Reliability Engineering and System Safety 278 (2027) 113612



Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress



Short communication

Multi-class bearing faults: Label-guided unsupervised ensemble with a One-vs-Rest benchmark of isolation forest methods

Yoshiyasu Takefuji

SciencePark Corporation, 3-24-9 Iriya-Nishi, Kanagawa, Zama-shi, 252-0029, Japan

ARTICLE INFO

Keywords:
Unsupervised learning
Bearing fault detection
Isolation forest
Label prediction accuracy
Anomaly detection comparison

ABSTRACT

Reliability Engineering & System Safety has published numerous studies on unsupervised label prediction, yet most rely on a single model without comparative benchmarking, leaving model choice effects on prediction quality unexamined. This paper addresses that gap by evaluating eight unsupervised algorithms—Isolation Forest (IF), Local Outlier Factor (LOF), Kernel Density Estimation (KDE), Gaussian Mixture Model (GMM), Unsupervised Random Forest (URF), Mean Shift (MS), DBSCAN, and OPTICS—using a publicly available bearing fault dataset comprising 2300 samples, 9 features, and 10 fault states. A One-vs-Rest strategy adapted binary anomaly detectors for multi-class classification. Results show IF achieved the highest overall accuracy (0.857), outperforming the closest baseline (GMM, 0.848) by less than one percentage point, while maintaining consistently high per-class F1 scores across nearly all fault types; OPTICS also performed competitively, achieving the third-highest overall accuracy (0.800). Conversely, URF, DBSCAN, and MS exhibited complete detection failures ($F1 = 0$) for multiple fault categories—Mean Shift failing on as many as four classes—revealing sensitivity to class imbalance and non-convex distributions. Although GMM slightly outperformed IF on select individual classes (e.g., Ball_007_1, IR_014_1), IF's advantage lay primarily in its overall consistency and stability across fault severities rather than uniform superiority on every class. These findings establish label-prediction accuracy as both a data-quality metric and a practical guide for model selection in reliability-critical fault detection tasks.

1. Introduction

This paper addresses a critical gap in unsupervised fault diagnosis research by systematically evaluating and comparing the performance of multiple unsupervised models for label prediction, using accuracy and per-class F1 scores as evaluation metrics. The analysis is conducted on a publicly available and widely adopted bearing fault detection dataset comprising ten balanced classes (230 samples each), spanning three defect locations (ball, inner race, and outer race) across three severity levels (0.007″, 0.014″, and 0.021″ fault diameters), in addition to one baseline normal-condition class.

A review of the literature reveals that Reliability Engineering & System Safety has published 27 articles on label prediction using supervised and unsupervised models, underscoring the field's strong emphasis on identifying key influential factors and leveraging the label-independence inherent to unsupervised methods. This focus is well justified, as unsupervised approaches are especially valuable in real-world reliability contexts where labeled failure data is scarce or entirely unavailable. Notably, 11 of these 27 studies employed unsupervised models for label prediction based solely on extracted features, without relying on ground-truth labels. However, every one of these studies adopted only a single unsupervised model, and none measured or compared the performance of multiple unsupervised models against one another. This represents a significant and previously unaddressed gap: reliance on a single model offers no empirical basis for understanding how model choice influences prediction quality, robustness, or generalizability across fault types and severities.

Numerous peer-reviewed studies have established that, in unsupervised methods relying solely on features without labels, label-prediction accuracy is proportional to data quality; specifically, as the number of missing features increases, data quality degrades correspondingly across multiple domains [1–8]. This relationship arises because unsupervised models depend entirely on the structure and completeness of the input feature space to infer patterns that correspond to true labels; when critical features are missing, the underlying data structure becomes distorted, making it harder for the model to correctly separate classes. Because predicted labels can be directly compared against true labels, this comparison serves as a practical proxy for measuring data quality; in

