Interpolative Bayesian Formulation to Improve Transfer Learning for Anomaly Detection in Rotating Machinery

Jia Liang[1], Rajesh Gupta[2], Huanyi Shui[3], Devesh Upadhyay[4], and Eric Darve[5]

[1,5] *Institute of Computational and Mathematical Engineering at Stanford University, Stanford, CA, 94305*
jialiang2015@gmail.com
darve@stanford.edu

[2,3,4] *Ford Motor Company, Dearborn, MI, 48126*
rgupta39@ford.com
huanyis@umich.edu
deveshu@gmail.com

ABSTRACT

Anomaly detection in rotating machinery is essential for reliable industrial operations, yet building accurate detectors remains difficult when fault labels in a new domain are scarce. Although transfer anomaly detection has been increasingly studied, most methods do not explicitly exploit the characteristic fault-frequency structure—i.e., the fact that only specific orders/frequency components are strongly diagnostic of emerging faults. Here, we extend our prior key-order transfer framework. In this framework, a key order is a spectral feature-weight vector that upweights diagnostically informative orders and downweights less relevant components when computing the anomaly score, and we adapt it to the realistic regime in which a small (but growing) number of labeled target anomalies becomes available over time. We estimate key orders in both source and target domains and fuse them using uncertainty-aware Bayesian combination as well as robust heuristic rules. Experiments on automotive transmission vibration data from two manufacturing sites show that adaptive fusion consistently outperforms source-only or target-only weighting in label-scarce settings. Overall, these results highlight the value of uncertainty-aware transfer for practical industrial anomaly detection under domain shift.

1. INTRODUCTION

Rotating machinery is essential across industries such as power generation, oil & gas, manufacturing, and automotive. Early fault identification and timely maintenance play a crucial role in industrial operations, safety, and cost-effectiveness (Yadav & Chawla, 2024). Issues in these machines often present as noise, vibration, and harshness (NVH) (Sarrazin et al., 2013), which can trigger cascading effects, leading to unintended vibrations that propagate throughout the system. Monitoring the health of these systems involves analyzing vibration data collected by sensors like accelerometers. It is standard practice to transform these vibration signals into order or frequency domain to facilitate more detailed analysis, as supported by studies such as (Mark, Lee, Patrick, & Coker, 2010; Sharma & Parey, 2017; Mey & Neufeld, 2022; Kim, Yun, & Park, 2021).

Not all frequencies are equally significant for diagnosing faults in rotating machinery. For example, fault diagnosis in rolling element bearings (REBs) relies on detecting characteristic defect frequencies (CDFs), which are determined by the bearing's geometry, rotational speed, and number of rolling elements. In the absence of defects, CDFs should not appear in the vibration spectrum; however, faults in the bearing races, rolling elements, or cage typically induce these frequencies. Furthermore, experts often rely on years of empirical experience and warranty data to identify the most diagnostically relevant frequencies for anomaly detection. When a new plant is established, domain experts often face a shortage of warranty data, making it difficult to construct a robust set of key diagnostic features for the new dataset.

In industrial practice, commissioning a new plant often entails a fundamental challenge for fault detection: labeled fault data are scarce or entirely unavailable. To mitigate this limitation, a growing body of recent work has adopted transfer learning, in which knowledge acquired from an established system (the *source domain*) is leveraged to improve anomaly detection in a newly deployed system (the *target domain*). This literature has expanded rapidly, and transfer-based anomaly detection methods are commonly categorized as *instance-based*

https://doi.org/10.36001/IJPHM.2026.v17i1.4626

transfer, *feature/representation transfer*, or *parameter transfer*. Despite this broad activity, relatively few studies explicitly target transfer anomaly detection in *rotating machinery*. Even when rotating machinery datasets are employed, existing transfer strategies rarely exploit domain-specific diagnostic structure—most notably the *characteristic defect frequencies* that underpin classical vibration-based fault diagnosis. The closest frequency-aware transfer (Wu, Mao, Zhang, Fan, & Zhong, 2024) focuses on invariance/transferability rather than diagnostic predictiveness. Furthermore, much of the prior work assumes a *static* transfer setting, treating the source and target datasets as fixed snapshots, and does not adequately address the practically important scenario in which the target system continues to *accumulate new anomalous data over time*.

To address the aforementioned challenge, our prior work (Liang, Shui, Gupta, Upadhyay, & Darve, 2025) proposed a transfer-learning framework for anomaly detection that operates on frequency-domain signatures. Following (Liang et al., 2025), we refer to the learned spectral feature-weight vector as key orders. Here, "order" follows common automotive NVH usage (frequency normalized by rotational speed), so key orders can be interpreted as weights over frequency components expressed in the order domain, with larger weights indicating higher diagnostic relevance. More generally, key orders define a feature-weight vector that emphasizes a subset of dimensions in the anomaly scoring process. This approach learns key orders from a source dataset by optimizing the f_β score and subsequently transfers them to the target dataset to amplify informative features while attenuating uninformative ones. However, this framework was developed for a setting in which the target dataset is assumed to be anomaly-free, and consequently, it cannot exploit labeled anomalies that may emerge over time in the target domain. It also assumes that an appropriate source–target pairing is known when multiple candidate datasets exist.

In this paper, we extend our prior work (Liang et al., 2025) to the practically relevant setting where labeled anomalies gradually accumulate in the target domain. Building on the same key-order definition and source-domain learning procedure, we introduce new mechanisms to incorporate emerging target anomalies, improve robustness across anomaly-scarce and anomaly-richer regimes, and handle cases where the source–target pairing is unknown. Our contributions are summarized as follows.

Specifically, as anomalous samples become available, we estimate a target-domain key order and fuse it with the source-derived key order to improve anomaly detection performance. A commonly adopted strategy for fusing information from multiple sources is the Bayesian framework. In our setting, its effectiveness relies on accurate estimation of uncertainty in the key order. However, we find that in extremely low-anomaly regimes—where the target domain only contains 1 or 2 anomalies—the uncertainty associated with the target key order is often underestimated, leading to diminished detection accuracy. To address this limitation, we propose a simple yet effective alternative: a weighted linear combination of the source and target key orders. This heuristic demonstrates improved robustness in settings with extremely limited anomalous data. As more labeled anomalies become available in the target domain, the Bayesian approach becomes increasingly reliable and eventually preferable for combining key order information across domains.

As anomalies progressively accumulate in the target dataset, it becomes desirable for the fused key order estimator to place increasing emphasis on the target key order. Standard Bayesian fusion—driven solely by the uncertainties in the source and target key orders—often fails to achieve this shift in weighting. To remedy this limitation, we introduce a standard discounting factor (Ibrahim & Chen, 2000), denoted by Θ, that explicitly reweights the contributions of the source and target key orders. We term the resulting procedure the **Bayesian-Θ method**. Empirical results demonstrate that Bayesian-Θ consistently outperforms the conventional Bayesian approach as the number of target domain anomalies grows.

Another key assumption in the existing framework is the availability of an optimal pairing between source and target datasets when multiple candidates exist in each domain. Identifying this optimal pairing is particularly challenging when the target dataset lacks anomalies. However, once anomalous samples begin to emerge in the target domain, it becomes feasible to infer a more reliable alignment between source and target datasets. To address scenarios where the optimal source-target pairing is unknown, we propose a heuristic selection method based on Kullback–Leibler (KL) divergence. This approach identifies the most relevant source dataset from a pool of candidates by measuring distributional similarity. Empirical results demonstrate that the proposed heuristic achieves high selection accuracy, correctly identifying the optimal source dataset in approximately 90% of cases with as few as two labeled anomalies in the target dataset.

We evaluate our proposed approach using multiple datasets, comparing various key order estimation strategies and anomaly detection algorithms. Our findings suggest that a hybrid strategy—using a linear combination for very small anomaly counts and Bayesian estimation for larger anomaly counts—yields the best performance. The contributions of this work are summarized as follows:

- We extend key-order transfer to a *dynamic target* setting by estimating a target-domain key order as labeled anomalies appear, enabling continual adaptation.
- We diagnose failure modes of uncertainty-based Bayesian fusion in the 1–2 anomaly regime and propose a

simple weighted linear-combination alternative that is more robust under extreme anomaly scarcity.

- We introduce **Bayesian-Θ**, a Bayesian fusion rule that incorporates a standard discounting factor (Ibrahim & Chen, 2000) to explicitly shifts weight toward the target as anomaly evidence accumulates, improving performance in higher-anomaly regimes.
- We propose a KL-divergence heuristic for source selection when the source–target pairing is unknown, and show it reliably identifies the most relevant source with very few labeled target anomalies.

The remainder of this paper is organized as follows. Section 2 discusses related work in transfer learning and anomaly detection. Section 3 introduces our extended transfer learning framework, key order estimation, and the proposed heuristic method. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes the paper and outlines future directions.

2. LITERATURE REVIEW

2.1. Transfer Learning for Anomaly Detection

Anomaly detection aims to identify patterns in data that deviate from expected behavior, which is essential for applications such as fraud detection, network security, and industrial monitoring. Traditional anomaly detection methods often assume that anomalies are rare and significantly different from normal instances. However, these methods can struggle when labeled data are scarce or when the data distribution shifts over time. Transfer learning addresses these limitations by transferring knowledge from a source domain with abundant labeled data to a target domain with limited or no labeled data. This approach is especially valuable in anomaly detection, where labeled anomalies are rare or expensive to obtain. Transfer learning in anomaly detection generally falls into three main categories: instance-based transfer, feature representation transfer, and parameter transfer.

Instance-based transfer learning reuses selected source domain data in the target domain by reweighting instances. Vercruyssen et al. (Vercruyssen, Meert, & Davis, 2017) introduced such an approach for time-series anomaly detection, transferring labeled source instances to an unlabeled target domain using two similarity measures: kernel density estimation (density-based) and k-means clustering (cluster-based). A 1-nearest neighbor classifier with Dynamic Time Warping (DTW) was used for detection. Building on this, Vincent et al. (Vincent, Wannes, & Jesse, 2020) proposed LOCIT, which selects source instances based on localized similarity using first- and second-order statistics, and scores anomalies using a semi-supervised nearest-neighbor method (SSKNNO). Both methods assume similar underlying distributions across domains, making them susceptible to negative transfer when domain shift is significant. To mitigate this risk, our approach employs a KL divergence–based heuristic to identify an optimal source–target pairing in scenarios where such a pairing is not explicitly available.

Feature representation transfer focuses on learning a robust feature representation for the target domain, encoding transferable knowledge into the learned representation to improve target task performance. The goal of feature representation transfer is to reduce feature discrepancies between source and target domains by minimizing the distance between the distributions of latent feature representations. Mao et al. (Mao, Zhang, Tian, & Tang, 2020) propose a robust early fault detection method for bearings using feature transfer learning, aligning normal-state data across domains via a domain-adaptive autoencoder and enabling accurate, low-false-alarm detection without target labels. Xu et al. (Xu, Wang, Zhang, & Li, 2021) propose a Cluster-based Deep Adaptation Network (CDAN) to improve anomaly detection in yarn spinning power consumption by mitigating domain mismatch in transfer learning. Their cluster-based adaptation layer aligns source and target features, outperforming standard methods in real-world scenarios.

