

Unsupervised Vibroacoustic Anomaly Detection for End-of-Line Testing of High-Variation Hydraulic Machinery

Maximilian Romeser

`Maximilian.Romeser@boschrexroth.de`

Technische Hochschule Ulm

Reinhold von Schwerin

Technische Hochschule Ulm

Research Article

Keywords: End-of-Line Testing, Anomaly Detection, Vibration Analysis, Fault Detection, Axial Piston Unit, Data-Driven Diagnostics

Posted Date: August 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10617384/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Unsupervised Vibroacoustic Anomaly Detection for End-of-Line Testing of High-Variation Hydraulic Machinery

Maximilian Romeser^{1,2*} and Reinhold von Schwerin¹

¹*Faculty for Informatics, Ulm Technical University of Applied Sciences, Prittwitzstraße 10, Ulm, 89075, Baden-Wuerttemberg, Germany.

²Development of Axial Piston Units, Bosch Rexroth AG, Glockeraustraße 2, Elchingen, 89275, Bavaria, Germany.

*Corresponding author(s). E-mail(s): Maximilian.Romeser@boschrexroth.de;
Contributing authors: reinhold.vonschwerin@thu.de;

Abstract

The shift towards Industry 4.0 drives mass individualization, resulting in high product variety and small batch sizes (High-Mix, Low-Volume or HMLV). This trend poses significant challenges for vibroacoustic fault detection in End-of-Line Testing (EoLT), particularly for customized rotating machinery like axial piston units (APUs). Data shifts caused by minor variations in test setups, product variants, and processes complicate reliable quality assurance. We propose Adaptive Spectral Decomposition (ASD)-based Batch-wise Outlier Detection (BOD) or ASD-BOD, a framework transferring unsupervised fleet-based condition monitoring methods to fault detection in EoLT. It enhances a hierarchical clustering-based approach for HMLV requirements by introducing weighted cophenetic correlation-based scoring. A second key innovation is the use of Robust Principal Component Analysis (RPCA) for ASD. This novel application suppresses process-induced background variations while enhancing fault-to-noise ratios in vibroacoustic frequency domain data. We demonstrate the framework's effectiveness for bearing fault detection using real-world measurement data from APUs. The results confirm ASD-BOD successfully addresses HMLV challenges, providing continuous, threshold-transferable, and interpretable outlier scores across evaluated batch sizes without product- or batch-specific re-tuning. Crucially, RPCA-based ASD significantly improves outlier detection, increasing the Area Under the Precision-Recall Curve from 0.4 to 0.8 on the real-world APU data.

Keywords: End-of-Line Testing, Anomaly Detection, Vibration Analysis, Fault Detection, Axial Piston Unit, Data-Driven Diagnostics

1 Introduction

Modern manufacturing systems are increasingly shaped by Industry 4.0-driven requirements for flexibility and customization [1]. Consequently, production environments are shifting towards high-mix, low-volume (HMLV) regimes with growing product diversity and decreasing batch sizes [2]. For manufacturers

of high-performance rotating machinery, this development substantially complicates quality assurance. In electric drive technology, modular product architectures rapidly expand the configuration space and increase product variance [3], while mobile hydraulics has long been characterized by high variance due to

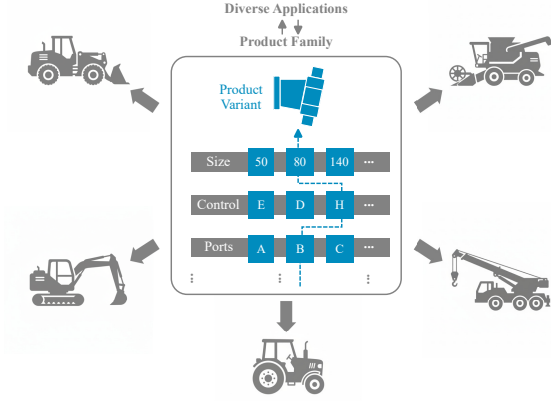


Fig. 1: Application-driven product portfolio complexity using the example of APUs for mobile machinery. Icons generated using Google’s Nano Banana model.

heterogeneous customer demands and diverse operating conditions requiring extensive customization [4]. Across both domains, newly introduced or infrequently produced variants offer little to no historical production or quality data, even though quality and reliability requirements remain unchanged.

Axial piston units (APUs) exemplify this challenge in industrial and mobile hydraulics. In particular, as key components in diverse mobile machinery, their high power density, versatility and efficiency make them indispensable [5]. As depicted in Figure 1, configuring APUs for diverse customers and operating conditions results in an immense range of variants, extending from housing size and displacement to control types, pressure settings and complex multi-unit assemblies. This product variance demands significant effort not only in engineering and manufacturing but explicitly in quality assurance, which must ensure efficiency and longevity in high-stakes applications such as construction [6], agriculture [7] and industrial processes [8].

1.1 End-of-Line Testing in High-Variance Production

Production testing of rotating machinery forms a crucial part of quality assurance, with End-of-Line testing (EoLT) as the most comprehensive inspection after full assembly [9] (in contrast to inline tests [10]). Despite inherent complexity from accumulated variance, EoLT remains indispensable. As exemplified for APUs in Figure 2, EoLT serves as the final gatekeeper, verifying integrity, functionality, and specification compliance before shipment, enabling clear distinction between fault-free (OK) and faulty

(nOK) products. The latter undergo detailed analysis essential for identifying systematic root causes. Traditionally, this relies on dedicated quality features compared against tolerance bands in statistical process control. However, high product variance considerably inflates complexity of this rule-based approach, while potential for deeper fault detection remains limited. Therefore, data-driven methods based on machine learning (ML), as suggested by Nunes et al. [11], offer significant potential.

This potential must be realized within demanding operational constraints. EoLT is governed by a fundamental trade-off between test duration and fault coverage. For complex products like APUs, this conflict is pronounced, often establishing EoLT as a production bottleneck where test cycles significantly exceed manufacturing cycle time [12]. Multi-phase test programs, composed of sequential steps evaluating units under different operating conditions (e.g., varying speed or pressure), drive these long durations [13]. The core challenge arises from high portfolio variance, since, instead of standardized sequences, vast libraries of variant-specific test programs are required [14]. This dynamic HMLV environment poses challenges for data-driven analysis methods, which must be robust and adaptive to cope with data distribution shifts occurring with product or test program changes.

1.2 Domain Shifts in Vibroacoustic End-of-Line Testing

Vibroacoustic (VA) analysis offers deep fault coverage, as APU operation, like other rotating machinery, depends on the integrity of rolling bearings, gears and tribological contacts. In condition monitoring (CM), data-driven VA methods are well-established for detecting incipient bearing faults [15]. Transitioning from CM to EoLT is challenging as CM tracks deviations from an individual machine’s baseline over time and EoLT must contend with inherent VA variations across populations of newly manufactured units due to manufacturing tolerances [16]. VA analysis has proven successful for other rotating machinery, serving as a direct quality indicator for quietness and an indirect diagnostic marker for mechanical faults in gearboxes [17], electric drive units (EDUs) [13], and bearing defects [18].

For hydraulic systems such as APUs, VA diagnostics are more challenging than for electric machines due to higher operational complexity. Strong, operation-dependent background noise from turbulent flow and pressure pulsations [19] easily masks subtle fault signatures [20], a problem also seen in reciprocating engines [21]. Beyond this loud

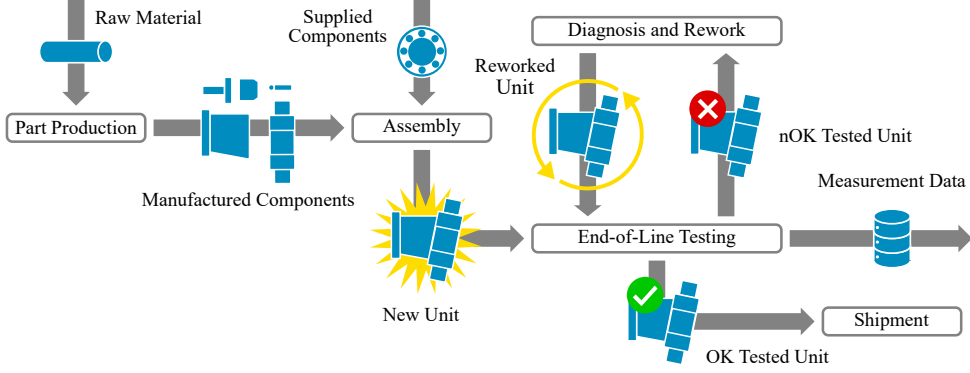


Fig. 2: Schematic overview of the EoLT within the value stream for APU manufacturing

background, diagnostics must contend with permissible series scatter from manufacturing tolerances and process-related variations, test rig coupling effects, hydraulic irregularities such as air entrainment and temperature shifts [22], and test rig configuration differences [23]. This is compounded by high sensitivity of VA measurements to structural variations between product variants.

These converging challenges manifest as domain shifts, i.e. systematic changes in the statistical properties of measured VA features between groups of tested units, even when all units are fault-free and tested under nominally identical conditions (e.g., between production batches). Root causes are manifold:

- **Product evolution:** specification changes alter structural dynamics and excitation patterns, giving each variant distinct VA characteristics [24]
- **Infrastructure variation:** different test rigs affect data via mounting resonances, sensor positioning, and load application [23]
- **Procedural differences:** test procedure variations alter operating conditions and measurement contexts [14]
- **Temporal drift:** gradual or abrupt shifts from environmental changes, tool wear, or process modifications [25]

These shifts critically threaten ML-based fault detection. A One-Class Classifier (OCC) trained on OK units from one domain (e.g., batch A) learns a normalcy representation that becomes invalid under shifted domains. As illustrated in Figure 3, fault-free units from batch B or C fall outside the learned decision boundary, yielding high rates of pseudo anomalies. Training across many heterogeneous domains simultaneously would instead require learning far more complex normalcy concepts, as heterogeneous

feature spaces make distinguishable decision boundaries harder to establish [24], potentially reducing sensitivity to real faults. Transfer or meta-learning techniques (e.g., [26]) could address multi-domain scenarios but require substantial multi-domain labeled data, often unavailable in HMLV production, and sophisticated MLOps environments [27].

Thus, applying VA analysis to APU EoLT is defined by converging challenges motivating adaptive approaches:

- **C1. Weak Fault-to-Noise Ratio and Permissible Scatter:** dominant hydraulic noise masks subtle fault signatures (e.g., bearing defects), while manufacturing and process scatter overlap with fault indicators, complicating detection sensitivity and threshold definition.
- **C2. Variant-Specific Vibroacoustic Characteristics:** high product variety causes distinct structural dynamics per variant, preventing universal diagnostic models.
- **C3. Limited Fault Data Availability:** scarcity of actual faults in HMLV production restricts balanced, labeled training data for supervised approaches.
- **C4. Domain Shift Sensitivity:** successive shifts from variant changes, test rig differences, procedural variations, and temporal drift degrade trained models rapidly, producing pseudo anomalies and requiring continuous adaptation.

1.3 Research Question, Contribution and Scope

These challenges, particularly domain shift prevalence, render (semi-)supervised models fragile in high-variance EoLT environments, leading to the central research question:

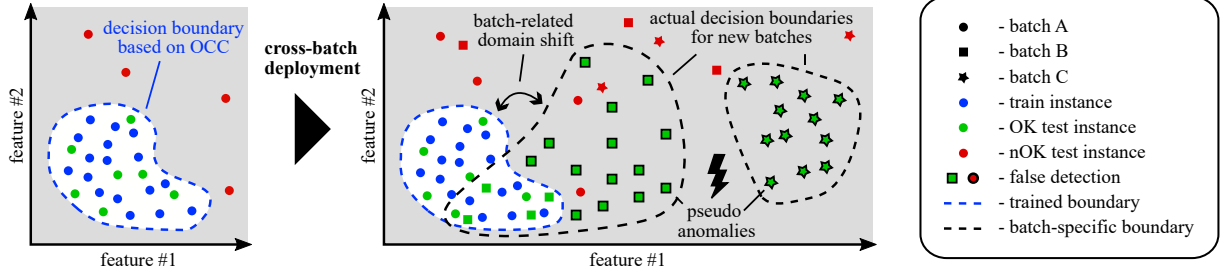


Fig. 3: The effect of domain shifts on an OCC in a two-dimensional feature space of a batch-wise production scenario where each data instance refers to a tested unit. The decision boundary, learned from the distribution of batch A, fails to generalize to batch B and C. Consequently, in-distribution instances from batch B and C are classified as an anomaly, resulting in a high rate of pseudo-anomalies.

How can subtle, localized vibroacoustic anomalies caused by rolling bearing defects be detected in EoLT scenarios considering an HMLV setting?

We propose a paradigm shift from training-based to unsupervised, batch-wise fault detection, grounded in the core EoLT assumption that while substantial distribution shifts occur between batches, conditions within a single batch remain comparatively stable. Units typically share the same variant, test program, and test bench, minimizing setup changes and short-term drift. This enables a reformulation where each batch is treated as an independent, localized domain rather than requiring a single global model robust to all shifts, analogous to divide-and-conquer strategies [28].

This perspective parallels fleet-based condition monitoring, where faults are identified as outliers within ensembles of units under comparable conditions [29], an approach Hendrickx et al. [30] successfully extended to EoLT of wind turbine generators. By analogy, we assume faulty units within a batch appear as rare outliers against a largely homogeneous healthy ensemble. This unsupervised outlier detection (OD) paradigm relies on two assumptions: faults are rare (addressing C3), occurring stochastically and infrequently, and faults are heterogeneous, varying in mode, severity, and localization. A key practical advantage is immediate applicability during ramp-up phases without historical data.

Based on this paradigm, our central contribution is ASD-BOD, a data-driven framework for vibroacoustic fault detection in EoLT under HMLV conditions and domain-shift-variant data (Figure 4). Built on hierarchical clustering, it follows an unsupervised, batch-wise OD strategy and introduces Adaptive Spectral Decomposition (ASD) to identify abnormal spectral components per batch, addressing

C1, without requiring prior domain knowledge about background noise or fault modes. RPCA-based preprocessing enhances variant-agnostic fault detection within batches, enabling robust, transparent detection under successive domain shifts without product- or batch-specific tuning.

ASD-BOD comprises four modules extending fleet-based CM frameworks [29] to noisy, HMLV EoLT: (1) data aggregation and preprocessing, (2) ASD, (3) dissimilarity indication, (4) clustering-based OD. Since the framework operates at batch level, as shown in Figure 4, it assumes complete batches are tested before final quality assessment. This batch-wise evaluation complements, rather than replaces, (semi-)supervised approaches for individual fault detection alongside rule-based tolerances, remaining aligned with established industrial QA workflows.

We validate ASD-BOD on a real-world APU dataset mimicking EoLT with faulty bearings under severe disturbances and distribution shifts from varying operating conditions [31]. Bearing faults were chosen as subtle, difficult-to-detect faults in hydraulic machinery, where their signatures are easily masked [20]. The key contributions are:

1. **Unsupervised HMLV Framework:** ASD-BOD extends fleet-based CM frameworks [29] with an ASD stage, addressing C2 and C4 through batch-wise OD without historical training data or labeled faults (C3), applicable from the first production batch.
2. **RPCA for ASD:** First application of Robust Principal Component Analysis (RPCA) [32] to adaptively decompose vibroacoustic spectra ensembles into shared background noise and anomalies, addressing C1 in EoLT.

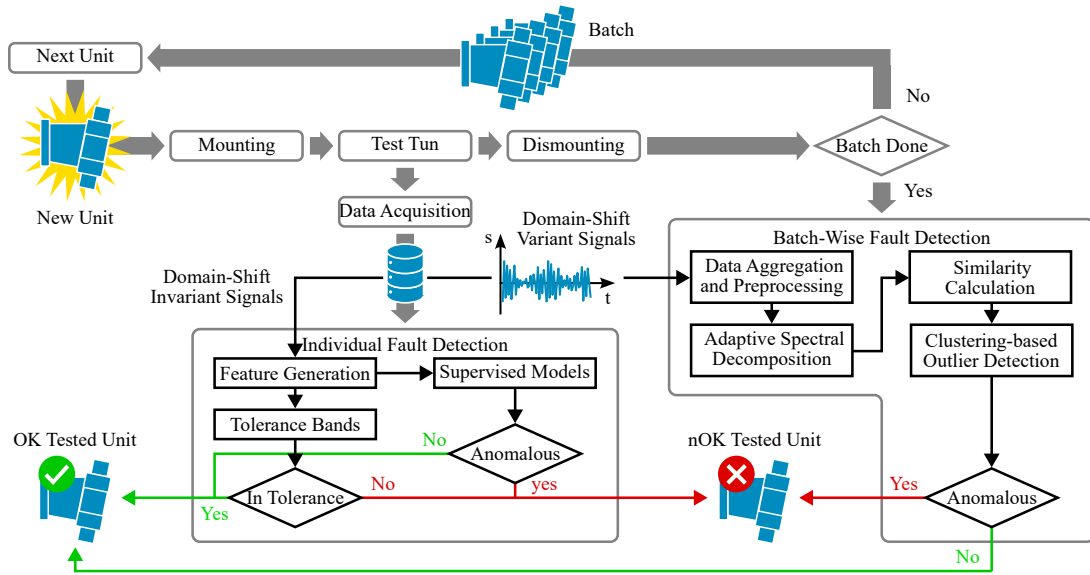


Fig. 4: Schematic overview of the ASD-BOD framework integrated into the EoLT process for APUs. Unsupervised batch-wise fault detection can be implemented in addition to conventional individual fault detection based on rule-based or (semi-)supervised ML approaches.