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.ress.2026.113612
Received 13 July 2026; Received in revised form 1 October 2026; Accepted 3 October 2026
Available online 3 October 2026
0951-8320/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

other words, label prediction accuracy itself becomes a quantifiable indicator of how complete and reliable the underlying data is. Building on this relationship, when constructing a new dataset, the discrepancy between predicted and true labels can be used to systematically identify which features are missing, since lower prediction accuracy signals the likely presence of missing features [9,10], while higher accuracy suggests a more complete feature set. This diagnostic use is particularly valuable during early-stage dataset design, where practitioners often lack ground-truth confirmation of feature sufficiency and instead must rely on indirect signals such as prediction performance. Therefore, label-prediction accuracy functions not only as a data-quality metric but also as a diagnostic tool for detecting and localizing missing features during dataset construction.

This paper evaluates the quality of label prediction, measured through accuracy and per-class F1 scores, across multiple unsupervised models: binary unsupervised models such as Isolation Forest (IF), Local Outlier Factor (LOF), Kernel Density Estimation (KDE), multi-class unsupervised models including Gaussian Mixture Model (GMM), Unsupervised Random Forest (URF), Mean Shift (MS), DBSCAN, and OPTICS. These models were selected for their diverse algorithmic foundations—density-based, distance-based, probabilistic, and ensemble-based—enabling a broader assessment of how different unsupervised paradigms perform under identical conditions.

For inherently binary classification algorithms such as IF and LOF, a One-vs-Rest strategy was applied to accommodate multi-class classification. To clarify the label-guided unsupervised ensemble built around a One-vs-Rest (OvR) benchmarking procedure for Isolation Forest, it is important to distinguish between the training objective of each model and the experimental design used for multi-class evaluation. First, note that the OvR strategy was applied only to those algorithms that are inherently binary in nature—namely IF, LOF, and KDE—which natively produce a normal-versus-anomalous decision rather than a multi-way class assignment.

Accordingly, when these binary unsupervised models (IF, LOF, KDE) are extended to multi-class classification through OvR, they are more accurately described as label-guided (or label-partitioned) unsupervised methods, since class labels are used to partition the data and define each one-vs-rest training subset, even though the underlying models themselves remain unsupervised. While Gao et al. [11] and Liu et al. [12] evaluated bearing fault diagnosis with unsupervised models, this paper utilizes diverse unsupervised models including binary unsupervised models (IF, LOF, KDE) are extended to multi-class classification through OvR.

For these algorithms, the dataset was partitioned on a per-class basis: for each of the ten known fault categories, an 80/20 train-test split was performed, yielding what is referred to as the label-guided ensemble with OvR. After splitting the data into "one class" versus "the rest" for each category, an unsupervised model was fit on each resulting subset.

By contrast, GMM, URF, MS, DBSCAN, and OPTICS are natively capable of producing multiple clusters or components within a single unsupervised fit and were therefore not subjected to OvR decomposition. Each of these algorithms was instead fit once on the full unlabeled dataset, and the resulting clusters or components were subsequently mapped to the ten fault categories solely for evaluation purposes, without any label information being used during the fitting process itself. In other words, each per-class binary model (IF, LOF, KDE) learns the intrinsic geometric or density-based structure of its assigned subset—via isolation depth, local density deviation, or kernel density estimation, respectively—without reference to whether its output aligns with the true class label. The inherently multi-class models (GMM, URF, MS, DBSCAN, OPTICS), on the other hand, learn the structure of the entire unlabeled feature space directly, in a single pass.

For the OvR-based algorithms, once all ten per-class binary models have been fit on their respective training subsets X (using no label information during fitting), the prediction process proceeds as follows. For each class c, an Isolation Forest is trained using only the training samples belonging to class c. This means each model learns what "normal" looks like specifically for its assigned class.

