Generalizing Bearing Fault Diagnosis to Physically Unseen Bearings: Adversarial Representation Learning and Bearing-Identity Analysis

Rashwan Ammon

Mechatronics Engineering Department, Latakia University, Latakia, Syria

rashwan.ammon@latakia-univ.edu.sy

Abstract

Reliable bearing fault diagnosis requires generalization to physical components unseen during model development. This study evaluates such generalization using strict 32-fold Leave-One-Bearing-Out (LOBO) testing on the Paderborn University Bearing Dataset. All four diagnostic classes are retained, and each physical bearing is treated as the unit of generalization. The input combines 53 dynamic vibration-derived features and 30 physics-informed features. A classical machine-learning baseline, a deep model without adversarial learning (No-GRL), and a class-conditional Gradient-Reversal Learning (GRL) model with four bearing-domain adversarial heads are compared under the same protocol. At the bearing level, the classical model achieves 81.25% Accuracy, 74.81% Balanced Accuracy, and 77.82% Macro-F1; No-GRL achieves 78.12%, 74.81%, and 77.75%; and GRL achieves 81.25%, 76.89%, and 79.85%, respectively. Relative to No-GRL, GRL yields numerical gains of 3.13, 2.08, and 2.10 percentage points. However, only one held-out bearing changed from an incorrect No-GRL prediction to a correct GRL prediction, with no observed regression; the exact two-sided McNemar test gives p=1.000. A post-hoc Bearing-Identity Representation Audit shows that mean linear Bearing-ID accuracy decreases from 99.22% for No-GRL to 86.27% for GRL across the folds. This remains well above the 31-class chance level, indicating reduced rather than eliminated bearing-specific information. Thus, GRL is associated with modest aggregate diagnostic gains and lower linear recoverability of bearing identity, but the paired results do not establish a statistically significant general improvement in fault classification.

Rashwan Ammon. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

https://doi.org/10.36001/IJPHM.2026.v17i2.5264

1. Introduction

Rolling-element bearings are among the most critical components in rotating machinery, and their degradation can directly affect machine availability, maintenance cost, and operational reliability. Consequently, intelligent fault diagnosis (IFD) of rolling bearings has received substantial attention, particularly with the development of machine-learning (ML) and deep-learning (DL) methods capable of learning discriminative representations from condition-monitoring signals (Hoang & Kang, 2019; Lei et al., 2020; Zhao et al., 2020). The Paderborn University (PU) Bearing Dataset has become an important benchmark for data-driven bearing diagnosis because it provides experimentally characterized bearing conditions and multiple measurements suitable for systematic investigation of bearing-fault classification (Lessmeier et al., 2016).

Deep-learning methods have demonstrated strong diagnostic capabilities by learning task-relevant representations from vibration and other condition-monitoring signals, reducing the dependence on manually designed signal-processing pipelines (Hoang & Kang, 2019; Zhao et al., 2020). However, high classification performance under conventional sample-wise, segment-wise, or recording-wise data partitions does not necessarily imply that a model can generalize to a physically unseen bearing. In particular, multiple recordings or signal segments originating from the same physical bearing may share persistent characteristics that are unrelated to the underlying fault mechanism. If data from the same physical bearing are present in both training and test partitions, a model may exploit such bearing-specific characteristics rather than learning fault-related information that transfers to a new physical component (Vieira et al., 2026).

This issue is closely related to the broader problem of domain generalization (DG), whose objective is to learn representations that remain useful when the model encounters previously unseen domains (Muandet et al.,

2013). In bearing fault diagnosis, recent DG studies have primarily investigated shifts associated with operating conditions, loads, speeds, machines, or other changes in the data-generating environment (Chen et al., 2022; Cheng et al., 2024; Li et al., 2023; Shi et al., 2023; Zhu et al., 2024). Adversarial representation-learning strategies have been particularly attractive in this context because they provide a mechanism for encouraging a shared representation to become less dependent on domain-specific information while preserving information useful for the primary diagnostic task (Ganin et al., 2016).

Nevertheless, an unseen operating condition and an unseen physical bearing should not be treated as identical generalization problems. In an unseen-condition setting, recordings from the same physical bearing may remain available during model development, while the operating condition changes between training and testing. By contrast, in an unseen-bearing setting, the physical component itself is absent from model development. The latter setting is particularly relevant to maintenance scenarios in which a trained diagnostic system may encounter a replacement bearing whose physical characteristics were not represented during training.

Recent methodological research has highlighted the importance of this distinction. Vieira et al. (2026) showed that data leakage is prevalent in bearing fault-diagnosis research and demonstrated that segment-wise and condition-wise partitions can introduce spurious correlations that inflate performance estimates. Their study advocates strict bearing-wise train-test separation, in which data originating from the same physical bearing are assigned exclusively to either training or testing, and evaluates this issue on several public bearing datasets, including the Paderborn University dataset (Vieira et al., 2026). These findings reinforce the need to define the physical bearing itself as the fundamental unit of generalization when the intended objective is to assess performance on previously unseen bearings.

Adversarial domain learning provides a natural framework for investigating this problem. The Gradient Reversal Layer (GRL), introduced within the Domain-Adversarial Neural Network (DANN) framework, enables a feature extractor to learn representations that are simultaneously useful for the primary classification task and less informative for domain discrimination (Ganin et al., 2016). Adversarial domain-generalization methods have subsequently been applied to bearing fault diagnosis under unseen operating conditions (Chen et al., 2022; Li et al., 2023). Physics-informed and causal approaches have also been investigated as complementary mechanisms for improving robustness under distribution shifts (Kim et al., 2024; Zhu et al., 2024). These developments suggest that reducing dependence on domain-specific information may improve diagnostic generalization; however, the specific domain selected for adversarial learning is crucial to the interpretation of the resulting model.

In the present study, the physical bearing is explicitly treated as the domain of generalization. A strict Leave-One-Bearing-Out (LOBO) protocol is therefore adopted, in which one complete physical bearing is held out in each outer fold and is excluded from training and model development. The study retains the four-class formulation of the PU dataset—Healthy, Outer-Race Fault, Inner-Race Fault, and Combined Fault—rather than excluding the combined-fault category because of its relatively small number of physical bearings. This design allows the practical difficulty of generalizing across physical bearings to be evaluated while explicitly reporting the uncertainty associated with the limited representation of the Combined Fault class.

The proposed learning framework combines dynamic vibration-derived features with physics-informed features and employs a recurrent shared representation based on a gated recurrent unit (GRU). A controlled comparison is performed between three model families: (i) a conventional machine-learning baseline, (ii) a deep-learning model without adversarial learning (No-GRL), and (iii) a class-conditional gradient-reversal model (GRL). In the GRL model, separate adversarial domain heads are associated with the four diagnostic classes. This class-conditional design is intended to encourage the shared representation to rely less on bearing-specific information within each fault category while preserving information useful for fault discrimination.

An important distinction is made between the mechanism of adversarial learning and the empirical evidence used to evaluate its effect. The GRL is theoretically intended to encourage domain-invariant representations (Ganin et al., 2016), but its use alone does not establish that physical-bearing identity has been eliminated from the learned representation. Therefore, in addition to the primary fault-diagnosis evaluation, the present study performs a post-hoc Bearing-Identity Representation Audit. In this analysis, the shared representations learned during the LOBO experiments are used to determine how readily the identity of the physical bearing can still be recovered using a separate linear classifier. The audit consequently provides representation-level evidence that complements the task-level fault-classification results.

The central research question is:

To what extent can class-conditional gradient-reversal domain-adversarial learning improve fault classification when the target physical bearing is completely unseen during model development?

Accordingly, the main contributions of this work are as follows:

1. Strict physical-bearing evaluation: A 32-fold LOBO protocol excludes the test bearing from model development in every fold. All four diagnostic classes, including Combined Fault, are retained, and bearing-level Accuracy, Balanced Accuracy, and Macro-F1 are reported.

