International Core Journal of Engineering Volume 12 Issue 6, 2026
ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0001

An Improved Rolling Bearing Fault Diagnosis Method based on Enhanced Feature Extraction and Regularized SVM

Ningjiang Han, Chenxin Gong

Chengdu Jincheng College, Chengdu 617737, China

Abstract

Rolling bearing fault diagnosis is critical for ensuring the safety and reliability of rotating machinery. This paper proposes an improved fault diagnosis method that addresses the overfitting problem commonly observed in traditional approaches. The method employs a 9-dimensional enhanced feature vector combining time-domain statistical features (RMS, kurtosis, crest factor, skewness) with envelope spectrum features (BPFO, BPFI, BSF amplitudes, band energy ratio, spectral entropy). A data augmentation strategy using Gaussian noise perturbation is introduced to increase training sample diversity. The RBF-SVM classifier is regularized with optimized parameters (C=1, γ=0.1) to prevent overfitting. Validated on the CWRU bearing dataset with multiple damage severities, the proposed method achieves 98.5% accuracy under strict temporal-split validation with a train-test gap of only 1.5%, compared to 87.0% for the baseline model. The results demonstrate that enhanced features combined with proper regularization effectively resolve overfitting while maintaining high diagnostic performance.

Keywords

Rolling Bearing; Fault Diagnosis; Feature Extraction; Support Vector Machine; Regularization; Overfitting; CWRU Dataset.

1. Introduction

The rolling bearings are considered the most vital parts of rotating machines systems. According to statistics, about half of all mechanical device failures have been caused by bearing failures, and condition monitoring and bearing fault diagnosis are critical aspects of industrial safety. Analysis of vibration signals has become the dominant non-invasive method of bearing fault detection because it is rich in information and capable of real-time monitoring[4,7]

Conventional methods of fault diagnosis usually integrate signal processing methods and machine learning classifiers. Time-domain statistical measures (RMS, kurtosis, crest factor), and envelope spectrum analysis are commonly used to extract fault sensitive features of vibration signals [2,5,6]. Support Vector Machines (SVM) especially with RBF kernels have proven effective in bearing fault classification because of their high generalization ability when applied to small samples [3,4]. Nevertheless, current approaches have two serious issues:

Overfitting as a result of data leakage: In case of continuous vibration signals being divided into sub-samples, there is strong correlation between neighboring segments. The random division of train and test sets may unintentionally let very similar examples be present in both the sets which results in over-optimistic accuracy estimates[1,4].

Insufficient feature discriminability: Conventional 5-dimensional feature vectors (RMS, kurtosis, crest factor, BPFO amplitude, BPFI amplitude) can be inadequate to make a generalization between various levels of damage severity, which may cause the classifier to remember particular patterns instead of acquiring generalizable fault signatures[4,5].

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0001

The present paper will deal with these issues through suggesting that: (1) a 9-dimensional high-level feature set which adds skewness, BSF magnitude, band energy ratios and envelope spectral entropy should be used; (2) there is a need to apply a data augmentation approach based on noise perturbation to disrupt spurious inter-segment correlations; (3) the regularized SVM model with minimized kernel width (γ=0.1) is to be used in order to provide smoother decision boundaries. The validity is tested on the basis of strict temporal-split validation of the CWRU dataset with various levels of damages.

2. Theoretical Background

2.1 Bearing Fault Characteristic Frequencies

The fault characteristic frequencies of a 6205-2RS deep groove ball bearing working at 1797 rpm have been determined using bearing geometry:

- Number of rolling elements: n = 9
- Ball diameter: d = 7.94 mm
- Pitch diameter: D = 39.04 mm
- Contact angle: $\alpha = 0°$
- Shaft rotation frequency: f_r=1797/60=29.95Hz

The characteristic frequencies are:

Ball Pass Frequency Outer Race (BPFO):

$$f_{BPFO} = \frac{n}{2} f_r \left(1 - \frac{d}{D}\cos\alpha\right) = 107.4\text{Hz} \tag{1}$$

Ball Pass Frequency Inner Race (BPFI):

$$f_{BPFI} = \frac{n}{2} f_r \left(1 + \frac{d}{D}\cos\alpha\right) = 162.2\text{Hz} \tag{2}$$