The evaluation is conducted on a publicly available bearing fault dataset comprising 2300 samples and 9 features, spanning 10 distinct states: normal, small/medium/large ball defect, small/medium/large inner race defect, and small/medium/large outer race defect. By applying multiple models to this shared dataset, the paper enables a controlled comparison of predictive performance, offering insight into which unsupervised techniques are best suited for fault detection tasks of this nature. Although the analysis centers on a single dataset, the diverse set of unsupervised methods examined is broadly applicable to other domains and datasets. The ten original classes are balanced (230 samples each), but the One-vs-Rest conversion introduces binary imbalance (230 vs. 2070, ratio 1:9). Despite this, anomaly detection methods still perform well empirically, indicating robustness to this imbalance.

This study relies on a dataset addressing the scarcity of publicly available mechanical engineering data for applying Machine Learning in industrial environments, made available on Kaggle as a community contribution based on motor performance data originally provided by the Case Western Reserve University Bearing Data Center. The test bench used for motor performance assessment consists of a 2 HP motor, a torque transducer, a dynamometer, and associated control electronics. The test bearings support the motor shaft, and defects were introduced at a single point via EDM (electrical discharge machining), with defect diameters of 0.007 inches (0.178 mm), 0.014 inches (0.356 mm), and 0.021 inches (0.533 mm). A time series is available for each defect, located in one of three bearing components: the ball, the inner race, or the outer race. Telemetry measurements were collected using three accelerometers positioned at the drive end (DE), fan end (FE), and base (BA) of the system. The dataset used in this study corresponds to a 1 HP load applied to the motor, a shaft rotating speed of 1772 rpm, and an accelerometer sampling frequency of 48 kHz. For fault identification, nine features were calculated for each time segment of 2048 points (equivalent to 0.04 s at the 48 kHz sampling frequency): maximum, minimum, mean, standard deviation, RMS, skewness, kurtosis, crest factor, and form factor.

Per-class F1 scores across unsupervised algorithms are summarized in Table 1 and overall accuracy (mean ± standard deviation with 95% confidence intervals) for Isolation Forest and Unsupervised Random Forest is illustrated in Tables 2 and 3 respectively.

2. Methods

The dataset used in this study, feature_time_48k_2048_load_1.csv, contains time-domain features extracted under a fixed sampling rate of 48 kHz, a window length of 2048 samples and 9 features, and a load condition of 1 [13]. The target variable is a multi-class categorical label, with class names containing the substring "fault," representing different fault conditions—For example, Ball_007_1 and OR_007_6_1 represent a ball fault with a defect diameter of 0.007 inches (load condition 1) and an outer-race fault with a defect diameter of 0.007 inches at the 6 o'clock position (load condition 1), respectively. The shape of the dataset (i.e., the number of samples and number of features) and the distribution of the target classes were examined prior to modeling to characterize class balance and dataset dimensionality. No scaling, transformation, or normalization was applied to the feature values in order to preserve the original raw values throughout the analysis pipeline.

Because the target labels represent multiple fault categories, a set of purely unsupervised anomaly detection and clustering algorithms was employed to estimate class membership without using label information during training. The following eight algorithms were applied independently, without combining or ensembling their outputs, and each was run using its default hyperparameter settings as provided in the respective implementation—primarily scikit-learn.

Isolation Forest (IF) is an ensemble-based anomaly detection method

Table 1
label prediction accuracy and F1 per-class scores across unsupervised algorithms.
IF: Isolation Forest, LOF: Local Outlier Fact, KDE: Kernel Density Estimation, GMM: Gaussian Mixture Model, URF: Unsupervised Random Forest, DB: DBSCAN, MS: Mean Shift, OP:Optics.

