Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 8, No. 1, 2026

Noise in Rolling Bearing Fault Diagnosis: From Characterization to Robust Solutions

Jiajun Liu [a], Baishun Su [b, *]

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000, China

[a] 212409010018@home.hpu.edu.cn, [b] subaishun@hpu.edu.cn

Abstract: Rolling bearing fault diagnosis is essential for the reliability and safety of rotating machinery, yet its effectiveness in industrial practice is fundamentally challenged by noise. This narrative review argues that noise in bearing fault diagnosis should not be treated as a single additive disturbance, but rather as a collection of distinct phenomena — environmental and background signal noise, sensor and acquisition noise, operating-condition-induced domain shift, label noise, and compound noise scenarios — each affecting different stages of the diagnostic pipeline and demanding differentiated strategies. This review develops a noise taxonomy for rolling bearing fault diagnosis and systematically examines how each noise type degrades signal processing, feature extraction, and diagnostic decision-making. It surveys denoising strategies, noise-robust feature engineering approaches, and noise-aware diagnostic models through this noise-type lens, identifying which strategies are appropriate for which noise conditions. A standardized evaluation and benchmarking protocol is proposed to address the inconsistency that currently prevents meaningful comparison of noise-robustness claims across studies. The central synthetic contribution is a noise–method matching matrix that maps each noise category to appropriate combinations of denoising, feature, model, and evaluation strategies, providing a structured framework for method selection and future research design. The review concludes that robust bearing fault diagnosis cannot be achieved through a single universal model. The more realistic path forward requires noise-type-aware strategies, standardized multi-noise evaluation, and methods co-designed for industrial deployment constraints.

Keywords: Rolling bearing fault diagnosis, noise robustness, noise taxonomy, denoising, domain shift, label noise, benchmarking.

1. Introduction

1.1. Background: Bearing Fault Diagnosis in Rotating Machinery

Rolling element bearings are among the most failure-prone components in rotating machinery, with bearing-related faults accounting for a substantial fraction of unplanned downtime across industries including power generation, manufacturing, transportation, and resource extraction. The economic and safety consequences of bearing failures have driven decades of research into vibration-based condition monitoring and fault diagnosis.

The physical principle underlying vibration-based bearing diagnosis is well established. Localized defects on bearing raceways or rolling elements generate characteristic impulsive signatures each time a rolling element passes over the defect, repeating at defect-specific frequencies — the ball pass frequency outer race (BPFO), ball pass frequency inner race (BPFI), ball spin frequency (BSF), and their harmonics and modulation sidebands — determined by bearing geometry and rotational speed. The evolution of bearing diagnosis over the past two decades has tracked the broader machine learning trajectory: from signal processing combined with shallow classifiers such as support vector machines (SVMs), through deep convolutional neural networks (CNNs) and autoencoders, to transfer learning, domain adaptation (DA), and self-supervised approaches. Comprehensive reviews by Lei et al., Zhang et al., and Zhao et al. have documented this progression, establishing the foundations upon which the current review builds [1–3].

1.2. Noise as a Central Challenge in Bearing Fault Diagnosis

In industrial environments, the assumption that vibration signals are dominated by bearing dynamics rarely holds. The recorded vibration signal is a superposition of bearing-generated vibration and interference from multiple noise sources. Noise enters the diagnostic pipeline at multiple levels through fundamentally different mechanisms. Environmental noise — mechanical vibration from adjacent machinery, electromagnetic interference (EMI), structural resonances, and fluid turbulence — superimposes as additive corruption that can be broadband, colored, impulsive, or non-stationary. Sensor noise — thermal noise in transducer electronics, quantization error from analog-to-digital conversion (ADC), and drift from sensor aging — limits acquisition fidelity. Operational variability — changes in rotational speed, applied load, lubrication condition, and ambient temperature — shifts the distribution of vibration features. Label noise — errors in health-state annotations from human mistakes, severity boundary ambiguity, or misalignment between vibration segments and maintenance records — corrupts the supervisory signal itself.

These noise types are different in kind and affect different pipeline stages. Environmental and sensor noise corrupt the signal waveform; operational variability shifts the feature distribution; label noise contaminates the training objective. A strategy effective against one may be irrelevant to another: signal denoising cannot correct label errors, and domain adaptation cannot recover information lost to sensor quantization. In real deployments, multiple noise types coexist — a wind turbine bearing experiences environmental noise from gear meshing and blade passage, sensor noise from the data acquisition chain, operational variability from fluctuating wind speeds, and potential label noise from maintenance records — creating compound scenarios demanding integrated strategies.

Despite the centrality of noise to diagnostic reliability, the literature has treated it with uniformity that belies its diversity. The standard protocol — adding additive white Gaussian noise (AWGN) at one or two signal-to-noise ratio (SNR) levels — addresses one specific noise type under controlled conditions bearing little resemblance to industrial reality. Pancaldi et al. showed that diagnostic performance depends materially on the adopted noise model, including non-Gaussian and impulsive disturbances [4], yet few studies evaluate systematically across colored, impulsive, or non-stationary noise. Domain shift and label noise are often treated as separate topics rather than integral dimensions of the broader robustness problem. Compound-noise evaluation remains limited.

1.3. Limitations of Existing Studies and Reviews

The dominant organizing principle — classification by model architecture or methodological paradigm — facilitates comparison within method categories but can obscure the deployment-relevant question: given a specific disturbance or distribution shift, which strategy is appropriate? Bao et al. surveyed deep-learning-enabled bearing diagnosis in complex environments [6], while Chen et al. reviewed noise-robust rotating-machinery diagnosis across mechanism-based, data-driven, and hybrid approaches [7]. The present review instead adopts an explicitly noise-type-centric organization. Cross-cutting limitations include inconsistent evaluation protocols across studies, heavy dependence on laboratory datasets and synthetic AWGN, and weak reporting of deployment considerations such as inference latency, model size, data efficiency, and interpretability. Hendriks et al. further showed that commonly used CWRU domain-shift splits may place data from the same physical bearings in both training and testing sets, weakening the intended generalization test [5]. High classification accuracy and noise robustness should therefore be evaluated as distinct properties.

1.4. Scope and Contributions of This Review

This narrative review differs from existing surveys in its organizing principle: rather than classifying methods by architecture, it classifies noise by type and traces each type through the full diagnostic pipeline. Five principal contributions are made:

First, a noise taxonomy distinguishing environmental noise, sensor noise, operational variability (domain shift), label noise, and their compound effects. Second, a systematic analysis of how each noise type degrades signal processing and diagnostic decision-making. Third, a survey of denoising, feature engineering, and model-level strategies through the noise taxonomy lens. Fourth, a standardized evaluation and benchmarking protocol addressing the inconsistency preventing meaningful cross-study comparison. Fifth, a noise–method matching matrix integrating taxonomy, strategies, and evaluation into a unified decision framework.

The review focuses on vibration-based diagnosis of rolling element bearings, drawing primarily on 2016–2025 literature with emphasis on developments since 2022. It is a narrative rather than systematic review, emphasizing critical synthesis over exhaustive coverage.

1.5. Organization of the Review

Section 2 establishes the noise taxonomy serving as the organizing framework. Section 3 analyzes how noise propagates through conventional signal processing. Section 4 surveys signal-level denoising strategies. Section 5 examines noise-robust feature engineering. Section 6 discusses noise-aware diagnostic models. Section 7 addresses the evaluation crisis and proposes a standardized benchmarking protocol. Section 8 synthesizes preceding analyses, presenting the noise–method matching matrix. Section 9 concludes with principal findings and five future research directions.