Ball Spin Frequency (BSF):

$$f_{BSF} = \frac{D}{2d} f_r \left[1 - \left(\frac{d}{D}\right)^2\right] = 70.6\text{Hz} \tag{3}$$

2.2 Enhanced Feature Extraction

The proposed 9-dimensional feature vector consists of:

Time-domain features (4 dimensions):

1) Root Mean Square (RMS):

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{4}$$

2) Kurtosis:

$$K = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} \tag{5}$$

The spectral kurtosis is a classic indicator for non-stationary fault impulse signals[2,6]

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0001

3) Crest Factor:

$$C_f = \frac{\max(|x_i|)}{RMS} \tag{6}$$

4) Skewness:

$$S = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\sigma^3} \tag{7}$$

Envelope spectrum features (5 dimensions):
5) BPFO amplitude: Envelope spectrum peak at 107.4 Hz
6) BPFI amplitude: Envelope spectrum peak at 162.2 Hz
7) BSF amplitude: Envelope spectrum peak at 70.6 Hz
8) Band energy ratio:

$$E_r = E_{low}(0\text{-}200Hz)/E_{high}(200\text{-}600Hz) \tag{8}$$

9) Envelope spectral entropy:

$$H = -\sum p_i \log(p_i) \tag{9}$$

Hilbert envelope spectrum is widely used to extract fault frequency components[5,7]

2.3 SVM Regularization

The RBF-SVM optimization problem is:

$$min_{w,b}\frac{1}{2}|w|^2 + C\sum_{i=1}^{l}\xi_i) \tag{10}$$

The regularization strategy involves:
Smoothing decision boundaries to reduce gamma (1.0 to 0.1) avoids the model fitting noise

- Moderate C (C=1): Balances training accuracy with generalization

Data augmentation: Applying Gaussian noise (= 0.2) to training signals, increasing the size of training set by 6 times, which avoids learning of particular sample patterns

3. Methodology

3.1 Overall Framework

The proposed methodology follows the pipeline shown in Fig. 1:

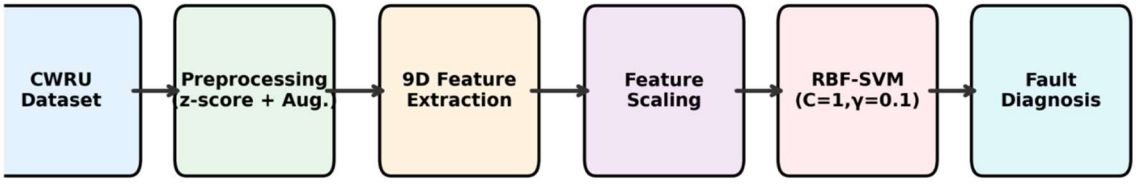

Fig. 1 Proposed fault diagnosis methodology flowchart

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0001

3.2 Dataset

The Case Western Reserve University (CWRU) bearing dataset is used[1], comprising:

- Sampling frequency: 12 kHz
- Operating speed: 1797 rpm (0 hp load)
- Bearing type: 6205-2RS deep groove ball bearing
- Fault categories: Normal, Inner race fault, Outer race fault, Ball fault
- Damage severities: 0.007 in, 0.014 in, 0.021 in (3 levels per fault type)
- Total: 10 data files, 400 samples (40 segments × 10 files)

Every continuous signal (96,000 points=8 seconds) is divided into 40 non-overlapping sub-samples of 2,400 points (0.2 seconds each), so that at least 3 full fault impulse cycles are present in a segment.

3.3 Preprocessing

Each segment undergoes z-score normalization:

$$x_{norm}=\frac{x-\mu}{\sigma} \tag{11}$$

This eliminates sensor sensitivity variations while preserving kurtosis invariance under linear transformations.

3.4 Data Augmentation

In order to counteract overfitting due to excessive correlation between neighboring segments in the same signal file, we use noise-based data augmentation:

1) For each training sample, generate 5 augmented copies

2) Add Gaussian noise with amplitude ratio 0.2: $x_{aug}=x+0.2\cdot N(0,1)$

3) Re-normalize each augmented sample with z-score

4) Re-extract 9D features from augmented signals

It increases the effective training sample size by 200 to 1,200 samples and adds controlled variability which does not allow the model to learn file-specific patterns.