algo	acc	F1_Ball_007_1	F1_Ball_014_1	F1_Ball_021_1	F1_IR_007_1	F1_IR_014_1	F1_IR_021_1	F1_Normal_1	F1_OR_007_6_1	F1_OR_014_6_1	F1_OR_021_6_1
IF	**0.857**	**0.779**	**0.710**	0.742	**0.989**	0.943	0.911	0.892	**1**	**0.700**	0.928
LOF	0.693	0.286	0.455	0.553	0.722	0.854	0.872	0.884	0.918	0.529	0.903
KDE	0.77	0.54	0.538	0.708	0.943	0.892	0.903	0.835	0.947	0.5	**0.938**
GMM	0.848	0.854	0.687	**0.784**	**0.989**	**1**	**0.923**	0.821	0.989	0.471	0.929
URF	0.559	0	0.265	0.433	0.575	0.857	0.772	0.857	0.747	0	0.562
DB	0.472	0.512	0.361	0.637	0.73	0.73	0.526	0	0	0	0
MS	0.348	0.259	0	0	0	0	0.725	0	0.821	0	0.804
OP	0.8	0.722	0.387	0.659	0.893	0.935	0.915	**0.909**	0.939	0.61	0.93

Table 2
Classification performance of Isolation Forest over 10 independent runs, reporting per-class F1 scores and overall accuracy (mean ± standard deviation, with 95% confidence intervals) CIs truncated to the valid range [0, 1].

class	F1_mean	F1_std	F1_CI_lower	F1_CI_upper
Ball_007_1	0.8526	0.0455	0.8200	0.8852
Ball_014_1	0.7386	0.0368	0.7123	0.7650
Ball_021_1	0.8473	0.0497	0.8117	0.8829
IR_007_1	0.9945	0.0078	0.9890	1.0000
IR_014_1	0.9357	0.0284	0.9154	0.9561
IR_021_1	0.9703	0.0262	0.9515	0.9890
Normal_1	0.9491	0.0286	0.9286	0.9696
OR_007_6_1	0.9933	0.012	0.9847	1.0000
OR_014_6_1	0.7515	0.0498	0.7159	0.7871
OR_021_6_1	0.9594	0.0168	0.9474	0.9714
OVERALL_ACCURACY	0.8965	0.0158	0.8852	0.9079

Table 3
Classification performance of Unsupervised Random Forest over 10 independent runs, reporting per-class F1 scores and overall accuracy (mean ± standard deviation, with 95% confidence intervals) CIs truncated to the valid range [0, 1].

class	F1_mean	F1_std	F1_CI_lower	F1_CI_upper
Ball_007_1	0.3936	0.2704	0.2002	0.5870
Ball_014_1	0.4074	0.2584	0.2225	0.5923
Ball_021_1	0.4223	0.1555	0.3111	0.5335
IR_007_1	0.7431	0.1043	0.6685	0.8177
IR_014_1	0.8692	0.0989	0.7985	0.9400
IR_021_1	0.8226	0.0918	0.7569	0.8883
Normal_1	0.8984	0.1785	0.7707	1.0000
OR_007_6_1	0.8228	0.1063	0.7468	0.8989
OR_014_6_1	0.0518	0.1095	0.0000	0.1301
OR_021_6_1	0.7348	0.1468	0.6298	0.8398
OVERALL_ACCURACY	0.6535	0.0782	0.5975	0.7095

that isolates observations through random recursive partitioning. Local Outlier Factor (LOF) is a density-based method that measures the local deviation of a sample with respect to its neighbors. Kernel Density Estimation (KDE) is a non-parametric density estimator used to identify low-density, anomalous regions of the feature space. Gaussian Mixture Model (GMM) is a probabilistic clustering approach assuming that the data are generated from a mixture of several Gaussian distributions. Unsupervised Random Forest (URF) is a random forest variant that constructs a synthetic dataset for self-supervised partitioning, enabling clustering via proximity matrices without label input. Mean Shift (MS) is a non-parametric clustering algorithm that iteratively shifts data points toward regions of higher density. DBSCAN (DB) is a density-based clustering method that groups points that are closely packed while marking sparse points as noise. OPTICS (OP) is a density-based method that extends DBSCAN by producing a reachability plot representing the clustering structure across varying densities.

Based on Breiman' s method [14], unsupervised Random Forest (URF) creates pseudo-labels by distinguishing real data (labeled "1″) from synthetic data (labeled "0") generated by independently shuffling each feature column, using scikit-learn's RandomForestClassifier. The classifier's leaf assignments, obtained via apply(), serve as a learned proximity measure between samples, which is then clustered using KMeans to produce the final unsupervised labels.