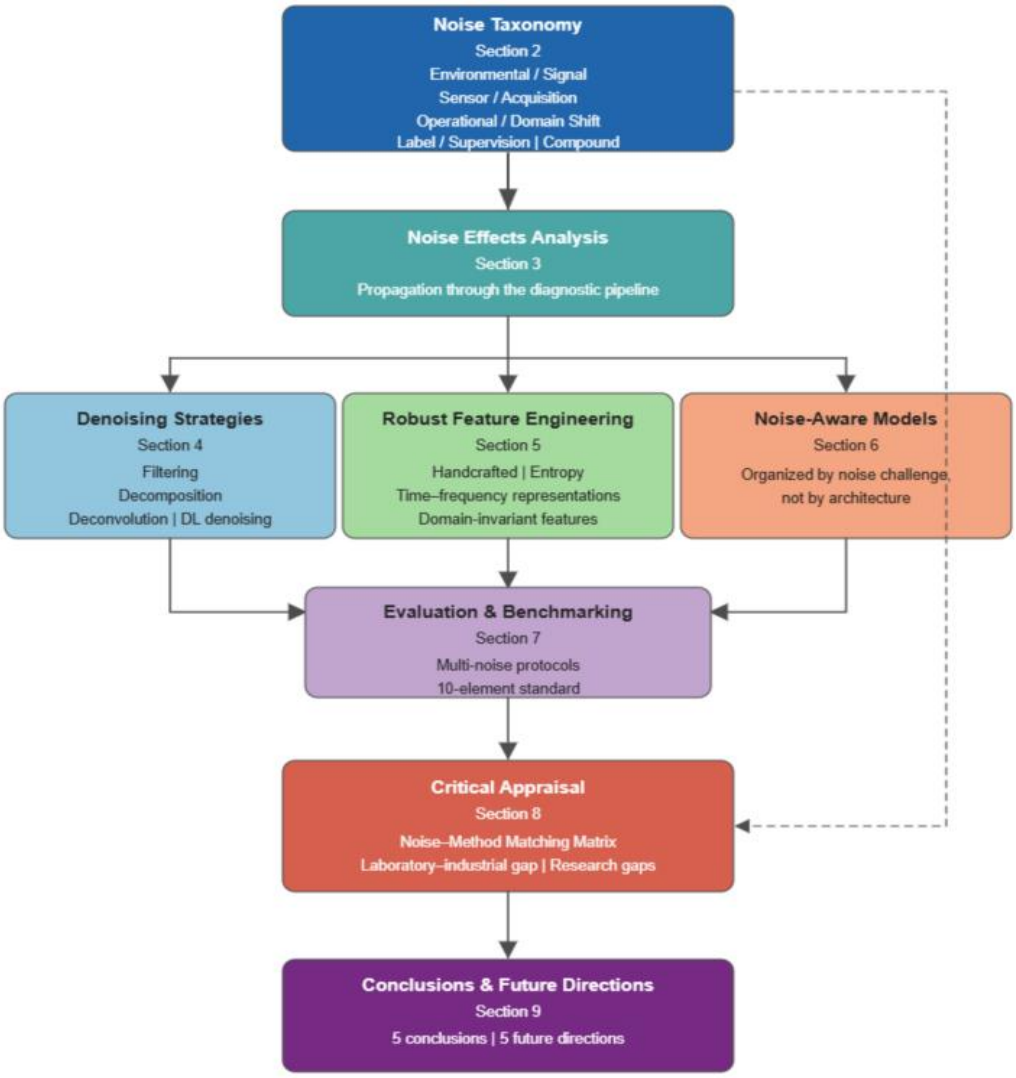

Figure 1. Overall framework of this narrative review

2. Noise Taxonomy in Rolling Bearing Fault Diagnosis

2.1. Why a Noise Taxonomy Matters

This review adopts a deliberately broad definition of "noise": any non-target disturbance or supervision contamination that reduces diagnostic reliability. The four categories are positioned at distinct pipeline levels: environmental noise (signal-level, external interference), sensor noise (acquisition-level), operational variability (distribution-level nuisance variability or domain shift), and label noise (supervision-level contaminating ground truth). Operational variability and label noise are not strictly physical signal noise but diagnostic confounding factors included under a broader noise-robust diagnosis framework, maintaining a unified vocabulary while avoiding the conceptual ambiguity of treating all perturbations as equivalent.

Treating all noise as AWGN with a single SNR obscures critical differences in how various noise types degrade performance. Pancaldi et al. compared multiple noise models and showed that the measured performance of bearing-fault algorithms is sensitive to the assumed disturbance model [4]. Recognizing these distinctions is the first step toward targeted mitigation strategies rather than generic denoising as a one-size-fits-all remedy.

2.2. Environmental Noise

Environmental noise originates from sources external to the bearing but internal to the industrial environment — adjacent machinery, electromagnetic fields, structural resonances, and fluid turbulence. Broadband mechanical interference is often modeled as additive Gaussian noise; impulsive sources produce non-Gaussian distributions with heavier tails. Hebda-Sobkowicz et al. compared informative-frequency-band selection methods for bearing diagnosis in the presence of non-Gaussian noise and showed that method effectiveness depends on the interference characteristics [8]. Real industrial noise is often colored, with concentrated energy in bands overlapping bearing characteristic frequencies, and non-stationary due to production schedules and load changes, violating stationary assumptions underlying many signal processing techniques.

2.3. Sensor Noise

Sensor noise is intrinsic to the measurement chain: thermal (Johnson-Nyquist) noise, shot noise, quantization noise from ADC, and transmission noise. The aggregate sensor noise is modeled as additive noise with a spectral density floor expressed in $\mu g/\sqrt{Hz}$. The effective SNR depends on both the sensor's intrinsic noise and the vibration signal amplitude; in small bearings at moderate speeds, fault-induced amplitudes can approach the sensor noise floor. Quantization noise from

16-bit or 24-bit ADCs imposes a fundamental amplitude resolution limit — sub-quantization-level variations in early degradation become invisible to any downstream algorithm, particularly relevant for low-cost edge devices. Sensor aging and calibration drift introduce systematic biases distinct from zero-mean random noise, potentially triggering false alarms or missed detections.

2.4. Operational Variability Noise

Operational variability refers to distribution shift between training and deployment conditions. Unlike additive signal corruption, it manifests as a change in the data-generating distribution: $p(x)$ changes between training and deployment while $p(y|x)$ may remain stable. Speed variations cause characteristic defect frequencies to shift along the frequency axis; load variations alter impulse response amplitudes and damping characteristics through nonlinear, interdependent transformations. A fault mechanism-based model addressing non-stationary conditions leverages bearing kinematics to predict fault signature transformations across operating conditions without requiring target condition samples [9]. Severity depends on context: negligible in constant-speed drives, critical in wind turbine gearboxes where rotational speed varies continuously.

2.5. Label Noise

Label noise refers to errors in the supervisory mapping, characterized by a noise transition matrix T where $T_{ij}=P(\tilde{y}=j|y=i)$. In bearing diagnostics, label noise is often asymmetric — incipient faults are more likely mislabeled as healthy than as severe faults, and adjacent severity levels are easily confused. In general intelligent fault diagnosis, Ma et al. proposed a noisy universal domain-adaptation framework with dual-stage selection of clean and reusable source samples [10]; direct validation for bearing-specific asymmetric annotation noise remains limited. The practical significance is amplified by deep learning adoption: deep networks can memorize noisy labels rather than learning generalizable patterns, a risk particularly pronounced with small training datasets typical of industrial fault diagnosis.

2.6. Cross-Coupling and Compound Noise Effects

The four noise categories do not operate in isolation. Key coupling patterns include: environmental noise compounding with sensor noise when external interference masks the sensor noise floor; variable operating conditions combined with low SNR pushing fault signatures into interference-dominated bands; operational variability interacting with label errors when run-to-failure data collected under changing conditions is annotated without accounting for those changes; and sensor drift accumulating alongside degradation, blurring the boundary between artifacts and genuine fault progression. In any real deployment, multiple noise types coexist. These interactions can be multiplicative: operational variability amplifies environmental noise effects; sensor noise negligible at high amplitudes becomes dominant under low-speed, light-load conditions. The current literature predominantly addresses noise types in isolation — denoising studies neglect speed variation, domain adaptation studies assume clean labels — representing a significant methodological gap motivating the noise–method matching framework of Section 8.2.

2.7. Summary: Noise Taxonomy Overview

Table 1. Noise taxonomy for rolling bearing fault diagnosis: sources, affected diagnostic stages, mathematical representations, primary consequences, and corresponding strategy categories.

Noise Category	Layer	Typical Source	Mathematical Model	Primary Effect	Common Simulation	Realism Limitation
Environmental	Signal	Adjacent machinery, EMI, structural resonance, fluid turbulence	Additive: Gaussian (broadband) or non-Gaussian (impulsive/ colored)	Masks fault characteristic frequencies; reduces effective SNR; introduces spurious spectral peaks	AWGN at fixed SNR levels	Ignores temporal non-stationarity, colored spectra, and impulsive components of real industrial noise
Sensor	Acquisition	Thermal noise, quantization, ADC resolution, transmission loss, mounting resonance	Additive: noise floor spectral density; quantization error (uniform)	Limits resolvable fault amplitude; introduces spurious resonances from mounting	White Gaussian noise added to clean signal	Neglects quantization effects, sensor aging drift, and frequency-dependent noise characteristics
Operational Variability	Distribution	Speed/load variation, lubrication change, temperature drift	Covariate shift: $p_{train}(x) \neq p_{deploy}(x)$; nonlinear frequency scaling	Shifts characteristic frequencies; alters feature distributions; causes class overlap	Fixed-speed training, single-condition testing	Underestimates severity of continuous speed/load variation in real applications
Label Noise	Supervision	Human annotation error, severity boundary ambiguity, run-to-failure misalignment	Noise transition matrix: $P(\tilde{y}=j|y=i)$; symmetric or asymmetric	Degrades training signal quality; induces model overfitting to erroneous labels	Clean labeled datasets (assumed)	Overlooks the prevalence of ambiguous severity boundaries and annotation errors in industrial data

Note: Table entries are descriptive and define the problem space. Mitigation strategies are addressed in Sections 4–8.