2. Class-conditional adversarial learning: A GRL architecture with four class-conditional bearing-domain heads is investigated using a unified representation of 53 dynamic and 30 physics-informed features.

3. Controlled comparison: Classical machine learning, deep No-GRL, and deep GRL models are compared under the same physical-bearing generalization protocol, allowing the effect associated with adding adversarial learning to be examined.

4. Task- and representation-level analysis: Paired bearing-level error analysis is combined with a post-hoc Bearing-Identity Representation Audit to assess both diagnostic outcomes and the linear recoverability of physical-bearing identity.

The remainder of this paper is organized as follows. Section 2 reviews conventional bearing fault diagnosis, domain generalization, adversarial representation learning, physics-informed approaches, and recent work on leakage-aware bearing-wise evaluation. Section 3 describes the Paderborn dataset, feature representation, strict LOBO protocol, model architectures, GRL formulation, training procedure, evaluation metrics, and post-hoc Bearing-Identity Representation Audit. Section 4 presents the experimental results and discusses the comparison between the classical, No-GRL, and GRL models, including bearing-level errors and representation-level analysis. Section 5 summarizes the main findings, limitations, and directions for future research.

2. Related Work

2.1. Conventional Machine Learning and Deep Learning for Bearing Fault Diagnosis

Intelligent fault diagnosis of rolling-element bearings has evolved from conventional signal-processing and machine-learning approaches toward deep learning models that learn diagnostic representations directly from condition-monitoring data. The Paderborn University (PU) Bearing Dataset has become an important benchmark for this research area because it provides experimentally characterized bearing conditions together with vibration and motor-current measurements, enabling systematic investigation of data-driven bearing diagnosis (Lessmeier et al., 2016). Early studies and benchmark investigations demonstrated the feasibility of machine-learning-based classification using this dataset, while subsequent research increasingly adopted deep neural architectures for automated feature extraction (Lei et al., 2020; Zhao et al., 2020).

Deep learning has substantially expanded the range of approaches available for bearing diagnosis. Convolutional neural networks (CNNs), recurrent neural networks, autoencoders, and related architectures have been extensively investigated for extracting discriminative representations from vibration signals (Hoang & Kang, 2019; Zhao et al., 2020). However, high classification performance under conventional random, segment-wise, or recording-wise partitions does not necessarily indicate that the learned representation will generalize to a physically unseen bearing. This distinction has become increasingly important as the field has moved from laboratory-oriented classification toward more realistic predictive-maintenance scenarios.

Consequently, the central challenge is no longer only how accurately a model can classify samples originating from a known population of bearings, but also whether the learned diagnostic information remains useful when the physical bearing itself has not been observed during model development.

2.2. Domain Generalization and Domain-Invariant Representation Learning

Domain generalization (DG) addresses the broader problem of learning representations from one or more source domains that remain useful in previously unseen domains. A fundamental objective of DG is to identify information that is predictive of the target task while reducing dependence on domain-specific characteristics (Muandet et al., 2013). In condition-monitoring applications, such characteristics may arise from operating conditions, machines, sensors, acquisition settings, or other factors unrelated to the underlying fault mechanism.

Several recent studies have therefore investigated domain-generalization strategies for rotating machinery and bearing fault diagnosis and related bearing prognostics under unseen conditions (Chen et al., 2022; Ding et al., 2023; Guo et al., 2025; Li et al., 2023; Pang et al., 2024; Shi et al., 2023). For example, Chen et al. proposed an adversarial domain-

invariant generalization framework that learns domain-invariant information from multiple available source conditions without requiring target-condition data during training. Similarly, Li et al. introduced adversarial domain augmentation to construct additional domains and improve generalization to unseen working conditions (Li et al., 2023). Shi et al. investigated domain augmentation generalization for real-time fault diagnosis under unseen working conditions (Shi et al., 2023). These studies illustrate the increasing use of adversarial representation learning and domain augmentation as mechanisms for reducing sensitivity to domain-specific variations.

Other approaches have incorporated physical knowledge or causal representations into domain-generalization frameworks. Kim et al. developed a single-source domain-generalizable and physically interpretable framework for bearing diagnosis under unseen working conditions, emphasizing the use of physical knowledge to improve generalization and interpretability. Zhu et al. proposed PhysiCausalNet, which combines causal and physics-driven mechanisms for cross-machine fault diagnosis under unseen domains. Pang et al. developed a fault-aware domain-generalization framework driven by a fault vibration model (Pang et al., 2024), while Cheng et al. investigated causal feature learning for domain generalization across mixtures of working conditions and machines (Cheng et al., 2024). In addition, Ding et al. investigated adversarial out-domain augmentation for bearing remaining useful life prediction under unseen conditions (Ding et al., 2023), and Guo et al. explored a condition-guided diffusion-model framework for bearing fault diagnosis under unknown working conditions (Guo et al., 2025). These studies collectively indicate that domain invariance, physical knowledge, causal representations, domain augmentation, and generative approaches are promising directions for improving robustness under distribution shifts. Nevertheless, the meaning of a "domain" in bearing diagnosis requires careful consideration because an unseen operating condition and an unseen physical bearing represent different sources of distribution shift.

2.3. Unseen Operating Conditions Versus Unseen Physical Bearings

A substantial portion of the bearing domain-generalization literature evaluates generalization across unseen operating conditions. In such settings, the physical machine or bearing may be represented in both training and test data, while the operating condition, load, speed, or related environmental factor changes between the domains (Chen et al., 2022; Kim et al., 2024; Li et al., 2023; Shi et al., 2023). This setting is practically important because industrial machinery commonly operates under varying conditions.

However, an unseen physical bearing constitutes a different and potentially more demanding generalization problem. When recordings from the same physical bearing appear in both training and testing, a model may exploit bearing-specific characteristics that are stable across recordings but unrelated to the intended fault class. Such characteristics may reflect persistent physical, structural, or measurement-related variability that is not the intended diagnostic target (Vieira et al., 2026).

Therefore, an evaluation in which the test bearing is completely excluded from model development provides a substantially different interpretation of generalization. In this setting, the physical bearing itself becomes the held-out domain, and all recordings belonging to that bearing must remain outside training and model-selection procedures.

Recent research has explicitly highlighted this issue. Vieira et al. demonstrated that commonly used segment-wise and condition-wise partitioning strategies can introduce data leakage and artificially inflate bearing-diagnosis performance. Their study advocates strict bearing-wise train-test separation and evaluates this issue across several public datasets, including the Paderborn University dataset (Vieira et al., 2026). Importantly, their analysis also emphasizes the role of the number and diversity of unique physical bearings in determining generalization capability.

This recent evidence strengthens the methodological motivation for treating the physical bearing—not merely the recording, operating condition, or signal segment—as the fundamental unit of generalization in the present study.

2.4. Adversarial Representation Learning and Gradient Reversal

Adversarial representation learning provides one mechanism for encouraging a model to rely less on domain-specific information. The Domain-Adversarial Neural Network (DANN) framework introduced the widely used Gradient Reversal Layer (GRL), which allows a feature extractor to be optimized simultaneously for the primary task and against a domain classifier (Ganin et al., 2016). During forward propagation, the GRL behaves as an identity mapping, whereas during backpropagation it reverses and scales the gradient received from the domain classifier.

The underlying objective is to learn representations that remain discriminative for the primary classification task while becoming less informative about the domain used by the adversarial classifier. In bearing diagnosis, this principle

has been adopted in several domain-adaptation and domain-generalization studies (Chen et al., 2022; Li et al., 2023).

Conditional adversarial strategies are particularly relevant when domain alignment is expected to interact with class structure (Zhang et al., 2021). Zhang et al. proposed a conditional adversarial domain-generalization method for bearing fault diagnosis in which domain discrimination is conditioned on fault-related information. Such approaches recognize that forcing complete domain invariance across all samples may be undesirable because diagnostic classes themselves can have different distributions.