Parameter transfer assumes that the source and target tasks share some model parameters or prior distributions of hyperparameters, with the transferred knowledge encoded into these shared parameters or priors. AnoTransfer (Zhang et al., 2022) is an unsupervised KPI anomaly detection framework that improves training efficiency by using transfer learning with parameter transfer, where base models are trained on representative KPIs and fine-tuned for target KPIs using either full or partial parameter transfer based on similarity. Pan et al. (Pan, Bao, & Li, 2023) propose a model-based inductive transfer learning approach using a CNN for structural health monitoring (SHM), where a model trained on one bridge's data is adapted to another by fine-tuning only the last layers. Unlike the above parameter transfer methods, this paper introduces a parameter transfer approach in which the transferred parameters are the weights of orders/frequencies in spectrum analysis. This method is specifically designed for rotating machinery, motivated by the observation that identifying important orders and filtering out irrelevant ones is crucial for effective anomaly detection in such systems.

2.2. Transfer Anomaly Detection in Rotating Machinery

Recent transfer anomaly detection work for rotating machinery (especially rolling bearings) is largely driven by the fact that normal-behavior models break under domain shift (changes in speed/load, operating regimes, sensors, and unit-to-unit variability). A major thread treats the problem as one-class/unsupervised domain adaptation, aiming to transfer healthy knowledge without requiring labeled target faults. Michau & Fink propose

an unsupervised transfer learning framework for one-class anomaly detection that transfers complementary healthy operating conditions to improve coverage under changing regimes (Michau & Fink, 2021). Mao et al. develop domain-adversarial one-class transfer learning explicitly for anomaly detection under distribution gaps (Mao, Wang, Kou, & Liang, 2023). In bearings, domain-adaptation-based anomaly detection has also been proposed to adapt detection across distributions for fault detection (Qin, 2024). Several works focus on online/early fault detection as a transfer-AD problem: for example, DTDA aligns temporal information across working conditions for online early bearing fault detection (Mao, Sun, & Wang, 2021). More recently, continuous test-time domain adaptation (e.g., TAAD) targets evolving operating conditions and enables robust anomaly/fault detection when deployment distributions drift beyond what was seen during training, demonstrated on pump monitoring data (Sun, Ammann, Giannoulakis, & Fink, 2024). Overall, these methods typically do not rely on labeled target anomalies; instead they adapt representations or decision rules using unlabeled (often assumed mostly-normal) target data.

A complementary line addresses the realistic case where only a few target fault/anomaly examples are available and rapid adaptation is needed. Zhang et al. cast bearing anomaly detection as few-shot learning using MAML so that a model can adapt to new bearing conditions with minimal labeled examples (Zhang, Ye, Wang, & Habetler, 2020). Wu et al. propose a few-shot transfer-AD approach combining domain-adversarial learning with contrastive objectives and time–frequency transferability analysis to make few-shot transfer more reliable (Wu et al., 2024). Beyond bearings, similar transfer-AD patterns appear in other rotating assets: Roelofs et al. study cross-turbine transfer for autoencoder-based anomaly detection (fine-tuning full AE vs decoder-only vs threshold-only) (Roelofs, Gück, & Faulstich, 2024), while fleet-based frameworks transfer reconstruction models (e.g., LSTM encoder–decoder) across a fleet when labels are scarce (Wang, Baraldi, Zio, Brown, & Gauthier, 2026).

Most existing transfer anomaly detection methods on rotating machinery do not explicitly leverage characteristic fault-frequency structure (i.e., the fact that specific orders/frequency bands are diagnostic because they align with known fault mechanisms). The closest related approach is Wu et al. (Wu et al., 2024), who introduce a Frequency Importance Metric to identify frequencies that are most transferable—i.e., most consistent between source and target under their hypersphere-matching objective. Their results suggest that lower-frequency bands tend to transfer more reliably across domains. While this line of work uses frequency information explicitly, its goal is fundamentally different from ours: Wu et al. prioritize cross-domain invariance, whereas we aim to transfer diagnostic frequency weights—the orders/frequencies that are most predictive for anomaly detection. Moreover, our prior key-order transfer work (Liang et al., 2025) leveraged fault-frequency intuition, but it did not address the practically important regime where a small number of labeled anomalies become available in the target domain. This paper fills that gap by incorporating characteristic fault-frequency structure into a transfer framework that adapts to emerging target anomalies, enabling more reliable detection under label-scarce, domain-shifted industrial conditions.

2.3. Bayesian Transfer Learning

The Bayesian framework is a statistical approach that uses probability to represent uncertainty in models and incorporates prior knowledge with new evidence to update beliefs. Recent work (Karbalayghareh, Qian, & Dougherty, 2018; Shwartz-Ziv et al., 2022; Xie, Huang, & Dubljevic, 2021; Zhu, Peng, Wang, & Huang, 2023) has begun to incorporate Bayesian methods into transfer learning. In (Karbalayghareh et al., 2018), the authors formulate the optimal bayesian classifiers in the target domain using both the joint prior knowledge and data from the source and target domains. Under their framework, transfer the source domain to the target domain is facilitated by a joint prior probability density function that represents the model parameters for the feature-label distributions in both domains. The modeling of joint prior densities enables a better understanding of the "transferability" between domains.

In (Shwartz-Ziv et al., 2022), instead of relying on initialization from pre-trained models, the authors propose using highly informative Bayesian posteriors from source tasks as priors for downstream tasks. The proposed method reshapes the loss surface of the downstream task, leading to significant performance gains and more data-efficient learning across various classification and segmentation tasks. Transfer Slow Feature Analysis (TSFA) (Xie et al., 2021) is a dynamic transfer learning technique that uses variational Bayesian inference to enhance predictive model performance in industrial processes by dynamically updating weighting functions to quantify transferability from the source domains to the target domain. In (Zhu et al., 2023), the authors present a Bayesian semi-supervised transfer learning framework with active querying for Remaining Useful Life (RUL) prediction across different machines with limited data. This framework integrates transfer learning (TL) and active learning within a Bayesian deep learning structure. One of the key components is leveraging uncertainty quantification from Bayesian Neural Networks (BNNs) with Monte Carlo Dropout Inference to select the most informative training data points. The Bayesian transfer learning methods above are formulated for classification problems, whereas the approach introduced in this work is specifically designed for anomaly detection tasks.

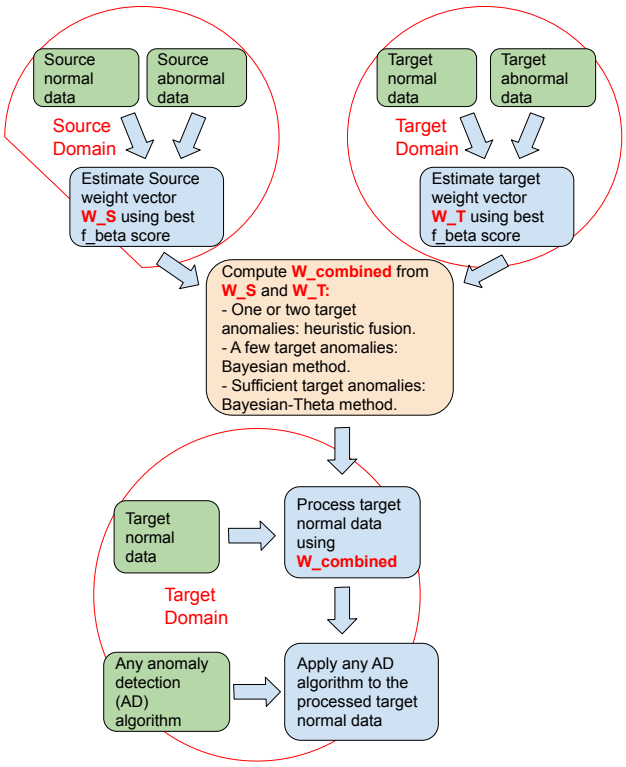

Figure 1. The proposed framework integrates feature importance information from both source and target domains to enhance anomaly detection under limited supervision. In the source domain, feature importance weights W_S are computed based on the best f_β-score for each feature. In the target domain, a corresponding weight vector W_T is estimated using the limited number of available anomalies. These source and target weights are then combined to produce a unified importance vector, W_{combined}. The fusion strategy is selected based on the quantity of target anomalies: a heuristic approach is employed when only fewer than three anomalies are observed, a Bayesian method is used when a few anomalies are available, and the Bayesian-Θ method is adopted when the target domain contains a sufficient number of anomalies. The resulting combined weights are used to reweight the normal target data, after which any unsupervised anomaly detection algorithm can be applied.

3. METHOD

The goal of this work is to combine information from both a related domain, known as the **source domain** ($\mathcal{D}_S$), and the **domain of interest**, known as the **target domain** ($\mathcal{D}_T$), to improve anomaly detection in $\mathcal{D}_T$. Specifically, we assume that the source domain $\mathcal{D}_S$ contains a sufficiently large set of both normal (X_S^N) and anomalous (X_S^A) samples, while the target domain $\mathcal{D}_T$ has access to a sufficient number of normal samples (X_T^N) but only a limited number of anomalous samples (X_T^A). The overall method is illustrated in Figure 1.

3.1. Key Order W

An essential component of fault detection in rotating machinery is the precise identification of critical orders that differentiate normal operation from anomalous behavior. It is standard practice to transform vibration signals from rotating machinery into the order or frequency domain. Deviations in specific orders or frequencies serve as clear indicators of faults, whereas variations in other orders may be attributed to noise. Traditionally, experts determine these critical orders based on extensive experience and warranty claim data. In this paper, we aim to use data-driven methods to determine these critical orders and adopt the term key orders to align with established industrial terminology.

Mathematically, we represent the importance of all features in detecting anomalies using a weight vector $W \in \mathbb{R}^M$, where M is the total number of features. Without loss of generality, we assume that each weight value $W[i]$ lies within the range $[0, 1]$ for all $i \in \{1, 2, \ldots, M\}$. A higher value of $W[i]$ signifies a greater importance of the i-th feature in anomaly detection. For instance, if $W[i] = 1$, the feature is highly significant in identifying anomalies, whereas if $W[i] = 0$, the feature is irrelevant to anomaly detection. In the specific context of anomaly detection in rotating machinery, the input features correspond to different orders. Thus, the weight vector W directly quantifies the importance of each order in detecting anomalies.

