The Fifth International Conference on Railway Technology
RAILWAYS 2022
22–25 August 2022
Montpellier, France

Proceedings of the Fifth International Conference on
Railway Technology:
Research, Development and Maintenance
Edited by J. Pombo
Civil-Comp Conferences, Volume 1, Paper 27.21
Civil-Comp Press, Edinburgh, United Kingdom, 2022, doi: 10.4203/ccc.1.27.21

Hybrid Feature Selection and Fault Identification of Train Bearings Based on Integrated Learning

Y.X. Li[1,2] S.C. Xie[1,2] R.D. Liu[1,2]

[1]**Key Laboratory of Traffic Safety on Track (Central South University), Ministry of Education, Central South University, Changsha, China**

[2]**Joint International Research Laboratory of Key Technology for Rail Traffic Safety, Central South University, Changsha, China**

Abstract

This Fast and accurate implementation of train bearing fault identification has been one of the key tasks of intelligent train health maintenance. In recent years, with the development of deep network technology, some bearing fault identification solutions based on deep learning have shown strong competitiveness. However, in the process of actual train application and maintenance, it has higher requirements for data volume and more complicated calculation. Therefore, an integrated learning-based bearing fault identification scheme is proposed. Overlapping sampling is performed considering the correlation before and after time series. Feature sets are constructed based on the characteristics of train fault signals and improved based on the XGBoost (Extreme Gradient Boosting) algorithm to achieve adaptive hybrid feature selection as well as fault identification. The effectiveness and superiority over the above method is verified by testing and evaluating two open-sources bearing datasets and laboratory bearing datasets.

Keywords: train bearing fault diagnosis, integrated learning, feature selection, XGBoost.

1 Introduction

As a key component of the travel section, the health conditioned monitoring and maintenance of train bearings have been the focus of research[1]. In the 21st century, along with the rise of deep learning, many sophisticated deep learning frameworks have been applied to time series regression and classification problems. These methods construct the feature engineering and prediction models after traditional fault identification with a large framework of feature self-selection based on networks, avoiding the subjectivity of the manual feature selection process and the dependence on historical experience of feature extraction in each domain[2,3].

However, in order to obtain deep learning models with high generalization ability and robustness, a large amount of data is needed for model training or expanding the sample size based on deep learning, such as GAN[5], domain adaptive[6], or semi-supervised learning[4]. However, among train bearing fault identification tasks, the number of negative samples and label acquisition are still more challenging. In practical engineering applications, complex networks often imply more demanding hardware facilities, which is more challenging for train health monitoring, especially for train in-transit detection.

Vibration signals are commonly used for fault diagnosis due to their better characterization capabilities. The characterization methods of its features are divided into two main categories physical features and virtual features. Virtual features are derived from deep learning feature extractors and may be better but less explanatory[7]. Physical features have some physical meaning [8], but a large number of physical features may cause a dimensional disaster, increase the training difficulty, and get an overfitting model.

Ensemble learning combines multiple learners to accomplish the task, and Boosting can upgrade weak learners to strong learners. XGBoost (Extreme Gradient Boosting)[7] is an improvement on GBDT that introduces the idea of regularization to reduce the complexity of decision trees and obtain a more optimal model.

In this paper, for the train bearing fault identification problem, we proposed to configure a stronger feature input processing structure combined with an integrated learning classification model based on the characteristics of the train bearing vibration signal[9], and apply it to the practical:

1. Pre-analysis of bearing vibration signals and selection of key features to form a feature set.
2. Based on XGBoost, scoring high-dimensional feature sets, dimension reduction, and identifying fault types.
3. Validation of the model using three data sets.

2 Methods

2.1 Overall Workflow

In this section, the proposed integrated learning based bearing fault identification method is presented in detail. As shown in Figure1, the proposed method consists of two main modules, the hybrid feature selection and the fault identification modules. The input to the feature selection module is a high-dimensional feature vector $\chi = \{x_i \mid x_i = (m_{(i,1)}, m_{(i,2)}, m_{(i,3)}, ..., m_{(i,\tau)})\}_{i=1}^{N}$ constructed based on time-frequency features, where N is the sample size and τ is the feature dimension of a single sample. In the classification module, the input features selection module uses the filtered low-dimensional features for the classification task.