The present study builds on this general principle but considers a more specific setting: the physical bearing is treated as the domain, and adversarial learning is performed conditionally within the four fault categories. The objective is therefore not to remove all bearing-related information from the representation. Instead, the adversarial mechanism is used to encourage the shared representation to rely less on bearing-specific information while retaining information useful for fault discrimination.

This distinction is important because complete elimination of bearing identity is neither assumed nor directly guaranteed by adversarial training. Accordingly, the present study complements the task-level LOBO evaluation with a separate post-hoc Bearing-Identity Representation Audit to quantify how readily physical-bearing identity can still be recovered from the learned representation.

Prior studies have investigated adversarial representation learning and conditional domain discrimination for bearing diagnosis (Chen et al., 2022; Ganin et al., 2016; Li et al., 2023; Zhang et al., 2021). The present work differs by defining the domain as the physical bearing and pairing strict bearing-held-out evaluation with a post-hoc linear Bearing-Identity Representation Audit. Its contribution is this combined empirical assessment of diagnostic performance and residual identity recoverability, rather than a claim to introduce gradient reversal itself.

2.5. Physics-Informed and Feature-Engineered Representations

Another direction in bearing fault diagnosis seeks to incorporate physical knowledge into the representation rather than relying exclusively on automatically learned features. Physics-informed or physically interpretable approaches can provide information related to characteristic fault frequencies and other physically meaningful mechanisms associated with fault generation (Kim et al., 2024; Zhu et al., 2024).

Recent domain-generalization studies have increasingly combined physical information with representation learning. Kim et al. demonstrated that physically informed representations can support generalization to unseen working conditions while also providing a degree of physical interpretability. Similarly, PhysiCausalNet integrates causal and physics-driven information to address cross-machine generalization (Zhu et al., 2024). Pang et al. investigated fault-aware domain generalization using a vibration-model-driven framework, while Cheng et al. explored causal feature learning for generalization across multiple working conditions and machines (Cheng et al., 2024). These studies motivate the inclusion of physics-informed descriptors as complementary information to data-driven vibration features.

In the present work, dynamic vibration-derived features and physics-informed features are therefore combined into a unified representation. This design is intended to expose the learning model to both signal-derived and physically motivated information. However, the study does not assume a priori that either feature family is individually superior. Their relative contribution is treated as an empirical question and is discussed separately from the principal domain-generalization comparison.

2.6. Limitations of Existing Evaluation Practices

Despite substantial progress in intelligent bearing diagnosis, several methodological limitations remain relevant to the interpretation of reported results.

First, random, segment-wise, or recording-wise partitions can place highly correlated observations from the same physical bearing in both training and test sets. In that case, test performance may partly reflect recognition of a bearing already represented during training, rather than generalization to a new physical component. Recent systematic evidence has shown that such leakage can materially affect reported results and has emphasized the importance of strict bearing-wise separation (Vieira et al., 2026).

Second, domain-generalization studies often define the domain by operating condition, machine, or dataset rather than by physical bearing. Although these studies address important forms of distribution shift, they do not necessarily establish generalization to a bearing that was entirely absent from model development.

Third, adversarial methods are commonly assessed using diagnostic performance as the main outcome (Chen et al., 2022; Li et al., 2023; Pang et al., 2024; Zhang et al., 2021). Better classification performance alone does not show whether physical-bearing identity remains recoverable from

the learned representation. These are related but distinct questions and should be evaluated separately.

Finally, differences in the number of physical bearings across diagnostic classes can make overall accuracy insufficient for describing class-specific performance. This is particularly relevant to the Combined Fault class, which contains only three bearings in the present dataset. Balanced Accuracy and Macro-F1 are therefore reported alongside Accuracy to provide a more class-sensitive assessment.

2.7. Research Gap and Positioning of the Present Study

The literature reviewed above addresses bearing diagnosis, domain generalization, adversarial representation learning, and the risks of bearing-level data leakage. However, these strands do not by themselves establish how class-conditional adversarial learning performs when the test domain is a physical bearing completely excluded from model development, or whether any change in diagnostic performance is accompanied by a reduction in recoverable bearing-identity information. This distinction is important in the four-class Paderborn setting, where the Combined Fault class is represented by only three physical bearings and therefore supports more limited class-specific conclusions. The present study investigates these questions using strict bearing-held-out evaluation and a post-hoc Bearing-Identity Representation Audit, without assuming in advance that GRL will improve classification or eliminate bearing-specific information.

3. Methodology

3.1. Dataset and Feature Representation

The experiments were conducted using the Paderborn University (PU) Bearing Dataset (Lessmeier et al., 2016), with the physical bearing treated as the fundamental unit of generalization. The experimental dataset comprised 32 physical bearings and 2,560 recordings, with 80 recordings associated with each bearing. The bearings were distributed across four diagnostic classes: Healthy, Outer-Race Fault, Combined Fault, and Inner-Race Fault. Specifically, bearings B01–B06 were Healthy, B07–B18 exhibited Outer-Race Faults, B19–B21 represented Combined Faults, and B22–B32 represented Inner-Race Faults. Thus, the four classes contained 6, 12, 3, and 11 physical bearings, respectively.

The feature representation combined dynamic condition-monitoring information with physics-informed descriptors. The dynamic representation contained 53 features, while the physics-informed representation contained 30 features, resulting in an 83-dimensional All83 representation. The same All83 feature representation was used for the principal classical, No-GRL, and GRL experiments to provide a controlled basis for model comparison.

Before model development, the correspondence between the recording-level feature data and the physical-bearing information was independently audited. The FileCode mapping was complete for all 2,560 recordings, with 639,053 out of 639,053 feature rows successfully matched to the corresponding physical-bearing records. No FileCode was found to correspond to more than one physical bearing. This mapping audit ensured that the physical-bearing identity used to define the LOBO folds was consistent with the underlying recording and feature structure.

The overall experimental framework adopted in this study is illustrated in Fig. 1. The framework links the dataset and feature construction stages with the strict physical-bearing LOBO evaluation, the classical and deep-learning model branches, and the subsequent fault-classification and Bearing-ID analyses.

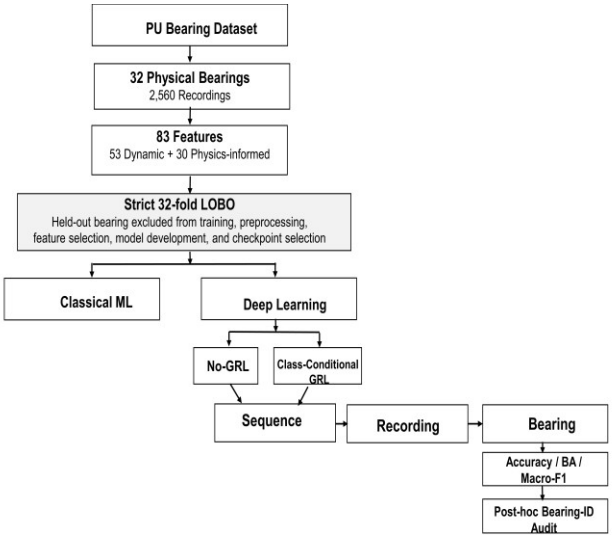

Figure 1. Overall experimental framework.

The framework illustrates the Paderborn dataset, the 32 physical bearings and 2,560 recordings, the 83-feature representation, strict 32-fold LOBO partitioning, the classical ML and deep-learning branches, the No-GRL and class-conditional GRL models, sequence-to-recording-to-bearing aggregation, and the final fault-classification and Bearing-ID analyses.

3.2. Strict Leave-One-Bearing-Out Evaluation

To evaluate generalization to completely unseen physical bearings, a strict leave-one-bearing-out (LOBO) protocol was adopted. The 32 physical bearings were used to define 32 outer folds. In each fold, one physical bearing was

designated as the held-out test bearing, while the remaining 31 bearings constituted the development set.

All model-development operations were restricted to the 31 development bearings. In particular, the held-out bearing was excluded from model training, preprocessing-statistic estimation, feature-selection procedures, checkpoint selection, and any other model-development decision. The resulting test evaluation therefore measured generalization to a physical bearing that was completely unseen during model development.