There exist multiple strategies for determining the feature weights W. Among these, our prior work (Liang et al., 2025) has shown that leveraging the best f_β-score for each feature provides the most effective weighting scheme for anomaly detection. To illustrate this process, we use the source dataset X_S as an example. For each feature in the source domain, we consider it independently. Let $X_S[:, i]$ denote the values of the i-th feature across all samples in X_S. An anomaly detection algorithm is then trained using only this feature, yielding a corresponding vector of outlier scores, denoted by $O_S[:, i]$. In high-dimensional settings, training an outlier detector for each feature may be computationally prohibitive. As a more efficient alternative, simple statistical measures—such as the standard deviation—can be employed to generate

outlier scores. Moreover, in scenarios where anomalies are right-skewed (i.e., associated with unusually high values), the raw feature values $X_S[:, i]$ themselves can serve directly as outlier scores; in this work, we use $X_S[:, i]$ as the anomaly scores. Once the outlier scores $O_S[:, i]$ are obtained, the performance of the i-th feature in identifying anomalies is evaluated by computing its best f_β-score. Concretely, for a given threshold τ, we binarize the scores by predicting an anomaly when $O_S[j, i] \geq \tau$, compute the corresponding $\mathrm{precision}(\tau)$ and $\mathrm{recall}(\tau)$ with respect to the ground-truth labels, and define

$$f_\beta(\tau) = \frac{(1+\beta^2)\,\mathrm{precision}(\tau)\,\mathrm{recall}(\tau)}{\beta^2\,\mathrm{precision}(\tau) + \mathrm{recall}(\tau)}. \quad (1)$$

We then take the best f_β score over all candidate thresholds τ (e.g., thresholds given by the unique values of $O_S[:, i]$). This score is used as the feature's importance weight, denoted $W_S[i]$. Formally, for each $i \in \{1, 2, \ldots, M\}$, we define:

$$W_S[i] = \max_\tau f_\beta(O_S[:, i], \text{label}), \quad (2)$$

where τ is a decision threshold.

Similarly, we can calculate the feature weight vector W_T from the target domain:

$$W_T[i] = \max_\tau f_\beta(O_T[:, i], \text{label}), \quad (3)$$

where $O_T[:, i]$ denotes the outlier scores for the i-th feature $X_T[:, i]$ in the target dataset.

Since the number of anomalies in the target dataset is limited, the estimated key order W_T may not be reliable. To address this, we incorporate key order information from the source dataset, W_S, to supplement W_T. Mathematically, we aim to compute a new weight W_{combined} by combining information from both W_S and W_T using a function $\mathbf{f}$, defined as:

$$W_{\text{combined}} = \mathbf{f}(W_S, W_T). \quad (4)$$

Once W_{combined} is obtained, we preprocess the target data X_T according to its corresponding weight vector:

$$X_T^{\text{weighted}} = X_T\,\mathrm{diag}(W_{\text{combined}}). \quad (5)$$

Here, each feature i in X_T is scaled by the corresponding weight $W_{\text{combined}}[i]$, effectively adjusting its significance based on the combined information from the source and target datasets. After preprocessing, any unsupervised anomaly detection method can be applied to X_T^{weighted} for improved anomaly identification.

In cases where the target dataset contains only normal data, W_T cannot be estimated, which aligns with the setting of our previous paper (Liang et al., 2025). In this scenario, we directly set $W_{\text{combined}} = W_S$. However, in the non-trivial case where the target dataset includes anomalous samples, allowing us to estimate W_T, we apply different methods to integrate the source and target key orders.

3.2. Computing W_{combined} Using Bayesian Method

Bayesian analysis is a well-established methodology for integrating information from multiple sources, particularly for updating prior beliefs based on newly acquired evidence. It provides a systematic framework for combining prior knowledge—representing initial assumptions—with observed data, yielding more informed and reliable decisions.

In the context of our problem, we aim to determine an optimal key order, denoted as W_{combined}, by integrating information from both the source key order, W_S, and the target key order, W_T. From a Bayesian perspective, the source key order, W_S, serves as the *prior belief* regarding the optimal key order for the target dataset. This assumption is justified by the fact that, in the absence of specific information about the target key order, the source dataset provides the most reasonable initial estimate. As anomalies begin to emerge in the target dataset, we can estimate a new target key order, W_T, which represents the *new evidence*. Bayesian inference then offers a principled approach to updating our prior belief, W_S, by incorporating this newly obtained evidence, W_T, resulting in a refined estimate of W_{combined}. This approach ensures that our model dynamically adapts as more information becomes available, leading to improved accuracy in key order estimation.

Before presenting the Bayesian formulation for combining key orders, we first recall *Bayes' Theorem*, which provides a principled approach for updating prior beliefs based on new evidence. It is expressed as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (6)$$

where:

- $P(A)$ represents the *prior probability*, which quantifies our initial belief about event A before incorporating any new data. In the context of this work, the source key order distribution serves as the prior distribution.
- $P(B|A)$ denotes the *likelihood*, the probability of observing evidence B given that A is true. Here, it corresponds to the distribution of the target key order conditioned on the source key order.
- $P(B)$ is the *marginal probability*, which ensures proper normalization by considering all possible occurrences of B.
- $P(A|B)$ is the *posterior probability*, the updated belief about A after incorporating the evidence B. In this study, it corresponds to the updated key order distribution obtained by combining the source and target key orders via Bayes' rule.

Now we introduce the notations $\mathbf{W_S}$ and $\mathbf{W_T}$ as the probability distributions of the source key order W_S and the target key order W_T respectively. To apply Bayes' Theorem, we regard $\mathbf{W_S}$, the probability distribution of source key order, as the prior probability distribution for the optimal key order W_{combined}. To simplify our analysis, we assume that $\mathbf{W_S}$ follows a multivariate normal distribution:

$$\mathbf{W_S} \sim \mathcal{N}(\mu_S, \Sigma_S), \tag{7}$$

where μ_S represents the mean of the prior distribution and Σ_S denotes its covariance. This choice allows for a closed-form Bayesian update, although alternative prior and likelihood models could be explored in future work. Specifically, $\mu_S = W_S$, which is calculated from Equation (2). Thus, Equation (7) can be rewritten as:

$$\mathbf{W_S} \sim \mathcal{N}(W_S, \Sigma_S), \tag{8}$$

where Σ_S is estimated using the bootstrap method. In a high-level, the bootstrap algorithm estimates the variance of the key order by repeatedly resampling the normal and anomalous data and computing the best f_β score for each feature in every bootstrap iteration. The variance of these best f_β scores across the bootstrap iterations is then used to construct a diagonal covariance matrix, which quantifies the uncertainty associated with each key order. Moreover, $\Sigma_S = \text{BOOTSTRAPPING}(n_{\text{boot}}, X_S^{nor}, X_S^{abn}, \beta)$, where n_{boot} is the number of bootstrap iterations, X_S^{nor} represents the normal data in the source domain, X_S^{abn} denotes the anomalies in the source domain, and β is the hyperparameter for the best f_β score. Specifically, in the bootstrap algorithm, we assume that there are a total of M features; the normal data X^{nor} contain N^{nor} samples; and the abnormal data X^{abn} contain N^{abn} samples. In each bootstrap iteration, we sample N^{nor} normal samples from X^{nor} with replacement. Similarly, we sample N^{abn} abnormal samples from X^{abn} with replacement. Using the resampled normal data for the i-th iteration, X_i^{nor}, and the resampled abnormal data for the i-th iteration, X_i^{abn}, we estimate outlier scores for the combined resampled dataset, denoted O_i. This estimation can be based on simple statistics (e.g., z-scores) or obtained from a simple anomaly detection model. Combining the known labels with the outlier scores for the j-th feature, $O_i[:, j]$, we then compute the best (maximum) f_β score over all candidate thresholds. This score is used as the current weight estimate for the i-th iteration and the j-th feature, i.e., $W[i, j]$. After completing all n_{boot} bootstrap iterations for all M features, we obtain the full weight matrix $W \in \mathbb{R}^{n_{boot} \times M}$. The variance of the j-th feature is then $\Sigma[j, j] = \text{var}(W[:, j])$. The overall procedure is summarized in **Algorithm 1**. Note that we further assume that Σ_S is a diagonal matrix, since the weight of each feature in W_S is calculated independently. We emphasize that bootstrapping is used here only to quantify the finite-sample uncertainty of the key-order estimator, while the point estimates W_S and W_T are still computed using all available labeled normal and anomalous samples in each domain. Because each component of the key order is obtained via a score-based optimization (maximizing an f_β-based criterion) rather than from a parametric likelihood, closed-form expressions for the variance are generally unavailable; the bootstrap therefore provides a distribution-free approximation of the sampling variability by repeatedly resampling the labeled normal and anomalous sets and recomputing the key order. The resulting empirical per-feature variances across bootstrap replicates form the diagonal covariance matrices Σ_S and Σ_T used in the Bayesian fusion update.

Algorithm 1 bootstrapping methods for variance

procedure BOOTSTRAPPING(n_{boot}, X^{nor}, X^{abn}, β)
  $M := len(X^{nor}[0, :])$
  $N^{nor} := len(X^{nor})$
  $N^{abn} := len(X^{abn})$
  label = concatenate $([0] * N^{nor}, [1] * N^{abn})$
  Randomly initialize W as $\mathbb{R}^{n_{\text{boot}} \times M}$
  Initialize Σ as $I^{M \times M}$
  for $i = 1, 2, 3, ..., n_{\text{boot}}$: **do**
    X_i^{nor} := Sampling N^{nor} normal samples from X^{nor} with replacement
    X_i^{abn} := Sampling N^{abn} anomalous samples from X^{abn} with replacement
    estimate O_i from $[X_i^{nor}; X_i^{abn}]$
    for $j = 1, 2, 3, ..., M$: **do**
      $W[i, j] := \max_\tau f_\beta(O_i[:, j], \beta, label)$
    end for
  end for
  for $j = 1, 2, 3, ..., M$: **do**
    $\Sigma[j, j] = var(W[:, j])$
  end for
  return Σ
end procedure

We further treat $\mathbf{W_T}|\mathbf{W_S} = \mathbf{w_s}$ as the likelihood, which also follows a multivariate Gaussian distribution:

$$\mathbf{W_T}|\mathbf{W_S} = \mathbf{w_s} \sim \mathcal{N}(\mathbf{w_s}, \Sigma_T). \tag{9}$$

Similarly, we set $\Sigma_T = \text{BOOTSTRAPPING}(n_{\text{boot}}, X_T^{nor}, X_T^{abn}, \beta)$, where n_{boot} is the number of bootstrap iterations, X_T^{nor} denotes the normal data in the target domain, and X_T^{abn} denotes the anomalous data in the target domain. Σ_T is also assumed to be a diagonal matrix.