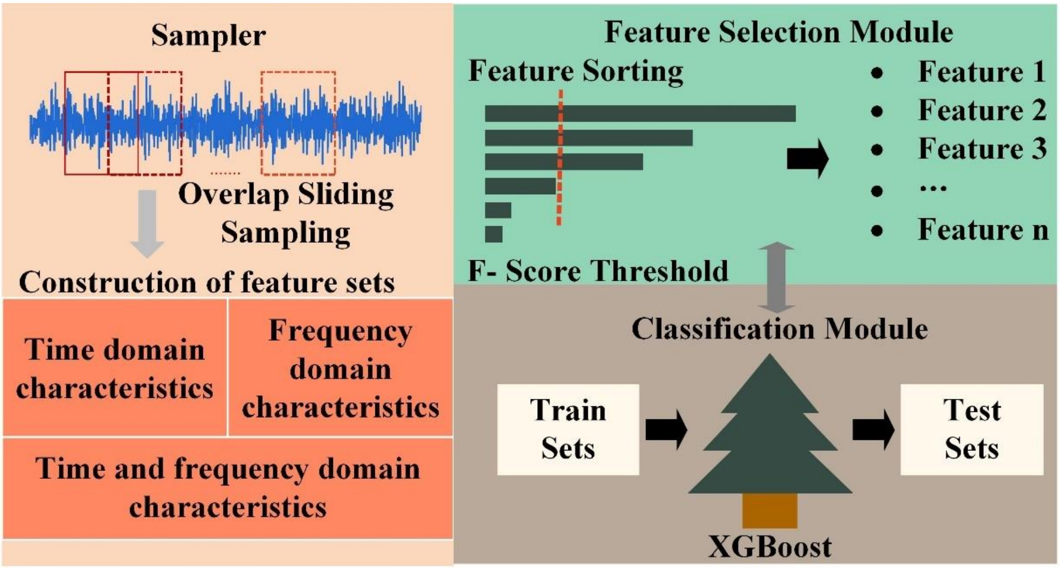

Figure 1: Schematic diagram of the technical approach.

2.2 Features Engineering

Feature set construction. The features are extracted in the time domain, frequency domain and time-frequency domain from the vibration signal. In the time domain, statistical features are mainly extracted, such as variance, margin factor. In the frequency domain, the Fourier transform is performed on the data to extract the center of gravity frequency, frequency variance, and other frequency domain features.

The spectrograms of each state were observed, and it was found that the amplitude of the bearings on different states differed greatly in the characteristic frequencies, and the amplitude features were extracted from the characteristic frequency range (Figure 2). Welch spectral estimation method is performed to analyse the extracted features, which combines two methods of adding window smoothing and averaging period to reduce the variance and make the spectrum smoother. A rectangular window function is selected to obtain the periodogram and then the Welch power spectral work is calculated and the energy amplitude of the feature range is extracted as the eigenvalue (Figure 3).

2.3 Integrated Learning-based Fault Classification

XGBoost[10] is an improved algorithm based on GBDT. XGBoost is used as a classifier to train K trees, and the $\Omega(f_k)$ is complexity of the first K trees.

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{1}$$

Where j is the leaf node, T the number of leaf nodes, ω_j is the node value of the leaf, γ and λ is the hyperparameter.

The objective function is：

$$G_i = \sum_{i \in I_j} g_i \tag{2}$$

$$H_i = \sum_{i \in I_j} h_i \tag{3}$$

$$obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i \in I_j} g_j)^2}{\sum_{i \in I_j} h_j + \lambda} + \gamma T = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_i^2}{H_j + \lambda} + \gamma T \tag{4}$$

g_i and h_i, respectively, denote the 1st and 2nd order derivatives of the loss function of the k-1th decision tree.