Several of the selected algorithms—including IF, LOF, and KDE in binary anomaly detection mode—inherently produce binary outputs, such as normal versus anomalous. To adapt these binary unsupervised models for the multi-class fault classification task, a One-vs-Rest strategy was employed. For each fault class, the corresponding algorithm was trained to distinguish that specific class from all remaining classes combined, producing one binary decision boundary per class. This process resulted in a separate classifier for each fault type, with each classifier specialized in recognizing its own class while treating all other classes as a single, unified negative category. The final predicted class

for each sample was determined by aggregating the outputs from all One-vs-Rest classifiers and selecting the class whose classifier produced the highest confidence score. In cases where an algorithm generated an anomaly score rather than a probability, this score was used as the basis for comparison instead, ensuring a consistent decision rule across all algorithm types.

Software: scikit-learn version [1.6.1], Python [3.9.18] and default settings (random state=42)

Algorithm	Parameter	Default Value
Isolation Forest	n_estimators	100
	max_samples	'auto'
	contamination	'auto'
	max_features	1.0
	bootstrap	False
Local Outlier Factor	n_neighbors	20
	algorithm	'auto'
	metric	'minkowski'
	contamination	'auto'
Kernel Density Estimation	kernel	'gaussian'
	bandwidth	1.0
	algorithm	'auto'
	metric	'euclidean'
Gaussian Mixture Model	n_components	1 (default) — set to number of classes in this study
	covariance_type	'full'
	max_iter	100
	n_init	1
Unsupervised Random Forest*	n_estimators	100
	max_depth	None
	min_samples_split	2
Mean Shift	bandwidth	None (estimated automatically via estimate_bandwidth)
	bin_seeding	False
DBSCAN	eps	0.5
	min_samples	5
	metric	'euclidean'
OPTICS	min_samples	5
	metric	'minkowski'
	cluster_method	'xi'
	xi	0.05

New program (isof.py and urf.py) was developed for Isolation Forest and Unsupervised Random Forest on confidence intervals and standard deviations. The results are as follows with 95% CI. To address this concern quantitatively, we have added Tables 2 and 3, which reports the performance of Isolation Forest and Unsupervised Random Forest over 10 independent runs with different random seeds (affecting both the train/test split and the model's internal stochastic sampling). For each run, we recorded the per-class F1 score and overall accuracy, and computed the mean, standard deviation, and 95% confidence interval (CI) across runs. The results are summarized below:

For reproducibility and transparency, Python code, labelpredfault.py is publicly available at GitHub [15].

3. Results

Table 1 summarizes the overall cross-validatoin accuracy and per-class F1 scores achieved by IF and seven traditional clustering/anomaly detection baselines (LOF, KDE, GMM, URF, DB, MS, and OP) across the ten fault conditions. IF achieved the highest overall accuracy (0.857), slightly outperforming all baseline methods, the closest of which (GMM) trailed by less than one percentage point (0.848), while the majority of the remaining methods fell well below 0.800. IF and GMM show comparable performance (0.857 vs. 0.848), and that GMM was evaluated on a single run only.

Beyond aggregate accuracy, IF also demonstrated markedly more consistent and robust performance across individual fault categories. It attained the highest or near-highest F1 scores in several conditions, including IR_007_1 (0.989), IR_014_1 (0.943), IR_021_1 (0.911), Normal_1 (0.892), OR_007_6_1 (1.0), and OR_021_6_1 (0.928), indicating reliable discrimination of both healthy and faulty states regardless of fault type or severity. Notably, GMM slightly outperformed IF on several classes, including Ball_007_1 (0.854 vs. 0.779), IR_014_1 (1.0 vs. 0.943), and OR_007_6_1 (0.989 vs. 1.0, comparable), suggesting that IF's advantage lies primarily in overall consistency rather than uniform superiority on every individual class.

