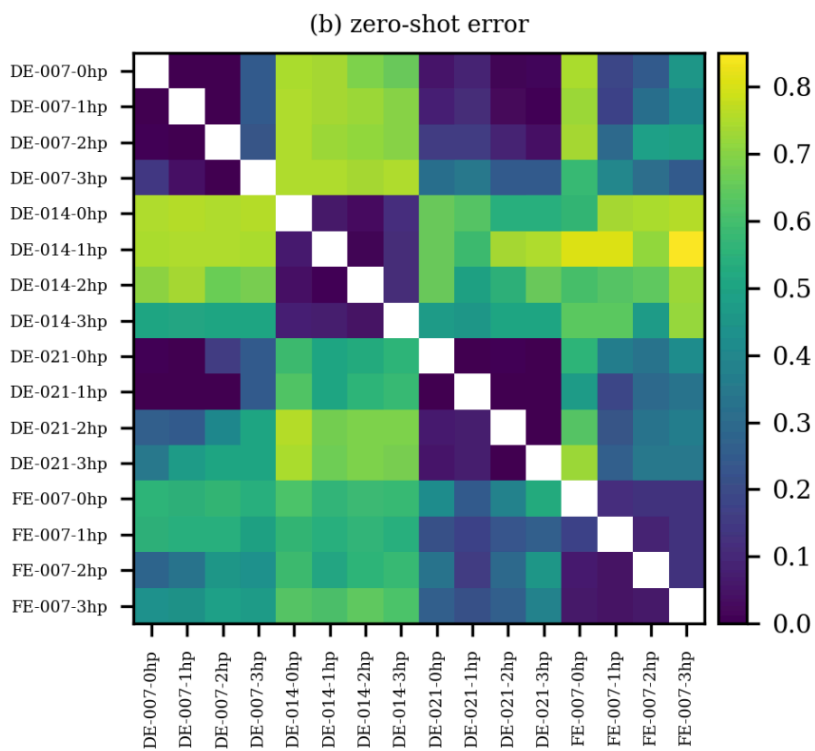


Figure 4

CWRU pairwise results ($K = 1$): (a) DSI and (b) zero-shot error; rows are source domains, columns target domains (sensor–diameter–load).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Appendices.docx](#)