World Journal of Engineering Research
ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3014

FAULT DIAGNOSIS METHOD OF AERO-ENGINE ROLLING BEARINGS BASED ON TIME-FREQUENCY ANALYSIS AND MACHINE LEARNING

YeQi Jin
College of Aerospace Engineering, Shenyang Aerospace University, Shenyang 110136, Liaoning, China.
Corresponding Email: 18624306466@163.com

Abstract: As a crucial component of aircraft, aero-engine bearings operate under extreme conditions such as high temperature, high pressure, and high rotational speed, making them highly prone to failure, which seriously affects aviation safety. Traditional bearing fault diagnosis methods suffer from problems such as low diagnostic accuracy and poor real-time performance, and it is difficult to meet the requirements of modern aviation industry for high reliability and safety of engines. With the development of machine learning technology, this paper proposes a fault diagnosis method for aero-engine bearings based on machine learning. Firstly, time-domain and frequency-domain features of vibration data are extracted, and dimensionality reduction processing is carried out through principal component analysis (PCA) to reduce data complexity and retain key information. Subsequently, machine learning models such as logistic regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and decision tree are used for fault prediction, and a comparative analysis is conducted with deep learning models. The experimental results show that the Support Vector Machine (SVM) performs best in the fault classification task, with an accuracy rate of 99%. This research provides an efficient and accurate solution for aero-engine bearing fault diagnosis and has important practical application value.
Keywords: Aero-engine bearings; Fault diagnosis; Machine learning; PCA; SVM; Time-domain and frequency-domain features

1 INTRODUCTION

As the core component of an aircraft, the reliability and safety of an aero-engine's operation are directly related to the normal operation of air transportation and the safety of passengers[1].Aero-engine bearings are crucial and indispensable parts in the engine. They continuously operate under extreme working conditions such as high temperature, high pressure, and high rotational speed, bear complex and variable loads, and are prone to failure, which can cause serious accidents. Therefore, in-depth research on the fault diagnosis of aero-engine inter-shaft bearings, and using advanced technical means to detect bearing fault hidden dangers in a timely and accurate manner, can not only ensure the safe and stable operation of aero-engines, but also provide strong support for the design optimization and life prediction of aero-engines. It is of great significance for promoting the high-quality development of the aviation industry[2].

At present, traditional bearing fault diagnosis methods often have problems such as low diagnostic accuracy, poor real-time performance, and difficulty in effectively identifying early-stage faults when dealing with aero-engine bearing faults. They cannot meet the strict requirements of modern aviation for the high reliability and safety of engines[3]. With the development of emerging technologies such as machine learning, new opportunities have emerged for aero-engine bearing fault diagnosis[4]. Zhang Jian and Qian Haiting[5] used three common classifiers, namely Support Vector Machine (SVM), decision tree, and random forest, to classify and learn bearing vibration data, and evaluated their performance in bearing vibration data classification. Cheng Xiang[6], in order to solve problems such as complex background noise of bearings in industrial environments, small amounts of fault data acquisition, and difficulty in detailed analysis of fault states, adopted a signal denoising algorithm combined with a machine learning algorithm to monitor bearing vibration signals for faults. Cai Zhengyin[7] conducted research around traditional signal processing and machine learning, proposed an Adaptive Variational Mode Decomposition(IVMD) method and an improved scheme combining multiple technologies to solve problems in the application of traditional signal processing and Support Vector Machine (SVM), and verified the effectiveness of the method through multiple datasets.

In summary, this paper proposes a bearing fault diagnosis method that combines time-domain and frequency-domain feature extraction, and reduces the dimensionality of feature data through Principal Component Analysis (PCA) to retain key information to the greatest extent while reducing the data dimension. This method can not only effectively handle the complexity of high-dimensional data, but also improve the computational efficiency and prediction performance of subsequent fault diagnosis models. Secondly, this paper uses traditional machine learning algorithms (such as logistic regression, KNN, SVM) to explore the applicability, advantages, and disadvantages of different models in aero-engine bearing fault diagnosis, providing a reference for future fault diagnosis technologies. Finally, for the optimization of model hyperparameters, this paper adopts a combination of grid search and 5-fold cross-validation to finely adjust hyperparameters, significantly improving the classification accuracy and demonstrating the importance of parameter adjustment for model performance. Through these innovations, this paper not only provides an efficient and

accurate solution for aero-engine bearing fault diagnosis, but also provides theoretical support and practical guidance for future research in related fields.