3. **Threshold-Transferable Outlier Scoring:** A weighted cophenetic hierarchical procedure generates bounded, interpretable anomaly scores. Its transferability across batch sizes is evaluated against the original Hendrickx framework and generic outlier detectors.
4. **Real-World Industrial Validation:** Validation on an experimental APU dataset with authentic bearing faults under realistic disturbances demonstrates practical applicability and robustness.

The paper is organized as follows. Section 2 reviews related work, Section 3 details the ASD-BOD framework, Section 4 presents the experimental results on real-world data, Section 5 discusses results and limitations, and Section 6 concludes.

2 Related Work

The state-of-the-art review is structured from an application-oriented and a methodology-oriented perspective. Research specifically addressing data-driven EoLT of APUs is exceptionally scarce. Vibroacoustic fault detection for APUs in CM is relatively mature. General data-driven methods for EoLT of rotating machinery have been reviewed in prior work [31]. That work compared sensor placements and signal processing techniques such as envelope spectrum, cepstrum and empirical mode decomposition. It combined these with semi-supervised anomaly detection

methods such as One-Class SVM, Isolation Forest and Local Outlier Factor for a single, fixed ensemble of APUs under varying operating conditions. The present work addresses a fundamentally different setting. It focuses on unsupervised batch-wise OD under batch-to-batch domain shifts, as encountered in HMLV production with changing product variants and test configurations. The most relevant EoLT literature from Gaugel et al. [12, 14] focuses on test time reduction using hydraulic data. This leaves vibroacoustic fault detection largely unexamined. Given this gap, the application-oriented review is extended to vibroacoustic production testing of rotating machinery more broadly. The methodology-oriented review focuses on RPCA.

2.1 Data-Driven Approaches for Vibroacoustic End-of-Line Testing

Data-driven methods for vibroacoustic production testing are typically classified as supervised, semi-supervised or unsupervised. Supervised methods require labeled OK and nOK data. In manufacturing, this often suffers from scarcity of faulty units due to quick root cause elimination during production. Del Rosso et al. [33] use time and frequency domain features with Logistic Regression and k-Nearest Neighbor

to identify faults in induction motors during production testing. Mlinarić et al. [34] employ tree-based ensembles such as random forest and gradient boosting for fault classification in brushless electric motors, using broadband and narrowband spectral features. They later extended this with a CNN-BiGRU architecture classifying Mel-frequency spectrograms [35]. Shaw et al. [21] and Borriello et al. [36] highlight the importance of feature engineering for ML diagnostic performance, in internal combustion engines and electric pumps respectively. Nonetheless, the inherent scarcity of nOK examples in complex rotating machinery continues to limit supervised methods. This drives greater adoption of semi-supervised anomaly detection.

Semi-supervised methods such as OCCs are trained primarily on OK unit data. They are often explored where faulty units are rare or unavailable. Blickensdorff et al. [37] investigate isolation forest and angle-based OD for production testing of EDUs. In more complex scenarios, autoencoder-based techniques are increasingly used. Sun et al. [38] propose a deep belief network-based autoencoder for accelerometer anomaly detection. Principi et al. [39] demonstrate superior performance using CNNs and LSTMs on log-Mel coefficients from vibroacoustic signals. Wißbrock et al. [40] explore envelope analysis, spectrograms and psychoacoustic features for anomaly detection on geared electric motors during EoLT, applying both OCC and isolation forest. Bonart [13] introduces a statistical approach computing coincidence and correlation matrices of spectrogram data during electric drive ramp-ups, scoring anomalies based on deviation from batch averages. However, semi-supervised methods rely on a stable, well-learned normalcy concept. This makes them sensitive to domain shifts and problematic in HMLV production.

Challenges of distribution shifts and high product variability in HMLV testing have driven growing interest in unsupervised approaches. These methods require no labeled data or predefined normalcy concepts. This makes them more adaptable to complex manufacturing environments. Muhr et al. [28] implement a batch-wise divide and conquer strategy, partitioning data by domain such as test bench or product type before applying local anomaly detection methods like LOF, OCSVM and isolation forest. Hendrickx et al. [30] extend this to a fleet-based framework analyzing motor currents from wind turbine generators within localized batches. However, these early methods often lack robust preprocessing to suppress significant noise sources like hydraulic pressure pulsations. Frameworks such as Hendrickx et al.'s [30] also rely on batch-size dependent fixed thresholds. This complicates industrial deployment with varying batch

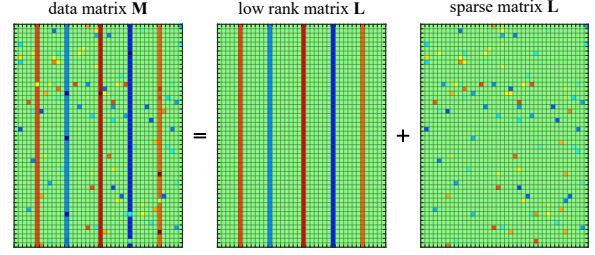


Fig. 5: The principle of low-rank decomposition via RPCA

configurations. A fully unsupervised framework must therefore integrate both adaptive preprocessing and scalable decision mechanisms. This gap is addressed here by ASD-BOD.

2.2 Robust Principal Component Analysis in Fault Detection

Traditional multivariate techniques such as Principal Component Analysis (PCA) are widely applied for fault detection and dimensionality reduction in industrial monitoring. PCA identifies low-dimensional subspaces and detects anomalies as deviations from them. However, PCA's sensitivity to outliers limits its use in noisy real-world environments [41]. RPCA resolves this by decomposing the observed data matrix $\mathbf{M}$ into a low-rank matrix $\mathbf{L}$ representing background behavior and a sparse matrix $\mathbf{S}$ representing outliers or anomalies [32].

$$\mathbf{M} = \mathbf{L} + \mathbf{S}. \quad (1)$$

This isolates faults and anomalies in $\mathbf{S}$ from noise and other low-rank components constituting normalcy, as shown in Figure 5.

RPCA has been successfully applied across industrial settings. Gharibnezhad et al. [41] show its superiority over standard PCA in suppressing faulty outliers during structural health monitoring. Tharault et al. [42] use it to generate structured residuals, enhancing fault isolation in process monitoring. For vibroacoustic fault detection, Li et al. [43] adopt a non-convex RPCA modification for bearing diagnostics. Ge et al. [44] introduce local RPCA to enhance fault feature extraction by segmenting vibration data. Wang et al. [45] apply RPCA to bearing diagnostics under variable speed, separating time-varying fault features from noise without prior knowledge of operating conditions.

RPCA's success in CM and fault detection is well documented. Still, these methods predominantly focus

on individual time-series or isolated data instances. Its application to batch-oriented production testing remains largely unexamined. Here, data from multiple units must be analyzed collectively to suppress shared background noise and extract batch-specific anomalies. The ASD-BOD framework presented here extends RPCA methodology to address this gap within HMLV EoLT. It provides robust noise-reduction preprocessing and enables accurate batch-level fault detection.

3 Proposed Framework

The ASD-BOD framework addresses challenges C1 to C4 identified in Section 1 through unsupervised, batch-wise fault detection. Unlike global models, it treats each production batch as a statistically independent ensemble. The framework also meets industrial requirements, including transparency for decision-making [46], automation for minimal manual recalibration, and robustness to varying batch sizes common in HMLV production.

At its core, the framework uses RPCA to decompose batch-level vibroacoustic data into low-rank background noise and sparse fault-induced anomalies. Batch-level homogeneity allows shared disturbances, such as pressure pulsations and structural resonances, to be suppressed while rare fault features are emphasized. RPCA eliminates the need for prior knowledge of system-specific dynamics, enabling adaptation to variant changes, test rig setups, and operating conditions. The framework builds on fleet-based CM methods but adapts them to withstand HMLV-associated variance.

As depicted in Figure 6, ASD-BOD operates in four sequential stages.

1. **Data aggregation and preprocessing.** Vibroacoustic signals are normalized, filtered, and converted to spectral representations for batch-wise comparison.
2. **ASD.** RPCA identifies shared low-rank noise and isolates sparse fault signals for analysis.
3. **Dissimilarity indication.** Dissimilarity between the sparse frequency-domain representations is computed.
4. **Clustering-based OD.** A hierarchical clustering algorithm uses weighted cophenetic coefficients to detect outliers independently of batch size.

This design enables batch-specific fault detection without historical training data and without per-product or per-batch parameter re-tuning. Remaining hyperparameters, such as minimum cluster size fraction and clustering threshold range, stay fixed across all products and batch sizes considered here, making

the framework scalable and suitable for HMLV EoLT environments.

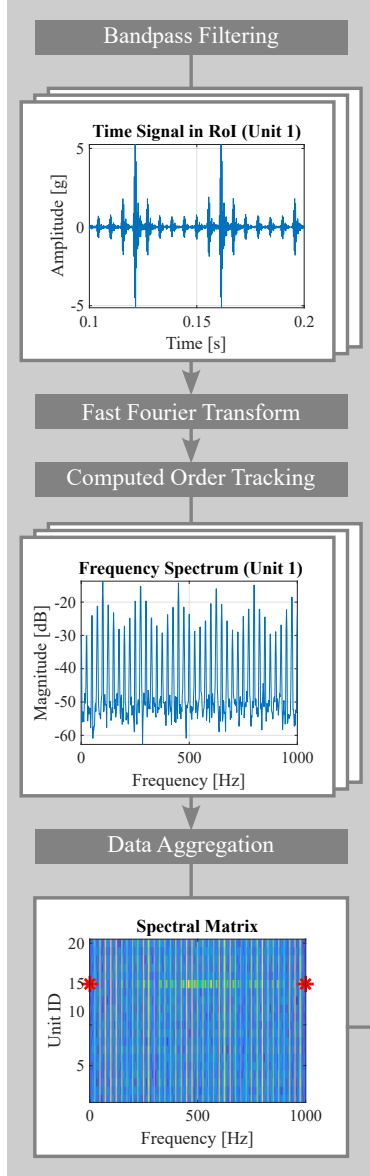
3.1 Data Aggregation and Preprocessing

The first stage of the ASD-BOD framework transforms raw vibroacoustic time-series data $x^{(i)}(t)$ from the i -th unit into a batch-level spectral matrix $\mathbf{M}$. Operational stationarity is assumed, ensuring constant operating conditions such as speed and pressure during testing, which is typically guaranteed in standardized EoLT programs [12]. While non-stationary conditions (e.g., run-ups or speed sweeps [13]) require time-frequency spectrograms or order-domain analysis, this paper focuses on stationary conditions. Under stationary conditions, faults often manifest as distinct features in the frequency domain, enabling compact spectral representations. To counter spectral smearing caused by minor speed variations, order tracking can normalize frequencies. This can be achieved through encoder-driven angular resampling or through computed order tracking using concurrent speed signals. Both methods ensure a meaningful, standardized frequency representation across all tested units.

The transformation into the frequency domain is performed using the Fast Fourier Transform (FFT), applied to windowed signal segments with a Hanning window to reduce spectral leakage. The focus is placed on a specific spectral region of interest (RoI), typically defined using domain knowledge to target frequency bands where faults are likely to occur, such as regions sensitive to high-frequency bearing faults. Fault events are isolated using a band-pass filter F_b , yielding filtered signals $x_{\text{filt}}^{(i)}(t)$. From these, FFT-based magnitudes $A_{\text{filt}}^{(i)} \in \mathbb{R}^m$ are computed for subsequent analysis. Compared to Welch's Power Spectral Density method, which averages noise over segments, the FFT is favored here for its ability to preserve transient fault signatures in individual components [15]. Envelope analysis, often used to highlight bearing fault repetition rates and enhance the signal-to-noise ratio (SNR) [15], is another alternative. However, to maintain general applicability across fault types, the framework employs filtered spectra, even though envelope analysis could augment SNR in specific cases.

Each spectral vector $A_{\text{filt}}^{(i)}$ contains m magnitude values corresponding to frequency bins $[f_1, f_2, \dots, f_m]$ within the RoI. These are aggregated for n units by

Data Preprocessing and Aggregation



Adaptive Spectral Decomposition

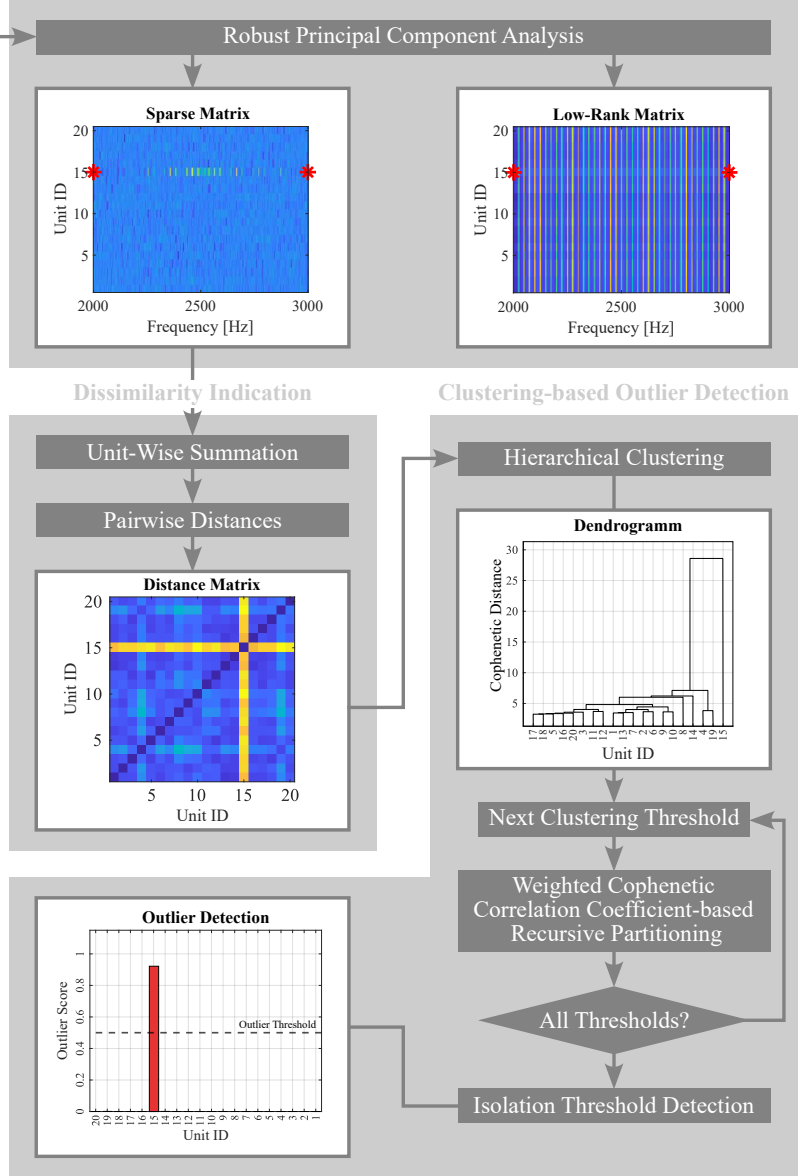


Fig. 6: Detailed overview of the individual processing stages of ASD-BOD

stacking them row-wise into the batch spectral matrix:

$$\mathbf{M} = \begin{bmatrix} A_{\text{filt}}^{(1)} \\ \vdots \\ A_{\text{filt}}^{(n)} \end{bmatrix} = \begin{bmatrix} a_1^{(1)} & a_2^{(1)} & \cdots & a_m^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{(n)} & a_2^{(n)} & \cdots & a_m^{(n)} \end{bmatrix} \in \mathbb{R}^{n \times m}. \quad (2)$$

Here, each row of $\mathbf{M}$ represents a unit's spectrum, while each column corresponds to a specific frequency

bin shared across all units. This spectral matrix serves as the basis for the subsequent stages of the ASD-BOD framework.

3.2 Adaptive Spectral Decomposition via RPCA

The core of the ASD-BOD framework is the separation of batch-common spectral structures from unit-specific vibroacoustic anomalies using RPCA [32]. As shown in Figure 6, RPCA decomposes the batch spectral matrix $\mathbf{M}$ into a low-rank component $\mathbf{L}$, representing shared background noise, and a sparse component $\mathbf{S}$, isolating localized anomalies, such that $\mathbf{M} = \mathbf{L} + \mathbf{S}$. The low-rank component captures batch-shared features such as pressure pulsations, turbulent flow noise, and structural resonances, while sparse outliers represent rare fault-specific patterns. RPCA’s data-driven nature requires no prior knowledge of fault frequencies, kinematic models, or product-specific parameters, making it ideal for HMLV production environments.

RPCA addresses the limitations of traditional model-based preprocessing methods such as comb filtering [47], which are effective at suppressing deterministic harmonics but struggle with variable broadband noise. Comb filters need precise parameter tuning for harmonics, frequencies, and filter properties, which are challenging to maintain in dynamic production settings involving product variants and varying operational conditions. Similarly, techniques like wavelet denoising, baseline correction, or spectral subtraction [47, 48] require extensive parameterization and manual adjustment, which is not scalable for constantly changing batches. In contrast, RPCA automatically infers the dominant background structure and stochastic envelopes from batch data, removing the need for manual configuration and enabling robust noise suppression. These advantages make RPCA a compelling choice for ASD in HMLV settings.

APU spectral data exemplifies the need for such approaches. Compared to relatively clean spectra from EDUs, APU spectra exhibit more complex backgrounds, including strong broadband noise, pressure pulsations, fluid dynamics, and cavitation [19]. This noise is further complicated by batch-to-batch variabilities caused by tolerances in assembly, fluid properties, or mounting conditions. At the same time, deterministic peaks in APU spectra corresponding to harmonics of the pump’s piston frequency and sidebands are critical components of normal operation, even though their magnitudes may vary due to production tolerances. RPCA’s optimization process suppresses the broadband envelope and preserves these deterministic features while isolating sparse anomalies, such as bearing fault signatures, in a data-driven manner.