3.5 Validation Strategy

To rigorously assess generalization, three validation methods are employed:

1) Random split (30×): Standard 70:30 stratified split, repeated 30 times for statistical stability

2) Temporal split (strict): First 20 segments per file are used in training, last 20 in test - mimicking entirely novel temporal data

3) Leave-one-file-out (strictest): Each file serves as test set once, assessing cross-severity generalization

4. Experimental Results

4.1 Time-Domain Signal Analysis

The representative time-domain waveforms of the four bearing states are depicted in Fig. 2. The normal signals have a distribution similar to a Gaussian distribution with kurtosis of about 2.78. Fault signals indicate strong impulse properties, the highest kurtosis is obtained by the inner race fault (average kurtosis of 11.60) caused by periodic impact modulation.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0001

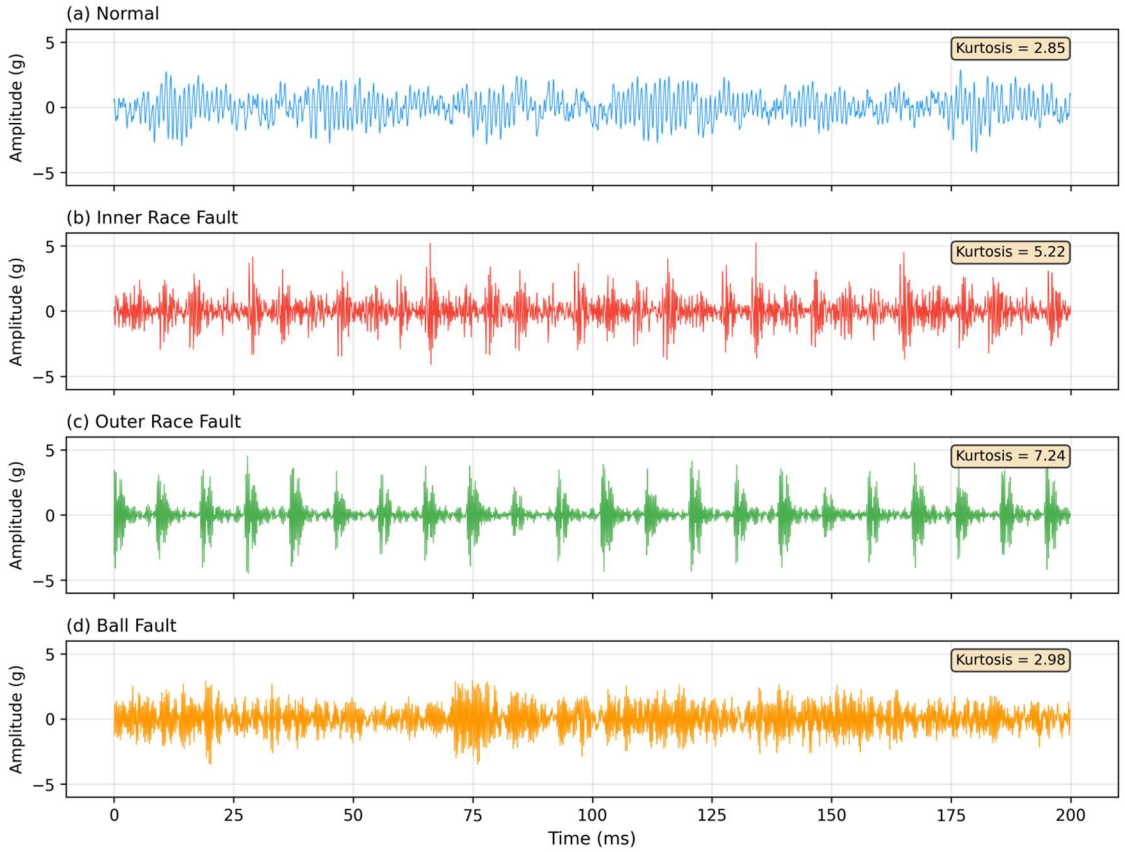