3. Noise Effects on Signal Processing in Bearing Fault Diagnosis

3.1. Propagation of Noise Through the Diagnostic Pipeline

A bearing fault diagnosis system operates sequentially: signal acquisition, preprocessing, feature extraction, and decision-making. Noise introduced at any stage propagates forward, often with amplification. Environmental and sensor noise corrupt the raw waveform; when this corrupted signal passes through preprocessing, noise is transformed along with the signal — narrowband interference becomes a dominant spectral peak after FFT; a transient sensor glitch masquerades as an impulsive fault signature after envelope detection. Operational variability shifts the statistical relationship between signal and health state, moving feature vectors outside the learned manifold. Label noise propagates in reverse: erroneous labels become training targets, and the learned misassociation contaminates every deployment prediction. These propagation pathways reveal why addressing noise at a single stage is insufficient — a perfect denoising algorithm cannot correct label errors, and a flawless domain adaptation method cannot recover information lost to sensor quantization.

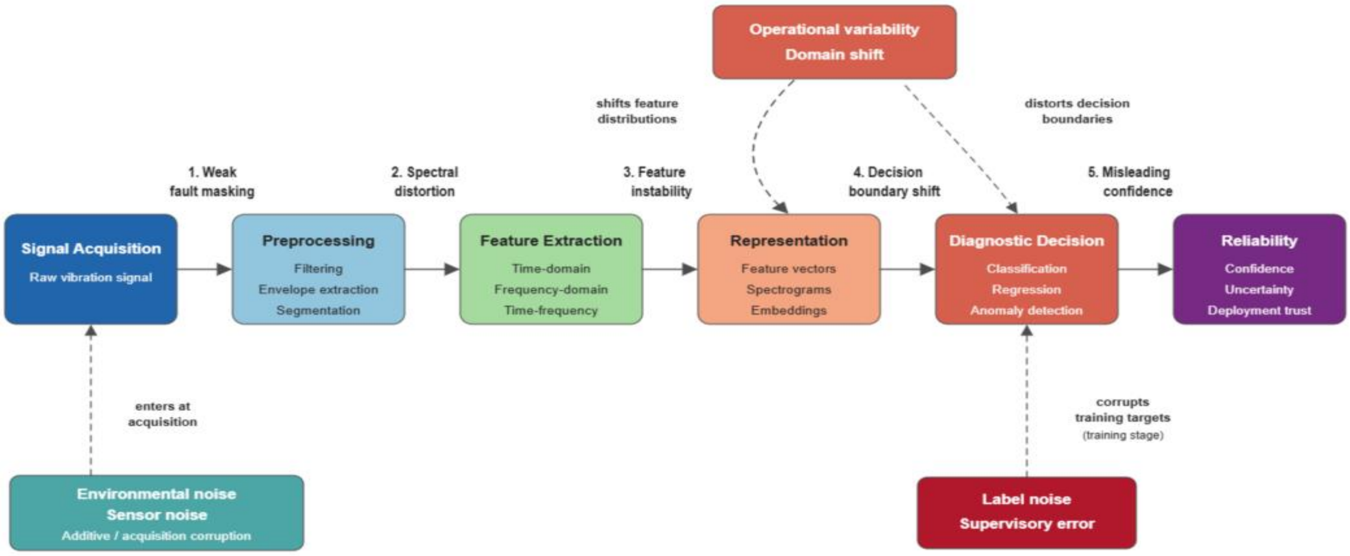

Figure 2. Propagation chain of noise effects through the diagnostic pipeline

3.2. Effects on Time-Domain Indicators

Time-domain statistical indicators — RMS, peak value, kurtosis, crest factor, and impulse factor — form the simplest and most widely used feature set. RMS is relatively robust to moderate zero-mean Gaussian noise but vulnerable to impulsive transients that inflate energy measures. Kurtosis, celebrated for sensitivity to fault-induced impulsiveness, produces false positives under impulsive environmental noise and delayed detection under heavy-tailed sensor noise. Crest factor and impulse factor amplify individual extreme values, making them hypersensitive to noise spikes. In practice, these indicators exhibit high variance even under nominally constant conditions, exacerbated under variable speed or load.

3.3. Effects on Frequency- and Envelope-Domain Analysis

Broadband environmental noise elevates the spectral noise floor, submerging characteristic defect-frequency peaks. In a gear-fault study, Schmidt et al. formulated generalized envelope-spectrum-based signal-to-noise objectives for filter optimization under time-varying speed [11]. This work supports defining an objective in the diagnostic representation itself, but its transfer to rolling-bearing diagnosis requires separate validation. Narrowband interference within the selected resonance band can create peaks that resemble fault harmonics in machines with dense, overlapping spectra. Impulsive noise presents a distinctive challenge because an envelope detector cannot inherently distinguish fault-induced from interference-induced impulses. For weak bearing faults, Liu et al. proposed resonance demodulation based on the amplitude Z-score of the envelope spectrum [12]. Cepstrum analysis has a related vulnerability because the logarithmic operation can propagate background fluctuations into the cepstral domain.

3.4. Effects on Time-Frequency Representations

The STFT faces the uncertainty-principle trade-off: short windows provide good time resolution but unreliable spectral estimation under low SNR; long windows average noise but smear transient fault signatures. Wavelet decomposition suffers from dyadic frequency partitioning misalignments with bearing characteristic frequencies. EMD-based methods experience mode mixing under noise; EEMD mitigates this through noise-assisted decomposition but introduces residue noise. VMD offers improved separation through variational optimization but requires parameter prespecification that can be unreliable under strong noise. A hybrid AGWO-PSO strategy has been used to optimize VMD-related parameters in a strong-noise bearing-diagnosis pipeline [15]. Deep deconvolution can integrate filter learning with classifier-driven optimization [14], but the reported gains remain tied to the evaluated datasets and noise settings.

3.5. Effects of Operational Variability

Speed variation shifts characteristic frequencies proportionally to shaft speed; fixed-frequency peak detection misses off-nominal signatures. Spectral smearing occurs when speed varies within an FFT frame, broadening peaks and reducing effective SNR. Load variation alters contact angle and impulse response amplitude — heavier loads produce stronger fault impacts, lighter loads produce weaker ones masked by noise. The combined effect is multiplicative, and the resulting feature distribution shift means feature vectors from different conditions are not directly comparable without domain adaptation.

3.6. From Signal Distortion to Diagnostic Uncertainty

The cumulative propagation produces three diagnostic uncertainty types: detection uncertainty (inability to confirm fault presence), identification uncertainty (ambiguous spectral patterns from selectively masked harmonics), and severity estimation uncertainty (unreliable amplitude-based quantification). Compound noise interactions amplify all three, motivating the multi-layered strategies in Sections 4–8.

4. Denoising Strategies for Noisy Bearing Fault Signals

4.1. Role of Denoising in Noise-Robust Diagnosis

Denoising operates between signal acquisition and feature extraction, suppressing non-informative signal components so that fault-related features become more identifiable. Denoising can address signal-level and acquisition-level noise — environmental interference and sensor noise — where corruption is additive and signal structure remains recoverable. It cannot address distribution-level noise from operational variability (which alters signal structure) or supervision-level noise from label errors (which are in the mapping, not the signal). This scope boundary must be kept in view throughout.

4.2. Conventional Filtering and Signal Enhancement

Band-pass filtering around bearing structural resonances is the standard preprocessing for envelope analysis; the critical challenge is passband selection, which depends on bearing geometry, mounting structure, sensor placement, and operating speed. Adaptive filtering (LMS, RLS) and Wiener filtering adjust coefficients in response to changing signal statistics but require spectral knowledge rarely available in field applications. The fundamental limitation is reliance on spectral separability — when fault signatures and noise occupy overlapping frequency bands, no linear filter can separate them.

4.3. Decomposition-Based Adaptive Denoising

Wavelet thresholding exploits the localization of fault impulses in time and frequency; the choice of mother wavelet, decomposition level, and thresholding rule all influence outcomes. Wavelets resembling bearing fault impulses (e.g., Morlet wavelet) provide compact representations, but this advantage diminishes under variable-speed conditions. EMD and ensemble variants perform data-driven separation into IMFs, but mode mixing and IMF selection criterion degradation under noise compromise this separation. VMD offers improved separation through variational optimization; a hybrid AGWO-PSO framework addresses parameter selection under strong noise [15]. The common thread is dependence on decomposition model validity under specific noise conditions.