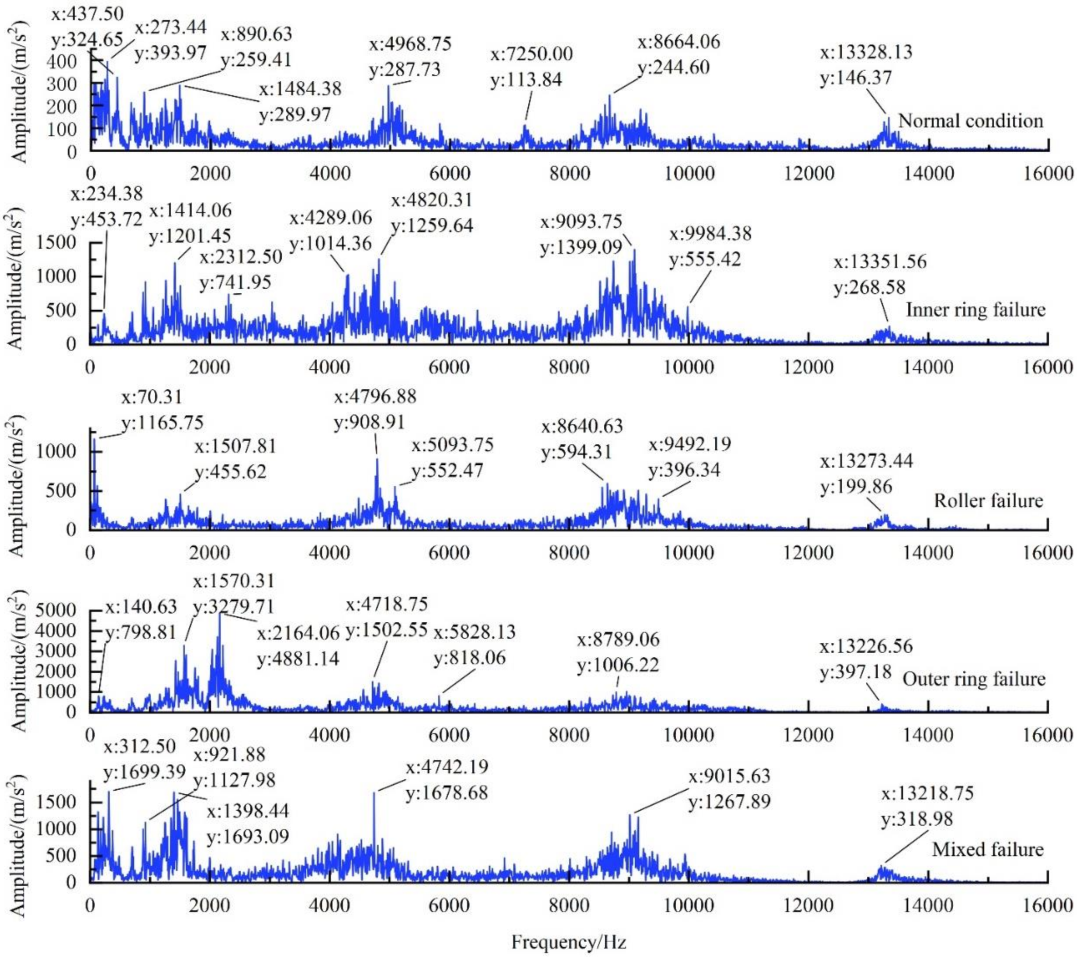

Figure 2: Frequency domain analysis based on FFT to extract the maximum amplitude of the feature frequency ranges as features（Spectrum of samples of CSU datasets）.