2 PRINCIPLE AND MODEL BUILDING

2.1 Time-domain Feature and Frequency-domain Feature Extraction

When differentiating the fault types of aero-engine bearings, time-domain features and frequency-domain features are usually combined as the fault evaluation criteria. The time-domain features include mean value, variance, peak value, root mean square (RMS) value, root amplitude, margin, kurtosis index, waveform factor, impulse value, and peak factor. The frequency-domain features, on the other hand, consist of mean frequency, centroid frequency, root mean square frequency, standard deviation frequency, and kurtosis frequency. Through a comprehensive analysis of these features, the fault types can be judged more accurately[8].

2.2 RobustScaler Outlier Handling

2.2.1 Algorithm features, advantages, and application scenarios

RobustScaler is a data standardization method based on statistical characteristics, mainly used to handle datasets containing outliers. Its main feature lies in its robustness to outliers. Different from traditional standardization methods (such as Z-score standardization), RobustScaler uses the median instead of the mean as the central value and the interquartile range (IQR) instead of the standard deviation as the scaling scale. Moreover, it does not change the overall distribution shape of the data during the scaling process, making it particularly suitable for non-normally distributed data. RobustScaler is applicable to various data types and makes few assumptions about the data distribution. The advantages of RobustScaler are mainly reflected in its robustness to outliers and wide applicability. Due to the use of the median and IQR, RobustScaler is insensitive to extreme values and can effectively reduce the impact of outliers on data scaling. During the scaling process, RobustScaler can preserve the original structure of the data, making it suitable for machine-learning tasks that require maintaining data characteristics, such as clustering and classification. It is one of the important tools in data pre-processing.

2.2.2 Calculation formula

Median: The median is the middle value after the data is sorted. For a datasets with an odd number of data points, the median is the middle value. For a datasets with an even number of data points, the median is the average of the twomiddle values. Its mathematical definition is as follows:

$$Median(X) = \begin{cases} X_{\frac{n+1}{2}} & if\ n\ is\ odd \\ \dfrac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} & if\ n\ is\ even \end{cases} \tag{1}$$

Inter-Quartile Range (IQR): The IQR is the difference between the 75th percentile $Q_3(X)$ and the 25th percentile $Q_1(X)$, which is used to measure the degree of data dispersion. Its mathematical definition is:

$$IQR(X) = Q_3(X) - Q_1(X) \tag{2}$$

Given a datasets $X = \{x_1, x_2, ..., x_n\}$ where x_i is a feature vector, the standardization process of RobustScaler can be divided into the following two steps.

(1) Centering:

Center the data using the median. For each feature , calculate its median $Median(X_j)$, And subtract the median from the data：

$$x_{ij}^{centered} = x_{ij} - Median(X_j) \tag{3}$$

Here, x_{ij} represents the j-th feature value of the i-th sample.

(2) Scaling:

Scale the data using the Inter-Quartile Range (IQR). For each feature j, calculate its IQR:

$$IQR(X_j) = Q_3(X_j) - Q_1(X_j) \tag{4}$$

where $Q_3(X_j)$ and $Q_1(X_j)$ are the 75th percentile and the 25th percentile of the j-th feature respectively. Then divide the centered data by the IQR.

$$x_{ij}^{scaled} = \frac{x_{ij}^{centered}}{IQR(X_j)} \tag{5}$$

Finally, the standardized data x^{scaled} can be expressed as:

$$x^{scaled} = \frac{X - Median(X)}{IQR(X)} \tag{6}$$

2.3 KNN

The K-Nearest Neighbors (KNN) algorithm is an instance-based non-parametric classification algorithm. Its core idea is to calculate the distances between the sample to be classified and the known samples, find the K closest neighbors, and then determine the class of the sample to be classified based on the classes of these neighbors. The problem of engine fault classification involves analyzing the sensor signals collected under various working conditions and determining the fault type according to their characteristic patterns.