Thus, the posterior distribution of the optimal key order, W_{combined}, is $\mathbf{W_S}|\mathbf{W_T} = W_T$, where W_T is calculated from Equation (3) , and is of the form:

$$\mathbf{W_S} \mid \mathbf{W_T} = W_T \sim \mathcal{N}\Big((\Sigma_T^{-1} + \Sigma_S^{-1})^{-1} (\Sigma_T^{-1} W_T + \Sigma_S^{-1} W_S),\ (\Sigma_T^{-1} + \Sigma_S^{-1})^{-1}\Big). \tag{10}$$

Specifically, W_{combined} is the expectation of $\mathbf{W_S}|\mathbf{W_T} =$

W_T, i.e.,

$$\begin{aligned} W_{\text{combined}} &= \mathbb{E}[\mathbf{W_S} \mid \mathbf{W_T} = W_T] \\ &= (\Sigma_T^{-1} + \Sigma_S^{-1})^{-1}(\Sigma_T^{-1} W_T + \Sigma_S^{-1} W_S). \end{aligned} \tag{11}$$

After applying some basic linear algebra, the above equation can be rewritten as:

$$\begin{aligned} W_{\text{combined}} &= \mathbb{E}[\mathbf{W_S} \mid \mathbf{W_T} = W_T] \\ &= W_S + K\,(W_T - W_S). \end{aligned} \tag{12}$$

where K is the gain matrix and can be represented as following:

$$K = \Sigma_S(\Sigma_S + \Sigma_T)^{-1}. \tag{13}$$

It is important to note that the results presented in Equations (11) to (13) correspond to the standard Kalman filter update equations. Moreover, the gain matrix K is diagonal, as both Σ_S and Σ_T are assumed to be diagonal covariance matrices. Thus, for $i = 1, 2, 3, \ldots, M$,

$$K[i,i] = \frac{\Sigma_S[i,i]}{(\Sigma_T[i,i] + \Sigma_S[i,i])}. \tag{14}$$

Dividing both the numerator and denominator of Equation (14) by $\Sigma_S[i,i]$, we obtain:

$$K[i,i] = \frac{1}{(\Sigma_T[i,i]/\Sigma_S[i,i] + 1)}. \tag{15}$$

From Equation (15), we can see that the exact values of $\Sigma_S[i,i]$ and $\Sigma_T[i,i]$ may not be critical. It is the ratio between $\Sigma_S[i,i]$ and $\Sigma_T[i,i]$ that ultimately determines the gain.

In general, the gain matrix K computed here is also referred to as the Kalman gain matrix in the Kalman filter literature. It represents the update step between the source and target domains: a small value of K indicates that the posterior remains closer to the source domain, whereas a large value implies greater influence from the target domain. From Equation (15), we observe that a small ratio of $\Sigma_T[i,i]$ to $\Sigma_S[i,i]$ leads to a larger gain $K[i,i]$, meaning the i-th component of the posterior relies more heavily on the target information, reflecting higher confidence in the target relative to the source.

3.2.1. Discounting Factor Θ

Notice that depending on how strongly we believe in the validity of the source data (prior information) when combining the source and target key orders, we can discount the importance of the source key order by modifying its variance when calculating the value of the Kalman gain K in Equation (13). This approach is theoretically grounded in the Power Prior framework (Ibrahim & Chen, 2000), where historical or source information is down-weighted to prevent it from overwhelming the current target data.

Specifically, we can adjust the variance of the source key order by dividing it element-wise by a discounting factor matrix Θ. We denote the modified source variance as $\Sigma_{S/\Theta}$, where

$$\Sigma_{S/\Theta} := \Sigma_S \odot \Theta^{-1}. \tag{16}$$

Equation (16)[1] represents the element-wise product between Σ_S and Θ^{-1}. As noted above, Σ_S and Σ_T are both diagonal matrices. Thus, we only need to estimate the diagonal elements of the Θ matrix. If we possess domain knowledge about both the source and target datasets, we can set the value of Θ accordingly. However, we can also estimate Θ in a data-driven manner by maximizing the likelihood. The details for estimating Θ in a data-driven way are shown in the Section 5.1. Finally, we denote the estimation of K using the discounting factor Θ as $K_{Bayesian,\Theta}$, and it is calculated as

$$K_{Bayesian,\Theta} = \Sigma_{S/\Theta}(\Sigma_{S/\Theta} + \Sigma_T)^{-1}. \tag{17}$$

3.3. Compute W_{combined} Using Heuristic Method

The effectiveness of the Bayesian formulation depends on accurate estimation of Σ_S and Σ_T. Specifically, both source and target key orders' variances are estimated using the bootstrap method shown in **Algorithm 1**. In particular, for the target distribution, the estimation of Σ_T can become inaccurate when fewer than three anomalous samples are available in the target dataset. In this special situation, we propose combining W_S and W_T using a simple heuristic method:

$$W_{\text{combined}} = W_S + \max(n/K_0, 1)(W_T - W_S), \tag{18}$$

where n is the number of anomalies in the target dataset used to estimate the key order W_T, and K_0 is a hyperparameter.

Note that this equation has the same form as Equation (12), and we can rewrite it using the Kalman gain matrix formulation:

$$W_{\text{combined}} = W_S + K_{\text{linear}}(W_T - W_S), \tag{19}$$

where

$$K_{\text{linear}} = \max(n/K_0, 1). \tag{20}$$

3.4. Overall Method Summary

The overall method can be summarized as follows (also illustrated in Figure 1):

1. **Compute weights:** Calculate the source and target weights W_S and W_T using Equation (2) and Equation (3), respectively.
2. **Combine weights:**
   - If the number of anomalies in the target dataset is very small (e.g., only 1 or 2), compute the

[1]Note that Θ is a diagonal matrix, and Θ^{-1} is the element-wise inverse on the diagonal element of Θ

combined weights as:

$$W_{\text{combined}} = W_S + K_{\text{Linear}}(W_T - W_S)$$

- With a few anomalies (e.g. greater than or equal to 3 but less than 8 in our specific setting), use the Bayesian combination:

$$W_{\text{combined}} = W_S + K_{\text{Bayesian}}(W_T - W_S)$$

- With a sufficient number of anomalies (e.g. greater than or equal to 8 in our specific setting.), use the Bayesian-Θ combination:

$$W_{\text{combined}} = W_S + K_{\text{Bayesian},\Theta}(W_T - W_S)$$

where K_{Linear}, K_{Bayesian}, and $K_{\text{Bayesian},\Theta}$ are defined in Equation (20), Equation (13), and Equation (17), respectively.

3. **Preprocess data:** Apply the combined weights to the target normal data:

$$X_T^{\text{weighted}} = X_T \ \text{diag}(W_{\text{combined}})$$

4. **Train model:** Train an unsupervised anomaly detection algorithm on the preprocessed data X_T^{weighted}.

4. EXPERIMENTAL RESULTS

4.1. Data

To evaluate our proposed method, we use vibration data collected during standard end-of-line (EOL) tests for automotive transmissions. Accelerometers mounted on the transmission transfer case record vibration signals under a fixed test protocol. We analyze data from two manufacturing sites, denoted Plant A and Plant B. Plant A corresponds to earlier transmission designs, while Plant B represents more recent models.

The data are organized into multiple datasets defined by the transmission model variant and the EOL test type. Plant A contains 16 datasets formed by four transmission models (trans_a, trans_b, trans_c, trans_d) crossed with four tests (test_1–test_4), i.e., $4 \times 4 = 16$. Plant B is organized analogously, with four (different) transmission models (trans_I, trans_II, trans_III, trans_IV) and the same four tests (test_1–test_4), again yielding $4 \times 4 = 16$ datasets. Note that the Plant A models (lowercase letters) and Plant B models (Roman numerals) are distinct designs, although the test protocol is shared across plants. Each dataset is split into training and test sets; sample counts are provided in Tables 1 and 2.

4.2. Key Order

For each target dataset, we assume the existence of a corresponding source dataset. The key order derived from the source dataset is denoted as W^{src}. In the target dataset, we assume the presence of n anomalous samples, where n ranges from 1 to 20 in our benchmark evaluations. We constrain $n < 20$ to align with our problem setting, which assumes a very limited number of anomalies in the target dataset.

Using Equation (3), we compute the key order for the target dataset. This estimation assumes access to 3000 normal samples and n anomalous samples in the target training set. Specifically, we denote the estimated key order for the target dataset with n anomalies as W_n^{tgt}. Additionally, we employ **Algorithm 1** to estimate the variance of W^{src} and W_n^{tgt}.

We integrate information from both W^{src} and W_n^{tgt} to obtain a refined key order for the target dataset. This is achieved through Equation (12), which combines W^{src} and W_n^{tgt}. The resulting key order is denoted as $W_{n,K_{\text{method}}}^{\text{combined}}$, where the parameter K is defined in Equation (12). Specifically, we define several variations of W^{combined} based on different formulations of K:

- $W_{n,K_{\text{Bayesian}}}^{\text{combined}}$: Here, K follows the standard Bayesian formulation, where K_{Bayesian} is determined by Equation (13). Here, n is the number of anomalies used to estimate W_n^{tgt}. Note that $W_{n,K_{\text{Bayesian}}}^{\text{combined}}$ should be used when there are a few anomalies in the target data (i.e., $n \geq 3$) according to our method.

- $W_{n,K_{\text{Bayesian},\Theta}}^{\text{combined}}$: A modified Bayesian approach that incorporates a discounting factor matrix Θ to adjust the variance of the source dataset in estimating K. The value of $K_{\text{Bayesian},\Theta}$ is determined in Equation (17). Again, n is the number of anomalies used to estimate W_n^{tgt}. Note that $W_{n,K_{\text{Bayesian},\Theta}}^{\text{combined}}$ should be used when there is a sufficient number of anomalies, i.e., n ≥ 8.

- $W_{n,K_{\text{Linear}}}^{\text{combined}}$: A heuristic method where K is estimated using a linear approach, as given in Equation (20). Note that n is the number of anomalies assumed to be known in the target training dataset. $W_{n,K_{\text{Linear}}}^{\text{combined}}$ should be used when n is extremely small (i.e., $n = 1$ or 2) according to our proposed method.

- $W_{n,K_{\text{Geometric}}}^{\text{combined}}$: Another heuristic approach for cases where n is extremely small, in which K is estimated geometrically as follows:

$$K_{\text{Geometric}} = \frac{1/N_{\text{src,anomaly}}}{1/n + 1/N_{\text{src,anomaly}}}, \qquad (21)$$

where $N_{\text{src,anomaly}}$ represents the number of anomalies in the source dataset, and n is the number of anomalies assumed to be known in the target training dataset.

Finally, we introduce W^{gt} (the ground truth key order), which is computed by Equation (3) using all normal and anomalous samples in the target test set. This key order serves as a reference representing the best achievable performance.