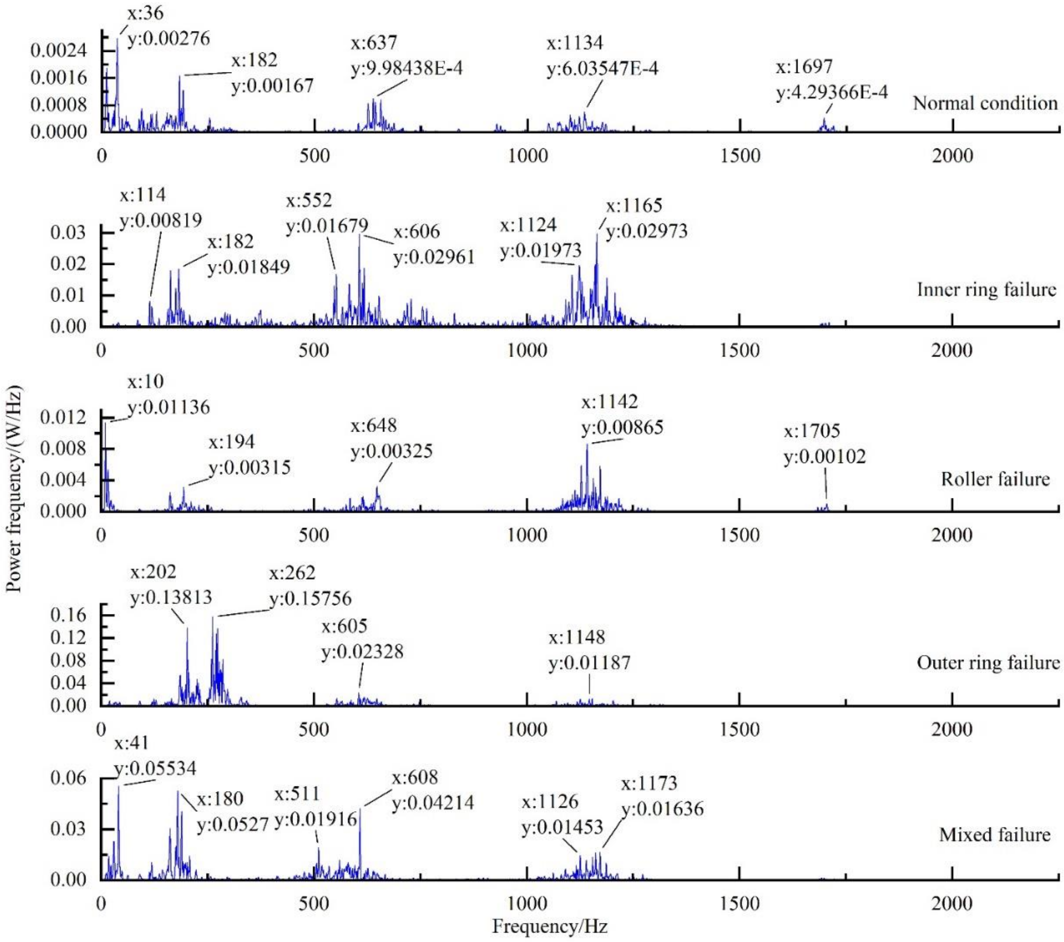

Figure 3: Feature extraction based on Welch's spectral estimation method to extract the maximum magnitude of the feature frequency range (CSU datasets sample power spectrum).

3 Results

This section tests and evaluates the above methods based on three datasets, and also compares the classical machine learning models GBDT, SVM and the more popular deep learning frameworks currently available.

3.1 Experimental Setup

Datasets. To evaluate the effectiveness of this method, bearing failure data from two public datasets CWRU and XJTU [11], and data from the Key Laboratory of the Ministry of Education for Rail Transit Safety of Central South University (CSU) were used for testing. The sample sizes and length of each dataset are shown in Table 1.

Dataset	Sample	Length	Class	Window Length
CWRU	640	3000	4	1024
XJTU	1920	1024	15	512
CSU	5440	4096	5	2048

Table 1: Statistics of Datasets

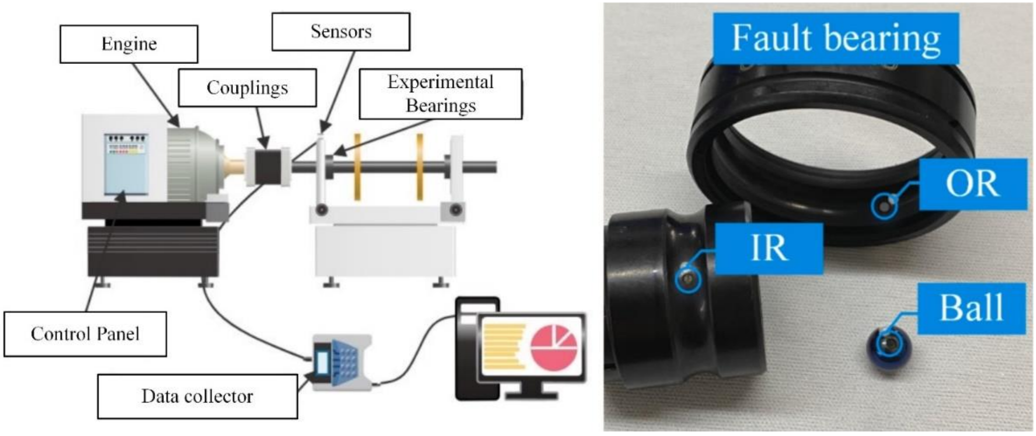

Figure 4: Test stand and bearing used for CSU data set (Bearing Type: normal condition, inner ring failure, outer ring failure, roller failure, mixed failure)