4.4. Deconvolution and Sparse Enhancement

Bearing fault diagnosis has the distinctive advantage that fault impulses exhibit known statistical structure — periodicity, impulsiveness, and cyclostationarity. MED and MCKD exploit this by optimizing filters for maximum impulsiveness or correlated kurtosis. Maximum L-kurtosis deconvolution was proposed to improve robustness when classical kurtosis objectives are distorted by large interference impulses [13]. Deep network-based maximum correlated kurtosis deconvolution couples deconvolution-filter learning with backpropagation and reports robustness to random impulsive noise in the tested settings [14]; broader generalization still requires validation. Sparse representation and low-rank decomposition separate fault impulses (sparse) from harmonic interference (low-rank), but non-fault impulsive noise can violate the assumed decomposition.

4.5. Data-Driven and Deep-Learning-Based Denoising

Denoising autoencoders (DAEs) and convolutional variants (CDAEs) learn mappings from noisy to clean signals. Classifier-guided neural blind deconvolution provides one physics-informed denoising module for bearing diagnosis under noisy conditions [16]. Conditional diffusion has also been explored for unknown bearing-fault diagnosis through out-of-distribution error guidance [35], but that study does not constitute direct evidence for generic waveform denoising. The critical limitation is the training-data assumption: networks trained on synthetic AWGN may not transfer to colored, impulsive, or non-stationary industrial noise and may introduce diagnostically misleading artifacts. SNR improvement after denoising does not necessarily imply diagnostic improvement.

4.6. Method Selection under Different Noise Types

Environmental noise (broadband/colored) is the natural domain of frequency-selective methods; impulsive environmental noise demands deconvolution with safeguards against noise-impulse hijacking. Sensor noise is addressed by Wiener/adaptive filtering, but quantization noise resists linear treatment — calibration and sensor selection are often more effective. Operational variability exposes the fundamental limit that denoising cannot restore consistency when signal structure itself changes. Label noise is entirely unaddressable at the signal level. Compound noise requires coordinated denoising and domain adaptation, integrated through the matching framework of Section 8.2.

4.7. Limitations of Denoising-Centric Strategies

Key structural limitations include: aggressive denoising can suppress weak incipient fault features; SNR improvement does not equal diagnostic improvement; the gap between laboratory denoising experiments and industrial reality is substantial; denoising methods require parameters difficult to set without clean reference signals in unsupervised deployment; and signal denoising cannot address operational variability, label noise, or domain shift.

Table 2. Overview of denoising strategies for bearing fault signals: method categories, representative methods, suitable noise conditions, strengths, limitations, and connections to the noise taxonomy.

Method Category	Representative Methods	Suitable Noise Type	Strengths	Limitations	References
Conventional Filtering	Band-pass, Wiener, adaptive LMS/RLS	Environmental (known frequency band)	Interpretable, real-time, no training data	Requires band knowledge; fails with spectral overlap; non-stationary noise	—
Decomposition-Based	Wavelet thresholding, EMD/EEMD, VMD, SSD	Environmental (broadband, colored)	Data-driven modes; suitable for non-stationary signals	Mode mixing; parameter sensitivity; mode selection criteria degrade under noise	[20], [15]
Deconvolution & Sparse	MED, MCKD, MOMEDA, sparse coding, low-rank+sparse	Environmental (impulsive noise separation)	Exploits fault structure; physically interpretable	Sensitive to outlier noise impulses; periodicity assumption; parameter tuning	[13], [14]
Deep-Learning-Based	DAE, CDAE, GAN, diffusion, physics-informed NN	Environmental, Sensor (learned from data)	Learns complex noise distributions; no explicit noise model needed	Lab-to-field generalization gap; training data bias; SNR ≠ diagnostic gain	[16], [17], [18], [19]

Note: Operational variability and label noise require feature-level and model-level strategies (Sections 5–6). Abbreviations: LMS = Least Mean Squares; RLS = Recursive Least Squares; EMD = Empirical Mode Decomposition; VMD = Variational Mode Decomposition; SSD = Singular Spectrum Decomposition; MED = Minimum Entropy Deconvolution; MCKD = Maximum Correlated Kurtosis Deconvolution; DAE = Denoising Autoencoder; CDAE = Convolutional DAE; GAN = Generative Adversarial Network; NN = Neural Network.

5. Noise-Robust Feature Engineering for Bearing Fault Diagnosis

5.1. From Denoised Signals to Robust Fault Representations

Signal denoising improves waveform quality but does not guarantee stable, discriminative, and transferable features. Noise affects features through three mechanisms: inflated within-class variance, reduced between-class separation, and spurious correlations driven by noise structure rather than fault structure. Robust feature engineering pursues three complementary objectives — invariance to irrelevant variation, preservation of fault-discriminative structure, and stability across operating conditions — which are often in tension.

5.2. Handcrafted Features under Noisy Conditions

Time-domain statistical features exhibit a spectrum of noise sensitivity. RMS is relatively stable under moderate zero-mean Gaussian noise but its averaging property becomes a liability under impulsive noise. Kurtosis sensitivity to impulsiveness is valuable for incipient fault detection but produces false positives under impulsive environmental noise. Crest factor and impulse factor share this vulnerability. Frequency-domain features offer specificity by targeting known fault-related spectral locations but become liabilities under speed variation: a feature defined at a fixed frequency bin misses the fault signature entirely under off-nominal speed. Order-domain features address this by normalizing the frequency axis by instantaneous shaft speed but require accurate speed measurement.

Time-frequency features — wavelet packet node energies, spectrogram sub-band energies — capture joint temporal-spectral structure. The challenge is dimensionality: most elements are noise-dominated, making feature selection essential (Section 5.4). The overarching limitation is that handcrafted features are designed based on simplified physical models. When actual noise deviates from the assumed model, features optimized under the simplified model may perform worse than generic features.

5.3. Statistical, Entropy-Based, and Complexity Features

Entropy-based features — sample entropy, permutation entropy, fuzzy entropy, and multiscale entropy — quantify regularity and complexity in vibration signals. He et al. used refined time-shift multiscale fractional-order fuzzy dispersion entropy with hybrid kernel ridge regression for train-bearing diagnosis in a noisy environment [21]. This supports entropy-based feature extraction as one practical option, but it does not establish that entropy features are universally more robust than amplitude-based indicators. Complexity measures such as Lempel–Ziv complexity and the Higuchi fractal dimension capture complementary structural information. A consistent vulnerability is parameter sensitivity: entropy calculation requires choices such as embedding dimension, time delay, and scale range, whose useful values depend on sampling rate, bearing geometry, and disturbance characteristics.

5.4. Feature Selection and Dimensionality Reduction

Filter methods (mutual information, ReliefF, mRMR) rank features independently of the classifier but their ranking reflects the training data distribution and may not generalize. Wrapper and embedded methods integrate selection with classifier training at the risk of overfitting; stability selection provides some protection. PCA preserves variance that may be noise-driven rather than fault-driven; LDA is label-quality-sensitive. The critical underexplored issue is cross-condition stability — a feature selected under AWGN may be noise-sensitive under impulsive noise, yet cross-noise generalization is rarely tested.

5.5. Domain-Invariant and Noise-Invariant Feature Representations

Domain-invariant feature learning seeks representations whose distribution depends on health state more than on operating condition, using mechanisms such as distribution matching or adversarial training. Source-free domain adaptation removes the need to retain source data during target adaptation in cross-machine fault diagnosis [22]; causal factorization has been studied for cross-machine bearing domain generalization [23]. In rotating-machinery experiments, time-frequency-perception-guided multi-level contrastive learning was used to improve representation consistency across conditions [24]. Prior-knowledge-enhanced self-supervised learning has also been evaluated for machinery diagnosis with small labeled datasets [25]. The central tension is between invariance and discriminability: excessive invariance can suppress fault-relevant variation correlated with operating condition.

5.6. Limitations of Feature-Centric Robustness

Robust features may lack physical interpretability. Handcrafted features exhibit inconsistent stability across noise types, and systematic multi-noise-type evaluations are absent. Feature selection performed on a single dataset embeds dataset-specific biases. Robustness is a property of the feature-model combination, not of features alone.

6. Noise-Aware Diagnostic Models for Bearing Fault Diagnosis

6.1. From Robust Features to Noise-Aware Diagnostic Decisions

Noise manifests at the model level as decision boundary distortion and overfitting to noise-specific patterns. A model trained on clean data learns boundaries tight around the training distribution; under deployment noise, shifted distributions, or erroneous labels, these boundaries fail even when individual features are reasonably robust. The goal is stable diagnostic performance under low SNR, domain shift, label uncertainty, and compound noise — requiring robustness designed into architecture, training, or adaptation mechanisms.