This protocol differs from recording-wise or sample-wise random splitting, in which recordings originating from the same physical bearing may occur in both training and testing. Such a split can allow bearing-specific characteristics to be shared across the two sets and can therefore produce an overly optimistic estimate of generalization to genuinely unseen bearings (Vieira et al., 2026). In the present study, the physical bearing itself was consequently treated as the domain over which generalization was evaluated.

3.3. Sequence Construction

For the deep-learning experiments, the feature observations were organized into fixed-length temporal sequences. Each sequence contained L=16 consecutive observations, with a sliding step of S=8. Thus, an input sequence was represented as $X_i \in R^{83\times16}$, where the 83 rows correspond to the All83 feature dimensions and the 16 columns correspond to consecutive observations.

Sequence construction was performed independently within each FileCode. Consequently, no sequence was allowed to cross a recording boundary. This restriction prevented observations from different recordings from being combined into a single temporal input and maintained the correspondence between each sequence and its original recording.

The sequence-level predictions were subsequently aggregated at the recording and physical-bearing levels. This multi-level aggregation allowed the experiments to distinguish the behavior of the model at the sequence level from its final performance at the physical-bearing level, which was the primary evaluation unit in this study.

3.4. Model Architectures and Baselines

Three principal model configurations were evaluated under the same strict LOBO framework: a classical machine-learning baseline, a deep-learning model without domain-adversarial training (No-GRL), and a deep-learning model incorporating class-conditional gradient-reversal learning (GRL).

3.4.1. Classical Machine-Learning Baseline

The classical baseline followed a bearing-level nested domain-generalization ensemble protocol under the same 32-fold LOBO framework used for the deep-learning experiments. Each physical bearing was represented by the median of each of its 83 features, producing one 83-dimensional bearing-level feature vector per physical bearing. In every outer fold, the held-out bearing was completely excluded from model development, while the remaining 31 bearings formed the outer-development set. Model and feature-selection decisions were performed using an inner bearing-level validation procedure within the outer-development set. One bearing from each diagnostic class was assigned to inner validation, while the remaining development bearings were used for inner training and feature ranking. Feature validity and feature ranking were determined exclusively from the inner-training bearings during model selection. The candidate feature-set sizes were $K \in \{5,10,15,20,25\}$, and four classifier families were considered: Linear Discriminant Analysis (LDA), linear Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Random Forest (RF). The KNN classifier used seven neighbors with squared-inverse distance weighting, while the Random Forest contained 100 trees with a minimum leaf size of five.

The baseline used a two-stage classification strategy. Stage 1 discriminated Healthy from Faulty bearings, where the latter category comprised Outer-Race, Inner-Race, and Combined Fault bearings. Candidate Stage-1 models were evaluated on the inner-validation bearings using Balanced Accuracy together with a false-positive-rate penalty, with the selection objective defined as

$$J_{S1} = BA - 0.10\,FPR \quad (1)$$

The three highest-ranked Stage-1 candidates were combined into a weighted ensemble, and the decision threshold was selected exclusively from the inner-validation results. Stage-1 ensemble probabilities were obtained from a weighted combination of mean, median, and binary-vote information.

For bearings classified as Faulty by Stage 1, Stage 2 performed three-class discrimination among Outer-Race Fault, Inner-Race Fault, and Combined Fault. Candidate Stage-2 models were ranked primarily by Macro-F1 and secondarily by Balanced Accuracy using the inner-validation bearings. The three highest-ranked candidates were retained and combined using descending ensemble weights.

After the inner model-selection process was completed, the selected model configuration was refitted using the complete outer-development set. Feature rankings for the

final outer fit were recomputed using only these 31 development bearings, while the selected feature-set size and model configuration remained those determined by inner validation. The held-out bearing was then evaluated only once as the outer test bearing.

This nested procedure provided a classical non-deep-learning reference that incorporated bearing-level aggregation, nested feature selection, model selection, threshold selection, and ensemble construction without using information from the held-out physical bearing. Under the resulting strict 32-fold LOBO evaluation, the classical baseline achieved 81.25% Accuracy, 74.81% Balanced Accuracy, and 77.82% Macro-F1.

3.4.2. Deep No-GRL Model

The No-GRL model served as the deep-learning reference without domain-adversarial training. Each input sequence was represented by an 83×16 feature matrix. The shared feature extractor consisted of a fully connected layer with 128 units followed by a ReLU activation and dropout with a rate of 0.20. The resulting representation was processed by a GRU layer with 64 hidden units.

The GRU output was then passed through a fully connected layer with 64 units followed by a ReLU activation, producing a shared 64-dimensional embedding. This embedding was connected to the fault-classification head responsible for predicting the four diagnostic classes.

The No-GRL model therefore provided a controlled reference for assessing whether the addition of the class-conditional adversarial branch produced a measurable change in unseen-bearing fault classification.

3.4.3. Class-Conditional GRL Model

The GRL model retained the same shared feature-extraction architecture as the No-GRL model. The resulting 64-dimensional embedding was connected to two branches: a primary fault-classification branch and an adversarial bearing-domain branch.

The adversarial branch consisted of a gradient-reversal layer followed by four class-conditional bearing-domain classifiers corresponding to the four diagnostic classes: Healthy, Outer-Race Fault, Inner-Race Fault, and Combined Fault. Thus, the physical-bearing identity was modeled conditionally on the diagnostic class rather than through a single unconditional bearing-domain discriminator.

This design was intended to encourage the shared representation to retain information useful for fault classification while reducing information that allowed physical-bearing identity to be discriminated within each diagnostic class. The complete deep-learning architecture and the relationship between the No-GRL and class-conditional GRL branches are illustrated in Fig. 2.

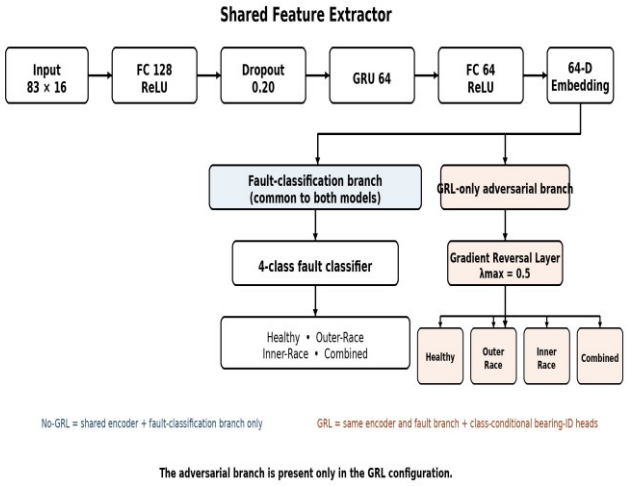

Figure 2. Proposed deep-learning architecture and model branches.

The figure shows the common 83×16 input and shared feature extractor, consisting of the fully connected and GRU layers followed by the 64-dimensional embedding. The No-GRL configuration connects the shared embedding directly to the four-class fault classifier, whereas the GRL configuration retains the same fault-classification branch and additionally applies a Gradient Reversal Layer followed by four class-conditional bearing-domain classifiers.

3.5. Class-Conditional Gradient-Reversal Learning

Let z_i denote the shared representation of sample i, y_i its diagnostic class, and d_i its physical-bearing identity. The primary diagnostic task was optimized using a fault-classification loss denoted by L_{fault}. In parallel, the class-conditional bearing-domain classifiers attempted to predict the physical-bearing identity from the shared representation, producing an adversarial loss denoted by L_{adv} .