Feature Engineering. To save calculation time, sampling and Welch spectrum estimation sliding window lengths are shown in Table 1. The feature sets are constructed according to the method described in the previous chapter. Feature scoring is performed based on XGBoost, and an adaptive threshold picker is set. When the score of the next digit is lower than 1/3 of the previous digit, the value is set as the threshold.

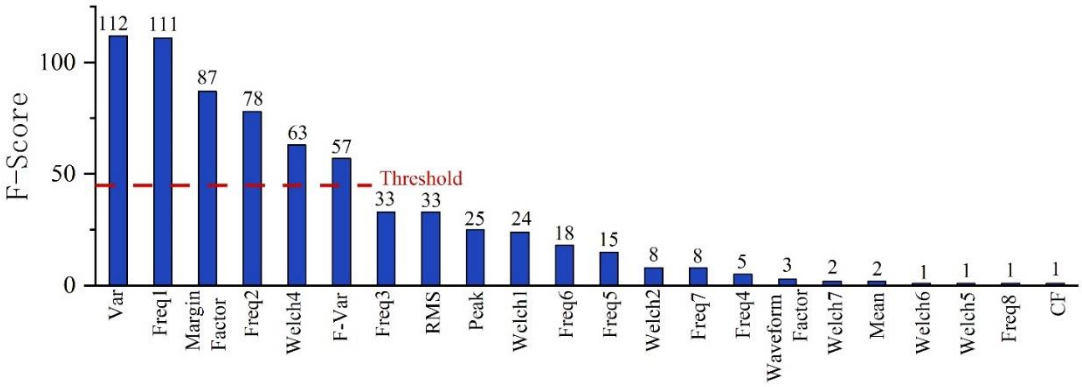

Figure 5: Feature score of one of the samples in the datasets CSU

3.2 Test results

In the classification test, each dataset was divided into training and test sets at 10%, 30%, 50%, 70%, and 90%. The classification accuracy and speed of SVM, GBDT, XGBoost (Featureless selection), XGBoost, MLP were compared for the classification task.

As shown in Table 1 and Table 2, the accuracy rate is used as the first evaluation criterion, and when the accuracy rate is consistent, the computation time was borrowed for ranking. It can be seen from Table 1 and Table 2 that the SVM’s training time is long, especially in the data set with large sample size MLP's performs better in classification tasks with less variety (CRWU,CSU), but the recognition accuracy is not high in XJTU. The accuracy of the classifier based on the XGBoost algorithm with the addition of feature selection is reduced compared to the previous one but the comprehensive speed-up effect is evaluated and the method is better than the former.

Test Ratio	Datasets	SVM	GBDT	XGBoost (Featureless selection)	XGBoost	MLP
10%	CWRU	100	100	100	100	100
30%		100	100	100	99.48	100
50%		96.88	100	100	99.69	100
70%		93.30	99.77	99.77	99.78	100
90%		93.92	23.78	23.78	99.83	100
10%	XJTU	8.97	90.63	95.31	95.83	29.69
30%		8.54	89.41	93.31	94.97	27.43
50%		8.85	88.54	93.85	91.35	25.83
70%		9.64	86.16	91.07	89.14	21.21
90%		0	82.18	75.64	76.85	24.77
10%	CSU	92.85	99.81	100	99.82	99.08
30%		98.46	99.93	100	99.94	98.47
50%		99.03	99.96	99.69	99.89	99.19
70%		98.46	99.92	99.89	99.84	98.47
90%		92.85	99.57	99.61	99.95	92.85