Before decomposition, spectral magnitudes are converted to a logarithmic (decibel) scale to compress

dynamic ranges and ensure anisotropic noise variance is approximately equalized across the spectrum [15]. The decomposition is achieved by solving the following optimization problem:

$$\min_{\mathbf{L}, \mathbf{S}} \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 \quad \text{subject to} \quad \mathbf{M} = \mathbf{L} + \mathbf{S}. \quad (3)$$

Here, the nuclear norm $\|\mathbf{L}\|_*$ minimizes the rank of $\mathbf{L}$ (capturing shared background), while the ℓ_1 -norm $\|\mathbf{S}\|_1$ ensures sparsity in $\mathbf{S}$ (isolating anomalies). The regularization parameter λ is adaptively set as $\lambda = 1/\sqrt{\max(n, m)}$ [32], which dynamically balances the trade-off between low-rank and sparse components without manual tuning.

An important requirement is that the number of frequency bins m significantly exceeds the batch size n ($m \gg n$). Contrary to classical methods, RPCA benefits from this so-called curse of dimensionality, as physical systems exhibit low-dimensional subspaces within high-dimensional data. This ensures that shared spectral structures concentrate in the low-rank $\mathbf{L}$, while sparse anomalies such as fault signatures remain distinguishable. Such characteristics make RPCA particularly effective for automated fault isolation and robust detection under the complexities of HMLV production [19, 32].

3.3 Dissimilarity Indication

Following the decomposition, the sparse matrix $\mathbf{S} \in \mathbb{R}^{n \times m}$ contains the residual spectral information within the RoI, where each row $S_i \in \mathbb{R}^m$ represents the spectral deviation of the i -th unit from the learned low-rank structure. The subsequent stage quantifies the abnormality of these residuals without relying on feature extraction, directly analyzing the full residual spectra. This avoids the loss of fault-relevant information and the need for domain-specific tuning often required in feature-based approaches, making it highly suited for unsupervised EoLT in HMLV settings.

To characterize overall sparsity, a summation across the frequency bins within each residual is used:

$$SI_i = \sum_{j=1}^m |s_{ij}|, \quad (4)$$

where s_{ij} denotes the element at the i -th row and j -th column of $\mathbf{S}$. Under ideal decomposition, $\mathbf{L}$ captures all batch-common patterns, leaving $\mathbf{S}$ with only unit-specific deviations. Healthy units yield $SI_i \approx 0$, while anomalous units exhibit elevated values due to significant non-zero entries at fault-specific frequencies. This scalar sparsity indicator (SI) is computationally efficient and interpretable, serving as a robust measure of

general abnormality. The dissimilarity between units is then computed using absolute differences of their SI values:

$$d_{ij}^{\text{SI}} = |\text{SI}_i - \text{SI}_j|, \quad (5)$$

generating a distance matrix $\mathbf{D}^{\text{SI}}$ (see Figure 6), which serves as input for the clustering stage described in Section 3.4.

While the sparsity indicator effectively reduces spectral data to a scalar, it may lose information about the morphology of anomalies with similar total energy but different distributions. To address this, the Wasserstein distance is introduced as an alternative dissimilarity measure:

$$d_{ij}^{\text{W}} = W_1(S_i, S_j), \quad (6)$$

quantifying the minimal cost of transforming one spectral distribution into another. Unlike simpler measures such as cosine similarity, the Wasserstein distance accounts for the physical proximity of spectral components and is robust to local frequency shifts. By preserving full spectral morphology, it provides superior sensitivity to fault patterns that may not be captured by scalar measures like the sparsity indicator. Both d_{ij}^{SI} and d_{ij}^{W} are evaluated in this work to balance sparsity and morphological characteristics in fault detection.

3.4 Clustering-Based Outlier Detection

To enable transparent and interpretable OD in contrast to black-box models from deep learning, Hendrickx et al. [29] discuss clustering algorithms for fleet-based CM frameworks. Their approach represents outliers as isolated or small clusters. Methods like K-Means require a cluster number, while DBSCAN relies on density parameters and lacks an equally direct hierarchical visualization, motivating the use of hierarchical clustering. Direct unsupervised detectors such as LOF, OCSVM, and Isolation Forest have also been used for batch-wise EoLT [28]. Because their score scales and dependence on ensemble composition differ, their threshold transferability must be evaluated rather than assumed. Hierarchical clustering additionally provides an inspectable dendrogram for traceable decisions.

While hierarchical clustering provides the preferred framework for transparent OD, partitioning the dendrogram into distinct clusters presents challenges. The hierarchy does not inherently specify unique cluster assignments and requires appropriate cutting strategies. Static thresholds at fixed cophenetic heights prove unsuitable for vibroacoustic data in HMLV EoLT scenarios due to varying distance

scales across test configurations. The inconsistency coefficient [50] evaluates links based on relative deviation but requires specification of depth parameters and thresholds lacking direct physical meaning.

The approach by Hendrickx et al. [29] uses recursive partitioning based on the cophenetic correlation (CC) coefficient:

$$r_{\text{CC}} = \frac{\sum (d_{ij} - \bar{d})(c_{ij} - \bar{c})}{\sqrt{\sum (d_{ij} - \bar{d})^2} \sqrt{\sum (c_{ij} - \bar{c})^2}} \quad (7)$$

The CC quantifies the linear correlation between cophenetic distances c_{ij} in the dendrogram and the original pairwise distances d_{ij} , with $\bar{c}_{ij}$ and $\bar{d}_{ij}$ denoting the corresponding mean values. High CC values (close to 1) indicate faithful preservation of distance structure. Outliers inflate this coefficient as isolated points create distinct branches. The recursive algorithm computes CC and splits the dendrogram at the top level into subtrees if CC exceeds threshold thr_{CC} , repeating recursively until CC falls below threshold in all clusters. However, this approach has limitations for HMLV EoLT settings, necessitating modifications that constitute a key contribution of this work.

The following paragraphs detail the substantial methodological adaptations: weighted CC coefficient for batch-size-invariant decisions, modified recursive partitioning with relative minimum cluster size, and threshold variation-based outlier scoring. In accordance with Figure 6, the proposed OD approach operates through sequential steps after hierarchical clustering:

1. **Dendrogram construction:** Hierarchical clustering with average linkage produces the dendrogram, encoded as linkage matrix $\mathbf{Z}$, from the pairwise distance matrix $\mathbf{D}$. $\mathbf{Z}$ records the sequence of merges and serves as input to the recursive partitioning described below.
2. **Threshold variation:** Clustering threshold thr_{CC} is varied across range $[\text{thr}_{\min}, \text{thr}_{\max}]$.
3. **Recursive partitioning:** Modified algorithm evaluates weighted CC coefficient ($r_{w\text{CC}}$) while enforcing minimum cluster size constraints.
4. **Cluster evolution tracking:** Cluster assignments are recorded for each threshold level.
5. **Isolation threshold detection:** Highest threshold at which each instance resides in isolated cluster is identified as raw outlier score.
6. **Score normalization:** Raw scores are normalized to $[0, 1]$ and instances exceeding $\text{thr}_{\text{outlier}}$ are classified as anomalous.

Weighted Cophenetic Correlation Coefficient for Batch-Size-Invariant Partitioning

For varying batch sizes, a critical challenge arises regarding transferability to EoLT. With larger batches, outliers exert a diminishing influence on the CC coefficient in Equation 7. Consequently, a fixed threshold cannot be robustly applied across varying batch sizes, as identical outlier distances may yield different coefficient values. This dependency stems from an imbalance in the number of pairwise comparisons. For p inliers and o outliers, the standard CC implicitly weights all $\binom{p+o}{2}$ pairwise distances equally, such that more frequent intra-cluster pairs dominate the correlation. In particular, the number of intra-cluster pairs $\binom{p}{2} = \mathcal{O}(p^2)$ grows quadratically with p , while the number of inter-cluster pairs $(p \cdot o)$ grows only linearly. For large p and small o , intra-cluster correlations dominate, thereby diluting the contribution of inter-cluster distances that encode the relevant structural deviations.

To address this limitation, we extend the metric to a weighted variant ensuring balanced representation regardless of cluster cardinalities. At each step of the recursive partitioning (see below), a candidate split into two clusters C_1, C_2 is evaluated. Let $|C_1|, |C_2|$ denote their cardinalities within the current subtree of $n_{sub} = |C_1| + |C_2|$ instances. Pairwise distances d_{ij} and cophenetic distances c_{ij} are weighted with factors w_{ij} to ensure equal representation of intra-cluster pairs (within C_1 or C_2) and inter-cluster pairs (C_1-C_2). For any instance $i \in C_u$ with $u \in \{1, 2\}$, we define an inverse size weight $p_i = \frac{1}{|C_u|}$. The preliminary weight for a pair of distinct instances is defined as $w'_{ij} = p_i \cdot p_j$. To eliminate the aforementioned quadratic-versus-linear imbalance, we introduce a scaling factor β_{ij} that neutralizes the dependency on cluster size for intra-cluster pairs:

$$\beta_{ij} = \begin{cases} \frac{2 \cdot |C_u|}{|C_u| - 1} & \text{if } i, j \in C_u \text{ (intra-cluster)} \\ 1 & \text{if } i \in C_1, j \in C_2 \text{ (inter-cluster)} \end{cases} \quad (8)$$

This targeted rescaling ensures that the combined weight of all intra-cluster pairs within C_u sums to a constant, perfectly matching the aggregate weight of the inter-cluster pairs. Note that if $|C_u| = 1$, no distinct intra-cluster pairs exist, rendering the first case obsolete. The final pairwise weights $w_{ij} = w'_{ij} \cdot \beta_{ij}$ are subsequently normalized such that $\sum w_{ij} = 1$. The weighted CC coefficient for the candidate split is then

given by:

$$r_{wCC} = \frac{\sum w_{ij}(d_{ij} - \bar{d}_w)(c_{ij} - \bar{c}_w)}{\sqrt{\sum w_{ij}(d_{ij} - \bar{d}_w)^2} \sqrt{\sum w_{ij}(c_{ij} - \bar{c}_w)^2}}, \quad (9)$$

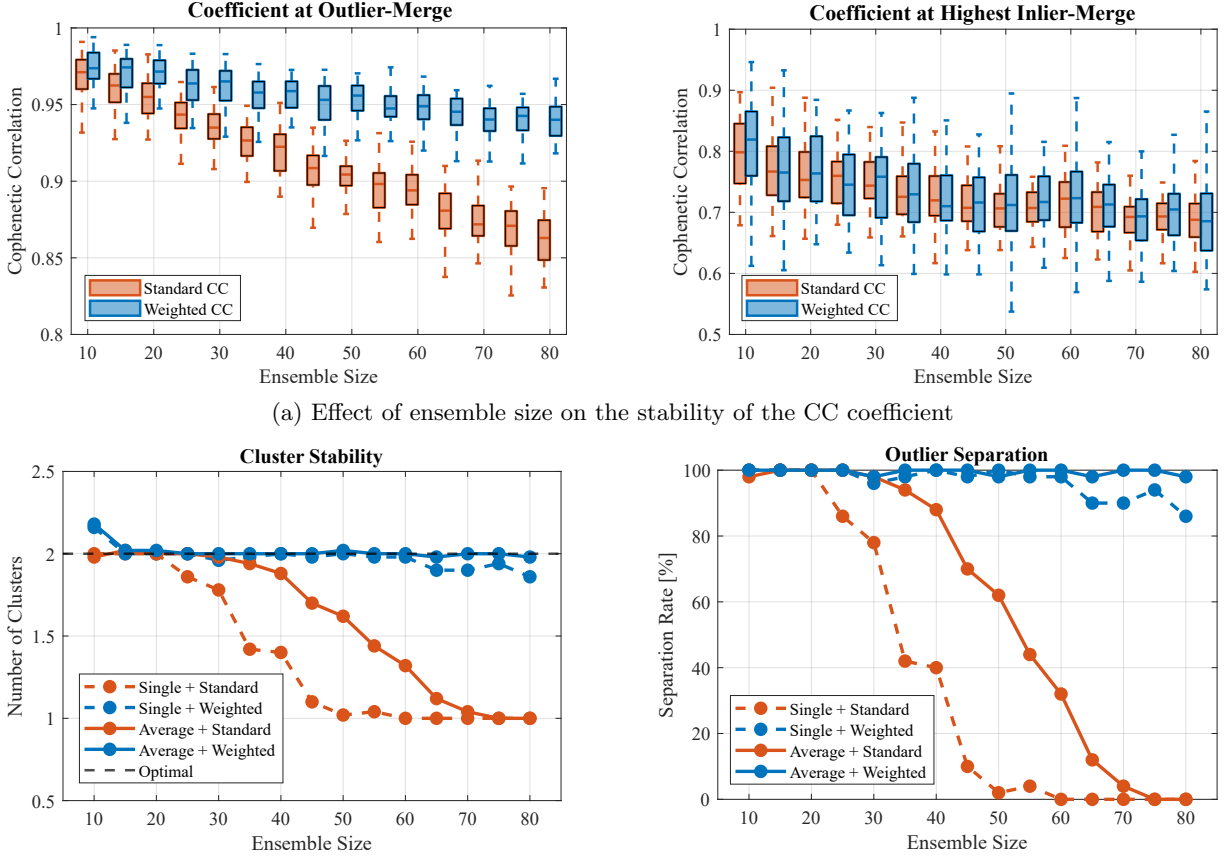
where $\bar{d}_w = \sum w_{ij}d_{ij}$ and $\bar{c}_w = \sum w_{ij}c_{ij}$ denote the weighted means of the pairwise and cophenetic distances, respectively. Note that r_{wCC} is recomputed at every recursion step for the current binary candidate split ($K = 2$), not once for the final set of clusters.

Figure 7a demonstrates the increased stability of the weighted CC compared to the standard coefficient using synthetic data with Gaussian-distributed inliers and a single outlier at a constant distance (z-score = 4). While the standard coefficient exhibits a pronounced dependence on the ensemble (batch) size, the weighted CC shows a largely batch-size-invariant behavior across all evaluated batch sizes. This increased robustness, however, comes at the cost of higher dispersion in the merging of inliers. Figure 7b shows the resulting cluster counts obtained via recursive partitioning with $thr_{CC} = 0.85$. Due to the higher dispersion, weighted CC-based clustering produces a more fine-grained clustering at an ensemble size of 10. Nevertheless, it yields highly stable clustering and outlier separation performance for ensemble sizes above 25 and does not degrade when using a constant clustering threshold. Although single linkage isolates outliers more readily (see [51]), it is more sensitive to batch-size variations. Therefore, average linkage is employed.

Modified Recursive Partitioning with Minimum Cluster Size

While the weighted CC coefficient addresses batch-size invariance, the recursive partitioning procedure requires modification for HMLV EoLT. The algorithm by Hendrickx et al. [29] uses CC as sole splitting criterion regardless of resulting cluster sizes. This leads to excessive fragmentation into statistically invalid clusters. We introduce an explicit minimum cluster size constraint during recursive partitioning: $min_cluster_size = \lceil min_cluster_size_fraction \cdot n \rceil$, where n denotes the total number of instances in the current batch. Clusters with a local cardinality $n_{sub} < min_cluster_size$ are classified as outlier candidates without further partitioning. The relative formulation maintains consistent statistical representativeness across varying batch sizes.

The modified algorithm (Algorithm 1) implements two validation mechanisms. First, before evaluating splitting criteria, it verifies whether current cluster size falls below minimum threshold (Lines 6-8). Second, even if $r_{wCC} > thr_{CC}$ suggests heterogeneity, splitting only executes if at least one resulting



(b) Averaged effect of ensemble size on clustering and outlier separation based on the recursive clustering algorithm from Hendrickx et al. [29] ($thr_{CC} = 0.85$)

Fig. 7: Comparison of standard CC coefficient and proposed weighted extension in OD scenario. Synthetic two-dimensional data with Gaussian-distributed inliers and single outlier at fixed distance (z-score = 4). Results averaged over 50 repetitions.

subcluster satisfies the size requirement (Lines 10-12):

$$\text{Split} \Leftrightarrow \max(|C_1|, |C_2|) \geq \min_{cluster_size} \quad (10)$$

The modifications formalize distinction between representative clusters and anomaly candidates through explicit relative size threshold during recursive splitting rather than post-hoc assessment as in the original framework. Split feasibility checks prevent fragmentation into exclusively non-representative subclusters. The relative formulation of $\min_{cluster_size_fraction}$ combined with r_{wCC} provides dual robustness through batch-size-invariant split decisions and adaptive representativeness requirements.

Outlier Score based on Variable Clustering Threshold

While relative sizes of identified clusters can serve as outlier scores [29], this approach is independent of actual distance from main population. Distance-sensitive scoring through exponential normalization [52] introduces additional hyperparameter α lacking direct physical interpretation. We propose systematic variation of clustering threshold thr_{CC} . Strong outliers become isolated at high thr_{CC} values near 1.0, while weak outliers remain grouped until lower thr_{CC} values. Tracking at which threshold each instance first becomes isolated yields continuous distance-sensitive outlier scores.