θ_f θ_y θ_d The class-conditional adversarial objective follows the standard gradient-reversal formulation (Ganin et al., 2016). Let denote the parameters of the shared feature extractor, those of the fault classifier, and those of the bearing-domain classifiers. Conceptually, the optimization can be expressed as

$$\min_{\theta_f,\theta_y} \max_{\theta_d} [L_{fault}(\theta_f,\theta_y) - \lambda L_{adv}(\theta_f,\theta_d)] \quad (2)$$

The bearing-domain classifiers therefore seek to minimize their physical-bearing classification loss, whereas the shared

feature extractor receives the reversed adversarial gradient and is encouraged to reduce the discriminative information associated with the physical-bearing domain while preserving information useful for fault classification.

In implementation, the gradient-reversal layer leaves the forward activation unchanged. During backpropagation, however, the gradient propagated from the adversarial branch to the shared feature representation is reversed and scaled according to

$$\frac{\partial \mathrm{L_{GRL}}}{\partial \mathrm{z}} = -\lambda \frac{\partial \mathrm{L_{adv}}}{\partial \mathrm{z}} \tag{3}$$

The maximum gradient-reversal coefficient was set to λ_max=0.5, with a five-epoch warm-up period during which the adversarial contribution was gradually introduced. The total training duration was 30 epochs.

Four class-conditional adversarial heads were used so that bearing-domain discrimination was performed separately within each diagnostic class. The adversarial losses contributed to model optimization but were not used as an independent criterion for selecting the final checkpoint.

Importantly, the GRL formulation does not imply that all physical-bearing information is completely removed from the learned representation. Rather, it provides a mechanism for opposing bearing-domain discrimination during representation learning. Whether bearing identity remained linearly recoverable after GRL was therefore examined separately using the post-hoc Bearing-Identity Representation Audit described in Section 3.8.

3.6. Training, Preprocessing, and Checkpoint Selection

All deep-learning models were trained using the same principal training configuration to maintain a controlled comparison between the No-GRL and GRL models. The batch size was 64, the learning rate was 10^-3, the total number of training epochs was 30, and the dropout rate was 0.20. The GRU contained 64 hidden units and the shared embedding dimension was 64. Gradient clipping with a maximum norm of 1.0 was applied during optimization.

Preprocessing parameters were estimated exclusively from the development bearings in each LOBO fold. Non-finite or invalid feature values were handled using statistics derived from the development data. Feature standardization was then performed using the development-set mean and standard deviation. The corresponding parameters were subsequently applied unchanged to the held-out test bearing. Therefore, no statistics estimated from the held-out bearing were used during preprocessing.

Checkpoint selection was based on the minimum mean training fault-classification loss observed over the fixed 30 training epochs. The held-out bearing was not used for checkpoint selection. Consequently, the checkpointing procedure did not use test-bearing performance and did not introduce test-bearing information into model development.

Although this strategy is leakage-free, it does not explicitly optimize the checkpoint for unseen-bearing generalization. This limitation is therefore acknowledged when interpreting the results.

3.7. Evaluation Metrics and Prediction Aggregation

Performance was evaluated using Accuracy, Balanced Accuracy (BA), and Macro-F1. Accuracy was calculated as

$$\mathrm{Accuracy} = \frac{\mathrm{N_{correct}}}{\mathrm{N_{total}}} \tag{4}$$

Balanced Accuracy was calculated as the mean recall across the four diagnostic classes:

$$\mathrm{BA} = \frac{1}{C}\sum_{C=1}^{C} \mathrm{Recall}_C \tag{5}$$

where C=4 is the number of diagnostic classes. Macro-F1 was calculated as

$$\mathrm{Macro-F1} = \frac{1}{C}\sum_{C=1}^{C} \mathrm{F1}_C \tag{6}$$

Balanced Accuracy and Macro-F1 were included because the four diagnostic classes were not equally represented at the physical-bearing level, particularly the Combined Fault class, which contained only three physical bearings.

For the deep-learning models, predictions were first obtained at the sequence level. Sequence predictions belonging to the same recording were aggregated independently, and recording-level predictions were subsequently aggregated to the physical-bearing level using majority voting. The three prediction granularities—sequence, recording, and physical bearing—were retained for analysis, while bearing-level performance was treated as the primary result because the physical bearing was the fundamental unit of generalization.

3.8. Post-Hoc Bearing-Identity Representation Audit

To examine whether the learned representations retained linearly recoverable physical-bearing information after adversarial training, a post-hoc Bearing-Identity

Representation Audit was performed for both the No-GRL and GRL models.

The audit was conducted independently within each of the 32 LOBO folds. The held-out test bearing was excluded from the identity-probe dataset, leaving 31 development bearings as the Bearing-ID classes. For each of these 31 bearings, 60 recordings were used to train the identity probe and 20 recordings were used for its evaluation. The same recording-level split was used for both representations to provide a controlled paired comparison. The embeddings were extracted using the previously trained No-GRL and GRL encoders, whose training and preprocessing-statistic estimation had used recordings from all 31 development bearings. Thus, the 20 probe-test recordings per development bearing were excluded from fitting the identity probe but were not held out from encoder training or preprocessing-statistic estimation.

Sequence-level embeddings produced by the trained deep models were aggregated to the recording level using the median representation. A linear one-versus-one support vector machine was then trained to predict physical-bearing identity from these recording-level representations.

The Bearing-ID audit was evaluated using Accuracy, Balanced Accuracy, and Macro-F1 Because each of the 31 bearing identities contributed the same number of test recordings, Accuracy and Balanced Accuracy are equal under this balanced evaluation design.

The Bearing-ID probe was applied post hoc and was not used to train the fault-classification models, select checkpoints, modify the representations, or determine any hyperparameters. Its purpose was to quantify the extent to which physical-bearing identity remained linearly recoverable from the learned representations. Accordingly, this audit is interpreted as a descriptive assessment of linear identity recoverability rather than as an evaluation on recordings held out from encoder training.

A reduction in Bearing-ID performance after GRL is interpreted as reduced linear recoverability of bearing identity rather than complete elimination of bearing-specific information. In particular, the linear probe provides evidence about information accessible to a linear classifier and cannot by itself establish that all nonlinear or otherwise encoded bearing information has been removed.

3.9. Experimental Integrity and Reproducibility Checks

Several integrity checks were applied before interpreting the final results. First, the mapping between the recording-level feature data and physical-bearing identity was verified for all 2,560 recordings and 639,053 feature rows. Second, all 32 physical bearings were used exactly once as the held-out bearing in the outer LOBO evaluation. Third, the diagnostic-class distribution was maintained as 6 Healthy, 12 Outer-Race Fault, 11 Inner-Race Fault, and 3 Combined Fault physical bearings.

Fourth, the pooled prediction vectors used for the final sequence-, recording-, and bearing-level analyses were checked for completeness and correspondence with the 32 held-out bearings. Finally, the No-GRL and GRL models were evaluated using the same set of 32 held-out physical bearings, enabling a paired bearing-level comparison of their predictions.

These procedures were designed to ensure that the reported differences between the No-GRL and GRL configurations reflect differences in the learned representations and adversarial training strategy rather than differences in the physical-bearing test population or inadvertent information leakage.

4. Experimental Results and Discussion

4.1. Overall Bearing-Level Results

The primary objective of the experimental evaluation was to determine whether class-conditional gradient-reversal learning (GRL) can improve fault classification when the target physical bearing is completely unseen during model development. All three models were therefore evaluated under the same strict 32-fold leave-one-bearing-out (LOBO) protocol, with each physical bearing used once as the held-out test bearing. The bearing-level results are summarized in Table 1

Model	Accuracy (%)	Balanced Accuracy (%)	Macro-F1 (%)
Classical ML	81.25	74.81	77.82
Deep No-GRL	78.12	74.81	77.75
Deep GRL	81.25	76.89	79.85

Table 1. Overall bearing-level performance under strict physical-bearing LOBO.

The classical machine-learning baseline achieved 81.25% Accuracy, 74.81% Balanced Accuracy, and 77.82% Macro-F1. The matched deep model without adversarial learning obtained 78.12%, 74.81%, and 77.75%, respectively. The GRL model achieved 81.25%, 76.89%, and 79.85%, respectively.