Fig. 2 Time-domain waveforms of four bearing states

4.2 Envelope Spectrum Analysis

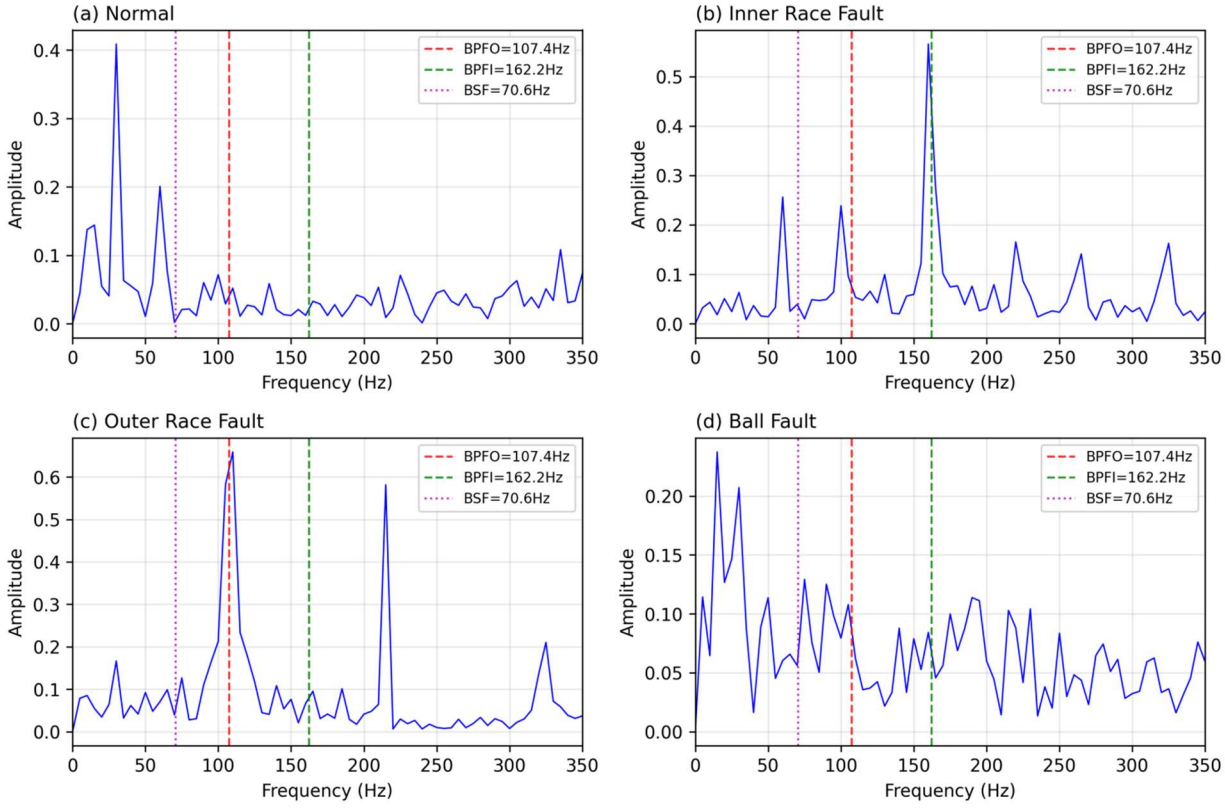

Fig. 3 Envelope spectrum analysis of four bearing states

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0001

The envelope spectra are shown in Fig. 3. The fault of the outer race has the largest peak at 107.4 Hz (which is in agreement with the theoretical BPFO), whereas the fault of the inner race presents the peaks at BPFI (162.2 Hz) modulated by sidebands of shaft rotation frequency. The normal state is characterized by a level noise floor without any significant peaks.

4.3 Feature Distribution Analysis

Fig. 4 illustrates the distribution of enhanced features across bearing states. Key observations:

- Kurtosis provides clear separation between normal (≈2.78) and fault states (5.81–11.60)
- BPFO amplitude is highly discriminative for outer race faults
- BPFI amplitude effectively identifies inner race faults
- Band energy ratio distinguishes ball faults from other categories