In contrast, the traditional clustering-based approaches exhibited pronounced instability and, in several cases, complete failure to detect specific fault classes. URF and DB both produced F1 scores of 0 for at least one category (URF: Ball_007_1 and OR_014_6_1; DB: Normal_1, OR_007_6_1, OR_014_6_1, and OR_021_6_1), reflecting a systematic inability to separate these classes from the rest. MS performed particularly poorly, with F1 scores of 0 in as many as four fault categories (Ball_014_1, Ball_021_1, IR_007_1, and OR_007_6_1), underscoring its sensitivity to class imbalance and non-convex fault distributions. Even the comparatively stronger baselines, KDE and LOF, showed considerable variability, with F1 scores dropping below 0.55 for challenging classes such as Ball_007_1 (KDE: 0.54; LOF: 0.286) and OR_014_6_1 (KDE: 0.5; LOF: 0.529).

Overall, these results highlight that while traditional clustering algorithms are prone to inconsistent performance—often failing entirely on specific fault types (particularly URF, DB, and MS)—Isolation Forest not only achieved the highest overall accuracy but also maintained relatively stable performance across most fault categories—with a few notable exceptions, such as Ball_007_1 ($F1 = 0.779$) and OR_014_6_1 ($F1 = 0.700$), where accuracy declined—whereas competing methods such as GMM, Mean Shift, and DBSCAN exhibited failures on specific fault classes, producing F1 scores of zero in several instances. Notably, OP also performed competitively, achieving the third-highest overall accuracy (0.8) and relatively strong F1 scores across most classes (e.g., IR_014_1: 0.935; Normal_1: 0.909; OR_007_6_1: 0.939), though it still trailed IF and GMM in overall accuracy.

4. Discussion

The comparative results across eight unsupervised models reveal a substantial disparity in robustness that single-model studies in the prior literature could not have detected—Isolation Forest not only achieved the highest overall accuracy but also sustained this advantage consistently across nearly every fault category, whereas competing methods such as GMM, Mean Shift, and DBSCAN exhibited failures on specific fault classes, producing F1 scores of zero in several instances.

This pattern underscores a critical limitation of relying on a single unsupervised model when constructing or validating a dataset, since the apparent adequacy of a chosen model may mask systematic blind spots that only become visible through comparative evaluation. The observed instability in density- and clustering-based methods—particularly their sensitivity to class imbalance and non-convex fault distributions—suggests that these approaches are structurally ill-suited to the geometry of bearing fault data, where fault severities may produce overlapping or irregularly shaped feature distributions that violate the assumptions underlying methods like Gaussian mixtures or density-reachability clustering. In contrast, the ensemble-based, recursive-partitioning mechanism of Isolation Forest appears inherently more resilient to such irregularities, likely because it does not depend on assumptions of cluster convexity or uniform density, allowing it to isolate anomalous or minority fault classes more reliably under the One-vs-Rest framework.

These findings directly support the paper's central argument that label-prediction accuracy functions as a diagnostic proxy for data quality and model suitability—when accuracy is low or highly variable across classes, this signals either missing or insufficiently discriminative features, or a mismatch between the chosen algorithm's assumptions and the true underlying data structure, or both simultaneously. Consequently, practitioners designing new reliability datasets should not treat

unsupervised model selection as an afterthought but should instead conduct comparative benchmarking similar to that performed here, since the choice of model can determine whether specific fault types are detected at all rather than merely affecting marginal accuracy gains.

Although this study examined a single dataset, the consistent superiority of Isolation Forest observed in this context suggests that an ensemble of unsupervised models may offer a promising candidate than probabilistic or density-based alternatives, particularly in real-world industrial settings where labeled failure data remains scarce and ground-truth validation is unavailable. This practical implication is noteworthy given that recursive partitioning within an ensemble framework appears to be more robust to the absence of labeled data than single probabilistic or density-based models. Nevertheless, this observation should be regarded as a preliminary indication rather than a general recommendation, and further validation across diverse datasets and operating conditions is needed to confirm its broader applicability.