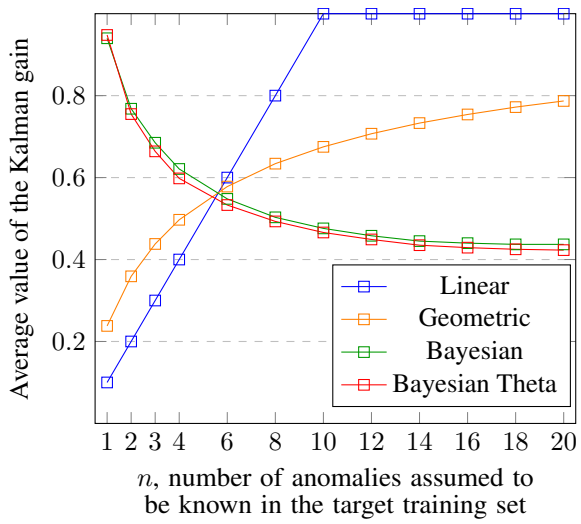

Figure 2. Average value of the Kalman gain across all orders vs number of anomalies assumed to be known in the target dataset.

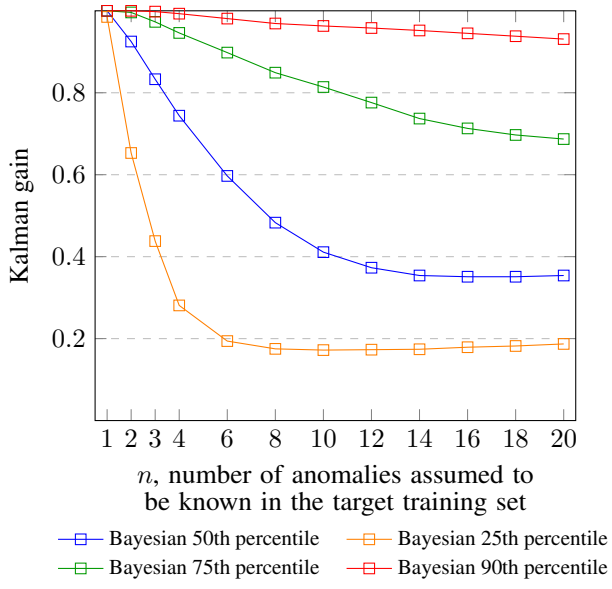

Figure 3. Kalman gain percentiles across all orders vs. number of anomalies in the target.

4.3. Variance Estimation and Kalman Gain for Cases With Very Few Observed Anomalies

Note that both $W^{\text{combined}}_{n,K_{\text{Bayesian}}}$ and $W^{\text{combined}}_{n,K_{\text{Bayesian},\Theta}}$ rely on **Algorithm 1** to estimate the variance of the source key order, Σ_{src}, and the variance of the target key order (when n anomalies are in the target training set), $\Sigma_{\text{tgt},n}$. As defined in Equations (12) and (15), a larger value of $K[i,i]$ indicates greater reliance on the target key order W^{tgt}_n for the i-th feature when combining source and target key orders, whereas a smaller value indicates greater reliance on the source. Intuitively, when the number of anomalies n used to estimate the target key order is small, our confidence in W^{tgt}_n should be low. Consequently, we would expect the corresponding weights $K[i,i]$ to also be small in this regime, as reflected in Equation (13). However, we observe the opposite: when n is small, the average value of $K[i,i]$ remains high for both $W^{\text{combined}}_{n,K_{\text{Bayesian}}}$ and $W^{\text{combined}}_{n,K_{\text{Bayesian},\Theta}}$, as shown in Figure 2. In contrast, the weighting schemes used in $W^{\text{combined}}_{n,K_{\text{Linear}}}$ and $W^{\text{combined}}_{n,K_{\text{Geometric}}}$ behave as expected: the average $K[i,i]$ remains small when n is small, reflecting lower confidence in the target key order.

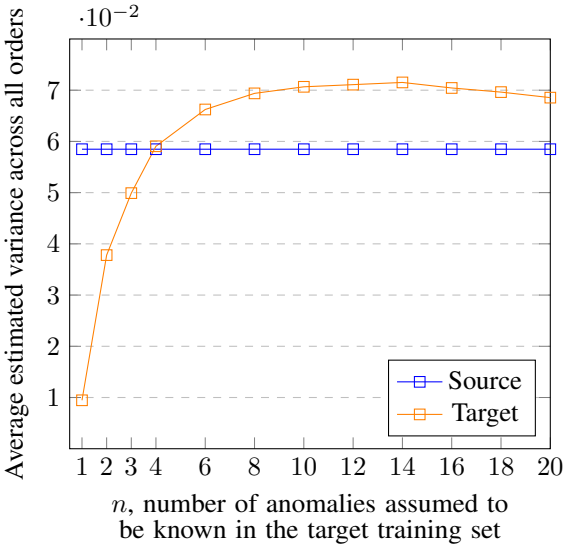

Figure 4. Average variance vs number of anomalies assumed to be known in the target dataset.

A closer analysis of the variances of W^{src} and W^{tgt}_n, denoted by Σ_{src} and $\Sigma_{\text{tgt},n}$ respectively, explains why the average values of K_{Bayesian} and $K_{\text{Bayesian},\Theta}$ tend to be large when n is extremely small. Specifically, when n is extremely small, the bootstrap-estimated variance $\Sigma_{\text{tgt},n}$ is underestimated, whereas Σ_{src}, also obtained via bootstrap, remains stable, as shown in Figure 4. This effect arises from the bootstrap procedure in **Algorithm 1**, where sampling (with replacement) is performed over the anomaly set. For instance, if the dataset contains only one anomaly, each bootstrap iteration will always select the same sample, leading to zero variance in the anomaly component. The only source of variation in the

key order then stems from resampling the 3000 normal samples. As n increases, variability in the anomaly subset grows—since different anomalies can be sampled across bootstrap iterations—resulting in a more accurate estimate of $\Sigma_{\text{tgt},n}$. Thus, the artificially low variance when n is extremely small leads to inflated values of K_{Bayesian} and $K_{\text{Bayesian},\Theta}$.

As shown in Figure 4, when n—the number of anomalies in the target training set—is very small (i.e., $n \leq 3$), the average estimated variance across target key orders increases rapidly with n. Beyond this point, the variance begins to plateau, indicating stabilization of the estimates. Interestingly, the estimated average variance even shows a slight decrease at the upper end of the range (i.e., n = 14 to 20). A similar trend is observed in Figure 2 for the average values of the Bayesian K-step (Kalman gain) parameters, K_{Bayesian} and $K_{\text{Bayesian},\Theta}$, which exhibit a sharp initial decrease when $n \leq 3$, but begin to stabilize as n increases beyond 3. This behavior suggests that when the number of anomalies in the target dataset is extremely small (i.e., n is 1 or 2), the bootstrap-based variance estimates are unreliable. As n increases, these estimates become more accurate. Consequently, in the extremely low n regime ($n \leq 3$), it is preferable to use the heuristic method K_{linear}, while the Bayesian-based methods K_{Bayesian} and $K_{\text{Bayesian},\Theta}$ become more appropriate once $n > 3$, as supported by our proposed framework.

However, it is important to note that the values in Figure 2 and Figure 4 represent averages across all orders (we have 256 orders in the order spectrum). In particular, Figure 2 may falsely suggest that the Kalman gains K_{Bayesian} and $K_{\text{Bayesian},\Theta}$ decrease consistently (for all orders) as n—the number of anomalies assumed to be known in the target dataset—increases. While this trend holds on average, focusing solely on the mean value of K does not provide a complete picture. To capture the variability across all orders, Figure 3 also reports the 25th, 50th (median), 75th, and 90th percentiles of K_{Bayesian} as n varies. Interestingly, although the average K value decreases rapidly with increasing n, the upper percentiles—particularly the 75th and 90th—remain relatively high. This indicates that a subset of key order $W^{\text{combined}}_{n,K_{\text{Bayesian}}}$ are always derived primarily from the target key order, regardless of the value of n. Given the superior performance observed in subsequent experimental results, this behavior suggests that $W^{\text{combined}}_{n,K_{\text{Bayesian}}}$ is capable of identifying and prioritizing the most informative orders/frequencies from the target when combining source and target information, especially when n is large. A similar conclusion holds for $W^{\text{combined}}_{n,K_{\text{Bayesian},\Theta}}$.

Taken together, these plots offer practical guidance for selecting an appropriate strategy to combine key orders from the source and target domains. When the number of known target anomalies n is extremely small (i.e., $n \leq 3$), both the variance and the average Kalman gain K for the Bayesian and Bayesian-Θ methods exhibit rapid changes. This indicates that, in this extremely low-n region, heuristic approaches such as $W^{\text{combined}}_{n,K_{\text{Geometric}}}$ or $W^{\text{combined}}_{n,K_{\text{Linear}}}$ are preferable to $W^{\text{combined}}_{n,K_{\text{Bayesian}}}$ or $W^{\text{combined}}_{n,K_{\text{Bayesian},\Theta}}$. As n increases, both the variance and Kalman gain estimates begin to stabilize, making the Bayesian-based methods more appropriate. This decision-making process is analogous to the elbow method commonly used to determine the optimal number of clusters in K-means clustering.

4.4. Experimental Setup

4.4.1. Source Data: Plant A; Target Data: Plant A

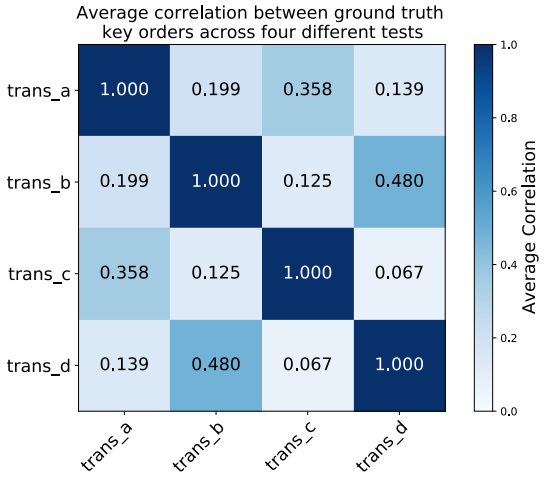

Figure 5. Average correlation across all four tests between different ground-truth key orders in Plant A.

In this experimental setup, we treat each of the sixteen datasets in Plant A as the target dataset in turn. For each target dataset, we use one of the datasets from Plant A as the corresponding source dataset. Specifically, transmission a is used as the source data for transmission c, and transmission c is used as the source data for transmission a. The pairing of transmissions a and c is due to the high correlation between their key orders. A similar pairing is applied to transmission b and transmission d. The complete list of pairings between source and target datasets is shown in Table 1. The average correlation between key orders for all transmissions can be observed in Figure 5.