Relative to the No-GRL model, the GRL model improved Accuracy by 3.13 percentage points, Balanced Accuracy by 2.08 percentage points, and Macro-F1 by 2.10 percentage points. Thus, GRL improved all three aggregate metrics relative to No-GRL, including both class-sensitive metrics, Balanced Accuracy and Macro-F1. However, the magnitude and statistical interpretation of these gains must be considered alongside the paired bearing-level results.

These results should be interpreted in the context of the evaluation protocol. Because the physical bearing is the held-out unit, the test bearing contributes no recordings to model development, preprocessing estimation, or model selection. This setting is substantially more demanding than random sample-level partitioning and is consistent with recent work emphasizing the importance of bearing-wise evaluation for avoiding leakage and obtaining more realistic estimates of bearing fault diagnosis performance (Vieira et al., 2026).

4.2. Confusion-Matrix Analysis

The bearing-level confusion matrices provide additional information about how the three approaches behaved across the four diagnostic classes. For the classical model, the confusion matrix was

$$C_{\text{Classical}} = \begin{bmatrix} 3 & 3 & 0 & 0 \\ 1 & 11 & 0 & 0 \\ 1 & 0 & 10 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix}$$

where the rows correspond to Healthy, Outer-Race Fault, Inner-Race Fault, and Combined Fault, respectively.

The No-GRL deep model produced

$$C_{\text{No-GRL}} = \begin{bmatrix} 4 & 1 & 1 & 0 \\ 0 & 9 & 3 & 0 \\ 1 & 0 & 10 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix}$$

whereas the GRL model produced

$$C_{\text{GRL}} = \begin{bmatrix} 4 & 1 & 1 & 0 \\ 0 & 10 & 2 & 0 \\ 1 & 0 & 10 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix}$$

The corresponding class-wise recall values are summarized in Table 2.

The GRL model therefore improved the Outer-Race Fault recall relative to No-GRL, from 75.00% to 83.33%, while the Healthy, Inner-Race Fault, and Combined Fault recalls remained unchanged. Consequently, the improvement in Balanced Accuracy was not uniformly distributed across all classes. In particular, the Combined Fault class remained difficult for all three approaches.

Diagnostic class	Classical ML (%)	Deep No-GRL (%)	Deep GRL (%)
Healthy	50.00	66.67	66.67
Outer-Race Fault	91.67	75.00	83.33
Inner-Race Fault	90.91	90.91	90.91
Combined Fault	66.67	66.67	66.67

Table 2. Class-wise recall at the bearing level.

The persistence of the Combined Fault limitation is important because this class contains only three physical bearings in the dataset. Therefore, its bearing-level performance should be interpreted cautiously, and a change of one bearing can have a relatively large effect on the estimated class recall.

The corresponding bearing-level confusion matrices are shown in Fig. 3, providing a direct comparison of the error distribution across the four diagnostic classes.

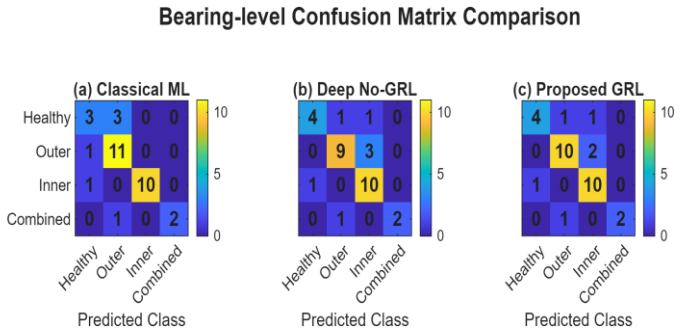

Figure 3. Bearing-level confusion matrix comparison.

The figure presents the bearing-level confusion matrices for the Classical ML, Deep No-GRL, and Deep GRL models using the same class order—Healthy, Outer-Race Fault, Inner-Race Fault, and Combined Fault—and identical axis conventions.

4.3. Paired Bearing-Level Comparison of No-GRL and GRL

Because the No-GRL and GRL models were evaluated on exactly the same 32 held-out bearings, their predictions can be compared on a paired basis. This analysis provides a more direct assessment of whether the aggregate difference between the two models corresponds to broad changes in bearing-level decisions.

The paired comparison showed that both models correctly classified the same 25 bearings and incorrectly classified the same 6 bearings. The remaining bearing, B08, was misclassified by No-GRL as Inner-Race Fault instead of its true Outer-Race Fault class, whereas GRL correctly classified it as Outer-Race Fault. Thus, No-GRL correctly classified 25 of the 32 bearings, compared with 26 for GRL. No bearing was correctly classified by No-GRL but incorrectly classified by GRL.

The paired bearing-level outcomes are summarized in Fig. 4, which shows the correct or incorrect decision for each of the 32 held-out physical bearings under the two deep-learning configurations.

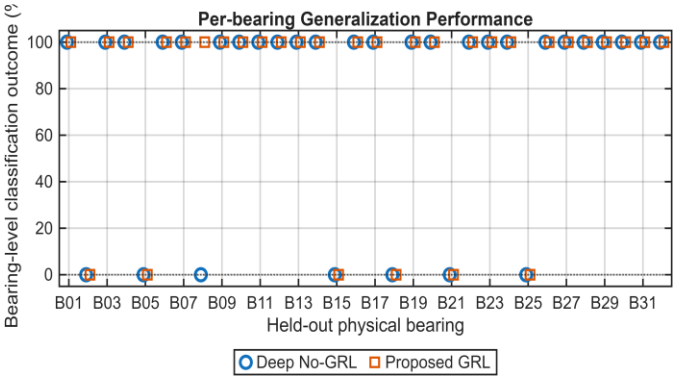

Figure 4. Per-bearing generalization outcome under strict physical-bearing LOBO.

Each marker indicates whether the held-out physical bearing was correctly classified by the Deep No-GRL and proposed GRL models. The two models produced the same decision for 31 of the 32 bearings; B08 was correctly classified by GRL but incorrectly classified by No-GRL, while six bearings remained incorrectly classified by both models.

An exact two-sided McNemar test yielded p=1.000. Accordingly, the observed aggregate improvement should not be interpreted as statistically significant evidence of a general improvement across the 32 physical bearings. Rather, the results indicate a modest empirical improvement that was concentrated in a single corrected bearing and involved no observed regression.

4.4. Persistent Bearing-Level Errors

The bearing-level predictions also revealed a set of persistent difficult cases. Bearings B02, B05, B15, B18, B21, and B25 were misclassified by both the No-GRL and GRL models.

The persistence of these errors indicates that adversarial representation learning did not resolve all challenges associated with unseen physical-bearing generalization. Importantly, no causal explanation is assigned to these individual failures because the present experiments do not provide sufficient evidence to attribute them to a particular mechanical, structural, or measurement characteristic.

Instead, the persistent errors demonstrate that physical-bearing variability remains a substantial challenge even when the representation-learning process is explicitly encouraged to reduce bearing-specific information.

4.5. Comparison with the Classical Machine-Learning Baseline

The classical machine-learning baseline provides an additional reference point for interpreting the deep-learning results. The GRL model and the classical baseline achieved the same bearing-level Accuracy of 81.25%. However, the GRL model achieved higher Balanced Accuracy (76.89% versus 74.81%) and higher Macro-F1 (79.85% versus 77.82%).

Thus, the results do not support a claim that the GRL model uniformly outperformed the classical approach. Rather, the comparison is metric-dependent: the two approaches achieved identical Accuracy, while the GRL model produced higher class-balanced performance according to Balanced Accuracy and Macro-F1.

4.6. Interpretation of the Effect of GRL

The purpose of GRL in the present study was not to create a fundamentally new adversarial-learning mechanism, but to investigate whether class-conditional adversarial representation learning could improve fault classification under a strict unseen-physical-bearing setting. The gradient-reversal mechanism encourages the shared representation to remain useful for the primary fault-classification task while reducing its ability to encode information that is useful to the adversarial bearing-domain classifiers (Ganin et al., 2016).