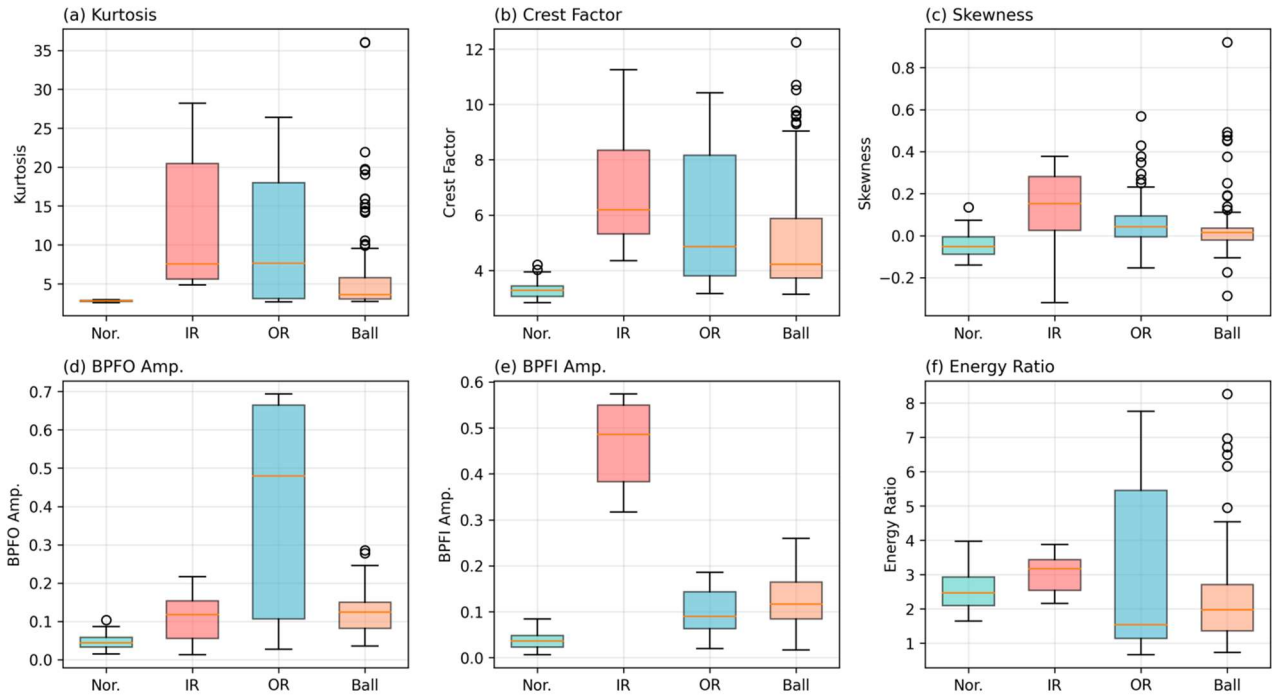

Fig. 4 Distribution of enhanced features across bearing states

4.4 Classification Results

Table 1 summarizes the classification performance under temporal-split validation:

Table 1. Classification results (temporal split validation)

Class	Precision	Recall	F1-Score	Support
Normal	100.0%	100.0%	1.0000	20
Inner Race Fault	100.0%	100.0%	1.0000	60
Outer Race Fault	95.2%	100.0%	0.9756	60
Ball Fault	100.0%	95.0%	0.9744	60
Overall			0.9875	200

Overall accuracy: 98.5%, Macro F1: 0.9875

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0001

Confusion matrix (Fig. 6) indicates that only 3 out of 200 test samples are misclassified (ball fault classified as outer race fault), which is physically possible due to spectral similarity between BSF and BPFO harmonics.