Algorithm 1 Modified recursive partitioning with relative minimum cluster size

```
1: procedure PARTITIONMODIFIED( $\mathbf{Z}, \mathbf{D}, thr_{CC}, min\_cluster\_size\_fraction, n$ )
2:    $n_{sub} \leftarrow$  number of samples in current subtree  $\mathbf{Z}$  ▷ Local size
3:    $min\_cluster\_size \leftarrow \lceil min\_cluster\_size\_fraction \times n \rceil$  ▷ Global size constraint
4:    $leaves_1, leaves_2 \leftarrow$  children of root in  $\mathbf{Z}$ 
5:    $n_1 \leftarrow |leaves_1|, n_2 \leftarrow |leaves_2|$ 
6:   if  $n_{sub} < min\_cluster\_size$  or  $n_{sub} \leq 2$  then
7:     return  $\mathbf{0}_{n_{sub}}$ 
8:   end if
9:   if  $n_1 < min\_cluster\_size$  and  $n_2 < min\_cluster\_size$  then
10:    return  $\mathbf{0}_{n_{sub}}$ 
11:   end if
12:    $r_{wCC} \leftarrow$  weighted_cophenetic_correlation( $\mathbf{Z}, \mathbf{D}$ )
13:   if  $r_{wCC} > thr_{CC}$  then
14:      $\mathbf{D}_1, \mathbf{D}_2 \leftarrow$  submatrices of  $\mathbf{D}$  restricted to  $leaves_1, leaves_2$ 
15:      $\mathbf{Z}_1 \leftarrow$  linkage( $\mathbf{D}_1$ ),  $\mathbf{Z}_2 \leftarrow$  linkage( $\mathbf{D}_2$ )
16:      $labels_1 \leftarrow$  PARTITIONMODIFIED( $\mathbf{Z}_1, \mathbf{D}_1, thr_{CC}, min\_cluster\_size\_fraction, n$ )
17:      $labels_2 \leftarrow$  PARTITIONMODIFIED( $\mathbf{Z}_2, \mathbf{D}_2, thr_{CC}, min\_cluster\_size\_fraction, n$ )
18:     return  $labels_1, labels_2$  with offset  $\max(labels_1) + 1$  added to  $labels_2$ 
19:   else
20:     return  $\mathbf{0}_{n_{sub}}$ 
21:   end if
22: end procedure
```

The procedure (Algorithm 2) constructs an evolution matrix $\mathbf{E} \in \{0, 1, \dots\}^{n_{thr} \times n}$ by performing recursive partitioning across a discrete range of thresholds $thr_{CC} \in [thr_{min}, thr_{max}]$ (see Figure 8). Each entry $e_{k,i}$ represents the cluster label assigned to instance i at threshold $thr_{CC}^{(k)}$. For any given threshold index k , the cluster containing instance i is defined by the set $C_{k,i} = \{j : e_{k,j} = e_{k,i}\}$. To derive the outlier score, instances are initially assigned a baseline score of thr_{min} . Assuming the thresholds are indexed in ascending order ($thr_{CC}^{(1)} < \dots < thr_{CC}^{(n_{thr})}$), the algorithm identifies the highest threshold at which instance i still resides in an isolated cluster. This raw score is determined by evaluating the cluster cardinality $|C_{k,i}|$:

$$raw_score_i = thr_{CC}^{(k^*)}, \quad k^* = \max\{k \mid |C_{k,i}| < min_cluster_size\} \quad (11)$$

Raw scores are normalized to $[0, 1]$:

$$outlier_score_i = \frac{raw_score_i - thr_{min}}{thr_{max} - thr_{min}} \quad (12)$$

An instance is classified as an outlier if $outlier_score_i > thr_{outlier}$. This approach offers three advantages over cluster-size-based scoring. First, it is distance-sensitive reflecting how distinctly instances separate from main population. Second,

it provides continuous scores across $[0, 1]$ range. Third, it avoids additional hyperparameters beyond $min_cluster_size_fraction$.

4 Experimental Studies

This section evaluates ASD-BOD through two controlled comparisons on the real APU data. First, the ASD and dissimilarity stages are ablated while retaining the proposed clustering backend. Second, this RPCA-Sparsity front-end is held constant to assess whether anomaly-score thresholds remain useful when the batch size changes. The latter is not intended as a general OD ranking, but as a deployment-oriented threshold-transfer test.

4.1 Evaluation Methodology and Benchmarking Design

To evaluate ASD-BOD rigorously, a multi-step methodology is employed, as illustrated in Figure 9. Its foundation is statistical validation of results by analyzing performance across multiple dataset configurations. This standardized evaluation process is applied to the proposed framework and all benchmark approaches, ensuring fair comparison. The first crucial step is use of multiple data acquisitions, where repetitions correspond to distinct data batches, so that each batch provides an independent outcome under specific defined conditions.

Algorithm 2 Outlier score computation via threshold variation

```

1: procedure COMPUTEOUTLIERScores( $\mathbf{D}, thr\_range, min\_cluster\_size\_fraction$ )
2:    $thr_{min}, thr_{max} \leftarrow thr\_range$ 
3:    $n \leftarrow$  total number of instances implied by  $\mathbf{D}$  ▷ Global size
4:    $thresholds \leftarrow discretize(thr_{min}, thr_{max}, step)$ 
5:    $n_{thr} \leftarrow |thresholds|$ 
6:    $min\_cluster\_size \leftarrow min\_cluster\_size\_fraction \cdot n$ 
7:    $cluster\_matrix \leftarrow zeros(n_{thr} \times n)$ 
8:   for  $k = 1$  to  $n_{thr}$  do
9:      $\mathbf{Z} \leftarrow linkage(\mathbf{D})$ 
10:     $labels \leftarrow PARTITIONMODIFIED(\mathbf{Z}, \mathbf{D}, thresholds[k], min\_cluster\_size\_fraction, n)$ 
11:     $cluster\_matrix[k, :] \leftarrow labels$ 
12:  end for
13:   $raw\_scores \leftarrow ones(n) \times thr_{min}$ 
14:  for  $i = 1$  to  $n$  do
15:    for  $k = n_{thr}$  down to  $1$  do ▷ Search high to low
16:       $cluster\_id \leftarrow cluster\_matrix[k, i]$ 
17:       $cluster\_size \leftarrow count(cluster\_matrix[k, :] == cluster\_id)$ 
18:      if  $cluster\_size < min\_cluster\_size$  then
19:         $raw\_scores[i] \leftarrow thresholds[k]$ 
20:        break
21:      end if
22:    end for
23:  end for
24:   $outlier\_scores \leftarrow \frac{raw\_scores - thr_{min}}{thr_{max} - thr_{min}}$ 
25:  return  $outlier\_scores$ 
26: end procedure

```

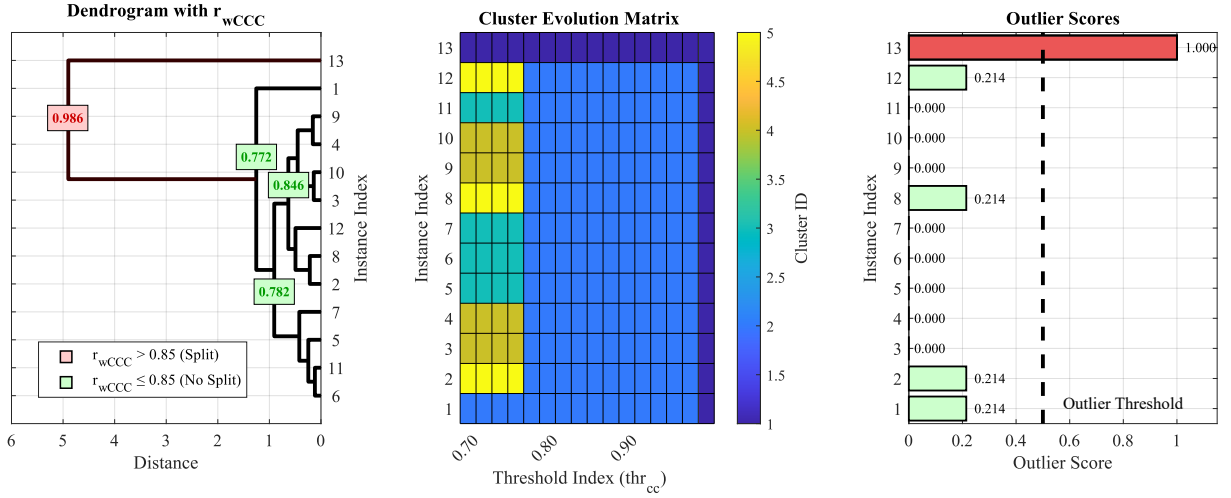


Fig. 8: Derivation of outlier score based on threshold variation of the weighted CC coefficient, illustrated using exemplary dendrogram for synthetic two-dimensional data. Recursive partitioning is shown for $thr_{CC} = 0.85$ in the dendrogram. The relative minimum cluster size is set to $min_{cluster_size_fraction} = 0.25$ and the outlier decision threshold to $thr_{outlier} = 0.5$.

As depicted in Figure 9, a central paradigm is that outlier score for any data point is calculated based

solely on information within its respective batch. However, for comprehensive evaluation, a global perspective is necessary regarding outlier threshold. Instead of

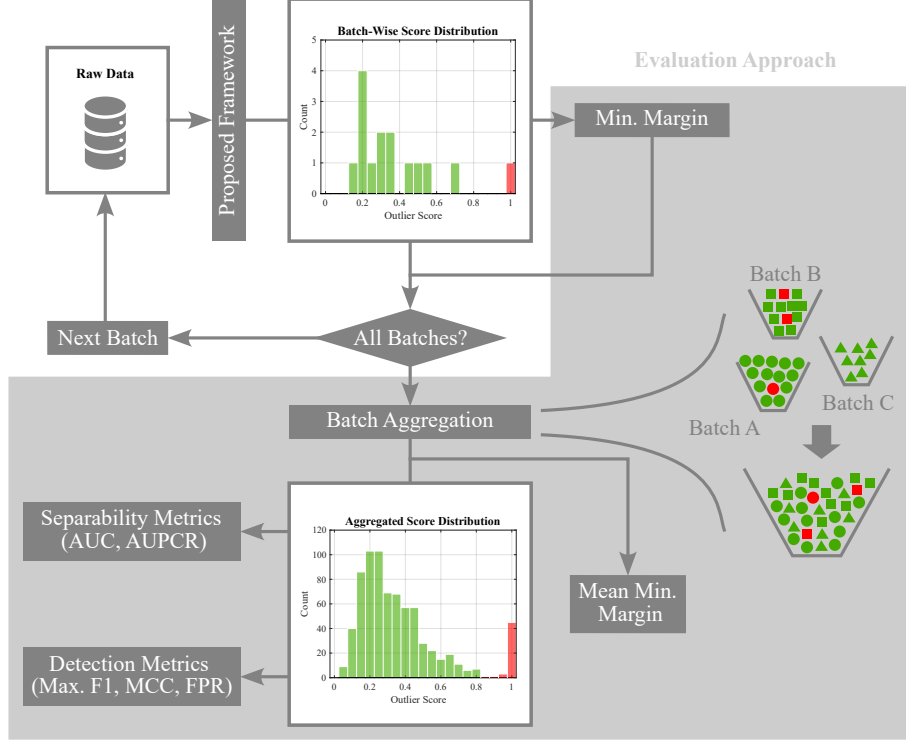


Fig. 9: Overview of the evaluation approach. The framework processes raw data one batch at a time, generating outlier scores independently. For robust assessment, scores from all batches are aggregated, since a single global outlier threshold is necessary for real-world application.

applying $thr_{outlier}$ to each batch individually (impractical in a real-world setting), outlier scores from all N_{reps} batches are aggregated into single combined distribution. This evaluation strategy allows holistic assessment of pipeline quality based on score separability. By pooling results, globally consistent performance metrics from this aggregated distribution are computed, summarized in Table 1.

4.1.1 Performance Metrics

The primary objective is achieving high separability between faulty instances and healthy instances. The most direct measure is separation margin. For each j -th repetition or batch, minimum margin is:

$$\text{Min. Margin}_j = \text{outlier_score}_{j,\text{faulty}} - \max(\text{outlier_score}_{j,\text{healthy}}) \quad (13)$$

To visualize general trends, Mean Min. Margin across repetitions is calculated. While margin evaluates separation to nearest neighbor, threshold-independent metrics capture quality of entire ranking. The Area Under the ROC curve (AUC) measures probability

that a randomly chosen faulty instance receives higher score than a randomly chosen healthy instance. For highly imbalanced data, Area Under the Precision-Recall Curve (AUPRC) is more telling. Precision is fraction of true faults among instances classified as outliers, while Recall is fraction of detected faults among all actual faults.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \quad (14)$$

Here, TP denotes True Positives, FP False Positives, and FN False Negatives. AUPRC summarizes performance across all possible thresholds.

In practice, a fixed decision threshold $thr_{outlier}$ must be chosen. To assess maximum potential, an optimal threshold is determined by maximizing F1-Score on aggregated scores:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (15)$$

At this optimal point, Matthews Correlation Coefficient (MCC) and False Positive Rate (FPR) are also

computed. MCC is robust on imbalanced data:

$$\text{MCC} = \frac{\text{TP TN} - \text{FP FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})}} \times \frac{1}{\sqrt{(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (16)$$

4.1.2 Benchmarking and Ablation Design

To systematically validate contribution of RPCA-based ASD, we employ a structured ablation study isolating performance impact of individual processing stages. This addresses two questions:

1. Does RPCA-based decomposition provide quantifiable improvements in fault-background separability compared to simpler preprocessing alternatives or no preprocessing?
2. How does choice of dissimilarity measure (morphology-preserving Wasserstein distance vs. sparsity indicator) affect detection performance on decomposed spectra?

We deconstruct the framework along the ASD and dissimilarity indication stage while holding clustering-based OD stage constant across all configurations. The raw configuration establishes absolute baseline where no preprocessing is applied. ResMean represents common practice in vibroacoustic EoLT [13], where per-frequency mean across all units is subtracted. ResMed provides robust alternative by subtracting per-frequency median. RPCA constitutes the proposed core method. Sequential combinations ResMean-RPCA and ResMed-RPCA test whether preliminary removal of global spectral trends simplifies optimization landscape for RPCA.

Two dissimilarity quantification approaches are evaluated: Sparsity Indicator (Sparsity), which reduces each residual spectrum to its total ℓ_1 -norm energy, and Wasserstein Distance (Wass), which preserves full spectral morphology. Systematic combination of five decomposition variants with two dissimilarity measures yields twelve benchmark pipelines shown in Table 2. Pipeline naming follows convention Decomposition-Indicator.

In a complementary backend benchmark, the RPCA sparsity indicator from Equation (4) is held fixed and six OD methods are compared. The pairwise SI distances follow Equation (5). This retains the same RPCA based input and batch compositions for every backend. The proposed weighted CC method from Section 3.4 is evaluated together with StandardCC,

kNN-Distance, Isolation Forest, RobustZ and the original Hendrickx procedure. The selection represents distinct principles for assigning an anomaly score.

The proposed method combines the weighted coefficient in Equation (9) with the threshold variation score from Section 3.4. StandardCC replaces this coefficient with the unweighted CC in Equation (7). It retains the same recursive partitioning, threshold sweep and bounded score mapping. It therefore isolates the effect of cardinality weighting within the proposed hierarchical approach. kNN-Distance scores a unit by the mean Euclidean distance to its nearest neighbours in the SI distance matrix. It is included as a transparent proximity based detector whose threshold remains in native SI units. Isolation Forest directly partitions the SI values using random trees and provides a nonlinear generic OD reference. RobustZ measures the absolute deviation from the batch median normalized by the median absolute deviation. It tests whether robust population relative standardization alone is sufficient for threshold transfer. The score is dimensionless but unbounded and has no cluster representation. The Hendrickx baseline applies the unweighted CC in Equation (7) with the original fixed $thr_{CC} = 0.9$. It scores a unit by the fraction of the batch outside its assigned cluster. It is the literature baseline from [29] and differs from StandardCC, which is an internal ablation of the proposed method.

4.1.3 Implementation and Hyperparameter Settings

The evaluation was conducted using MATLAB R2024b. For transformation into frequency domain, FFT is applied after Hanning windowing. For real-world APU data, computed order tracking accounts for speed fluctuations. Raw acceleration signals are resampled to uniform angular increments using speed measurements, with angular resolution $N_{\text{res}} = 2048$ samples per revolution. The order axis is converted to normalized frequencies via $f = o \cdot (n_{\text{ref}}/60)$ and interpolated onto common linear frequency axis. Sixth-order Butterworth bandpass filter is applied prior to order tracking.

The RPCA-based ASD stage introduces three hyperparameters. Regularization parameter λ is set to $\lambda = 1/\sqrt{\max(n, m)}$ as proposed by Candès et al. [32]. Penalty parameter μ follows adaptive scheme by Lin et al. [53], gradually increasing by factor $\rho > 1$ across iterations. Convergence tolerance is $\epsilon_{\text{tol}} = 10^{-7}$.