6.2. CNN-Based and Representation-Learning Models

One-dimensional CNNs provide inherent noise robustness through hierarchical feature learning: low-level incoherent noise is progressively filtered out while coherent fault patterns are preserved across scales. Multi-scale and multi-domain architectures with attention mechanisms compensate for noise-induced degradation in any single domain [26]. Anti-noise CNN variants incorporate denoising as a learnable preprocessing stage [17] or exploit cross-channel redundancy to suppress uncorrelated noise [27]. The critical limitation is vulnerability to dataset-specific bias — when trained and tested on the same rig (predominantly CWRU), CNNs may exploit rig-specific acoustic properties, sensor characteristics, and noise patterns that do not generalize.

6.3. Sequential, Attention-Based, and Transformer Models

LSTMs and GRUs can track fault-related temporal patterns across intervals containing noise-dominated segments, particularly valuable under intermittent noise conditions. Transformer architectures, through self-attention, attend to fault-bearing segments while down-weighting noise-dominated segments. A time-frequency dual-channel Transformer processes both temporal and spectral representations through parallel attention streams [28]. Practical limitations include large training data requirements conflicting with industrial data scarcity, quadratic computational complexity, and unreliable attention interpretability.

6.4. Domain Adaptation and Domain Generalization

Adversarial domain adaptation aligns feature distributions through a minimax game, while metric-based distribution matching provides an alternative. Source-free DA avoids retaining source data during target adaptation and has been evaluated for cross-domain machine fault diagnosis [22]. DG addresses unknown target domains; representative studies include causal factorization for cross-machine bearing diagnosis [23] and curriculum learning for cross-domain fault diagnosis with category shift [30]. Because these studies span machine-level and bearing-specific tasks, their conclusions should not be treated as interchangeable without task-level validation. Distribution alignment may also suppress diagnostically relevant variation.

6.5. Label-Noise-Aware, Weakly Supervised, and Semi-Supervised Models

Robust losses and co-teaching are general strategies for learning with corrupted annotations, but the two cited bearing studies address related rather than identical problems. Chen et al. proposed a semi-supervised prototype network for noisy data with limited labels [32], whereas Liu et al. estimated pseudo-label uncertainty in source-free domain adaptation under limited-sample conditions [33]. Neither paper, by itself, validates robustness to controlled asymmetric ground-truth label flips. The domain-specific challenge is therefore to distinguish signal-level noisy samples, uncertain pseudo-labels, and genuine annotation errors in both method design and evaluation.

6.6. Self-Supervised, Contrastive, Generative, and Foundation-Model-Oriented Approaches

Self-supervised learning can pretrain feature extractors before limited-label fine-tuning; examples include machinery diagnosis with time-frequency invariance [25] and high-speed-train bearing diagnosis under limited labels and speed variability [34]. Contrastive learning has also been studied for rotating-machinery diagnosis [24], but that evidence is not bearing-specific. Diffusion-related models have been explored for unknown bearing-fault recognition [35] and speed-invariant bearing diagnosis [36]; these studies target out-of-distribution faults or operating-speed variation rather than demonstrating general signal-noise removal. LLM-TSFD is a human-in-the-loop industrial time-series diagnosis method [37], not a bearing-specific noise-robust model, so its relevance here is prospective and cross-domain.

6.7. Limitations of Model-Centric Robustness

Pervasive limitations include: validation exclusively under synthetic AWGN; accuracy as an insufficient robustness metric; rare cross-dataset, cross-machine, and cross-sensor validation; systematic underestimation of model complexity and deployment cost; conflation of noise types in evaluation; and insufficient interpretability and physical consistency checks. These limitations motivate the standardized evaluation framework and noise–method matching matrix of Sections 7 and 8.

Table 3. Noise-aware diagnostic models and their robustness mechanisms: model categories, targeted noise challenges, mechanisms, strengths, limitations, and representative references.

Model Category	Targeted Noise	Robustness Mechanism	Strengths	Limitations	References
CNN-Based (1D/2D, multi-scale, anti-noise, physics-informed)	Signal noise; operating-condition variation	Hierarchical feature learning; multi-scale processing; learned denoising or physical constraints	Automatic feature learning; flexible signal representations	Dataset-specific bias; lab-to-field gap	[20], [17], [26], [38], [18]
Attention / Transformer	Non-stationary, intermittent noise	Self-attention; temporal context integration	Long-range dependency capture; adaptive focus	Large data requirements; quadratic complexity	[28]
DA / DG (machine and bearing studies)	Cross-condition and cross-machine domain shift	Adaptation without retained source data; causal factorization; curriculum learning; staged transfer	Directly targets distribution shift	Task and target-domain assumptions; possible negative transfer	[22], [23], [31], [30]
Noisy-/Limited-Label and Pseudo-Label-Aware	Noisy samples; limited labels; uncertain pseudo-labels	Semi-supervised prototype learning; pseudo-label uncertainty estimation	Uses unlabeled or uncertain target samples	Not direct evidence for asymmetric ground-truth label noise	[32], [33]
Self-Supervised / Contrastive	Limited labels; operating-condition variation	Pretext-task pretraining; cross-contrastive or multi-level contrastive learning	Reduced dependence on labeled data	Task transfer and positive-pair design; mixed machinery scope	[25], [34], [24]
Diffusion / LLM-Oriented (Emerging)	Unknown faults; speed variation; industrial time-series reasoning	Conditional diffusion; diffusion transformation; human-in-the-loop LLM	New mechanisms for OOD and cross-condition diagnosis	Heterogeneous tasks; q13 is not bearing-specific; no generic noise-robustness evidence	[37], [35], [36]

Note: Abbreviations: DA = Domain Adaptation; DG = Domain Generalization; SVM = Support Vector Machine; RF = Random Forest; k NN = k -Nearest Neighbors; LLM = Large Language Model.

7. Evaluation and Benchmarking for Noise-Robust Bearing Fault Diagnosis

7.1. Why Evaluation Matters in Noise-Robust Diagnosis

Translation of noise-robust diagnostic methods is bottlenecked by the absence of a shared evaluation language. Across the reviewed literature, the same problem is studied under different noise assumptions, datasets, splits, SNR definitions, and metrics, making cross-study comparison impossible. High accuracy on clean test sets does not imply noise robustness — a model may achieve near-perfect clean accuracy while degrading to random performance under moderate noise. The gap between laboratory noise simulation and industrial reality further complicates evaluation: AWGN-only validation demonstrates a necessary but not sufficient condition.

7.2. Common Benchmark Datasets and Their Limitations

The CWRU dataset is a widely used benchmark, but Hendriks et al. identified a specific flaw in a common domain-shift construction: the same physical bearings can appear in both training and testing sets, allowing models to learn bearing-specific characteristics rather than generalize to independent bearings [5]. Other public datasets provide different rigs, operating conditions, and degradation settings, yet no single widely adopted benchmark spans the noise taxonomy used in this review. Consequently, marginal accuracy gains on one CWRU split should not be interpreted as broad evidence of noise robustness or deployment readiness.

7.3. Evaluation under Signal Noise

An adequate signal-noise evaluation protocol requires multiple noise models (minimum AWGN plus one non-Gaussian type), multi-level SNR from mild (5–10 dB) to severe (-5 to -10 dB), and explicit documentation of training/testing noise relationship (clean, matched, or mismatched). Reporting accuracy at a single SNR provides no information about the robustness degradation profile; accuracy-SNR curves revealing the shape of the robustness function are essential.

7.4. Evaluation under Operating-Condition Shift

Domain shift evaluation requires cross-condition protocols: train on source, test on held-out target conditions differing in speed, load, machine, or sensor. Within this, source-only, domain adaptation, and domain generalization protocols define increasing difficulty levels. Leave-one-condition-out cross-validation provides systematic generalization assessment. The critical concern is target-domain leakage — hyperparameter tuning based on target performance overestimates generalization ability. An application-oriented perspective on DG benchmarking emphasizes realistic deployment scenarios over arbitrary dataset splits [29].

7.5. Evaluation under Label Noise

Label noise evaluation introduces controlled errors into cleanly labeled datasets, measuring performance degradation as a function of error rate. Symmetric noise dominates evaluations but poorly represents industrial reality — asymmetric noise (incipient-to-healthy confusion, adjacent severity confusion) is more realistic but rarely tested.

Instance-dependent noise, where ambiguity depends on the specific sample, is almost entirely absent. For methods including label detection mechanisms, label correction accuracy and downstream diagnostic accuracy should be reported separately.