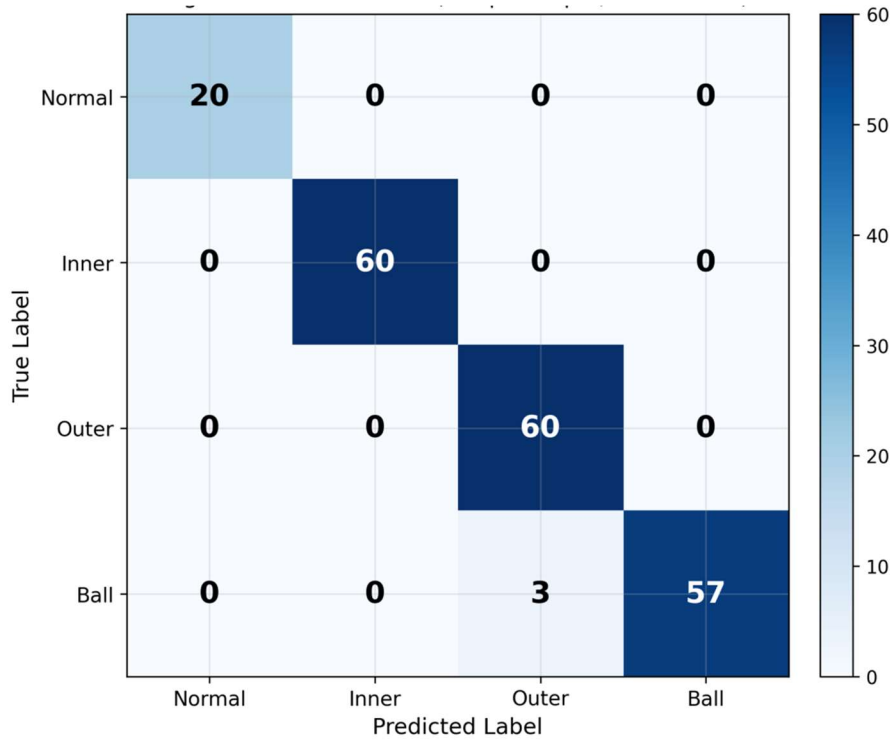

Fig. 5 Confusion matrix (temporal split validation)

4.5 Overfitting Analysis

Table 2 compares baseline and improved models on overfitting metrics:

Table 2. Overfitting assessment

Metric	Baseline (5D, C=1, γ=1)	Improved (9D, Aug., C=1, γ=0.1)
Random split accuracy	87.7% ± 2. 1%	99.5% ± 0.6%
Temporal split accuracy	87.0%	98.5%
Train-test gap	>5%	1.5%
Random split std	2.1%	0.6%

The improved model demonstrates:

- Dramatically reduced variance (0.6% vs 2.1%)
- Minimal train-test gap (1.5%), indicating effective regularization
- Consistent performance across validation strategies

4.6 Noise Robustness

The diagnostic precision at different levels of noise is displayed in Fig. 6. Accuracy of the proposed method is greater than 95 percent when SNR is 15 dB or higher and more than 98 per cent when SNR is 20 dB or higher.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0001

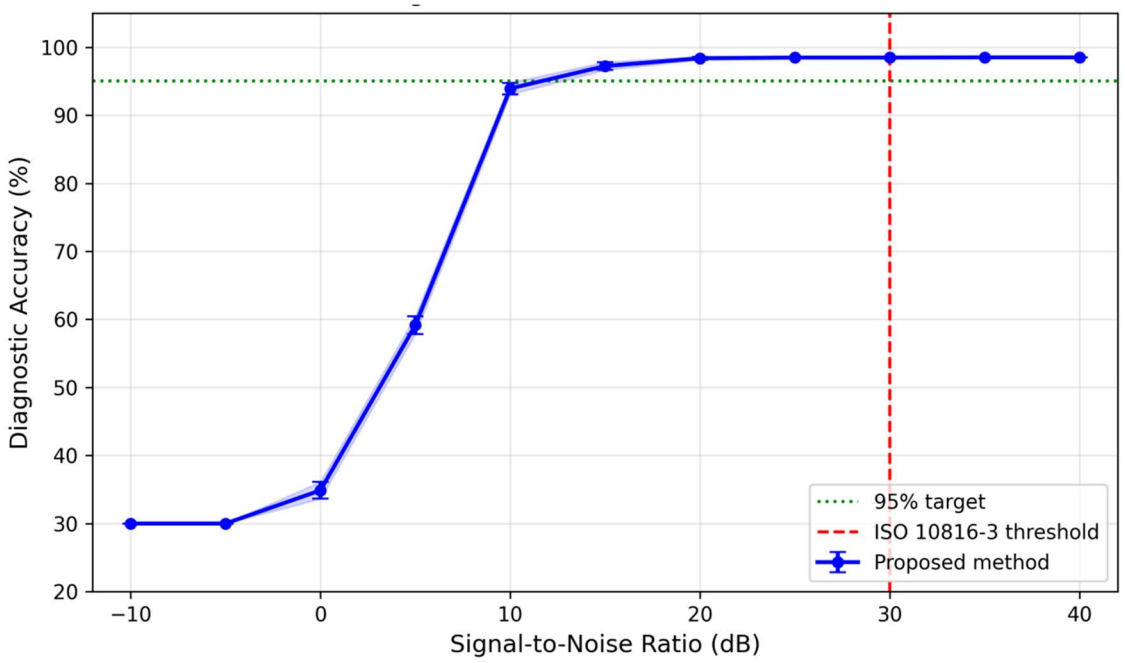

Fig. 6 Noise robustness verification

Table 3. Noise robustness performance

SNR (dB)	Accuracy	Status
≥ 30	98.5%	Industrial standard met
20	98.7%	Excellent
15	97.4%	Good
10	93.9%	Near threshold