For clustering-based OD stage, minimum cluster size fraction is $\text{min}_{\text{cluster_size_fraction}} = 1/3$, providing conservative balance between accommodating multiple faults and preventing nominal population

Table 1: Summary of performance evaluation metrics

Category	Metric	Purpose / Definition	Interpretation
Separability	AUC	Overall ranking quality across all thresholds	Higher is better (1.0 = perfect)
Separability	AUPRC	Ranking quality, sensitive to imbalanced data	Higher is better (1.0 = perfect)
Separability	Mean Min. Margin	Average score gap between faulty and best healthy unit	Higher is better (> 0 means separation)
Detection	Max F1-Score	Maximum achievable F1-Score at optimal threshold	Higher is better (1.0 = perfect)
Detection	MCC (@ Max F1)	Robust detection quality at optimal F1 point	Higher is better (+1 = perfect)
Detection	FPR (@ Max F1)	False alarm rate at Max. F1-Score	Lower is better (0 = no false alarms)

Table 2: Overview and rationale of the 12 benchmark pipelines

Preprocessing	Indicator	Pipeline Name	Rationale
—	Wass	Raw-Wass	Baseline: morphology-aware without decomposition
—	Sparsity	Raw-Sparsity	Absolute baseline: no decomposition, sparsity-based
ResMean	Wass	ResMean-Wass	Simple mean-based global suppression
ResMean	Sparsity	ResMean-Sparsity	Mean-based suppression with sparsity indicator
ResMed	Wass	ResMed-Wass	Robust median-based global suppression
ResMed	Sparsity	ResMed-Sparsity	Median-based suppression with sparsity indicator
RPCA	Wass	RPCA-Wass	Core framework: RPCA + morphology preservation
RPCA	Sparsity	RPCA-Sparsity	Core framework: RPCA + sparsity-based indicator
ResMean-RPCA	Wass	ResMean-RPCA-Wass	Potential enhancement: mean preprocessing + RPCA
ResMean-RPCA	Sparsity	ResMean-RPCA-Sparsity	Mean preprocessing variant with sparsity indicator
ResMed-RPCA	Wass	ResMed-RPCA-Wass	Potential enhancement: robust preprocessing + RPCA
ResMed-RPCA	Sparsity	ResMed-RPCA-Sparsity	Robust preprocessing variant with sparsity indicator

fragmentation. For outlier scoring algorithm, threshold range from $thr_{cc,min} = 0.70$ to $thr_{cc,max} = 1.0$ is divided into $n_{steps} = 20$ discrete values using exponential scaling factor $\alpha_{cc} = 3$:

$$thr_{CC}(k) = thr_{cc,min} + (thr_{cc,max} - thr_{cc,min}) \cdot \left(1 - \left(1 - \frac{k-1}{n_{steps}-1}\right)^{\alpha_{cc}}\right), k \in 1, n_{steps} \quad (17)$$

This concentrates higher density of threshold values near $thr_{cc,max}$, providing finer resolution where strong outliers transition from isolation to merging.

4.2 Real-World Case Study on APU Data

To validate ASD-BOD on experimental data and demonstrate its applicability under more realistic conditions, this section introduces a real-world case study. Since no real-world EoLT data from production was available, a dedicated dataset was generated during an extensive test campaign with APUs, explained in a more detailed way in [31]. This laboratory-based approach offers the significant advantage of allowing for the deliberate introduction of faults and their systematic investigation under controlled and repeatable conditions. Otherwise, this task would be infeasible

in a real production environment due to the rarity of such faults.

The following subsections detail the dataset’s origin, structure, and characteristics, followed by the experimental evaluation. Results are interpreted against known APU unit characteristics, validating diagnostic outputs against experimental ground truth. This includes batch size impact analysis through random ensemble sampling and demonstration of the framework’s integrated visual analytics capabilities for intuitive result exploration.

4.2.1 Experimental Test Campaign

The dataset was generated using seven APUs on a laboratory test rig. To simulate assembly-related variations from series production, drive shafts were systematically swapped between unit housings, resulting in 30 distinct healthy unit combinations. Faulty units were created by installing bearings intentionally damaged via abrasive machining, introducing a realistic spectrum of fault severities and types. This paper focuses on 10 specific nOK units exhibiting distinct vibroacoustic signatures, summarized in Table 3. An overview of the setup is illustrated in Figure 10.

The experimental procedure subjects each unit to a series of varying operating conditions. It is a relevant subset of the extensive test campaign detailed in [31]. Based on the findings regarding fault detectability in that study, this work focuses on 12 of the most informative load points, defined by three levels of rotational speed and four levels of high pressure (100, 200, 300, 400 bar), as shown in Table 4. While the proposed framework is broadly intended to address domain shifts arising from batch-to-batch variations in component specifications, this case study utilizes a dataset where such shifts are primarily induced by varying these operating conditions. This approach is nevertheless highly suitable for validation. From the perspective of ASD-BOD, the specific origin of the distribution shift, e.g., from component variations or changing test parameters, is of secondary importance. The fundamental challenge of distinguishing fault-induced anomalies from shifts in the distribution of healthy data remains identical. Furthermore, it is common in real-world EoLT that different production batches may also be subjected to different test programs. Thus, this dataset serves as a valid and representative test case to demonstrate the framework’s capability.

Beyond systematic domain shifts between load points, inherent variance exists within each operating condition due to manufacturing tolerances, assembly variations, and process fluctuations. A primary

contributor is unavoidable fluid temperature variation during testing. These fluctuations constitute realistic process noise mimicking industrial conditions (detailed analysis in [22]).

Vibroacoustic data was captured using a triaxial accelerometer on the test rig flange. Based on prior investigations, the y-axis signal (see Figure 10) is used due to favorable signal quality. Signals were acquired at 50 kHz sampling rate with anti-aliasing filtering, yielding 500,000 samples per 10-second acquisition interval.

4.2.2 Dataset Preprocessing and Exploration

The final dataset comprises 480 frequency domain spectra from 40 distinct APUs (30 healthy and 10 manipulated, faulty units), each measured under 12 load points. Following preprocessing from Section 3.1 and implementation from Section 4.1.3, angular resampling and Hanning windowing were applied before FFT. Ground truth is known for validation purposes, but the subsequent analysis is designed for an unsupervised scenario where such labels are unavailable in practice. The RoI is constrained to 2 kHz to 3 kHz based on empirical knowledge, as this spectral window exhibits strong resonant responses to bearing fault impulses, providing favorable Fault-to-Noise Ratio. Given the 7-piston configuration and 1600 rpm operating speed, piston harmonics are expected at approximately 2053 Hz, 2240 Hz, 2427 Hz, 2613 Hz, 2800 Hz, and 2987 Hz.

As an example, data from load point 8 is examined in linear scale. The heatmap visualization in Figure 11 reveals highly consistent underlying structure across most units, defined by a deterministic grid of spectral lines corresponding to primary piston harmonics (e.g., 2240 Hz) and associated sidebands modulated by the fundamental rotational frequency of 26.67 Hz. This harmonic grid constitutes the baseline vibroacoustic signature of normal APU operation. In stark contrast, faulty units exhibit distinct, high-energy peaks at inharmonic frequencies not associated with this deterministic grid, representing direct manifestations of induced bearing faults. Logarithmic scale inspection, chosen for subsequent RPCA computation, exposes a second significant variation source, which is broadband noise floor variations affecting both healthy (e.g., units 34 and 37) and faulty examples. This manifests as near-uniform vertical shifts in spectral energy, particularly pronounced above 2.5 kHz. Its presence across different health states suggests it is a systemic, process-induced artifact likely stemming from sensor coupling or test rig dynamics variations rather than unit condition indicators.

Table 3: Abrasively machined faults in large tapered rolling bearings used to generate faulty data instances.

Unit ID	Fault description
2	Multiple damages to outer ring raceway
4	Small damage to rolling element
5	Two near damages to outer ring raceway
9	Small damage to outer ring raceway
17	Multiple small damages to rolling elements and outer ring raceway
21	Small damages to multiple rolling elements
25	Damages to multiple rolling elements
28	Wide damage to outer ring raceway
30	Damage to outer ring raceway
39	Multiple damages to rolling elements and outer ring raceway

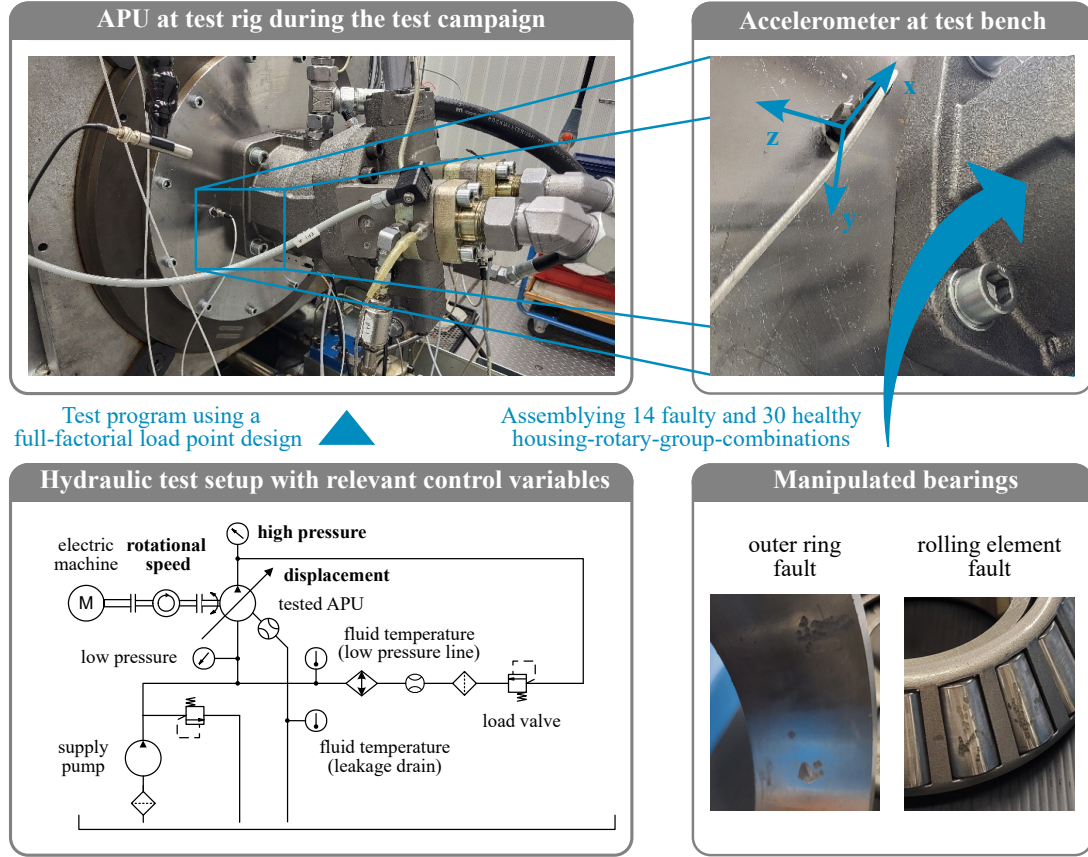


Fig. 10: Overview of the setup and design for the experimental test campaign, adapted from [22]. Note that not all data from the test campaign is used for this study (in contrast to [31]), as the analysis is restricted to a single displacement and three rotational speed levels. Moreover, only 10 faulty units are considered.

To model the healthy state without contamination, quantitative analyses are performed exclusively on 30 known healthy units. This prevents the normalcy model from being initially corrupted by faulty

unit spectral components. First, inter-frequency bin correlation was computed for this healthy-only subset. The resulting correlation matrix in Figure 11 shows distinct high-correlation blocks aligning with

Table 4: The 12 relevant load points of the testing procedure considered in this paper.

Load Point No.	1	2	3	4	5	6	7	8	9	10	11	12
Rotation Speed [rpm]	1200	1200	1200	1200	1600	1600	1600	1600	2000	2000	2000	2000
High Pressure [bar]	100	200	300	400	100	200	300	400	100	200	300	400

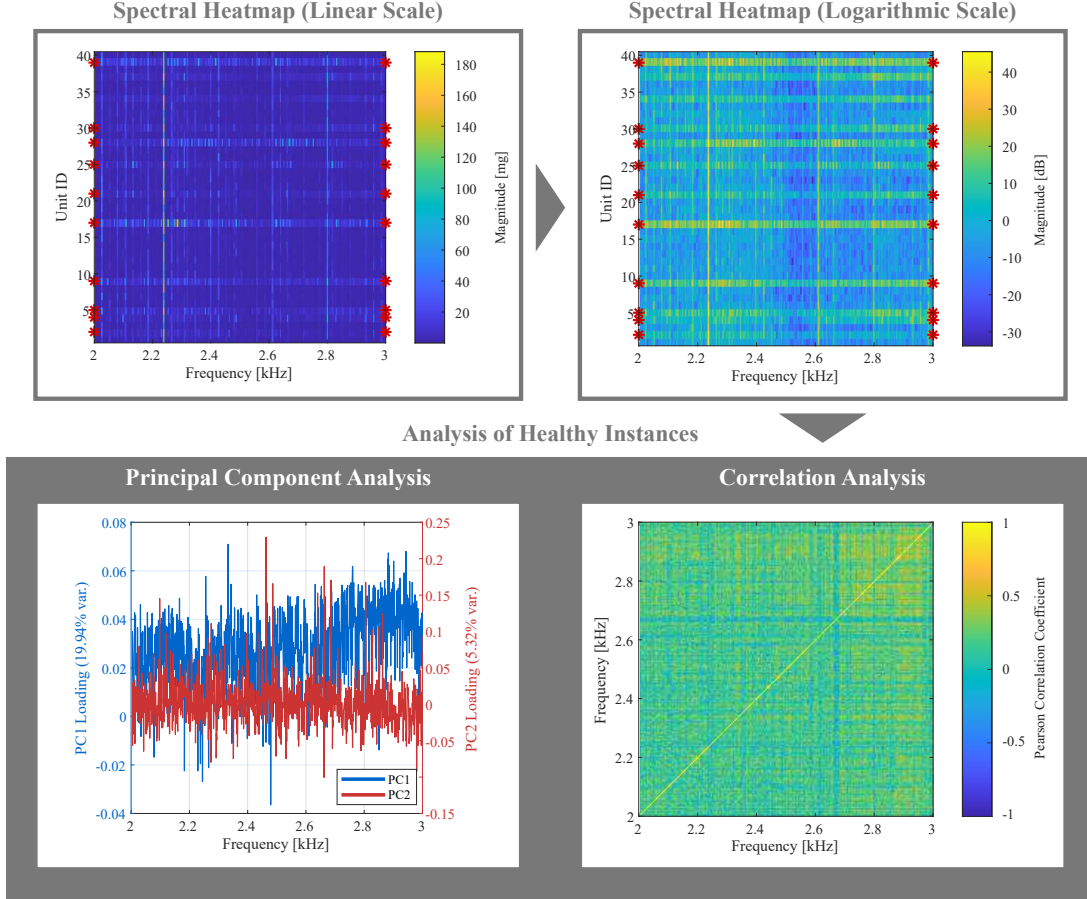


Fig. 11: Exploration of the processed real-world case study data set in the frequency domain

noise floor variation frequency bands under 2.7 kHz, confirming these spectral shifts represent structurally low-rank components rather than stochastic noise. Second, PCA conducted on this healthy-only dataset reveals persistent heterogeneity even without faulty units. PC1 captures disproportionately large variance (19.94%), followed by sharp decrease for PC2 (5.32%). PC loading examination confirms PC1 is almost exclusively characterized by broadband noise artifacts, while PC2 captures variance from deterministic piston harmonics and sidebands. This focused analysis demonstrates that even the healthy-only dataset is

heterogeneous, comprising information-rich harmonic signatures and low-rank process-induced noise.

Figure 12 examines load point influence on individual units using spectral energy (sum of FFT magnitudes) as a characteristic feature, analogous to the sparsity indicator from RPCA decomposition. Notably, faulty units 9, 17, and 39 exhibit comparatively high spectral energy across all load points, with induced faults generating vibroacoustic anomalies of such severity that they manifest distinctly in overall signal intensity. Conversely, faulty units 2, 21, and 30 are much more difficult to distinguish from healthy

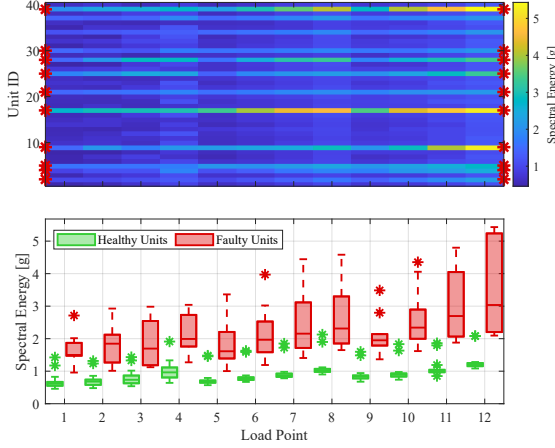


Fig. 12: Load point analysis based on spectral energy for the real-world dataset with a heatmap relating each unit across all load points (top) and a boxplot comparing the distributions (bottom)

units based solely on spectral energy, particularly considering baseline noise variations in healthy units 34 and 37.

The distribution in Figure 12 reveals that spectral energy increases systematically with both rotational speed and pressure, which is physically plausible as vibroacoustic energy scales with mechanical power processed by the unit. However, this creates significant domain shifts confounding simple fault detection, as a faulty unit at low load can exhibit similar spectral energy to a healthy unit at high load. This illustrates the fundamental limitation of broadband features applied to raw data, as they aggregate signal energy over the entire frequency domain, rendering them insensitive to narrowband characteristics that are often key bearing fault indicators. Thus, utilizing a single model across different test conditions is unfavorable due to load-induced domain shifts fundamentally altering data distribution. Furthermore, spectral energy alone causes misclassification when bearing faults lack strongly pronounced spectral peaks and baseline noise levels fluctuate significantly across the nominally healthy population.

4.2.3 ASD and Dissimilarity Ablation Results

To evaluate fault detection performance under realistic conditions, a structured experimental design is employed. The objective is to enable statistically significant assessment of detection algorithm variance and repeatability by systematically controlling confounding variables through unique, fault-specific

batch templates. A distinct set of 100 batch templates is generated for each of the 10 faulty units, deliberately chosen to maximize healthy unit combination diversity and eliminate single-template bias. Within each set, batch size distribution emulates HMLV production scenarios, ranging from 8 to 30 units (upper limit of healthy units) partitioned into five equally probable bins of 20% each, ensuring structured and reproducible variety without favoring specific sizes.