7.6. Metrics beyond Accuracy

The accuracy monoculture obscures class-imbalance sensitivity, robustness degradation profiles, and fault-type-specific vulnerabilities. Balanced accuracy and macro-averaged F1-score address class imbalance. Accuracy-SNR curves, robustness degradation rate, area under the accuracy-SNR curve, and low-SNR limit provide quantitative robustness characterization. Per-domain accuracy for domain shift reveals condition-specific generalization failures. Deployment-oriented metrics — inference time, model size, training time, data efficiency — are rarely reported but decisive for industrial adoption. Statistical significance testing with multiple random initializations is essential.

7.7. Proposed Standardized Benchmarking Protocol

Figure 3. Proposed benchmarking pipeline for noise-robust bearing fault diagnosis

This review proposes a ten-element protocol: (1) noise type declaration — specify which noise types from the taxonomy are evaluated, without conflation; (2) noise parameter specification — fully document noise model, parameters, and injection method; (3) multi-level evaluation — report robustness curves rather than single-point metrics; (4) transparent data splitting — identify temporal, machine-instance, or condition dependencies; (5) clean and noisy condition reporting — distinguish inherently accurate from specifically noise-robust methods; (6) cross-condition, cross-dataset, cross-sensor validation; (7) label noise as a separate evaluation dimension with asymmetric and class-dependent noise; (8) fair baseline comparison including conventional methods; (9) deployment-oriented reporting alongside accuracy; (10) reproducibility — code, models, split indices, and random seeds publicly available.

7.8. Limitations of Current Benchmarking Practice

Systematic shortcomings include: CWRU overdependence; AWGN-only idealized noise; absence of public industrial noise data; train-test contamination; single-point accuracy reporting; conflation of noise types without ablation; and the neglect of label noise–domain shift co-evaluation despite their industrial co-occurrence. These create a publication landscape where methods appearing to advance the state of the art under narrow conditions may offer no improvement under broader industrial deployment conditions.

8. Discussion — Critical Appraisal and Noise–Method Matching

8.1. From Method Categories to Noise-Specific Strategy Selection

The conventional architecture-based organization of the literature obscures the clinical question: given a specific noise environment, which strategy is appropriate? Noise types differ in the diagnostic pipeline stage they affect, the same method performs differently across noise types, and compound noise — the norm in industrial deployment — cannot be addressed by any single method category. These observations motivate the matching framework below.

8.2. Noise–Method Matching Matrix

The noise–method matching matrix (Table 4) is the central synthetic contribution of this review. It integrates the noise taxonomy (Section 2) with denoising strategies (Section 4), feature engineering (Section 5), model-level designs (Section 6), and evaluation protocols (Section 7) into a unified decision framework. Each row describes a complete diagnostic strategy for a given noise scenario; each column reveals differential strategy application across noise types. A critical observation is the asymmetry of coverage: signal-level noise is the most extensively studied scenario; domain shift has a growing but less mature base, with DG remaining an open challenge; label noise is the least studied in bearing diagnosis specifically; and compound noise has almost no dedicated methodology.

Table 4. Noise–Method Matching Matrix: a decision framework mapping each noise category to appropriate combinations of denoising, feature engineering, model-level, and evaluation strategies for rolling bearing fault diagnosis.

Noise Type	Main Source	Affected Stage	Denoising Strategy	Feature Strategy	Model Strategy	Evaluation Protocol	Key Limitations
Environmental / Background Signal Noise	Adjacent machinery, EMI, structural resonance	Signal acquisition	Band-pass filtering, wavelet thresholding, VMD/EEMD, adaptive filtering	Multi-domain features, multi-scale entropy, spectral band energy	Anti-noise CNN, multi-scale CNN, attention-enhanced models, physics-informed NN	Multi-SNR (AWGN + non-Gaussian), accuracy-SNR curve, matched vs. mismatched noise	Non-Gaussian/non-stationary noise understudied; filter parameter selection in unsupervised settings
Sensor and Acquisition Noise	Thermal noise, quantization, ADC resolution, sensor aging	Signal acquisition	Wiener filtering, adaptive filtering, calibration, signal reconstruction	Sensor-fusion features, entropy-based features (amplitude-invariant)	Multi-channel CNN, sensor-fusion architectures, uncertainty-aware models	SNR variation, cross-sensor validation, sensor degradation simulation	Quantization noise is signal-dependent; sensor aging is systematic; calibration-dependent
Operating-Condition-Induced / Domain Shift	Speed/load variation, machine-to-machine variation	Feature distribution, decision boundary	Not primarily a denoising problem; order tracking mitigates speed effects	Domain-invariant features, causal features, contrastive representations	DA (adversarial, MMD/CORAL), DG (causal, curriculum), progressive transfer	Cross-condition, leave-one-condition-out, source-only vs. DA vs. DG; document target leakage	DG harder than DA; alignment may suppress fault-discriminative variation; unknown target domains
Label Noise / Supervision Noise	Annotation error, severity ambiguity, maintenance record inaccuracy	Model training (supervisory signal)	Not addressable by signal denoising	Stability selection across label perturbations	Robust loss, co-teaching, prototype networks, uncertainty-weighted pseudo-labeling	Multi-rate (symmetric+asymmetric+class-dependent), label correction accuracy vs. diagnostic accuracy	Asymmetric/instance-dependent noise harder; small-sample amplification; bearing-specific patterns understudied
Compound Noise (Multi-Source)	Co-occurrence of above noise types	Entire diagnostic pipeline	Combined: filtering + decomposition + order tracking	Multi-strategy: domain-invariant + entropy + multi-domain	Integrated: DA with robust loss, self-supervised pretraining with noise augmentation	Separate ablation per noise type; interaction effect analysis; cross-dataset + cross-condition + label perturbation	Almost no dedicated compound-noise methodology; evaluation protocols non-existent; computational cost high

Note: This matrix is the central synthetic contribution of this review and is not a reproduction of any existing framework. Abbreviations: TFR = Time-Frequency Representation; MMD = Maximum Mean Discrepancy; CORAL = Correlation Alignment.

8.3. Industrial Realism, Methodological Trade-Offs, and Deployment Constraints

A major limitation of current noise-robust bearing diagnosis research is the persistent laboratory–industrial gap. This gap is not caused by a single factor, but by the interaction of noise realism, dataset construction, and deployment constraints. Industrial disturbances can be multi-source and non-stationary, while many experimental studies use controlled laboratory data and simplified injected noise. Deployment-oriented work therefore increasingly emphasizes model compactness alongside robustness; for example, Zhang et al. proposed a lightweight framework for hydro-turbine main-shaft-bearing diagnosis under noise interference [39]. Progress requires not only improved algorithms, but also realistic industrial data collection and reporting of model size, latency, and hardware conditions.

These constraints expose several trade-offs across the diagnostic pipeline. At the signal level, denoising must balance noise suppression against preservation of weak fault signatures, especially for incipient faults whose characteristic components may be easily removed by aggressive filtering or decomposition. At the feature level, robustness requires invariance to nuisance variation, but excessive invariance may suppress discriminative fault-related information. At the model level, higher expressive capacity can improve fitting under controlled settings, yet simpler or more constrained models may generalize more reliably under domain shift and limited supervision. At the system level, highly specialized methods can perform strongly under a narrow noise scenario, whereas composable and modular methods may offer broader robustness but weaker single-scenario performance. These trade-offs suggest that robustness should be treated as a system-level design objective rather than as an isolated property of denoising, feature extraction, or classification.

Practical deployment further strengthens this point. Industrial adoption depends on whether a method can operate with feasible data acquisition settings, such as lower sampling rates, single-axis measurements, and existing plant instrumentation. It also depends on inference latency, model size, delayed or uncertain label availability, and interpretability for maintenance decision-making. In this sense, interpretability is not merely an academic preference, but an operational requirement: diagnostic outputs must be credible enough to support maintenance actions, downtime decisions, and risk assessment in safety-critical environments.

8.4. Evaluation Pitfalls, Reproducibility, and Benchmarking Needs

The current evaluation paradigm remains insufficient for assessing true noise robustness. The most consequential pitfalls include accuracy monoculture, single-point evaluation, contamination between training and test splits, and noise-type conflation without ablation. Reporting only clean-test accuracy or a single noisy accuracy value obscures how performance degrades as noise intensity changes, whether a method is robust to mismatched noise, and whether improvements arise from genuine robustness or from favorable data splitting. Similarly, evaluating compound noise without separating the contribution of each noise source makes it impossible to determine whether a method addresses environmental noise, domain shift, label uncertainty, or their interactions.

Reproducibility is another major barrier. Many studies do not release code, trained models, data splits, random seeds, or detailed noise-generation procedures, making it difficult to compare methods fairly or reproduce reported gains. This issue is especially serious in noise-robust diagnosis because small changes in preprocessing, sampling windows, SNR definition, or train–test partitioning can produce large differences in reported accuracy. Robustness claims therefore require transparent protocols, repeated trials, and explicit reporting of noise settings, model complexity, computational cost, and statistical variability.