The experimental results are consistent with the intended direction of the adversarial objective. Compared with No-GRL, the GRL model achieved higher Balanced Accuracy and Macro-F1 and corrected one additional bearing without introducing a corresponding regression. However, the paired analysis showed that the improvement was limited in terms of the number of changed bearing-level decisions.

The results should therefore not be interpreted as evidence that GRL completely removes bearing-specific information. Instead, they motivate a more direct representation-level diagnostic, described in Section 4.7, to determine whether physical-bearing identity remains linearly recoverable from the learned representation.

4.7. Post-Hoc Bearing-Identity Representation Audit

To directly examine the representation-level effect of GRL, a post-hoc Bearing-ID audit was performed using the learned shared embeddings. For every LOBO fold, the held-out bearing was excluded from the Bearing-ID probe, leaving 31 development bearings as the identification classes. For each development bearing, 60 recordings were used to train the probe and 20 recordings were used for probe testing. The same recording partition was used for the No-GRL and GRL representations.

Sequence-level embeddings were aggregated to the recording level using the median, and a linear one-versus-one SVM was trained to predict the physical-bearing identity. The resulting mean performance across the 32 LOBO folds is shown in Table 3.

Representation	Accuracy (%)	Balanced Accuracy (%)	Macro-F1 (%)
No-GRL	99.22	99.22	99.22
GRL	86.27	86.27	85.96
Difference (GRL − No-GRL)	−12.95	−12.95	−13.26

Table 3. Post-hoc physical-bearing identity recoverability of No-GRL and GRL representations.

The GRL representation therefore exhibited substantially lower linear recoverability of physical-bearing identity than the No-GRL representation. GRL produced lower Bearing-ID performance in all 32 LOBO folds. Because every probe-test bearing contributed the same number of recordings, Accuracy and Balanced Accuracy were numerically identical in this audit.

The fold-wise Bearing-ID results and their aggregate differences between the two representations are visualized in Fig. 5. The figure complements Table 3 by showing both the overall metric comparison and the consistency of the reduction in linear Bearing-ID recoverability across the 32 LOBO folds.

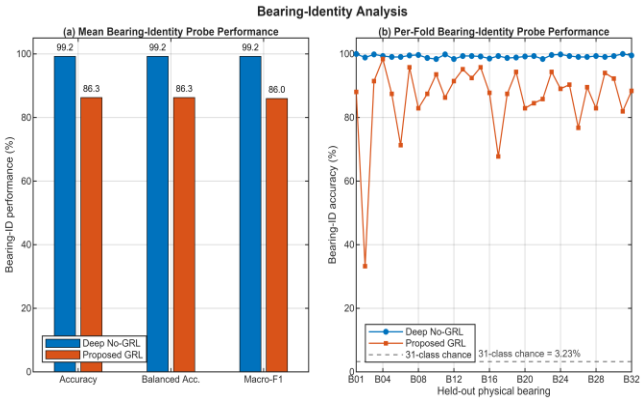

Figure 5. Bearing-Identity Analysis of the learned representations.

The figure compares the post-hoc linear recoverability of physical-bearing identity from the No-GRL and GRL representations across the 32 LOBO folds. Panel (a) summarizes Accuracy, Balanced Accuracy, and Macro-F1, while panel (b) shows the fold-wise Bearing-ID performance together with the 31-class chance level. The GRL representation exhibits substantially lower linear Bearing-ID recoverability than No-GRL across all folds, while remaining above the nominal chance level.

The GRL Bearing-ID accuracy of 86.27% remains far above the nominal uniform 31-class chance level of approximately 3.23%. Therefore, the audit does not support the claim that bearing identity was eliminated from the learned representation. Instead, it provides direct evidence that the linear recoverability of physical-bearing identity was substantially reduced by the GRL training procedure while considerable bearing-specific information remained.

This finding complements the fault-classification results. The GRL model simultaneously achieved higher bearing-level Balanced Accuracy and Macro-F1 than No-GRL and showed substantially lower linear Bearing-ID recoverability. The two observations are consistent with the intended direction of adversarial representation learning, although they do not establish that the reduction in Bearing-ID recoverability is the sole cause of the observed classification improvement.

The audit should also be interpreted within its methodological scope. Because the probe is linear, the results specifically measure the linear recoverability of physical-bearing identity from the learned representation. They do not establish that nonlinear classifiers would be unable to recover additional bearing-specific information.

4.8. Limitations

Several limitations should be considered when interpreting these findings.

First, the Combined Fault category contains only three physical bearings. Consequently, estimates of class-specific generalization for this category have substantially greater uncertainty than those for classes represented by more physical bearings.

Second, checkpoint selection was based on the minimum mean training fault-classification loss over the fixed 30-epoch training schedule. This procedure is leakage-free because held-out test-bearing performance was not used for checkpoint selection, but it does not directly optimize the checkpoint for unseen-bearing generalization.

Third, the GRL coefficient was fixed at a maximum value of 0.5 with a five-epoch warm-up. No systematic sensitivity analysis over multiple adversarial strengths was performed.

Fourth, the Bearing-ID audit used a linear SVM probe. It therefore measures linear recoverability of bearing identity and cannot rule out the presence of bearing-specific information that is recoverable by more expressive nonlinear probes.

Fifth, the experiments were conducted on the Paderborn University Bearing Dataset. Although the strict physical-bearing LOBO protocol provides a substantially more demanding test of generalization than random or recording-wise partitioning, validation on additional datasets and experimental platforms would be required to establish broader external generality.

Finally, six bearings remained incorrectly classified by both the No-GRL and GRL models. These persistent errors indicate that adversarial representation learning alone does not resolve all sources of physical-bearing variability.

5. Conclusions

This study investigated fault classification under a strict unseen-physical-bearing generalization setting using the Paderborn University Bearing Dataset. The physical bearing was treated as the fundamental unit of generalization, and a 32-fold leave-one-bearing-out protocol was used so that the target bearing was completely excluded from model development in each fold. Within this framework, a controlled comparison was conducted between a classical machine-learning baseline, a deep model without adversarial learning (No-GRL), and a class-conditional gradient-reversal learning (GRL) model.

The controlled comparison showed that the classical model achieved 81.25% Accuracy, 74.81% Balanced Accuracy, and 77.82% Macro-F1. The No-GRL deep model achieved 78.12%, 74.81%, and 77.75%, respectively, whereas the GRL model achieved 81.25% Accuracy, 76.89% Balanced Accuracy, and 79.85% Macro-F1. Relative to No-GRL, the GRL model therefore produced improvements of 3.13 percentage points in Accuracy, 2.08 percentage points in Balanced Accuracy, and 2.10 percentage points in Macro-F1. The GRL model also matched the classical baseline in Accuracy while obtaining higher Balanced Accuracy and Macro-F1.

However, the bearing-level paired analysis provides an important qualification to these aggregate results. The No-GRL and GRL models produced the same decision for 31 of the 32 held-out bearings: 25 bearings were correctly classified by both models and 6 were incorrectly classified by both. Only B08 changed from an incorrect prediction under No-GRL to a correct prediction under GRL, and no bearing changed from correct to incorrect. The exact two-sided McNemar test yielded $p=1.000$. Thus, the observed improvement should be interpreted as a modest empirical improvement under the present protocol rather than as statistically significant evidence of a general GRL advantage across physical bearings.

The representation-level analysis provided complementary evidence regarding the effect of adversarial learning. In the post-hoc Bearing-ID audit, a linear SVM was used to recover physical-bearing identity from the learned representations while excluding the held-out bearing from the probe. Across the 32 LOBO folds, the No-GRL representation achieved a mean Bearing-ID Accuracy of 99.22%, whereas the GRL representation achieved 86.27%, corresponding to a reduction of 12.95 percentage points. GRL produced lower Bearing-ID recoverability than No-GRL in all 32 folds. These results indicate that the GRL representation contained substantially less linearly recoverable physical-bearing identity than the No-GRL representation.