4.4.2. Source Data: Plant A; Target Data: Plant B

In this experimental setup, we treat each of the sixteen datasets in Plant B as the target dataset in turn. For each target dataset, we use one of the datasets from Plant A as the corresponding source dataset. Specifically, transmission_a is used as the source data for transmission IV, and transmission_b is used as the source data for transmission I, II, and III. The pairing between the source and target

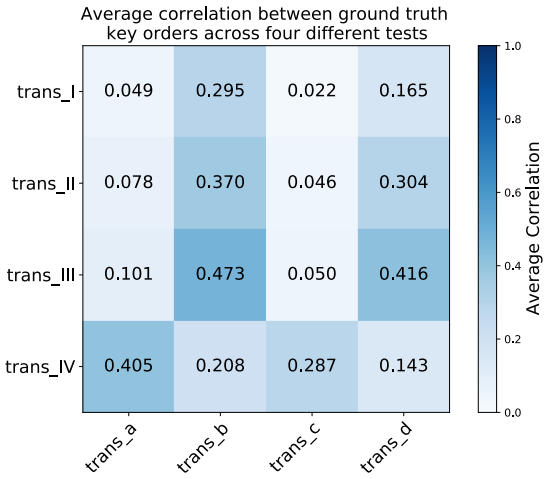

Figure 6. Average correlation across all four tests between ground-truth key orders in Plant B (target) and those in Plant A (source).

datasets is based on the high correlation between their key orders. The complete list of the source-target pairings is shown in Table 2. The average correlation between key orders for all transmissions can be observed in Figure 6.

4.4.3. Experimental Procedure

We consider 32 paired source–target transfer settings: (i) 16 pairs where both source and target come from Plant A, and (ii) 16 pairs where the source comes from Plant A and the target comes from Plant B. For each pair, we assume the target training set contains n *known* anomalous samples. We sweep

$$n \in \{1, 2, 3, 4, 6, 8, 10, 12, 14, 16, 18, 20\}.$$

For every source–target pair and each value of n, we compute the seven key orders defined in Section 4.2: W^{src}, W_n^{tgt}, $W_{n,K_{\text{Bayesian}}}^{\text{combined}}$, $W_{n,K_{\text{Bayesian},\Theta}}^{\text{combined}}$, $W_{n,K_{\text{Linear}}}^{\text{combined}}$, $W_{n,K_{\text{Geometric}}}^{\text{combined}}$, and W^{gt}. The source-only and target-only key orders, W^{src} and W_n^{tgt}, are computed directly using Equations (2) and (3). The β value in f_β is set to be 0.5. Precision and recall values for all the possible thresholds used for finding the best f_β can be efficiently computed using standard numerical libraries like scikit-learn, which offer highly optimized implementations.

To obtain the Bayesian combination weights K_{Bayesian} and $K_{\text{Bayesian},\Theta}$, we follow the bootstrap procedure in **Algorithm 1** and the Bayesian setup described in Section 3.2. We implement bootstrap resampling with NumPy and set $n_{\text{boot}} = 1000$. The discount factor Θ is computed numerically in Python/NumPy following the derivation in Section 5.1. For the linear combination $W_{n,K_{\text{Linear}}}^{\text{combined}}$, we found that choosing $K_0 \in [10, 20]$ in Equation (20) performs well across our datasets.

After computing the seven key orders for each source–target pair and each n, we preprocess the target-domain *normal* training data using Equation (5) with the corresponding key-order weights. This produces seven transformed versions of the target dataset (one per key order). On each transformed target dataset, we evaluate three anomaly detection algorithms—Local Outlier Factor (LOF), Principal Component Analysis (PCA), and Gaussian Mixture Model (GMM)—and report average performance. All anomaly detectors are implemented using the PyOD library (Zhao, Nasrullah, & Li, 2019). Moreover, each experiment is repeated five times with different random seeds.

4.5. Experimental Results

4.5.1. Source Data: Plant A; Target Data: Plant A

In this experiment, the evaluation is conducted over sixteen source–target dataset pairs. In total, this results in

$$16 \text{ pairs} \times 7 \text{ key orders} \times 3 \text{ algorithms} \times 12 \text{ values of } n \times 5 \text{ seeds} = 20{,}160$$

simulations.

In this experiment, all known anomalies and normal samples in the target test set are used for evaluation. The known anomalies in the target training set are randomly sampled from the target test set. The known anomalies in the target training set are mainly used to estimate the key order as illustrated in Figure 1. We following the experimental procedure in section 4.4.3 to first calculate key orders. After the target normal dataset is scaled by the estimated key orders as shown in Equation (5), unsupervised anomaly detection algorithms such as PCA, LOF, and GMM are applied. Then the best f_1 scores are calculated against the ground-truth target labels. In cases where the number of anomalies in the training set assumed to be known exceeds the actual number present in the target test set, all available anomalies in the test set are used both to estimate the key orders and to evaluate final performance. The average results are presented in Figure 7. As expected, anomaly detection algorithms using target data processed with W^{src} outperform those with W_n^{tgt} when n is extremely small—particularly for $n = 1$ and $n = 2$. However, as the number of known anomalies in the target dataset increases to $n = 3$, algorithms using W_n^{tgt} begin to outperform those relying on W^{src} on average.

As discussed in Section 4.3, when $n \leq 3$, the bootstrapping procedure in **Algorithm 1** tends to underestimate the target variance $\Sigma_{\text{tgt},n}$. As a result, the diagonal entries $K[i,i]$ of both K_{Bayesian} and $K_{\text{Bayesian},\Theta}$ become inflated, causing the combined weights $W_{n,K_{\text{Bayesian}}}^{\text{combined}}$ and

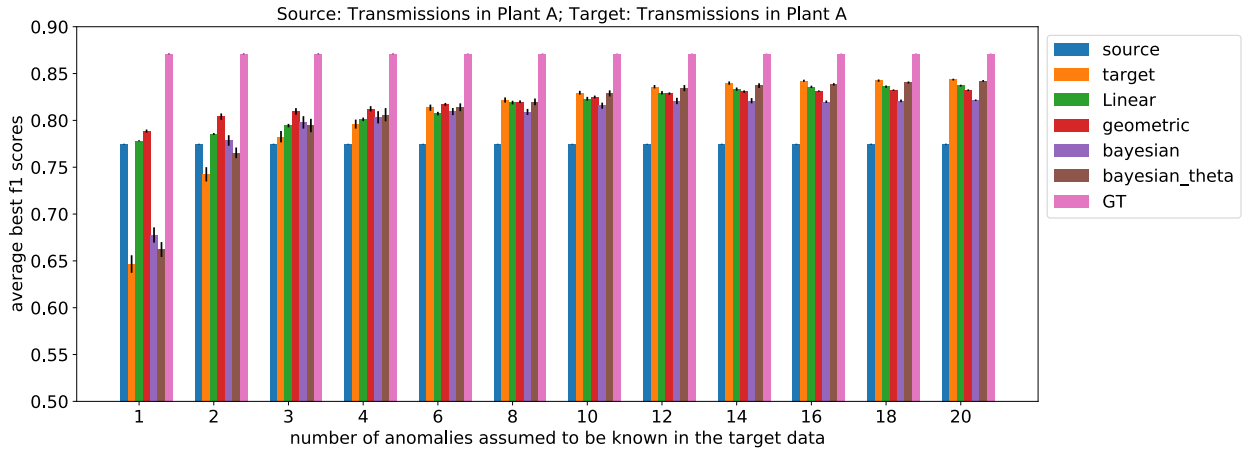

Figure 7. Average results using different key orders. Datasets from Plant A are used as both the source and target domains.

$W^{\text{combined}}_{n,K_{\text{Bayesian},\Theta}}$ to be dominated by the target weights W^{tgt}_n. Since earlier results show that anomaly detection performance suffers when relying on W^{tgt}_n with very few anomalies (i.e., $n \leq 3$), this explains the suboptimal performance of the Bayesian key order, $W^{\text{combined}}_{n,K_{\text{Bayesian}}}$, and Bayesian-$\Theta$ key order, $W^{\text{combined}}_{n,K_{\text{Bayesian},\Theta}}$, in this regime. In contrast, the heuristic methods using linear and geometric combination strategies—$W^{\text{combined}}_{n,K_{\text{Linear}}}$ and $W^{\text{combined}}_{n,K_{\text{Geometric}}}$—do not rely on variance estimation and are therefore unaffected by the underestimation of $\Sigma_{\text{tgt},n}$. When $n \leq 3$, these heuristic methods produce weights that remain close to the source key order W^{src}, resulting in comparable and slightly improved anomaly detection performance relative to using W^{src} alone, and clearly outperform the W^{tgt} baseline.

Once more than three anomalies are identified in the target dataset, the target key order variance $\Sigma_{\text{tgt},n}$ can be estimated more reliably via the bootstrap procedure. As a result, methods that depend on accurate estimation of target key order variance—such as $W^{\text{combined}}_{n,K_{\text{Bayesian}}}$ and $W^{\text{combined}}_{n,K_{\text{Bayesian},\Theta}}$—begin to perform effectively. Therefore, as the number of anomalies increases, it becomes appropriate to transition to Bayesian-based methods for combining source and target key orders.

Interestingly, this dataset exhibits a notable property: once the number of anomalies reaches as few as six, the target key order begins to perform exceptionally well—often outperforming or matching most of the combined methods. This behavior is not inconsistent with expectations in our experimental setup, where, beyond a certain threshold of available anomalies, incorporating information from the source dataset becomes unnecessary or even detrimental. However, as n grows larger, the combined key order using the original Bayesian method, $W^{\text{combined}}_{n,K_{\text{Bayesian}}}$, may struggle to fully transition to the target key order, W^{tgt}_n, since the variance of the target key order cannot reach zero. In contrast, the Bayesian-Θ method, which introduces a discounting factor Θ, is capable of learning to downweight the influence of the source dataset appropriately. As shown in the results, once $n > 8$, most combined methods begin to incorporate unnecessary information from the source and consequently underperform compared to the target key order alone. The proposed Bayesian-Θ method, however, continues to improve and nearly matches the performance of the target key order, demonstrating its ability to effectively discount irrelevant source information.

It is noteworthy that the method based on $W^{\text{combined}}_{n,K_{\text{Bayesian},\Theta}}$ consistently underperforms relative to $W^{\text{combined}}_{n,K_{\text{Bayesian}}}$ when $n \leq 3$. This underperformance can be attributed to the fact that estimating the additional parameter Θ requires more data—specifically, a larger number of anomalies in the target dataset. As a result, $W^{\text{combined}}_{n,K_{\text{Bayesian},\Theta}}$ is not as effective in low-n regimes but can match or exceed the performance of $W^{\text{combined}}_{n,K_{\text{Bayesian}}}$ as more anomalies become available. These findings support our strategy of adopting a heuristic method (to combine source and target key orders) when n is very small, transitioning to a Bayesian approach as n increases, and ultimately employing the Bayesian-Θ method once a sufficient number of target anomalies areavailable. In this regime, the Bayesian-Θ method leverages its adaptive weighting mechanism to outperform the standard Bayesian method.