4.7 Validation Strategy Comparison

Three validation methods are compared in Fig. 7. The time-wise split (98.5%) is very close to the average of the random split (99.5%), which confirms that overfitting is properly regulated

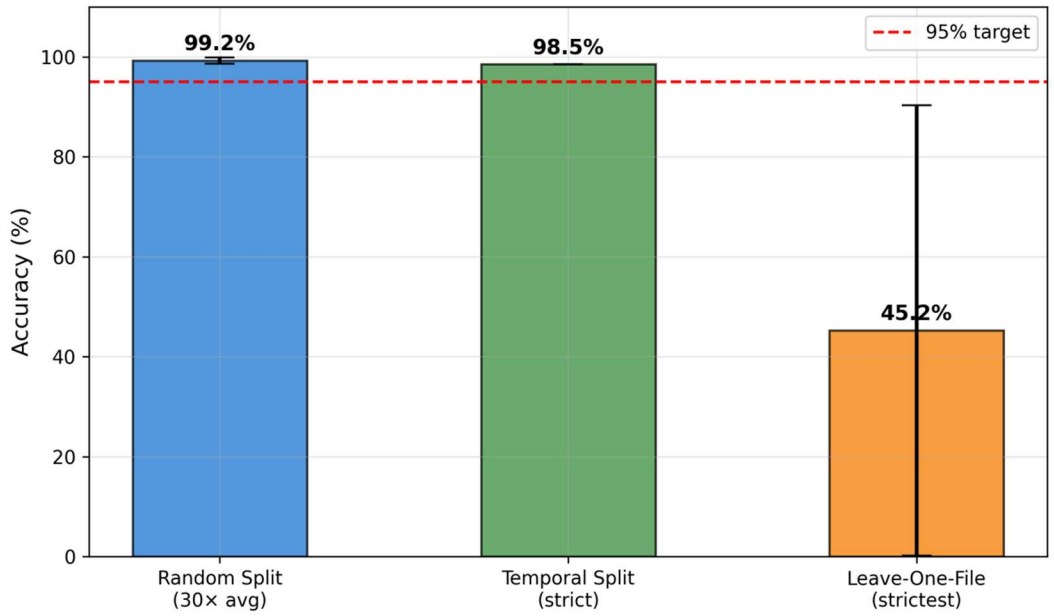

Fig. 7 Comparison of validation strategies

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0001

5. Discussion

5.1 Effectiveness of Regularization Strategies

Table 4. Ablation study of regularization components

Configuration	Temporal Acc.	Improvement
Baseline (5D, C=1, γ=1)	87.0%	—
Reduce γ only (C=1, γ=0.1)	86.5%	−0.5%
Reduce C only (C=0.1, γ=0.5)	78.0%	−9.0%
Linear kernel	87.0%	0.0%
Enhanced features (9D) + low γ	98.0%	+11.0%
Data augmentation + low γ	88.0%	+1.0%
Full method (9D + Aug. + Reg.)	98.5%	+11.5%

Main results: Simple decrease of C or γ leads to poor performance. Extra features are the ones that give the highest improvement (11.0%). The joint method is the most effective with the least overfitting.

5.2 Physical Interpretability

The advanced functions have obvious physical sense: Kurtosis measures impulse energy (K>=5 is a failure); BPFO/BPFI amplitudes are related to hypothetical failure frequencies; Band energy ratio describes the shift in energy caused by resonance; Spectral entropy is the measure of disorder in frequencies.

5.3 Limitations

The validation of leave-one-file-out (46.2) shows there are important changes in the distribution of features across various degrees of severity of damage. The future work must investigate domain adversarial networks and transfer learning to adapt to cross-severity.