At the same time, the Bearing-ID Accuracy of 86.27% remained far above the nominal 31-class chance level of approximately 3.23%. Therefore, the results do not support the claim that GRL completely removed bearing-specific information. Instead, they show that GRL substantially reduced the linear recoverability of physical-bearing identity while considerable bearing-specific information remained in the learned representation.

The combined experimental evidence therefore suggests that class-conditional GRL can be useful for unseen-physical-bearing fault diagnosis, but its effect should be characterized cautiously. The main finding is not that GRL universally improves diagnosis or completely removes bearing identity.

Rather, under the present strict LOBO protocol, GRL simultaneously produced higher aggregate class-balanced fault metrics and substantially lower linear physical-bearing recoverability in the learned representation. The classification improvement itself was modest and was concentrated in a single changed bearing.

The findings should be interpreted in light of the single-dataset evaluation, the small number of Combined Fault bearings, training-loss-based checkpoint selection, the fixed adversarial strength, and the use of a linear identity probe. These limitations motivate the future-work directions outlined below.

Despite these limitations, the study provides a controlled evaluation framework for examining adversarial representation learning under a particularly demanding generalization scenario. The combination of strict physical-bearing LOBO evaluation, a matched No-GRL control, class-conditional adversarial heads, bearing-level paired analysis, and a post-hoc Bearing-ID representation audit provides complementary evidence that cannot be obtained from aggregate classification accuracy alone.

5.1. Future Work

Several directions follow naturally from the present findings.

First, the proposed evaluation should be extended to additional bearing datasets and, where possible, independent experimental platforms. This would determine whether the observed relationship between adversarial learning, fault generalization, and bearing-identity separability is specific to the Paderborn dataset or persists across different sensors, machines, operating conditions, and bearing populations.

Second, future studies should systematically investigate the adversarial-strength parameter rather than relying on the fixed maximum value used in the present study. A controlled sensitivity analysis over different GRL coefficients and scheduling strategies could clarify the trade-off between preserving fault-discriminative information and reducing physical-bearing-specific information.

Third, future work could investigate validation strategies for selecting model checkpoints using development bearings only. Such strategies could help determine whether selecting models according to bearing-level validation performance improves the stability of generalization to unseen physical bearings, while keeping the held-out test bearing completely separate from model development.

Fourth, the present Bearing-ID audit could be extended beyond the linear SVM probe. Nonlinear probes, information-theoretic measures, or other representation-analysis techniques could determine whether bearing-specific information remains recoverable in nonlinear form after GRL.

Fifth, the Combined Fault category deserves dedicated investigation because it is represented by only three physical bearings in the current dataset. Additional physical examples of combined faults, if available in future datasets, would allow more reliable assessment of this practically important but statistically limited class.

Sixth, future research could investigate whether the persistent difficult bearings identified in the present study—B02, B05, B15, B18, B21, and B25—share measurable characteristics that distinguish them from more successfully generalized bearings. Such an analysis could help identify the types of physical variability that remain difficult even after adversarial representation learning, without assuming a causal explanation before it is experimentally established.

Finally, future work could investigate more structured approaches to combining physics-informed information with adversarial representation learning. The present results show that the feature space contains strong fault-related as well as bearing-specific information. Future models could therefore explore mechanisms that explicitly preserve physically meaningful fault information while reducing nuisance variability associated with the physical bearing, operating condition, or measurement environment.

References

Chen, L., Li, Q., Shen, C., Zhu, J., Wang, D., & Xia, M. (2022). Adversarial domain-invariant generalization: A generic domain-regressive framework for bearing fault diagnosis under unseen conditions. *IEEE Transactions on Industrial Informatics, 18*(3), 1790–1800. https://doi.org/10.1109/TII.2021.3078712

Cheng, L., Kong, X., Zhang, Y., Zhu, Y., Qi, H., & Zhang, J. (2024). A novel causal feature learning-based domain generalization framework for bearing fault diagnosis with a mixture of data from multiple working conditions and machines. *Advanced Engineering Informatics, 62*, Article 102622. https://doi.org/10.1016/j.aei.2024.102622

Ding, Y., Jia, M., Cao, Y., Ding, P., Zhao, X., & Lee, C.-G. (2023). Domain generalization via adversarial out-domain augmentation for remaining useful life prediction of bearings under unseen conditions. *Knowledge-Based Systems, 261*, Article

110199. https://doi.org/10.1016/j.knosys.2022.110199

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. S. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research, 17*(59), 1-35.

Guo, Z., Xu, L., Zheng, Y., Xie, J., & Wang, T. (2025). Bearing fault diagnostic framework under unknown working conditions based on condition-guided diffusion model. *Measurement, 242*, Article 115951. https://doi.org/10.1016/j.measurement.2024.115951

Hoang, D.-T., & Kang, H.-J. (2019). A survey on deep learning based bearing fault diagnosis. *Neurocomputing, 335*, 327–335. https://doi.org/10.1016/j.neucom.2018.06.078

Kim, I., Kim, S. W., Kim, J., Huh, H., Jeong, I., Choi, T., Kim, J., & Lee, S. (2024). Single domain generalizable and physically interpretable bearing fault diagnosis for unseen working conditions. *Expert Systems with Applications, 241*, Article 122455. https://doi.org/10.1016/j.eswa.2023.122455

Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing, 138*, Article 106587. https://doi.org/10.1016/j.ymssp.2019.106587

Lessmeier, C., Kimotho, J. K., Zimmer, D., & Sextro, W. (2016). Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. PHM Society European Conference, 3(1). https://doi.org/10.36001/phme.2016.v3i1.157

Li, Q., Chen, L., Kong, L., Wang, D., Xia, M., & Shen, C. (2023). Cross-domain augmentation diagnosis: An adversarial domain-augmented generalization method for fault diagnosis under unseen working conditions. *Reliability Engineering & System Safety, 234*, Article 109171. https://doi.org/10.1016/j.ress.2023.109171

Muandet, K., Balduzzi, D., & Schölkopf, B. (2013). Domain generalization via invariant feature representation. In Proceedings of the 30th International Conference on Machine Learning (Vol. 28, No. 1, pp. 10–18). PMLR

Pang, B., Liu, Q., Xu, Z., Sun, Z., Hao, Z., & Song, Z. (2024). Fault vibration model driven fault-aware domain generalization framework for bearing fault diagnosis. *Advanced Engineering Informatics, 62*, Article 102620. https://doi.org/10.1016/j.aei.2024.102620

Shi, Y., Deng, A., Deng, M., Xu, M., Liu, Y., Ding, X., & Bian, W. (2023). Domain augmentation generalization network for real-time fault diagnosis under unseen working conditions. *Reliability Engineering & System Safety, 235*, Article 109188. https://doi.org/10.1016/j.ress.2023.109188

Vieira, J. P., Bauler, V. A., Rosa, R. K., & Silva, D. (2026). Towards a more realistic evaluation of machine learning models for bearing fault diagnosis. *Mechanical Systems and Signal Processing, 258*, Article 114640. https://doi.org/10.1016/j.ymssp.2026.114640

Zhang, Q., Zhao, Z., Zhang, X., Liu, Y., Sun, C., Li, M., Wang, S., & Chen, X. (2021). Conditional adversarial domain generalization with a single discriminator for bearing fault diagnosis. **IEEE Transactions on Instrumentation and Measurement, 70**, Article 3514515. https://doi.org/10.1109/TIM.2021.3071350

Zhao, Z., Li, T., Wu, J., Sun, C., Wang, S., Yan, R., & Chen, X. (2020). Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. *ISA Transactions, 107*, 224–255. https://doi.org/10.1016/j.isatra.2020.08.010

Zhu, Y., Zi, Y., Li, J., & Xu, J. (2024). PhysiCausalNet: A causal- and physics-driven domain generalization network for cross-machine fault diagnosis of unseen domain. *IEEE Transactions on Industrial Informatics, 20*(6), 8488–8498. https://doi.org/10.1109/TII.2024.3369240