4.5.2. Source Data: Plant A; Target Data: Plant B

The source data for the following experiments still comes from Plant A. However, unlike the previous experimental setup, the target dataset is from Plant B. The rest of the experimental setup remains unchanged. Once again, we compare the average performance of three anomaly detection algorithms: LOF, PCA, and GMM. The same seven key orders are used for comparison.

The experimental results are presented in Figure 8. Consistent trends are observed across settings, where the

Figure 8. Average results using different key orders. The datasets from Plant A are used as the source domain, and those from Plant B are used as the target domain.

heuristic methods—$W^{\text{combined}}_{n,K_{\text{Linear}}}$ and $W^{\text{combined}}_{n,K_{\text{Geometric}}}$—perform well in the extremely low-anomaly regime (i.e., n $\leq$ 3). As the number of identified anomalies increases to the intermediate range ($3 < n < 8$), the Bayesian method begins to outperform the heuristics. Initially, the Bayesian-Θ variant underperforms slightly compared to the standard Bayesian method; however, as more anomalies are observed, the Bayesian-Θ approach surpasses the standard Bayesian method, as expected.

A key distinction from the earlier experiments—where both the source and target datasets originated from Plant A—is that, in the current setting, the target key order improves more slowly and takes longer to completely surpass the combined methods. Even at higher values of n, the target key order does not fully take precedence, suggesting that meaningful information from the source data continues to contribute. Consequently, anomaly detection methods based on the Bayesian-Θ approach outperform those relying solely on the target key order, particularly at higher values of n. This contrasts with the previous setup, where the performance of the Bayesian-Θ method closely matched that of the target-only method due to the rapid improvement of the target key order.

These results indicate that the proposed method performs better in scenarios where the target key order improves slowly. In such cases, the Bayesian-Θ method maintains superior performance at higher n, whereas in settings where the target key order improves quickly, its performance converges with that of the target-only method.

Combining the results presented in Figure 7 and Figure 8, our proposed strategy exhibits optimal performance across different anomaly regimes. The approach can be summarized as follows:

- When the number of anomalies n is extremely small (e.g., $n \leq 3$ in our specific datasets; the exact threshold can be determined using an elbow-like criterion, as discussed in Section 4.3), we recommend using the posterior key order based on K_{Linear} or $K_{\text{Geometric}}$ to mitigate the risk of underestimating the target variance $\Sigma_{\text{tgt},n}$.
- As n increases, we transition to using the posterior key order computed with K_{Bayesian}, which fully exploits the structure of our Bayesian framework.
- When n becomes sufficiently large (e.g., $n \geq 8$ in our specific datasets), we employ the posterior key order based on $K_{\text{Bayesian},\Theta}$ to address the limitation where the standard Bayesian approach may not discount the influence of the source key order quickly enough.

4.6. Optimal Source Data Selection

The inclusion of target anomalies and the estimation of the key order in the target dataset enable the identification of the optimal source data from a list of potential sources, even when an initial source-target pairing is not provided. Our previous study demonstrated that in Plant A, trans_a serves as the best source data for trans_c, and vice versa. Similarly, trans_b is the best source data for trans_d, and vice versa. When this initial pairing is unavailable, we have developed a simple heuristic for identifying the optimal source data. The process consists of the following steps:

1. **Initial key order estimation**: We estimate key orders for all source and selected target dataset across all four tests.
2. **Normalization**: Every estimated key order is normalized such that it satisfies $\sum_{i=0}^{2048} W[i] = 1$, treating it as a probability distribution.
3. **KL-divergence calculation**: For each test, we compute the Kullback-Leibler (KL) divergence between

the normalized target key order and the normalized source key order.

4. **Source selection**: By averaging the KL divergence across all four tests, we identify the source dataset with the lowest average KL divergence as the best match.

We evaluated the effectiveness of this heuristic using target datasets containing between 1 and 9 anomalies. For each specific number of anomalies, we conducted tests using 10 different random seeds. The percentage of correct matches (between source and target data), along with its standard deviation, is presented in Figure 9. The results indicate that our heuristic is highly effective. For instance, when the target dataset contains as few as 2 anomalies, the heuristic correctly identifies the best source dataset approximately 90% of the time.

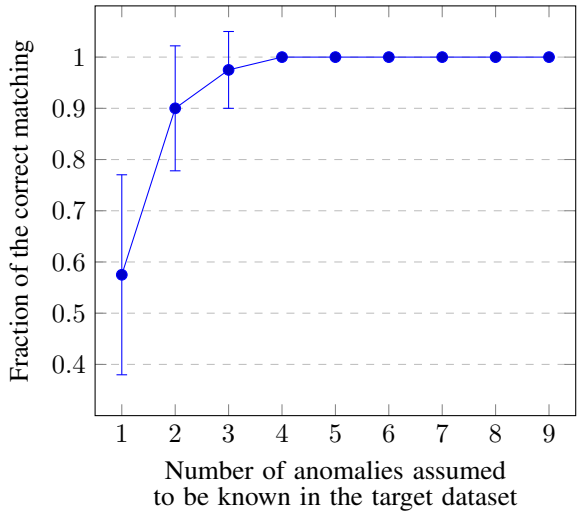

Figure 9. Effectiveness of our simple heuristics in finding optimal source data.

5. CONCLUSION

In this work, we extend our previously proposed transfer learning framework to account for scenarios where the target dataset contains anomalous data. This extension enables the estimation of the key order in the target dataset and facilitates the integration of source and target key orders to improve anomaly detection. One of the most effective approaches for combining information is the Bayesian framework, which relies on uncertainty estimation. However, we find that when the number of anomalies in the target dataset is extremely small ($n \leq 3$), the uncertainty in the target key order is often underestimated due to the limitations of the bootstrap method. This underestimation results in suboptimal performance when applying the Bayesian formulation in such cases. To mitigate this issue, we propose using alternative combination strategies, such as a weighted linear combination of the source and target key orders, when the number of anomalies is very low. As the number of anomalies in the target dataset increases ($n \geq 3$), the key order estimated using the Bayesian formulation becomes more reliable and outperforms both the individual source and target key orders, as well as simpler combination strategies.

Another potential limitation of the standard Bayesian method is its inability to rapidly discount information from the source dataset as the number of anomalies in the target dataset increases. To address this, we introduce a discounting factor Θ, which explicitly reduces the influence of the source key order. Our results show that when the number of known anomalies in the target dataset becomes sufficiently large (e.g., $n \geq 8$), the Bayesian-Θ method achieves strong performance and effectively mitigates this limitation.

Furthermore, we address the challenge of identifying the optimal source dataset when ground truth source-target pairings are unavailable. We introduce a heuristic method based on KL divergence, and our results demonstrate that this heuristic achieves high accuracy. Overall, our findings provide a robust framework for transfer learning in anomaly detection.

Although this study is motivated by rotating machinery, the proposed framework is broadly applicable beyond this specific transfer anomaly detection setting. Many anomaly detection problems exhibit a similar structure in which only a subset of features is consistently diagnostic, and performance depends on appropriately emphasizing these informative dimensions. In addition, the practical lessons from our study—such as using robust heuristic fusion when uncertainty estimates are unreliable, and employing a discount factor (e.g., the Θ "discounting" strategy) to temper source influence as target evidence accumulates—may translate to other transfer and domain-shift settings. More generally, our results underscore the effectiveness of uncertainty-aware transfer for improving industrial anomaly detection in the presence of distribution shift.

Despite the promising results, our approach has certain limitations. Although the proposed heuristic—analogous to the elbow method—provides a practical guideline for determining when to transition from heuristic strategies (e.g., linear or geometric combinations) to Bayesian-based methods, it lacks a rigorous mathematical formulation. Specifically, due to the limited amount of data, we are currently unable to establish formal criteria for when to switch from heuristic methods to the Bayesian method, or from the Bayesian method to the Bayesian-Θ variant. In future work, we aim to conduct more extensive simulations and theoretical analyses to better characterize these transition points and develop principled switching criteria.

ACKNOWLEDGMENT

This work was supported by the Ford Foundation.

REFERENCES

Ibrahim, J. G., & Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, *15*(1), 46–60.

Karbalayghareh, A., Qian, X., & Dougherty, E. R. (2018). Optimal bayesian transfer learning. *IEEE Transactions on Signal Processing*, *66*(14), 3724–3739.

Kim, M. S., Yun, J. P., & Park, P. (2021). An explainable neural network for fault diagnosis with a frequency activation map. *IEEE Access*, *9*, 98962–98972.

Liang, J., Shui, H., Gupta, R., Upadhyay, D., & Darve, E. (2025). Transfer learning for anomaly detection in rotating machinery using data-driven key order estimation. *IEEE Transactions on Automation Science and Engineering*, *22*, 13310–13326.

Mao, W., Sun, B., & Wang, L. (2021). A new deep dual temporal domain adaptation method for online detection of bearings early fault. *Entropy*, *23*(2), 162.

Mao, W., Wang, G., Kou, L., & Liang, X. (2023). Deep domain-adversarial anomaly detection with one-class transfer learning. *IEEE/CAA Journal of Automatica Sinica*, *10*(2), 524–546.

Mao, W., Zhang, D., Tian, S., & Tang, J. (2020). Robust detection of bearing early fault based on deep transfer learning. *Electronics*, *9*(2), 323.

Mark, W. D., Lee, H., Patrick, R., & Coker, J. D. (2010). A simple frequency-domain algorithm for early detection of damaged gear teeth. *Mechanical Systems and Signal Processing*, *24*(8), 2807–2823.

Mey, O., & Neufeld, D. (2022). Explainable AI algorithms for vibration data-based fault detection: Use case-adapted methods and critical evaluation. *Sensors*, *22*(23), 9037.

Michau, G., & Fink, O. (2021). Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer. *Knowledge-Based Systems*, *216*, 106816.

Pan, Q., Bao, Y., & Li, H. (2023). Transfer learning-based data anomaly detection for structural health monitoring. *Structural Health Monitoring*, *22*(5), 3077–3091.

Qin, L. (2024). Rolling bearing fault detection using domain adaptation-based anomaly detection. *International Journal on Artificial Intelligence Tools*, *33*(07), 2440003.

Roelofs, C. M., Gück, C., & Faulstich, S. (2024). Transfer learning applications for autoencoder-based anomaly detection in wind turbines. *Energy and AI*, *17*, 100373.

Sarrazin, M., Gillijns, S., Anthonis, J., Janssens, K., van der Auweraer, H., & Verhaeghe, K. (2013). Nvh analysis of a 3 phase 12/8 sr motor drive for hev applications. In *2013 world electric vehicle symposium and exhibition (evs27)* (pp. 1–10).

Sharma, V., & Parey, A. (2017). Frequency domain averaging based experimental evaluation of gear fault without tachometer for fluctuating speed conditions. *Mechanical Systems and Signal Processing*, *85*, 278–295.