To prevent sampling bias, a balanced selection mechanism is implemented during template generation. The algorithm maintains a usage count for every healthy unit and prioritizes selecting those used least frequently, ensuring all 30 healthy units appear with approximately equal frequency within each 100-template set. Each set is divided into 50 fault-free and 50 contaminated batches, the latter containing a randomized placeholder for the specific faulty unit being tested. This design reflects the rare fault assumption while facilitating focused analysis of individual fault detectability. Templates are systematically reused across all 12 load points. Although unit IDs remain identical within a template, the underlying spectral data differs for each load point, representing unique test cases. This methodologically sound approach enables direct assessment of algorithm robustness to changing operating conditions, while randomization across different fault-specific templates ensures overall results are not skewed by single compositions. The resulting comprehensive dataset comprises 12,000 unique batches (10 fault sets $\times$ 100 templates $\times$ 12 load points). It should be noted that these 12,000 batches are not independent observations in a statistical sense, as they are repeatedly drawn from the same pool of 30 healthy and 10 faulty physical units. The large batch count primarily ensures broad coverage of healthy-unit combinations and operating conditions.

The global aggregated detection performance, evaluated across all 12,000 generated batches, is summarized in Table 5, with corresponding ROC and Precision-Recall curves shown in Figure 13. A primary observation is the general superiority of sparsity-based methods over their Wasserstein-based counterparts. Based on metrics most relevant to imbalanced classification, the RPCA-Sparsity pipeline emerges as the best-performing approach. It achieves the highest AUPRC (0.887) and the highest F1-score (0.842) at an exceptionally low FPR of just 0.230%. In comparison, the ResMed-RPCA-Sparsity pipeline follows at considerable distance, with an AUPRC of 0.568 and an F1-score of 0.508, indicating that the combination of residual preprocessing with RPCA and a sparsity indicator is ineffective in this real-world case study.

A closer examination reveals a notable disparity in the magnitude of improvement provided by

Table 5: Globally aggregated performance metrics for the real-world dataset. Best (**bold**), second-best (underlined), and third-best (*italic*) performers are highlighted per column.

Pipeline	AUC	AUPRC	Opt. Thr.	Max. F1	MCC	FPR [%]
Raw-Wass	0.660	0.096	0.284	0.153	0.133	4.851
Raw-Sparsity	0.959	0.421	0.680	0.422	0.476	6.000
ResMean-Wass	0.695	0.109	0.403	0.195	0.174	3.519
ResMean-Sparsity	0.965	0.383	0.748	0.422	0.489	6.655
ResMed-Wass	0.712	0.122	0.403	0.203	0.183	3.682
ResMed-Sparsity	0.965	0.396	0.748	0.421	0.487	6.635
RPCA-Wass	0.725	0.137	0.403	0.221	0.201	3.154
RPCA-Sparsity	<u>0.973</u>	0.887	0.806	0.842	0.840	0.230
ResMean-RPCA-Wass	0.730	0.149	0.403	0.228	0.207	2.817
ResMean-RPCA-Sparsity	<i>0.972</i>	<i>0.513</i>	0.950	<i>0.466</i>	0.453	<i>1.810</i>
ResMed-RPCA-Wass	0.744	0.172	0.403	0.239	0.219	2.852
ResMed-RPCA-Sparsity	0.974	<u>0.568</u>	0.950	<u>0.508</u>	<u>0.496</u>	<u>1.269</u>

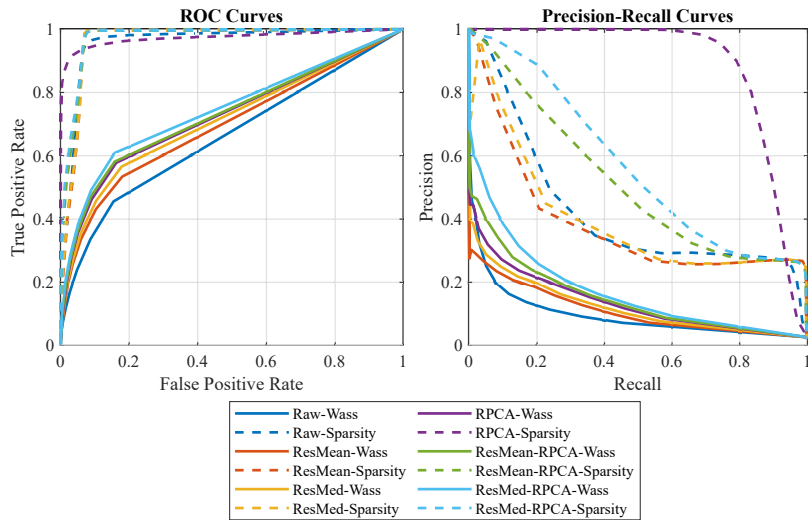


Fig. 13: ROC and Precision-Recall curves for the real-world dataset

RPCA-based methods when comparing threshold-independent metrics AUC and AUPRC. While the RPCA-Sparsity pipeline clearly excels in AUPRC, its advantage in AUC is comparatively small, especially when contrasted with methods such as ResMean-Sparsity and ResMed-Sparsity. In fact, ResMed-RPCA-Sparsity even achieves a slightly higher AUC. This outcome is a direct consequence of the substantial class imbalance inherent in the experimental design, where each contaminated batch contains only a single faulty unit. In such scenarios, AUC can appear misleadingly optimistic, as it is heavily influenced by the large number of correctly classified true

negatives. Consequently, AUPRC offers a more practical and informative measure of performance for this application.

Visual inspection of the performance curves in Figure 13 further illustrates this trade-off. The ROC curves show that all other sparsity-based pipelines require substantially higher FPR (exceeding 8%) to achieve TPR levels comparable to RPCA-Sparsity, which artificially inflates their overall AUC. A major underlying factor is the misclassification of units 34 and 37, which are healthy but deviate strongly from the remainder of the healthy ensemble due to high noise. Without RPCA, these units are incorrectly

flagged as anomalies, inflating the false positive rate. In contrast, the Precision-Recall curves more clearly highlight the practical superiority of RPCA-Sparsity, particularly in the low-recall regime corresponding to the low-FPR operating range most relevant for deployment. RPCA-Sparsity maintains high precision across a broad range of recall values, substantially outperforming all alternative pipelines. This sustained high precision directly explains its leading AUPRC and superior performance in threshold-dependent metrics such as F1-score. Overall, while all sparsity-based approaches achieve strong aggregate detection performance, the exclusive use of RPCA as a preprocessing step significantly enhances practical reliability.

To provide a disaggregated view of the global results under realistic conditions, the performance of each pipeline is evaluated using the single global optimal threshold determined previously (Table 5). By applying a fixed global threshold here, we approximate a more realistic deployment scenario in which a single calibration must operate robustly across all operating conditions. The resulting performance, broken down by batch size and load points, is illustrated in Figure 14.

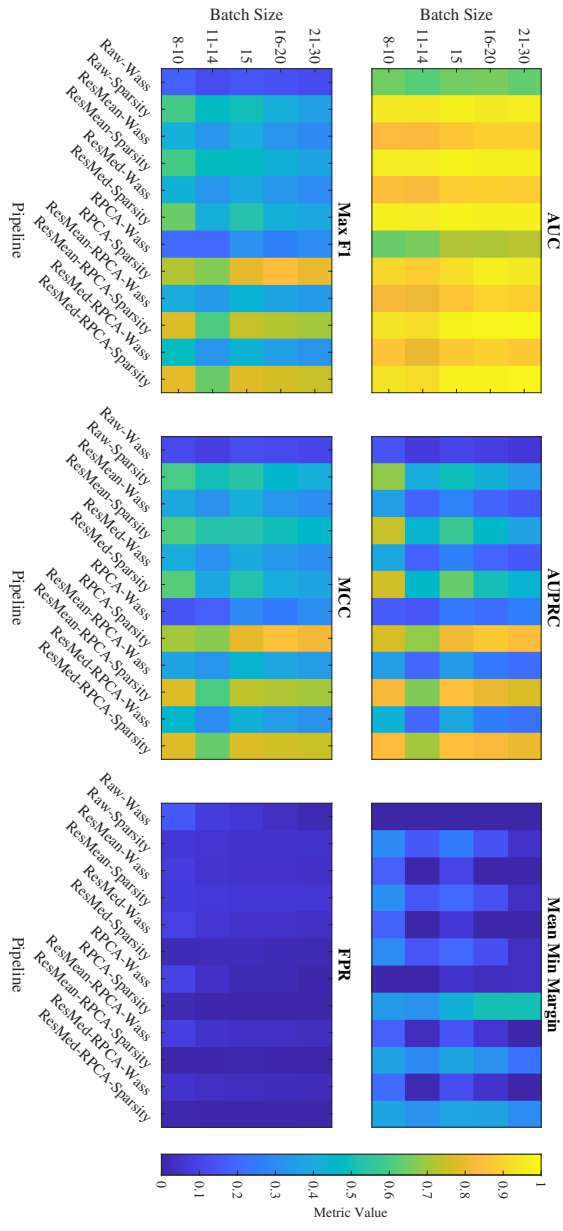
A first observation from this targeted analysis is the presence of mild variations in detectability across different batch sizes, as shown in Figure 14a. In general, based on AUPRC and maximum F1-score, fault detection tends to be better for smaller batch sizes. The notable exception is the RPCA-Sparsity pipeline, where consistent with expectations, larger batch sizes lead to more meaningful clustering and consequently more accurate anomaly scores. Considering that units 34 and 37 exhibit elevated noise levels, it is likely that performance degradation in other sparsity-based pipelines arises from bridging effects in clustering, where these two healthy but noisy units are grouped together and subsequently fused with the larger healthy ensemble, ultimately producing a single large cluster that obscures the true anomaly. This interpretation is strongly supported by the behavior of the mean minimum margin, which decreases consistently as batch size increases. When RPCA is applied, however, increasing batch size strengthens the decomposition process, producing cleaner low-rank structure and yielding more reliable detection outcomes.

Another pattern emerges in the load point analysis presented in Figure 14b, which examines the influence of operating conditions on performance under the fixed global threshold. When considering AUPRC, a clear trend emerges in which higher rotational speeds result in increased scores and improved detectability. For example, in the case of Raw-Sparsity, only at the highest speed level of 2000 rpm (load points 9 to 12) does AUPRC exceed 0.9 across all pressure levels.

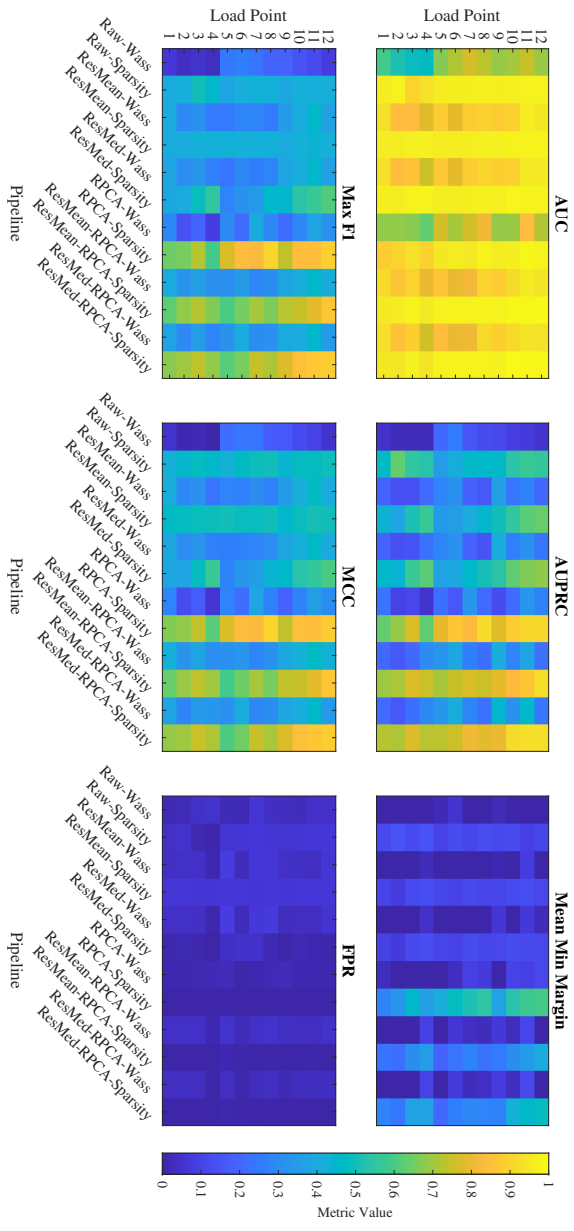
This is physically plausible, as higher speeds induce stronger structural excitation through impulsive loads generated during overrolling of bearing defects. However, for Raw-Sparsity, the mean minimum margin does not increase proportionally, indicating that discriminability of faulty units relative to healthy units does not improve to the same extent. Beyond rotational speed effects, the analysis also reveals clear pressure dependency, since higher pressures generally lead to higher AUPRCs across all sparsity-based methods. Conversely, at lower pressures such as 100 bar (load points 1, 5, and 9), lower AUPRC values are consistently observed. Load point 12 clearly provides the most favorable conditions for fault detection independent of the chosen method.

4.2.4 Interpretation with Visual Analytics

To obtain a deeper understanding of the pipelines' performance in fault detection within the present study, we show the visual analytics capability of the framework. For clarity of exposition, the analysis is restricted to simple sparsity indicators, as these exhibit substantially better performance than the Wasserstein distance on the real-world dataset. Moreover, we compare only the ResMean and RPCA pipelines against the baseline without any decomposition or residualization, in order to interpret the effect of RPCA. Two contaminated batch templates with distinct faulty units are examined, each with small batch sizes that are particularly challenging for OD. Batch 1 contains faulty unit 30 with Figure 12 indicating this case as especially difficult to detect, corresponds to load point 10, and has a batch size of 8, which is the minimum size considered here. Batch 2 contains faulty unit 4, corresponds to load point 12, and also includes 8 units. Visualizations are provided in Figure 15.

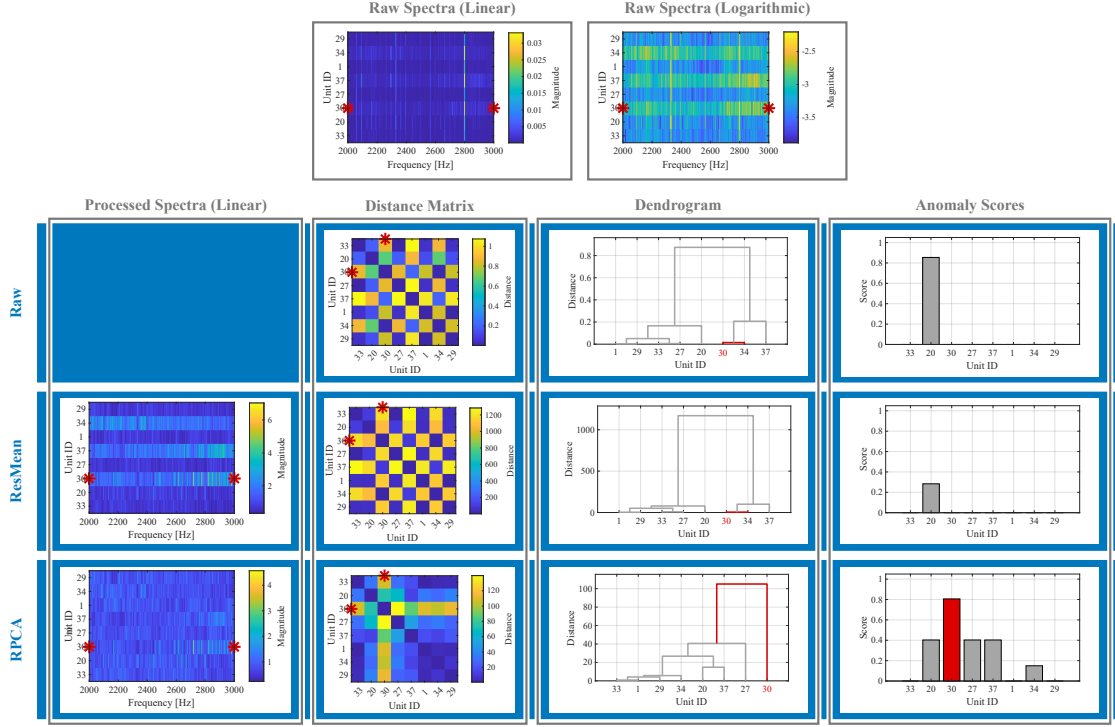


(a) Comparison of varying batch sizes

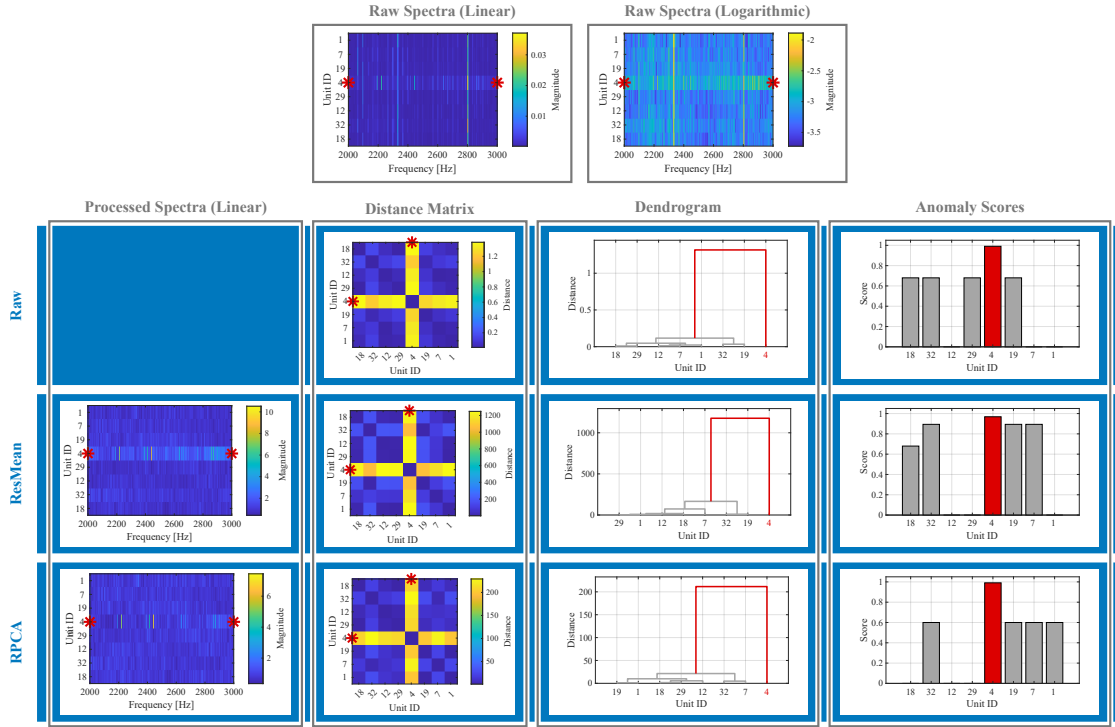


(b) Comparison of varying load points

Fig. 14: Conditional comparison of fault detection performance of the different pipelines for the real-world dataset



(a) Example 1 with faulty unit 30 and load point 10



(b) Example 2 with faulty unit 4 and load point 12

Fig. 15: Visual analytics-guided interpretation of two example batches based on the real-world dataset

Because Batch 1 (see Figure 15a) includes healthy units 34 and 37 with elevated broadband noise levels, the faulty unit is not immediately distinguishable from these units in the raw spectral heatmaps. Incipient rolling bearing-related narrowband peaks are therefore not detectable (either visually or downstream) by pipelines that do not employ RPCA. In this setting, the resulting distance matrices exhibit high similarity among healthy units 34 and 37 and faulty unit 30, producing a clearly separated cluster relative to the remaining five healthy units under hierarchical clustering. The automated clustering-based partitioning merges these two clusters only at relatively high cophenetic distance. However, because this subcluster contains only three units, no further partitioning is performed. In the remaining branch, especially in the raw pipeline, unit 20 stands out as an outlier and is consequently flagged as a false positive, as its outlier score clearly exceeds the optimal threshold determined in Table 5 (0.6797). In contrast, applying RPCA yields a marked improvement by fully decoupling process-induced broadband noise from the sparse component and absorbing it into the low-rank component. As a result, unit 30 appears as a singleton outlier in the dendrogram, and its clustering-based anomaly score (0.8059) equals the optimal threshold, which constitutes a correct detection based on the implementation. Units 27 as well as 20 and 37 emerge as mild outliers. However, their anomaly scores do not exceed the optimal threshold and therefore would not lead to false positives.