Theoretically, the superior performance of the Isolation Forest ensemble can be attributed to its recursive, axis-aligned spatial partitioning mechanism. Unlike parametric or density-based clustering models such as Gaussian Mixture Models, DBSCAN, or Mean Shift—which inherently rely on strong geometric assumptions like cluster convexity, spherical shapes, or uniform density—Isolation Forest isolates observations based on structural sparsity. This characteristic makes it exceptionally robust against the complex, overlapping, and non-convex feature distributions frequently generated across different bearing fault types and severity levels. Furthermore, within our One-vs-Rest framework, each class-specific model independently learns the localized inlier boundary of its designated state, and subsequent z-score normalization standardizes decision scores across all classes. This allows the ensemble to make balanced, scale-invariant multi-class predictions without being biased by global density disparities across different fault categories.

From a practical perspective, these findings indicate that tree-based anomaly ensembles offer a highly dependable, calibration-free default for real-world industrial condition monitoring, where labeled failure data is typically scarce and clustering methods are prone to detection failures ($F1 = 0$). Additionally, our results demonstrate that evaluating per-class prediction scores provides actionable diagnostic feedback, helping practitioners pinpoint which specific fault states lack sufficient discriminative features during early-stage dataset construction and sensor feature selection. We have fully incorporated these theoretical insights and practical considerations into the revised manuscript.

Future work should extend this comparative framework to additional datasets and loading conditions to assess whether Isolation Forest's advantage generalizes beyond the specific bearing fault dataset examined here, and should further investigate whether hybrid approaches combining ensemble-based detection with dimensionality reduction techniques can mitigate the class-specific failures observed in purely density-based pipelines.

5. Conclusion

This study systematically evaluated eight unsupervised anomaly detection and clustering algorithms for multi-class bearing fault diagnosis under a One-vs-Rest (OvR) benchmarking framework. The empirical results demonstrate that Isolation Forest (IF) achieved the highest overall accuracy (0.857) and exhibited superior consistency across diverse fault categories and severity levels, whereas traditional clustering methods (such as Mean Shift, DBSCAN, and Unsupervised Random Forest) suffered from severe instability and detection failures ($F1=0$) on multiple fault classes. These findings confirm that label-prediction accuracy serves not only as a quantifiable metric for data quality and missing-feature diagnosis, but also as an essential guide for model selection in reliability-critical applications.

Despite these promising results, several limitations should be acknowledged. First, the evaluation was conducted on a single publicly available bearing dataset under a specific load condition (1 HP, 1772 rpm), meaning that the observed superiority of Isolation Forest remains a preliminary indication and requires further validation across diverse operating regimes, noise levels, and machine architectures. Second, while IF demonstrated strong overall robustness, its performance was not uniformly superior across every individual category, showing reduced accuracy on select challenging classes (e.g., Ball_007_1 and OR_014_6_1) where localized density methods slightly outperformed it. Finally, the models were evaluated using default hyperparameter configurations without feature transformation or dimensionality reduction, which may not capture the optimal boundary for complex, non-convex fault distributions.

A limitation of this study is that robustness across fault severities and data quality was not systematically evaluated. In particular, severity-wise comparisons were not conducted, and these will be addressed in future work. The results should therefore not be interpreted as evidence of robustness in these respects.

Future work will address these limitations by validating the proposed framework across heterogeneous industrial datasets and loading conditions, while investigating hybrid architectures that integrate recursive-partitioning ensembles with dimensionality reduction to mitigate localized class-specific performance drops.

Funding

This research has no fund.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Code availability

The Python code supporting the findings of this study is openly available on GitHub (https://github.com/y-takefuji/bearing).

AI use

Not applicable.

According to ScholarGPS

Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 25th out of 1287,415 scholars in life sciences, 22nd out of 805,705 in COVID-19, and 1st out of 109,919 in environmental sciences.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

None.

Data availability

The dataset analyzed in this study is publicly available in the Kaggle repository (https://www.kaggle.com/datasets/brjapon/cwru-bearing-datasets).