Table 2: Test accuracy on different datasets

Test Ratio	Datasets	SVM	GBDT	XGBoost (Featureless selection)	XGBoost	MLP
10%	CWRU	394.53	1.90	1.10	0.86	6.27
30%		245.68	1.75	0.99	0.82	6.75
50%		132.50	1.60	0.88	0.75	4.92
70%		53.76	1.34	0.74	0.69	3.24
90%		10.12	1.25	0.66	0.59	1.05
10%	XJTU	451.33	20.09	15.21	9.72	45.11
30%		341.01	17.35	12.21	7.87	34.03
50%		272.44	13.79	8.99	5.99	29.78
70%		1992.10	10.42	5.73	6.00	18.77
90%		92.69	6.18	2.97	2.52	14.01
10%	CSU	267.24	9.19	5.91	7.99	97.44
30%		2114.47	7.36	4.86	6.38	103.61
50%		5701.43	5.77	3.84	5.14	80.03
70%		2117.63	4.28	2.80	3.55	110.99
90%		261.26	1.89	1.82	2.10	102.03

Table 3: Time required for classification testing

4 Conclusions and Contributions

In this study, we proposed a hybrid feature selection and classification method for bearing fault identification based on integrated learning XGBoost. Sliding overlap sampling was performed considering time series before and after correlations. An adaptive selector was added to the feature set evaluation to achieve feature dimension reduction. The improved feature input combined with an integrated learning classifier based on XGBoost was evaluated on two public bearing datasets and a laboratory measured dataset, respectively. The method was compared with the classical machine learning method SVM, the integrated learning algorithm GBDT, and the deep learning algorithm MLP. The evaluation was carried out in two aspects: recognition accuracy and recognition time. The experimental results demonstrate that well-designed features engineering, constructing effective inputs, combined with integrated machine learning methods can outperform not only classical machine learning but sometimes deep learning networks as well.

In train bearing operation and maintenance tasks, effective networks based on deep learning are costly to implement due to environment, hardware facilities, positive sample size, etc. The method in this paper is applied to train bearing fault monitoring, which can improve the threshold algorithm and input structure based on train bearing characteristics and reduce the difficulty of the task. In the future, the method can also be embedded into a deep learning framework such as ResNET, CNN to build a fault monitor widely used in railroad lines.

References

[1] G. LI, P.LU. "Data-driven Intelligent Operation and Maintenance Technology of High Speed Railway Signals", Railway Transport and Economy, 43(10): 61-67. 2021

[2] Z. LIU, B. LU, et.al. "A Stacked Auto-Encoder Based Partial Adversarial Domain Adaptation Model for Intelligent Fault Diagnosis of Rotating Machines", *IEEE T IND INFORM*, **17**:6798-809, 2021.

[3] ROY S, DEY S, CHATTERJEE S. "Autocorrelation Aided Random Forest Classifier-Based Bearing Fault Detection Framework", *IEEE SENS J*,**20**:10792-800, 2020.

[4] H.Y. Fan, F.B. Zhang, R.D. Wang. et al. "Semi-supervised time series classification by tempopal relation prediction", ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021.

[5] KUMAR K, KUMAR R, de BOISSIERE T, et al. "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis". 2019.

[6] REN Y, LIU J, WANG Q, ZHANG H. "HSELL-Net: A Heterogeneous Sample Enhancement Network With Lifelong Learning Under Industrial Small Samples", *IEEE T CYBERNETICS*.:1-13,2020.

[7] ELSAYED S, THYSSENS D, RASHED A, JOMAA HS, SCHMIDT-THIEME L. "Do We Really Need Deep Learning Models for Time Series Forecasting" 2021.

[8] LIU F, GAO J, LIU H. "The Feature Extraction and Diagnosis of Rolling

Bearing Based on CEEMD and LDWPSO-PNN", IEEE Access., 8: 19810-19819,2020.
[9] K.H. Zhu. "Research on Vibration Signal based Rolling Element Bearing Feature Extraction and Fault Diagnosis Method", Da Lian University of Technology, 2013.
[10] CHEN T, GUESTRIN C. "XGBoost: A Scalable Tree Boosting System", ACM, 785-94, 2016.
[11] Y. Lei, F. Jia, D. Kong, et al. "Opportunities and Challenges of Machinery Intelligent Fault Diagnosis in Big Data Era", Journal of Mechanical Engineering, 54(05): 94-104, 2018.