In Batch 2 (see Figure 15b), the beneficial effect of RPCA is less pronounced on visual inspection because the fluctuations among healthy units are less intensive. While all pipelines reliably detect the fault when using their respective optimal thresholds, the Raw-Sparsity and ResMean-Sparsity variants produce false positives. In the case of Raw-Sparsity, the anomaly scores of units 18, 32, 29, and 19 lie exactly at the optimal threshold. For ResMean-Sparsity, the scores of units 32, 19, and 7 clearly exceed this threshold. Although these units are far less conspicuous than the faulty unit 4, the dendrogram can be misleading. Under recursive partitioning, the faulty unit is isolated in the first split, and subsequent splits then classify healthy units as outliers within their respective pipelines, as the healthy ensemble exhibits slight heterogeneity. With RPCA, however, the healthy subset becomes considerably more homogeneous, since, for example, slightly elevated baseline noise in unit 32 is suppressed. As a result, separate subcluster formation for units such as 32 and 19 is avoided and the resulting subcluster with units 32 and 19 is less distinct.

4.2.5 Batch Size Transfer Benchmark Across OD Methods

For batch sizes $B \in \{5, 8, 12, 16, 20, 25, 30\}$, a controlled transfer benchmark is conducted. Here, B denotes the number of units in one batch. The RPCA sparsity indicator selected in Section 4.2.3 is held constant to isolate the score behavior of the OD backends. Thus, this experiment is distinct from the ASD and dissimilarity ablation in Section 4.2.3, which uses random batch sizes from 8 to 30.

For each faulty unit and each fixed batch size, 20 balanced templates are generated using the same usage-balanced healthy-unit sampling procedure as the ablation. Ten templates are fault free and ten contain one faulty unit at a random position. Each template is reused across the 12 load points, yielding 2,400 batches per size and 16,800 batches in total. All methods receive the same scalar input, so the comparison concerns the OD backend rather than pre-processing or feature dimensionality. kNN-Distance uses $k = \min(20, B - 1)$ nearest neighbours. Isolation Forest uses the MATLAB `iforest` defaults: 100 trees, $\min(B, 256)$ observations per tree, and a contamination fraction of zero. Since $B \leq 30$, every tree uses the complete batch. The global random seed is set to 42 before template generation and the benchmark loop.

For every method and calibration size B_{cal} , the pooled scores at B_{cal} determine an F1-optimal threshold. This threshold is then applied unchanged to the pooled scores at each evaluation size B_{eval} . The resulting transfer matrix quantifies whether a calibration at one ensemble size remains useful when production ensemble sizes vary.

The globally aggregated local detection performance is summarized in Table 6. Each metric is first calculated separately for every batch size and then averaged. These results quantify ranking and F1 performance after calibration at the respective size. They do not measure whether that calibration remains applicable at another size. Isolation Forest ranks first in all three columns. RobustZ ranks second for AUC and AUPRC, while kNN-Distance ranks second for F1 score. The fixed Hendrickx baseline has the lowest local performance in this comparison.

The optimum thresholds across batch sizes are shown in Figure 16. kNN-Distance decreases from 171.6 at $B = 5$ to 58.9 at $B = 30$, which confirms that its raw distance threshold depends strongly on ensemble composition. The fixed Hendrickx score rises from 0.60 to 0.93 and Isolation Forest also increases with batch size. The proposed method and StandardCC retain the common $[0, 1]$ score scale. RobustZ remains on an unbounded robust z score scale that is not directly comparable with this interval.

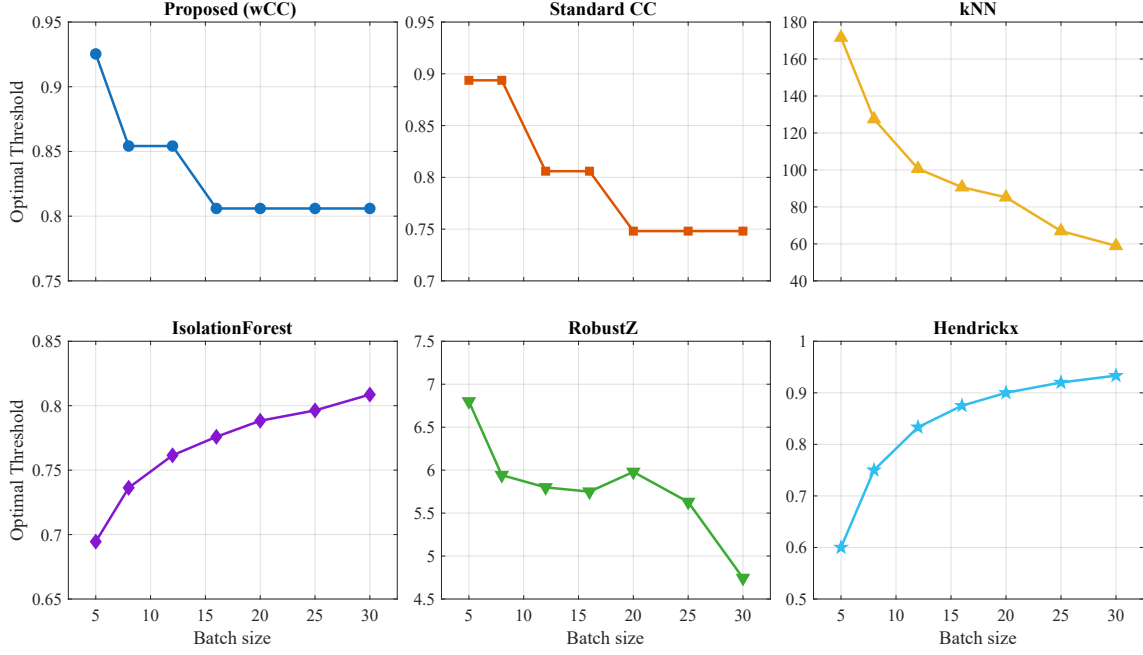


Fig. 16: F1 optimal thresholds in native score units. The scales are not comparable between methods.

Table 6: Mean detection performance at each batch size. Best (**bold**), second best (underlined) and third best (*italic*) results are highlighted per column.

Method	AUC	AUPRC	Opt. F1
Proposed (wCC)	0.968	0.881	0.834
StandardCC	<i>0.969</i>	<i>0.885</i>	0.843
kNN-Distance	0.962	0.874	<u>0.860</u>
Isolation Forest	0.985	0.931	0.881
RobustZ	<u>0.982</u>	<u>0.912</u>	<i>0.859</i>
Hendrickx (fixed thr_{CC})	0.911	0.774	0.776

The transfer results are reported in Figure 17. Each heatmap row uses the threshold calibrated at the indicated batch size and each column reports its F1 score at the evaluation size. RobustZ has the most consistent off diagonal results in this setting. The proposed wCC score also retains useful transfer across the tested range. For a threshold calibrated at $B = 30$ and evaluated at $B = 5$, the F1 score is 0.780 for RobustZ and 0.708 for wCC. Under the same transfer direction, Isolation Forest and Hendrickx both yield an F1 score of zero. Thus, the strongest local discrimination does not by itself provide a portable decision threshold.

5 Discussion

Based on the experimental results, key findings and their implications for the practical application of ASD-BOD are discussed below. While this study demonstrates the effectiveness of ASD-BOD for bearing fault detection in HMLV EoLT environments, several limitations warrant consideration and provide directions for future research.

5.1 Implications of the Findings

The central finding is that RPCA provides an adaptive spectral decomposition for vibroacoustic EoLT without manually specified fault frequencies, transfer functions or historical baselines. Its low rank component represents batch coherent background structure, while its sparse component retains localized deviations. On the APU data, RPCA raises AUPRC from 0.421 to 0.887 and reduces the false positive rate from 6.00% to 0.23%. Mean residuals are less reliable under heterogeneous operating noise because a faulty or noisy unit directly influences the reference spectrum.

RPCA also reduces sensitivity to batch coherent process variation. Residual methods can treat idiosyncratic noise as anomalous when their arithmetic reference is unrepresentative. The low rank component instead models correlated variation directly.

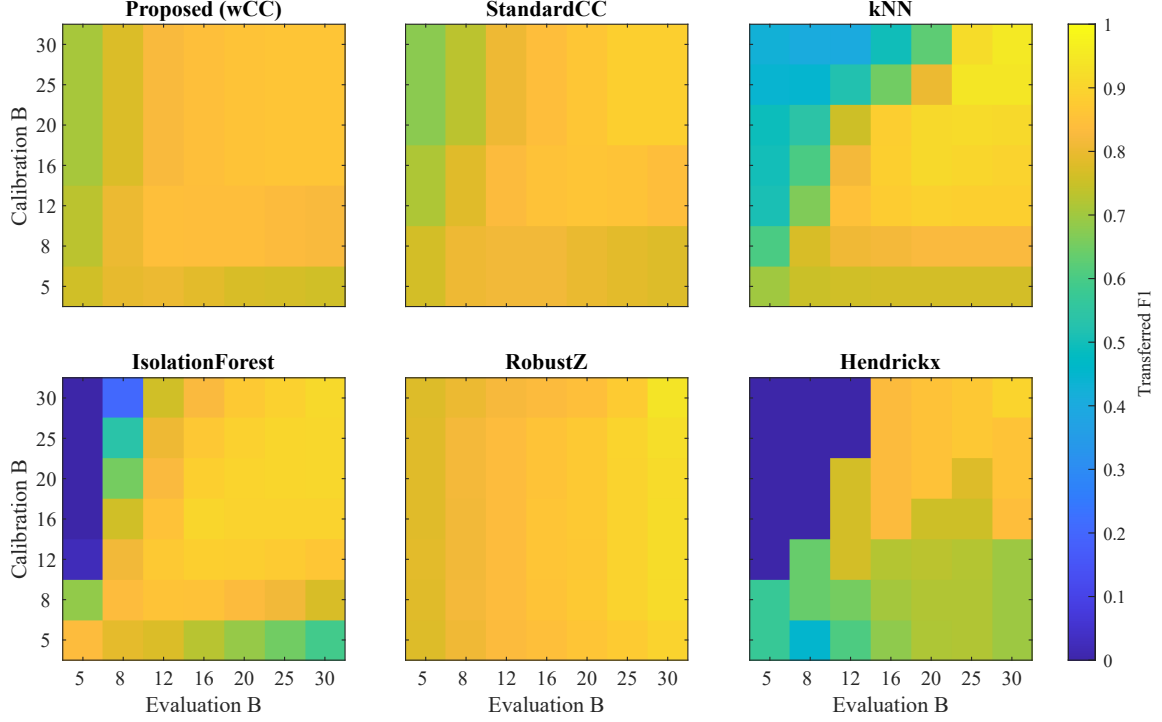


Fig. 17: F1 score after transferring a calibration threshold from the batch size in each row to the evaluation size in each column.

This can relax test environment conditioning requirements because changes in fluid temperature, ambient conditions or rig resonances are absorbed when they occur coherently within a batch.

The transfer benchmark separates local detection quality from threshold portability. Isolation Forest achieves the strongest local scores but can fail after a batch size change. RobustZ transfers best in the tested scalar input and one fault setting, which supports population relative standardization as a useful mechanism. kNN-Distance also detects faults well locally but depends on raw distance units. The proposed wCC score does not exceed RobustZ in this test. It provides a bounded score, an inspectable dendrogram and direct compatibility with pairwise dissimilarities.

The indicator ablation further supports the use of the sparsity indicator after RPCA. Wasserstein distance reaches an AUPRC of only 0.137, whereas RPCA sparsity reaches 0.887. In this dataset, total sparse energy is more robust to benign spectral morphology than a full spectrum distance. Together, RPCA and wCC provide an unsupervised basis for identifying localized bearing faults without labelled training data. The continuous score and sparse spectra also support risk based decisions and diagnostic inspection.

5.2 Limitations and Future Research

Fault Mode Specificity and Tribological Contacts: The framework was validated exclusively on rolling element bearing faults, which manifest as localized, narrowband spectral peaks naturally aligning with RPCA’s sparsity assumption. However, APUs exhibit various tribological contacts where faults may manifest as broadband spectral changes, potentially being absorbed into the low-rank structure. A cascaded approach could address this by first applying a conservative outlier detection pass on raw spectral data to identify broadband deviations, followed by RPCA-based decomposition to capture subtle, localized faults. For high-runner variants with sufficient data, transfer learning offers an alternative pathway to model such characteristics across domains.

Generalization and Production Contexts: ASD-BOD’s benefits are most pronounced in scenarios combining HMLV characteristics with significant vibroacoustic disturbances. In cleaner environments (e.g., electric motors), supervised approaches may prove more effective. Systematic evaluation across diverse EoLT contexts remains necessary to validate core assumptions (fault rarity, batch-wise domain shifts) beyond hydraulic systems. As illustrated in

Figure 4, the architecture supports parallel deployment alongside existing methods, enabling complementary quality assurance strategies rather than wholesale replacement.

Batch Size and Operational Assumptions:

The framework requires minimum ensemble sizes for meaningful clustering. The ASD and dissimilarity ablation uses batch sizes from 8 to 30, whereas the OD transfer benchmark extends the evaluation to $B = 5$. Ultra-small batches and particularly batch-size-one production [3] nevertheless fundamentally preclude ensemble-based analysis. Leveraging historical data aggregation by comparing new small batches against archived OK data represents a viable extension pathway, though at the cost of additional data mining complexity. The framework also assumes domain stability within batches; intra-batch variations (sensor mounting changes, test rig wear) can introduce systematic shifts. While RPCA provides inherent robustness by absorbing broadband disturbances, deployment scenarios with excessive internal heterogeneity may require batch-validity checks. Additionally, multiple faults with identical vibroacoustic signatures could become low-rank and escape detection, though heterogeneous faults are correctly isolated as individual sparse entries.

Benchmark Scope: The OD comparison uses the common one-dimensional RPCA-Sparsity input, artificial batches, one faulty unit per contaminated batch, and sizes from 5 to 30. RobustZ’s superior transfer is therefore demonstrated only in this favorable regime. Its behavior under multimodal healthy populations, multiple faults, and morphology-only changes remains to be evaluated, as does threshold transfer on full-spectrum dissimilarities. These limits also constrain extrapolation of the wCC result beyond the tested batch-size range.

Dataset Characteristics: Validation on real APU data necessitated artificial batch construction through random sampling, using only 7 raw APUs to imitate serial dispersion. Domain shifts were induced by test program variations rather than authentic product changes. While the framework’s unsupervised, per-batch operation mitigates cross-domain dependency, validation on authentic longitudinal EoLT datasets would strengthen generalizability claims. Experiments adhered to fault-rarity (one faulty unit per batch, minimum cluster size 1/3). The sparse matrix correctly isolates heterogeneous faults as individual entries, but contamination scenarios with multiple identical faults warrant systematic characterization, potentially through adapted RPCA formulations with cardinality constraints.

Feature Engineering and Computational

Constraints: The implementation employs domain-specific yet coarse feature engineering (manual RoI selection, magnitude spectra). The modular architecture accommodates alternative representations; synergies between advanced signal processing (envelope analysis, spectral kurtosis, time-frequency representations) and RPCA-based decomposition merit investigation. The asynchronous EoLT nature justifies computational intensity, particularly threshold-sweeping for outlier scoring, prioritizing batch-size-invariant automation over efficiency. Real-time fleet-based CM deployment would necessitate optimization through adaptive threshold selection or surrogate models.