References

[1] Baek S, Song K, Lee I. Single-cell foundation models: bringing artificial intelligence into cell biology. Exp Mol Med 2025;57(10):2169–81. https://doi.org/10.1038/s12276-025-01547-5.

[2] De Donno C, Hediyeh-Zadeh S, Moinfar AA, Wagenstetter M, Zappia L, Lotfollahi M, Theis FJ. Population-level integration of single-cell datasets enables multi-scale analysis across samples. Nat Methods 2023;20(11):1683–92. https://doi.org/10.1038/s41592-023-02035-2.

[3] Hanif A, Yıldız İ, Tian P, Kalkanlı B, Erdoğmuş D, Ioannidis S, Dy J, Kalpathy-Cramer J, Ostmo S, Jonas K, Chan RVP, Chiang MF, Campbell JP. Improved training efficiency for retinopathy of prematurity deep learning models using comparison versus class labels. Ophthalmol Sci 2022;2(2):100122. https://doi.org/10.1016/j.xops.2022.100122.

[4] Menshawi A, Hassan MM, Allheeib N, Fortino G. A hybrid generic framework for heart problem diagnosis based on a machine learning paradigm. Sensors (Basel) 2023;23(3):1392. https://doi.org/10.3390/s23031392.

[5] Rahal M, Ahmed BS, Szabados G, Fornstedt T, Samuelsson J. Enhancing machine learning performance through intelligent data quality assessment: an unsupervised data-centric framework. Heliyon 2025;11(4):e42777. https://doi.org/10.1016/j.heliyon.2025.e42777.

[6] Shaw D, Chen H, Jiang T. DeepIsoFun: a deep domain adaptation approach to predict isoform functions. Bioinformatics 2019;35(15):2535–44. https://doi.org/10.1093/bioinformatics/bty1017.

[7] Sumpena E, Cornelio A, Collazo A, Ting SJ, Kowalczyk T, Jiang X, Beiser A, Seshadri S, Kashani AH, Jones CK. Unsupervised quality assessment with generative adversarial networks for 3D OCTA microvascular imaging. Biomed Opt Express 2026;17(1):378–93. https://doi.org/10.1364/BOE.573843.

[8] Zhou M, Luo Z, Yin YH, Liu Q, Wang G, Zhao Y. scATD: a high-throughput and interpretable framework for single-cell cancer drug resistance prediction and biomarker identification. Brief Bioinformatics 2025;26(3):bbaf268. https://doi.org/10.1093/bib/bbaf268.

[9] Ehrlinger L, Wöß W. A survey of data quality measurement and monitoring tools. Front Big Data 2022;5:850611. https://doi.org/10.3389/fdata.2022.850611.

[10] Mohammed S, Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, Patzlaff H, Naumann F, Harmouch H. The effects of data quality on machine learning performance on tabular data. Inf Syst 2025;132:102549. https://doi.org/10.1016/j.is.2025.102549.

[11] Gao X, Mu Z, Gao H, Han H, Yang H, Qi Y. CMCA: a cluster matching based class-incremental unsupervised domain adaptation method for bearing fault diagnosis. Neurocomputing 2026;703:134717. https://doi.org/10.1016/j.neucom.2026.134717.

[12] Liu Y, Zhang L, Liu J, Zhang Q. An angular margin aware unsupervised domain adaptation network for cross domain rolling bearing fault diagnosis. Eng Appl Artif Intell 2026;182(Part 1):115862. https://doi.org/10.1016/j.engappai.2026.115862.

[13] Kaggle. feature_time_48k_2048_load_1.csv. https://www.kaggle.com/datasets/brjapon/cwru-bearing-datasets; 2026.

[14] Breiman L. Manual on setting up, using, and understanding random forests V3.1. Berkeley: University of California; 2003.

[15] GitHub. labelpredfault.Py, isof.Py and urf.Py. https://github.com/y-takefuji/bearing; 2026.