Shwartz-Ziv, R., Goldblum, M., Souri, H., Kapoor, S., Zhu, C., LeCun, Y., & Wilson, A. G. (2022). Pretrain your loss: Easy bayesian transfer learning with informative priors. *Advances in Neural Information Processing Systems*, *35*, 27706–27715.

Sun, H., Ammann, K., Giannoulakis, S., & Fink, O. (2024). Continuous test-time domain adaptation for efficient fault detection under evolving operating conditions. *Proceedings of the European Conference of the PHM Society*, *8*(1), 11.

Vercruyssen, V., Meert, W., & Davis, J. (2017). Transfer learning for time series anomaly detection. In *Proceedings of the workshop and tutorial on interactive adaptive learning@ ecmlpkdd 2017* (Vol. 1924, pp. 27–37).

Vincent, V., Wannes, M., & Jesse, D. (2020). Transfer learning for anomaly detection through localized and unsupervised instance selection. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 6054–6061).

Wang, B., Baraldi, P., Zio, E., Brown, J., & Gauthier, S. (2026). Fleet-based transfer learning for anomaly detection in industrial systems. *Mechanical Systems and Signal Processing*, *244*, 113725.

Wu, J., Mao, W., Zhang, Y., Fan, L., & Zhong, Z. (2024). A novel few-shot deep transfer learning method for anomaly detection: Deep domain-adversarial contrastive network with time-frequency transferability analytics. *IEEE Internet of Things Journal*, *11*(17), 28809–28823.

Xie, J., Huang, B., & Dubljevic, S. (2021). Transfer learning for dynamic feature extraction using variational bayesian inference. *IEEE Transactions on Knowledge and Data Engineering*, *34*(11), 5524–5535.

Xu, C., Wang, J., Zhang, J., & Li, X. (2021). Anomaly detection of power consumption in yarn spinning using transfer learning. *Computers & Industrial Engineering*, *152*, 107015.

Yadav, E., & Chawla, V. K. (2024). Role and significance of defect detection methods for rotating machines: An explicit literature review. *Journal of The Institution of Engineers (India): Series C*, *105*(5), 1293–1310.

Zhang, S., Ye, F., Wang, B., & Habetler, T. G. (2020). Few-shot bearing anomaly detection via model-agnostic meta-learning. In *2020 23rd international conference on electrical machines and systems (icems)* (pp. 1341–1346).

Zhang, S., Zhong, Z., Li, D., Fan, Q., Sun, Y., Zhu, M., ... others (2022). Efficient kpi anomaly detection through transfer learning for large-scale web services. *IEEE Journal on Selected Areas in Communications*, *40*(8), 2440–2455.

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, *20*(96), 1–7.

Zhu, R., Peng, W., Wang, D., & Huang, C.-G. (2023). Bayesian transfer learning with active querying for intelligent cross-machine fault prognosis under limited data. *Mechanical Systems and Signal Processing*, *183*, 109628.

APPENDIX

5.1. Estimation of Θ parameter

Recall that we assume that every dimension of the key order is independent from each other as shown in the Equation (15). Thus, the corresponding Θ in Equation (17) is diagonal, and $\Theta[i,i]$ is independent from $\Theta[j,j]\ \forall i \neq j$. Thus, we need to estimate $\Theta[i,i]$ given $\mathbf{W_S[i]}$ and $\mathbf{W_T[i]}|\mathbf{W_S[i]}$. To simplify the notation, let's use θ to represent $\Theta[i,i]$, X to represent $\mathbf{W_S}$, and Y to represent $\mathbf{W_T}$. Let's now treat θ as a random variable, and we will use the maximum likelihood method to estimate the optimal value of θ. Without loss of the generality, we first assume θ follows a uniform distribution in the range $(0,\alpha)$, for some $\alpha > 0$. Thus,

$$P(\theta) = \frac{1}{\alpha}, \forall \theta \in (0,\alpha). \tag{22}$$

Since the variance of X is discounted by θ, the conditional distribution of $X|\theta$ is

$$X|\theta \sim \mathcal{N}(\mu, \frac{1}{\theta}\tau^2) \tag{23}$$

Thus, the probability density function of $X|\theta$ is

$$P(X|\theta) = \frac{1}{\sqrt{2\pi\tau^2/\theta}} \exp\{-0.5\theta(x-\mu)^2/\tau^2\} \tag{24}$$

The joint distribution is $P(X,\theta)$ with the expression:

$$\begin{aligned} P(X,\theta) &= P(X|\theta)P(\theta) \\ &= \frac{1}{\alpha}\frac{1}{\sqrt{2\pi\tau^2/\theta}} \exp\Big\{-0.5\theta(x-\mu)^2/\tau^2\Big\} \end{aligned} \tag{25}$$

Moreover, $Y_i|X = x$ is independent of θ by the local markov property in Bayesian Network. Thus, $P(Y|X = x,\theta)$ is the following:

$$\begin{aligned} P(Y \mid X = x,\theta) &= P(Y \mid X = x) \\ &= \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left\{-\frac{0.5(y_i-x)^2}{\sigma^2}\right\} \end{aligned} \tag{26}$$

Combining everything together, the joint distribution of Y, X, θ together is

$$\begin{aligned} P(Y,X,\theta) &= P(Y \mid X,\theta)\, P(X,\theta) \\ &= \frac{1}{\alpha} \times \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \times \frac{1}{\sqrt{2\pi\tau^2/\theta}} \\ &\quad \times \exp\Bigg\{-\tfrac{1}{2}\Bigg(\frac{\sum_{i=1}^{n}(y_i-x)^2}{\sigma^2} \\ &\quad + \frac{\theta(x-\mu)^2}{\tau^2}\Bigg)\Bigg\} \end{aligned} \tag{27}$$

Notice that the marginal probability of Y is

$$P(Y) = \int_0^{\alpha}\int_{-\infty}^{\infty} P(Y,X,\theta)dxd\theta, \tag{28}$$

and $P(Y)$ is a constant term since Y is fixed and the term does not depend on the value of either of x or θ term. Thus,

$$P(X,\theta|Y) = \frac{P(X,\theta,Y)}{P(Y)} \tag{29}$$

Furthermore,

$$\begin{aligned} P(\theta \mid Y) &= \int_{-\infty}^{\infty} P(X,\theta \mid Y)\, dx \\ &= \text{const}\ \frac{1}{P(Y)}\int_{-\infty}^{\infty} P(X,\theta,Y)\, dx \\ &= \text{const}\ \frac{1}{P(Y)}\ \sqrt{\theta}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\tau^2}} \\ &\quad \times \exp\Bigg\{-\tfrac{1}{2}\Bigg(\frac{\sum_{i=1}^{n}(y_i-x)^2}{\sigma^2} \\ &\quad + \frac{\theta(x-\mu)^2}{\tau^2}\Bigg)\Bigg\}\, dx \end{aligned} \tag{30}$$

Now, we need to use either numerical or direct method to solve the following optimization problem,

$$\max_{\theta} P(\theta|Y) \tag{31}$$

The result is the optimal θ' we are looking for.

5.2. Data Train and Test Split

See Tables 1 and 2 for details on the train and test splits.

	Plant A Train		Test		
Target Data	Normal	Abnormal	Normal	Abnormal	Source Data (from Plant A)
trans_a_test_1	52293	n=1–20	13485	10	trans_c_test_1
trans_a_test_2	52291	n=1–20	13453	12	trans_c_test_2
trans_a_test_3	52292	n=1–20	13537	116	trans_c_test_3
trans_a_test_4	52291	n=1–20	13531	36	trans_c_test_4
trans_b_test_1	20593	n=1–20	4830	78	trans_d_test_1
trans_b_test_2	20592	n=1–20	4830	22	trans_d_test_2
trans_b_test_3	20592	n=1–20	4828	90	trans_d_test_3
trans_b_test_4	20591	n=1–20	4830	38	trans_d_test_4
trans_c_test_1	30560	n=1–20	7169	5	trans_a_test_1
trans_c_test_2	30558	n=1–20	7168	8	trans_a_test_2
trans_c_test_3	30559	n=1–20	7167	17	trans_a_test_3
trans_c_test_4	30558	n=1–20	7167	16	trans_a_test_4
train_d_test_1	5557	n=1–20	1303	29	trans_b_test_1
train_d_test_2	5557	n=1–20	1304	3	trans_b_test_2
train_d_test_3	5556	n=1–20	1303	21	trans_b_test_3
train_d_test_4	5556	n=1–20	1303	10	trans_b_test_4

Table 1. Train–test partition for Plant A. The number of anomalies in the target training set assumed to be known ranges from 1 to 20. Importantly, anomalous samples in the training data are randomly drawn from the test dataset and are used only for estimating the key orders. Additionally, only 3,000 normal training samples are employed for key order estimation. If n, the number of anomalies assumed to be known in the target data, exceeds the actual number of anomalies available in the target test dataset, then all available anomalies in the target test dataset are used for estimating the key orders.

	Plant B Train		Test		
Target Data	Normal	Abnormal	Normal	Abnormal	Source Data (from Plant A)
trans_I_test_1	3774	n=1–20	3775	18	trans_b_test_1
trans_I_test_2	4079	n=1–20	4079	61	trans_b_test_2
trans_I_test_3	4118	n=1–20	4118	16	trans_b_test_3
trans_I_test_4	4103	n=1–20	4104	12	trans_b_test_4
trans_II_test_1	4107	n=1–20	4108	37	trans_b_test_1
trans_II_test_2	7299	n=1–20	7300	61	trans_b_test_2
trans_II_test_3	7341	n=1–20	7342	24	trans_b_test_3
trans_II_test_4	7312	n=1–20	7313	17	trans_b_test_4
trans_III_test_1	8583	n=1–20	8584	7	trans_b_test_1
trans_III_test_2	8850	n=1–20	8851	100	trans_b_test_2
trans_III_test_3	8871	n=1–20	8871	39	trans_b_test_3
trans_III_test_4	8758	n=1–20	8758	48	trans_b_test_4
trans_IV_test_1	8801	n=1–20	8802	53	trans_a_test_1
trans_IV_test_2	13596	n=1–20	13597	40	trans_a_test_2
trans_IV_test_3	13652	n=1–20	13652	62	trans_a_test_3
trans_IV_test_4	13687	n=1–20	13688	10	trans_a_test_4

Table 2. Train–test partition for Plant B. The number of anomalies in the target training set assumed to be known ranges from 1 to 20. Importantly, anomalous samples in the training data are randomly drawn from the test dataset and are used only for estimating the key orders. Additionally, only 3,000 normal training samples are employed for key order estimation. If n, the number of anomalies assumed to be known in the target data, exceeds the actual number of anomalies available in the target test dataset, then all available anomalies in the target test dataset are used for estimating the key orders.