A standardized benchmark would substantially improve the field's ability to measure progress. Such a benchmark should include multiple datasets, multiple noise categories, predefined splits, and evaluation modes that distinguish clean-to-noisy, matched-noise, mismatched-noise, cross-condition, cross-dataset, and label-perturbation scenarios. It should also report multi-metric results, including accuracy, robustness degradation rate, area under the accuracy–SNR curve, inference latency, model size, and preferably interpretability or physical-consistency indicators. Without such benchmarking, the field risks optimizing for narrow laboratory performance rather than robust industrial diagnosis.

8.5. Research Gaps and Design Implications

Several structural research gaps define the frontier of noise-robust bearing fault diagnosis. The first and most fundamental is the absence of publicly available industrial noise data with documented noise sources. This gap limits both realistic method development and fair evaluation. Second, noise-type definitions remain unstandardized, making it difficult to compare studies or determine whether different papers are addressing the same problem. Third, the synthetic-to-real noise performance gap remains under-explored, especially for non-Gaussian, impulsive, colored, and non-stationary noise. Fourth, compound noise research is still rare, despite the fact that real industrial environments commonly involve multiple simultaneous noise sources. Fifth, label noise remains understudied in bearing diagnosis, particularly structured, asymmetric, and instance-dependent annotation errors caused by maintenance records, severity ambiguity, and delayed inspection.

Additional gaps concern validation scope and deployment realism. Cross-dataset, cross-machine, and cross-sensor validation remain uncommon; standardized noise-robustness benchmarks are absent; interpretability and physical consistency are rarely evaluated systematically; and the trade-off among robustness, lightweight design, and real-time performance remains insufficiently understood. Emerging approaches such as foundation models, LLM-assisted diagnosis, and diffusion-based methods may offer new opportunities, but their value for industrial bearing diagnosis remains unvalidated without realistic noise data and deployment-oriented evaluation.

These gaps imply four principles for future method design. First, noise-type awareness should be built into diagnostic frameworks from the outset, rather than added as a post-hoc evaluation condition. Second, evaluation protocols should match the claimed capability of the method: denoising methods should be tested across realistic signal noise, domain adaptation methods across condition and machine shifts, and robust training methods under structured label corruption. Third, generalization should become the primary measure of progress, rather than benchmark accuracy under narrow experimental assumptions. Fourth, deployment constraints such as latency, model size, data acquisition feasibility, interpretability, and continual adaptability should be treated as first-class evaluation criteria alongside diagnostic accuracy.

9. Conclusions and Future Directions

9.1. Main Conclusions

This narrative review leads to five principal conclusions. First, noise in bearing fault diagnosis is not monolithic; it comprises at minimum environmental noise, sensor and acquisition noise, operating-condition-induced domain shift, and label noise, each of which affects the diagnostic pipeline in a different way. Second, because these noise types influence different stages of diagnosis, they require differentiated strategies: signal-level noise calls for denoising and robust feature extraction, operational variability requires DA/DG and domain-invariant representation learning, and supervision noise requires robust training and uncertainty-aware learning. Third, no single class of methods is sufficient on its own. Effective robustness requires coordinated design across denoising, feature engineering, model architecture, training strategy, and evaluation protocol. Fourth, the dominant evaluation paradigm of single-dataset AWGN validation with accuracy-only reporting is structurally unable to assess industrial noise robustness. Fifth, the proposed noise taxonomy, evaluation protocol, and noise–method matching matrix provide a structured basis for noise-conditional strategy selection.

9.2. Future Directions

Future research should first prioritize realistic noise modeling and industrial noise data. The most consequential barrier remains the absence of open industrial vibration datasets with documented noise sources. Progress requires the collection and release of plant-level vibration data, the development of parametric noise models for impulsive EMI, colored mechanical background, and non-stationary process noise, and systematic comparison between synthetic-noise and real-noise performance. Such work would provide the empirical foundation needed to move beyond laboratory AWGN settings.

A second direction is the development of noise-type-aware diagnostic frameworks. Rather than assuming a fixed noise condition, future systems should identify active noise characteristics from the signal itself and adapt the diagnostic strategy accordingly. This creates a meta-diagnostic problem in which the system must jointly infer the fault condition and the noise environment. Modular architectures with swappable denoising modules, domain-invariant extractors, robust classification heads, and uncertainty-aware decision mechanisms represent a promising path beyond single-strategy methods.

A third direction is standardized robustness benchmarking. A curated multi-dataset, multi-noise benchmark with predefined splits and a public leaderboard would make it possible to compare methods under clean-to-noisy, matched-noise, mismatched-noise, cross-condition, and cross-dataset settings. Such a benchmark should report not only accuracy, but also robustness degradation rate, area under the accuracy–SNR curve, inference latency, model size, and computational cost. This would shift evaluation from isolated performance claims toward measurable and reproducible robustness.

A fourth direction is interpretable and physics-consistent robust diagnosis. Bearing-specific attention visualization, physics-informed regularization, and evaluation against known fault characteristic frequencies can improve trust and operational acceptance. Interpretability should be evaluated not only visually, but also through consistency with bearing kinematics and fault mechanisms. This is especially important in safety-critical settings where maintenance decisions must be justified.

Finally, future methods must become lightweight, deployable, and continually adaptable. Industrial systems require models that can operate under limited hardware resources, update from streaming data without catastrophic forgetting, and handle uncertain or delayed labels. Cost-sensitive evaluation is also needed because false alarms and missed detections have asymmetric operational consequences. Robustness, efficiency, adaptability, and decision cost should therefore be optimized jointly rather than treated as separate objectives.

9.3. Final Perspective

Noise in bearing fault diagnosis should not be treated merely as a disturbance to be removed, but as a condition to be characterized, understood, and designed for. The evidence reviewed in this paper suggests that noise-robust diagnosis cannot be achieved through a single universal model or a single preprocessing technique. A more realistic path forward requires noise-type identification, strategy selection, and evaluation protocols that reflect the complexity of industrial environments. The noise taxonomy, evaluation protocol, and noise–method matching matrix developed in this review are intended to support this transition from method-centered benchmarking toward noise-aware and deployment-oriented diagnostic design.

References

[1] 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, 106587. https://doi.org/10.1016/j.ymssp.2019.106587

[2] Zhang, S., Zhang, S., Wang, B., & Habetler, T. G. (2020). Deep learning algorithms for bearing fault diagnostics—A comprehensive review. IEEE Access, 8, 29857–29881. https://doi.org/10.1109/ACCESS.2020.2972859

[3] Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 115, 213–237. https://doi.org/10.1016/j.ymssp.2018.05.050

[4] Pancaldi, F., Dibiase, L., & Cocconcelli, M. (2023). Impact of noise model on the performance of algorithms for fault diagnosis in rolling bearings. Mechanical Systems and Signal Processing, 188, 109975. https://doi.org/10.1016/j.ymssp.2022.109975

[5] Hendriks, J., Dumond, P., & Knox, D. A. (2022). Towards better benchmarking using the CWRU bearing fault dataset. Mechanical Systems and Signal Processing, 169, 108732. https://doi.org/10.1016/j.ymssp.2021.108732

[6] Bao, Z., Liu, C., Yang, H., Zhang, J., & Li, Y. (2026). Deep learning-enabled intelligent diagnosis of bearing faults in complex environments: A comprehensive review. Engineering Applications of Artificial Intelligence, 163, 113068. https://doi.org/10.1016/j.engappai.2025.113068

[7] Chen, M., Zheng, J., Pei, H., Yang, L., Zhang, Q., Han, Q., & Luo, H. (2026). Advances and prospects in noise-robust fault diagnosis for rotating machinery: From mechanism-based approaches to data-driven models. Measurement Science and Technology, 37(19), 192002. https://doi.org/10.1088/1361-6501/ae65b7

[8] Hebda-Sobkowicz, J., Zimroz, R., Pitera, M., & Wyłomańska, A. (2020). Selection of the informative frequency band in a bearing fault diagnosis in the presence of non-Gaussian noise—Comparison of recently developed methods. Mechanical Systems and Signal Processing, 145, 106971. https://doi.org/10.1016/j.ymssp.2020.106971

[9] Sun, H., Gao, S., Ma, S., & Lin, S. (2022). A fault mechanism-based model for bearing fault diagnosis under non-stationary conditions without target condition samples. Measurement, 199, 111499. https://doi.org/10.1016/j.measurement.2022.111499

[10] Ma, Y., Yang, J., & Li, L. (2023). Meta Bi-classifier Gradient Discrepancy for noisy and universal domain adaptation in intelligent fault diagnosis. Knowledge-Based Systems, 276, 110735. https://doi.org/10.1016/j.knosys.2023.110735