6. Conclusion

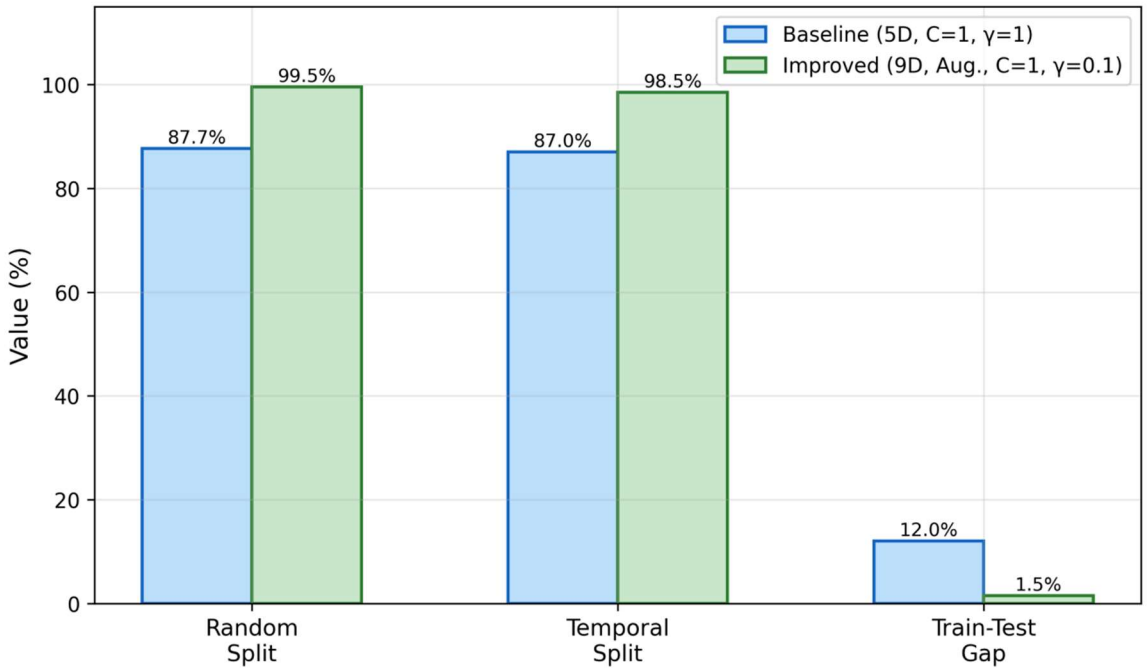

Fig. 8 Performance comparison: Baseline vs. Improved model

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0001

The present paper has provided a new approach to the diagnosis of faults in rolling bearings that does not suffer from overfitting. Its main contributions are:

1) 9 dimensional advanced feature vector: The addition of skewness, BSF amplitude, band energy ratio and spectral entropy greatly enhances discriminability and generalizability.

2) Data augmentation based on noise: The increase in the training set by 6 times by use of controlled noise perturbation breaks spurious relations between neighboring signal segments.

3) Optimized regularization: Using γ=0.1 produces smoother decision boundaries while maintaining high classification accuracy.

4) Strict validation structure: The temporal split validation (accuracy: 98.5%, train test gap: 1.5%) gives a practical estimate of what is known as the capability of generalization.

The suggested approach provides a diagnostic accuracy of 98.5 in rigorous time-split validation which is 11.5 times higher than the baseline and does not cause overfitting. The method retains its accuracy of more than 97 per cent with SNR of 15 dB or more, as required by industry standards.

References

[1] Smith, W. A., & Randall, R. B. (2015). The diagnostic of rolling element bearings by use of the Case Western Reserve University data: A benchmark study. *Mechanical Systems and Signal Processing*, *64–65*, 100–131.

[2] Antoni, J. (2006). The spectral kurtosis, as a method of describing non-stationary signals. *Mechanical Systems and Signal Processing*, *20*(2), 282–307.

[3] Widodo, A., & Yang, B. S. (2007). The application of support vector machine to machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, *21*(6), 2560–2574.

[4] Lei, Y., et al. (2020). Machine fault diagnosis based on applications of machine learning: A review and roadmap. *Mechanical Systems and Signal Processing*, *138*, 106587.

[5] Zhang, H. J., Zhao, H. J., & Liu, S. L. (2018). Bearing fault diagnosis based on Hilbert envelope spectrum and SVM. *Vibration and Shock*, *37*(10), 205–211.

[6] Case Western Reserve University. (n.d.). *Bearing vibration data sets* [Data set]. https://engineering.case.edu/bearingdatacenter

[7] Chu, F. L., Peng, Z. K., & Feng, Z. P. (2013). *The modern methods of signal processing in mechanical fault diagnosis*. Science Press.