Operational Integration: Practical deployment will likely involve integration within broader quality assurance ecosystems. The framework’s continuous outlier scores naturally support risk-based prioritization, enabling quality engineers to focus on units most likely harboring faults. Hybrid strategies where the batch-wise OD approach serves as first-stage screening for flagging candidates for deep-dive analysis could optimize cost-sensitivity trade-offs in quality assurance workflows.

6 Conclusion

The increasing individualization in manufacturing is establishing HMLV production as the predominant paradigm. This presents fundamental challenges to conventional, data-driven fault detection methods in vibroacoustic EoLT. This issue is particularly acute for hydraulic machinery such as APUs, where strong, process-induced disturbances from pressure pulsations and flow dynamics mask the subtle signatures of bearing defects. Simultaneously, high product variance prevents the development of universal diagnostic models, while data scarcity for new variants precludes the use of supervised learning approaches. In response to these challenges, ASD-BOD was developed. This framework transfers and adapts the paradigm of fleet-based CM to HMLV EoLT. The transfer leverages a fundamental analogy since both approaches identify faults as rare outliers within homogeneous ensembles. ASD-BOD performs batch-wise OD for units of identical specification tested consecutively. This approach shares the core assumptions of fault rarity and operational homogeneity with fleet-based CM and operates entirely unsupervised, without relying on labeled fault examples or historical training data. This makes it inherently suitable for HMLV scenarios where immediate applicability from the first production batch is crucial.

The methodological core of the framework is ASD via RPCA, which is applied for the first time to

the unsupervised, data-driven suppression of batch-specific background variations in vibroacoustic EoLT. By decomposing the spectral batch matrices into a common operational background (low-rank) and unit-specific anomalies (sparse), RPCA adaptively learns and removes variant-specific disturbances. This is achieved without prior knowledge of the hydraulic kinematics, the transfer functions of the test rig, or specific fault frequencies. Validation on real-world data from APUs confirms its practical effectiveness, as the RPCA-based decomposition improves the AUPRC from 0.421 to 0.887 and reduces the FPR from 6.00% to 0.23%. The controlled OD benchmark further shows that high local discrimination alone does not guarantee a portable threshold. Isolation Forest is strongest after per-size calibration, whereas RobustZ transfers best in the evaluated scalar, single-contaminant regime. The proposed wCC score provides a bounded, interpretable clustering alternative with useful transfer over batch sizes 5–30. Future work must test these trade-offs under multiple faults, multimodal nominal populations, and full-spectrum dissimilarities.

Declarations

Funding: This work was supported by Bosch Rexroth AG within the scope of an industrial PhD program. No external third-party funding or grant numbers apply.

Competing interests Financial interests: Maximilian Romeser is employed by Bosch Rexroth AG, the manufacturer of the axial piston units investigated in this study. Reinhold von Schwerin declares no financial interests. Non-financial interests: The authors declare no non-financial interests.

Author contribution: Maximilian Romeser: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Visualization, Investigation, Validation. Reinhold von Schwerin: Writing – review, Supervision.

References

- [1] Pech, M., Vrchota, J.: The Product Customization Process in Relation to Industry 4.0 and Digitalization. *Processes* **10**(3) (2022) <https://doi.org/10.3390/pr10030539>
- [2] Sanden, B., Verriet, J.: Performance engineering of high-mix low-volume production systems. Technical Report R10304, TNO, Delft, The Netherlands (March 2025). <https://publications.tno.nl/publication/34643957/wSPVOr5V/TNO-2025-R10304.pdf>
- [3] Wißbrock, P., Koschek, L., Ren, Z., Nejd, W.: Enhancing quality inspection of highly variant geared motors. *Applied Acoustics* **235** (2025) <https://doi.org/10.1016/j.apacoust.2025.110687>
- [4] Fassbender, D., Minav, T.: Finding the Perfect Match: Different Heavy-Duty Mobile Applications Call for Different Actuators. *Proceedings* **64**(1), 22 (2020) <https://doi.org/10.3390/IeCAT2020-08524>
- [5] Haug, S.: Ganzheitliche Optimierung einer Axialkolbenpumpe durch bedarfsangepasste Entlastung tribologischer Kontakte. Phd thesis, Karlsruhe Institute of Technology (2022). <https://doi.org/10.5445/KSP/1000140024>
- [6] Fan, Q., Fan, H.: Reliability analysis and failure prediction of construction equipment with Time Series Models. *Journal of Advanced Management Science* **3**(3), 203–210 (2015) <https://doi.org/10.12720/joams.3.3.203-210>
- [7] Katsitadze, J., Kutelia, G., Abuladze, I., Beridze, G.: IMPROVING THE RELIABILITY OF AGRICULTURAL MACHINERY, USING THE METHOD OF RESTORATION OF PARTS UNDER SUBMERGED ARC WELDING, AND ITS TECHNICAL AND ECONOMIC EVALUATION. *Machines. Technologies. Materials.* **11**(6), 307–310 (2017)
- [8] Yu, J., Zhang, Y.: Challenges and opportunities of deep learning-based process fault detection and diagnosis: a review. *Neural Computing and Applications* **35**(1), 211–252 (2023) <https://doi.org/10.1007/s00521-022-08017-3>
- [9] Hirsch, V., Reimann, P., Mitschang, B.: Data-Driven Fault Diagnosis in End-of-Line Testing of Complex Products. In: 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 492–503 (2019). <https://doi.org/10.1109/DSAA.2019.00064>
- [10] Bonart, J., Hümmer, P., Drossel, W.-G.: Enhancing End-of-Line Defect Classifications and Evaluating Early Testability for Inline Test Stands Using NVH Measurements. In: 2021 11th International Electric Drives Production Conference (EDPC) (2021). <https://doi.org/10.1109/EDPC53547.2021.9684212>
- [11] Nunes, C., Pires, E.J.S., Reis, A.: Machine Learning and Deep Learning applied to End-of-Line Systems: A review. *WSEAS TRANSACTIONS*

- ON SYSTEMS **21**, 147–156 (2022) <https://doi.org/10.37394/23202.2022.21.16>
- [12] Gaugel, S., Reichert, M.: Data-driven multi-objective optimization of hydraulic pump test cycles via wrapper feature selection. *CIRP Journal of Manufacturing Science and Technology* **50**, 14–25 (2024) <https://doi.org/10.1016/j.cirpj.2024.01.004>
 - [13] Bonart, J.: Entwicklung einer End-of-Line Körperschallprüfmethodik für hochintegrierte elektrifizierte Antriebstopologien. PhD thesis, Chemnitz University of Technology (2023)
 - [14] Gaugel, S., Reichert, M.: Industrial Transfer Learning for Multivariate Time Series Segmentation: A Case Study on Hydraulic Pump Testing Cycles. *Sensors* **23**(7) (2023) <https://doi.org/10.3390/s23073636>
 - [15] Randall, R.B., Antoni, J.: Rolling element bearing diagnostics—A tutorial. *Mechanical Systems and Signal Processing* **25**(2), 485–520 (2011) <https://doi.org/10.1016/j.ymssp.2010.07.017>
 - [16] Kolb, J.: Analyse von Fertigungstoleranzen in elektrischen Maschinen und deren Güteinstufung mittels maschinellen Lernens am Beispiel einer permanentmagneterregten Synchronmaschine. PhD thesis, RWTH Aachen University (2024)
 - [17] Bischof, T.: Root Cause Analysis of Acoustically Suspicious Transmissions During EOL Testing. In: 7th International Styrian Noise, Vibration and Harshness Congress: The European Automotive Noise Conference (2012). <https://doi.org/10.4271/2012-01-1528>
 - [18] Bertolini, T., Fuchs, T.: Schwingungen Und Geräusche. In: *Handbuch Elektrische Kleinantriebe*, Berlin, Germany, pp. 358–375 (2020)
 - [19] Nigel, D.: Measurement and prediction of the fluid borne noise characteristics of hydraulic components and systems. PhD thesis, University of Bath (1987)
 - [20] Xiao, C., Tang, H., Ren, Y., Xiang, J., Kumar, A.: A fault frequency bands location method based on improved fast spectral correlation to extract fault features in axial piston pump bearings. *Measurement* **171**, 108734 (2021) <https://doi.org/10.1016/j.measurement.2020.108734>
 - [21] Shaw, H., Vijay, S.: Diagnosis and Detection of IC Engine Fault at End of Line Engine Vibration Measurement System with Machine Learning Model. In: *Proceedings of the International Conference on Industrial Engineering and Operations Management*, Bangalore, India (2021). <https://doi.org/10.46254/IN01.20210070>
 - [22] Romeser, M., Schwerin, R.: Enabling Vibroacoustic End-of-Line Testing of APUs Across Varying Fluid Temperatures through ML-Based Conditional Anomaly Detection. In: *15th International Fluid Power Conference*, (2026). <https://doi.org/10.13052/rp-9788743808251A30>
 - [23] Bonart, J.A.E., Huemmer, P., Maurice, J.-B., Trautmann, F., Drossel, W.-G.: Transfer of NVH limits for highly-integrated electrical drives between end-of-line test benches. In: *2021 IEEE 13th International Symposium on Diagnostics for Electrical Machines, Power Electronics And Drives (SDEMPED)*, vol. 1, pp. 295–300 (2021). <https://doi.org/10.1109/SDEMPED51010.2021.9605505>
 - [24] Hirsch, V., Reimann, P., Treder-Tschechlov, D., Schwarz, H., Mitschang, B.: Exploiting domain knowledge to address class imbalance and a heterogeneous feature space in multi-class classification. *The VLDB Journal* **32**(5), 1037–1064 (2023) <https://doi.org/10.1007/s00778-023-00780-6>
 - [25] Berger, V., Wilbertz, A., Meyer, I.: Körperschall im End-Of-Line-Test von Doppelkupplungsgetrieben. *ATZproduktion* **3**(5), 24–27 (2010) <https://doi.org/10.1007/BF03224146>
 - [26] Mlinaric, J., Pregelj, B., Boskoski, P., Petrovic, J.a.D.G.: A transfer learning approach to machine learning-based end-of-line quality inspection. *Advances in Production Engineering and Management* **20**(2), 277–290 (2025) <https://doi.org/10.14743/apem2025.2.540>
 - [27] Kreuzberger, D., Kühl, N., Hirschl, S.: Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access* **11**, 31866–31879 (2023) <https://doi.org/10.1109/ACCESS.2023.3262138>
 - [28] Muhr, D., Tripathi, S., Jodlbauer, H.: Divide and conquer anomaly detection: A case study predicting defective engines. *Procedia Manufacturing* **42**, 57–61 (2019) <https://doi.org/10.1016/j.promfg.2020.02.090>

- [29] Hendrickx, K., Meert, W., Mollet, Y., Gyselinck, J., Cornelis, B., Gryllias, K., Davis, J.: A general anomaly detection framework for fleet-based condition monitoring of machines. *Mechanical Systems and Signal Processing* **139**, 106585 (2020) <https://doi.org/10.1016/j.ymssp.2019.106585>
- [30] Hendrickx, K., Meert, W.: Fleet-based health monitoring for end-of-production-line and operational testing. In: *The 29th International Conference on Noise and Vibration Engineering (ISMA2020)* (2020)
- [31] Romeser, M., Schwerin, R.: Data-Driven Serial Testing of Axial Piston Units: Integrating Signal Processing and Machine Learning for Roller Bearing Fault Detection. *International Journal of Fluid Power* **27**(1), 485–520 (2026) <https://doi.org/10.13052/ijfp1439-9776.2716>
- [32] Candes, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis. *J. ACM* **58**(3) (2011) <https://doi.org/10.1145/1970392.1970395>
- [33] Del Rosso, V., Andreucci, A., Boria, S., Corradini, M.L., Ranalli, A.: Mechanical fault detection for induction motors based on vibration analysis: a case study. In: *IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society*, pp. 1–6 (2021). <https://doi.org/10.1109/IECON48115.2021.9589189>
- [34] Mlinaric, J., Pregelj, B., Boskoski, P., Dolanc, G., Petrovic, J.: Optimization of reliability and speed of the end-of-line quality inspection of electric motors using machine learning. *Advances in Production Engineering and Management* (2024) <https://doi.org/10.14743/apem2024.2.500>
- [35] Mlinaric, J., Pregelj, B., Dolanc, G.: End-of-Line Quality Control Based on Mel-Frequency Spectrogram Analysis and Deep Learning. *Machines* **13**(7) (2025) <https://doi.org/10.3390/machines13070626>
- [36] Borriello, P., Tessicini, F., Ricucci, G., Frosina, E., Senatore, A.: A fault detection strategy for an ePump during EOL tests based on a knowledge-based vibroacoustic tool and supervised machine learning classifiers. *Meccanica* **59**(3), 279–304 (2024) <https://doi.org/10.1007/s11012-024-01754-w>
- [37] Blickensdorff, J., Degtiarev, A., Gulzar, K., Fleischmann, H., Grünke, R.: Production Quality Testing for Automotive Electric Drive Units with AI-Enabled Anomaly Detection based on NVH Data, Hamburg (2023)
- [38] Sun, J., Wyss, R., Steinecker, A., Glocker, P.: Automated fault detection using deep belief networks for the quality inspection of electromotors. *tm - Technisches Messen* **81**(5), 255–263 (2014) <https://doi.org/10.1515/teme-2014-1006>
- [39] Principi, E., Rossetti, D., Squartini, S., Piazza, F.: Unsupervised electric motor fault detection by using deep autoencoders. *IEEE/CAA Journal of Automatica Sinica* **6**(2), 441–451 (2019) <https://doi.org/10.1109/JAS.2019.1911393>
- [40] Wißbrock, P., Pelkmann, D., Richter, Y.: Discussion of Features for Acoustic Anomaly Detection under Industrial Disturbing Noise in an End-of-Line Test of Geared Motors. In: *2023 IEEE 21st International Conference on Industrial Informatics (INDIN)*, pp. 1–8 (2023). <https://doi.org/10.1109/INDIN51400.2023.10217989>
- [41] Gharibnezhad, F., Mujica, L.E., Rodellar, J.: Applying robust variant of Principal Component Analysis as a damage detector in the presence of outliers. *Mechanical Systems and Signal Processing* **50–51**, 467–479 (2015) <https://doi.org/10.1016/j.ymssp.2014.05.032>
- [42] Tharrault, Y., Mourot, G., Ragot, J.: Fault detection and isolation with robust principal component analysis. In: *2008 16th Mediterranean Conference on Control and Automation*, pp. 59–64 (2008). <https://doi.org/10.1109/MED.2008.4602224>
- [43] Li, C., Lu, P., Dong, G., Chen, G.: Bearing Fault Diagnosis via Robust PCA With Nonconvex Rank Approximation. *IEEE Sensors Journal* **24**(12), 19531–19542 (2024) <https://doi.org/10.1109/JSEN.2024.3395306>
- [44] Ge, M., Lv, Y., Zhang, Y., Yi, C., Ma, Y.: An Effective Bearing Fault Diagnosis Technique via Local Robust Principal Component Analysis and Multi-Scale Permutation Entropy. *Entropy* **21**(10) (2019) <https://doi.org/10.3390/e21100959>
- [45] Wang, R., Fang, H., Yu, L., Yu, L., Chen, J.: Sparse and low-rank decomposition of the time-frequency representation for bearing fault diagnosis under variable speed conditions. *ISA Transactions* **128**, 579–598 (2022) <https://doi.org/10.1016/j.isatra.2022.05.011>

[org/10.1016/j.isatra.2021.11.030](https://doi.org/10.1016/j.isatra.2021.11.030)

- [46] Alharbi, B., Liang, Z., Aljindan, J.M., Agnia, A.K., Zhang, X.: Explainable and Interpretable Anomaly Detection Models for Production Data. *SPE Journal* **27**(01), 349–363 (2022) <https://doi.org/10.2118/208586-PA> <https://onepetro.org/SJ/article-pdf/27/01/349/2627122/spe-208586-pa.pdf>
- [47] Yan, C.: A review on spectral data preprocessing techniques for machine learning and quantitative analysis. *iScience* **28**(7), 112759 (2025) <https://doi.org/10.1016/j.isci.2025.112759>
- [48] Schlenke, J., Hildebrand, L., Moros, J., Laserna, J.J.: Adaptive approach for variable noise suppression on laser-induced breakdown spectroscopy responses using stationary wavelet transform. *Analytica Chimica Acta* **754**, 8–19 (2012) <https://doi.org/10.1016/j.aca.2012.10.012>
- [49] Iglesias Vázquez, F., Marques, H.O., Zimek, A., Zseby, T.: What do anomaly scores actually mean? Dynamic characteristics beyond accuracy. *Data Mining and Knowledge Discovery* **39** (2025) <https://doi.org/10.1007/s10618-024-01077-0>
- [50] Zahn, C.T.: Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers* **20**(1), 68–86 (1971) <https://doi.org/10.1109/T-C.1971.223083>
- [51] Klutchnikoff, N., Poterie, A., Rouvière, L.: Statistical analysis of a hierarchical clustering algorithm with outliers. *Journal of Multivariate Analysis* **192** (2022) <https://doi.org/10.1016/j.jmva.2022.105075>
- [52] Hendrickx, K., Meert, W., Cornelis, B., Gryllias, K., Davis, J.: Similarity-based anomaly score for fleet-based condition monitoring. In: *Annual Conference of the PHM Society 2020*, (2020). <https://doi.org/10.36001/phmconf.2020.v12i1.1178>
- [53] Lin, Z., Chen, M., Ma, Y.: The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Mathematical Programming* **137**(1), 203–229 (2013)