[11] Schmidt, S., Wilke, D. N., & Gryllias, K. C. (2025). Generalised envelope spectrum-based signal-to-noise objectives: Formulation, optimisation and application for gear fault detection under time-varying speed conditions. Mechanical Systems and Signal Processing, 224, 111974. https://doi.org/10.1016/j.ymssp.2024.111974

[12] Liu, T., Wei, Z., Hu, L., Xie, X., Dong, X., Noman, K., & Li, Y. (2025). A resonance demodulation method based on amplitude Z-score of envelope spectrum for weak bearing fault detection. Mechanical Systems and Signal Processing, 238, 113265. https://doi.org/10.1016/j.ymssp.2025.113265

[13] Xu, H., & Zhou, S. (2025). Maximum L-Kurtosis deconvolution and frequency-domain filtering algorithm for bearing fault diagnosis. Mechanical Systems and Signal Processing, 223, 111916. https://doi.org/10.1016/j.ymssp.2024.111916

[14] Miao, Y., Li, C., Shi, H., & Han, T. (2023). Deep network-based maximum correlated kurtosis deconvolution: A novel deep deconvolution for bearing fault diagnosis. Mechanical Systems and Signal Processing, 189, 110110. https://doi.org/10.1016/j.ymssp.2023.110110

[15] Shen, J., Wang, Z., Wang, Y., Zhu, H., Zhang, L., & Tang, Y. (2025). AGWO-PSO-VMD-TEFCG-AlexNet bearing fault diagnosis method under strong noise. Measurement, 242, 116259. https://doi.org/10.1016/j.measurement.2024.116259

[16] Liao, J.-X., He, C., Li, J., Sun, J., Zhang, S., & Zhang, X. (2025). Classifier-guided neural blind deconvolution: A physics-informed denoising module for bearing fault diagnosis under noisy conditions. Mechanical Systems and Signal Processing, 222, 111750. https://doi.org/10.1016/j.ymssp.2024.111750

[17] Wang, Q., & Xu, F. (2023). A novel rolling bearing fault diagnosis method based on adaptive denoising convolutional neural network under noise background. Measurement, 218, 113209. https://doi.org/10.1016/j.measurement.2023.113209

[18] Kim, Y., & Kim, Y.-K. (2024). Physics-informed time-frequency fusion network with attention for noise-robust bearing fault diagnosis. IEEE Access, 12, 12517–12532. https://doi.org/10.1109/ACCESS.2024.3355268

[19] Wang, H., Liu, Z., Peng, D., & Cheng, Z. (2022). Attention-guided joint learning CNN with noise robustness for bearing fault diagnosis and vibration signal denoising. ISA Transactions, 128, 470–484. https://doi.org/10.1016/j.isatra.2021.11.028

[20] Sun, H., Cao, X., Wang, C., & Gao, S. (2022). An interpretable anti-noise network for rolling bearing fault diagnosis based on FSWT. Measurement, 190, 110698. https://doi.org/10.1016/j.measurement.2022.110698

[21] He, D., Zhang, Z., Jin, Z., Zhang, F., Yi, C., & Liao, S. (2025). RTSMFFDE-HKRR: A fault diagnosis method for train bearing in noise environment. Measurement, 239, 115417. https://doi.org/10.1016/j.measurement.2024.115417

[22] Zhang, Y., Ren, Z., Feng, K., Yu, K., Beer, M., & Liu, Z. (2023). Universal source-free domain adaptation method for cross-domain fault diagnosis of machines. Mechanical Systems and Signal Processing, 191, 110159. https://doi.org/10.1016/j.ymssp.2023.110159

[23] Jia, S., Li, Y., Wang, X., Sun, D., & Deng, Z. (2023). Deep causal factorization network: A novel domain generalization method for cross-machine bearing fault diagnosis. Mechanical Systems and Signal Processing, 192, 110228. https://doi.org/10.1016/j.ymssp.2023.110228

[24] Shu, Z., Peng, D., Wang, H., Mao, C., & Yu, Z. (2025). Time-frequency perception guided multi-level contrastive learning for rotating machinery fault diagnosis. Expert Systems with Applications, 293, 128664. https://doi.org/10.1016/j.eswa.2025.128664

[25] Tang, J., Xiao, J., Chen, W., Li, X., Wei, C., Ding, X., & Huang, W. (2024). A prior knowledge-enhanced self-supervised learning framework using time-frequency invariance for machinery intelligent fault diagnosis with small samples. Engineering Applications of Artificial Intelligence, 133, 108503. https://doi.org/10.1016/j.engappai.2024.108503

[26] Kim, Y., & Kim, Y.-K. (2023). Time-frequency multi-domain 1D convolutional neural network with channel-spatial attention for noise-robust bearing fault diagnosis. Sensors, 23(23), 9311. https://doi.org/10.3390/s23239311

[27] Zhang, W.-T., Liu, L., Cui, D., Ma, Y.-Y., & Huang, J. (2023). An anti-noise convolutional neural network for bearing fault diagnosis based on multi-channel data. Sensors, 23(15), 6654. https://doi.org/10.3390/s23156654

[28] Xu, Z., Jia, Z., Wei, Y., Zhang, S., Jin, Z., & Dong, W. (2024). A strong anti-noise and easily deployable bearing fault diagnosis model based on time-frequency dual-channel Transformer. Measurement, 236, 115054. https://doi.org/10.1016/j.measurement.2024.115054

[29] Zhao, C., Zio, E., & Shen, W. (2024). Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study. Reliability Engineering & System Safety, 245, 109964. https://doi.org/10.1016/j.ress.2024.109964

[30] Wang, Y., Gao, J., Wang, W., Yang, X., & Du, J. (2024). Curriculum learning-based domain generalization for cross-domain fault diagnosis with category shift. Mechanical Systems and Signal Processing, 212, 111295. https://doi.org/10.1016/j.ymssp.2024.111295

[31] Huang, K., Ren, Z., Zhu, L., Lin, T., Zhu, Y., Zeng, L., & Wan, J. (2025). A three-stage bearing transfer fault diagnosis method for large domain shift scenarios. Reliability Engineering & System Safety, 254, 110641. https://doi.org/10.1016/j.ress.2024.110641

[32] Chen, Y., Yue, J., Liu, Z., & Chen, J. (2025). A semi-supervised wise-attention weighted prototype network for rolling bearing fault diagnosis under noisy and limited labeled data conditions. Neurocomputing, 647, 130563. https://doi.org/10.1016/j.neucom.2025.130563

[33] Liu, R., Ma, W., Kuang, F., Guo, J., & Zhao, N. (2024). A source free robust domain adaptation approach with pseudo-labels uncertainty estimation for rolling bearing fault diagnosis under limited sample conditions. Knowledge-Based Systems, 304, 112443. https://doi.org/10.1016/j.knosys.2024.112443

[34] Li, Y., Wang, T., Xie, J., Yang, J., Pan, T., Niu, B., & Yu, X. (2025). A temporal cross-contrastive self-supervised learning framework for high-speed train bearing fault diagnosis: Addressing limited labeling and speed variability. Engineering Applications of Artificial Intelligence, 158, 111537. https://doi.org/10.1016/j.engappai.2025.111537

[35] Zhang, Y., Yang, B., & Zhao, Z. (2025). Differential out-of-distribution error guidance for unknown bearing fault diagnosis with conditional diffusion model. Engineering Applications of Artificial Intelligence, 158, 111350. https://doi.org/10.1016/j.engappai.2025.111350

[36] Liao, R., Wang, C., Peng, F., Liang, W., Zhang, Y., & Zhang, X. (2024). DTM-Bearing: A novel framework for speed-invariant bearing fault diagnosis based on diffusion transformation model (DTM). IEEE Access, 12, 8875–8888. https://doi.org/10.1109/ACCESS.2024.3351935

[37] Zhang, Q., Xu, C., Li, J., Sun, Y., Bao, J., & Zhang, D. (2025). LLM-TSFD: An industrial time series human-in-the-loop fault diagnosis method based on a large language model. Expert

Systems with Applications, 264, 125861. https://doi.org/10.1016/j.eswa.2024.125861

[38] Ni, Q., Ji, J. C., Halkon, B., Feng, K., & Nandi, A. K. (2023). Physics-informed residual network (PIResNet) for rolling element bearing fault diagnostics. Mechanical Systems and Signal Processing, 200, 110544. https://doi.org/10.1016/j.ymssp.2023.110544

[39] Zhang, H., Liu, Z., Si, H., Yu, K., Li, S., & Yan, Z. (2025). A lightweight fault diagnosis framework for hydro-turbine main shaft bearing under noise interference. IEEE Access, 13, 102133–102143. https://doi.org/10.1109/ACCESS.2025.3578682