Given a training sample set $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, where x_i represents the feature vector , and $y_i \in \{1,2,\cdots,C\}$ represents the sample class. For a sample to be classified x, the steps of KNN are as follows:

1. Distance calculation

Calculate the distances between x (the sample to be classified) and all the samples in the training set. The most commonly used distance metric is the Euclidean distance, and its formula is as follows:

$$d(x,x_i)=\sqrt{\sum_{j=1}^{m}(x_j-x_{ij})^2} \tag{7}$$

Where m is the number of feature dimensions, and x_j and x_{ij} are the feature values of the sample x and the x_i sample in the training set in the j-th dimension respectively.

2. Select the nearest neighbor samples

Sort the samples in ascending order of distance and select the K samples with the shortest distances $S \subseteq D$.

3. Classification decision

Count the number of samples of each class among the S. Adopt the principle of "the minority is subordinate to the majority" , and predict the class of the sample x to be the class that appears most frequently:

$$\hat{y}=arg\max_{c\in\{1,\ldots,C\}}\sum_{x_i\in S}\mathbb{I}(y_i=c) \tag{8}$$

Where $\mathbb{I}(\cdot)$ is an indicator function. If the condition holds, its value is 1; otherwise, it is 0.

2.4 SVM

The core idea of SVM is to find a classification hyperplane that maximizes the margin between different classes. For linearly separable data, the classification problem can be expressed as the following optimization problem:

$$\min_{\omega,b}\ \frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i \tag{9}$$

Constraints:

$$y_i(\omega\cdot\phi(x_i)+b)\geq 1-\xi_i,\ \xi_i\geq 0,\ i=1,\ldots,n \tag{10}$$

Among them, ω and b are the hyperplane parameter, ξ_i is the slack variable, C is the penalty coefficient,and $\phi(x_i)$ represents the kernel function mapping.

To deal with the non-linear distribution of fault signals, the Gaussian kernel function can be used here:

$$K(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{11}$$

Among them, σ controls the bandwidth of the Gaussian kernel function.

2.5 Decision Tree

A decision tree constructs a classification model by recursively splitting the sample space. Each split is based on a certain feature and its threshold to maximize the class purity of the samples after the split. The specific process of the algorithm is as follows:

1. Splitting criterion

At each node, the feature and its threshold that maximize the Information Gain or Gini Index are selected for splitting. The Information Gain is defined as follows:

$$IG(D,A)=H(D)-\sum_{v\in Values(A)}\frac{|D_v|}{|D|}H(D_v) \tag{12}$$

Among them, $H(D)$ is the information entropy of the node datasets D, A is the splitting feature, and D_v is the subset where the feature A takes the value v.

The Gini Index is defined as follows:

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2 \tag{13}$$

Among them, p_k represents the proportion of samples in class k.

2. Stopping condition

The splitting stops when all samples belong to the same class, the number of features is insufficient, or the number of samples at a node is lower than the preset threshold.

3. Prediction stage

When inputting features x, start from the root node and select a path layer by layer downward according to the feature values. Finally, reach a leaf node and output the predicted class.

3 RESULTS

3.1 Data Sources

The datasets selected in this paper is from the Bearing Data Center of Case Western Reserve University (CWRU). This datasets comprehensively records four main types of faults: inner-race faults, outer-race faults, rolling-element faults, and normal operating states. Some illustrations are shown in Figure 1. For each type of fault, the datasets provides samples with four different fault diameters (0.007 inches, 0.014 inches, 0.021 inches, and 0.028 inches respectively). All the data were collected at a sampling rate of 12 kHz.

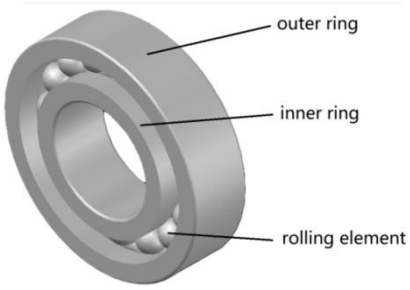


Figure 1 Bearing Fault Illustration

After the data collection work was completed, this paper used a variety of Python libraries to conduct a preliminary visual analysis of the obtained raw data. Due to the limitation of the article's length, only the visual graphs of partial data are presented here, as shown in Figure 2.

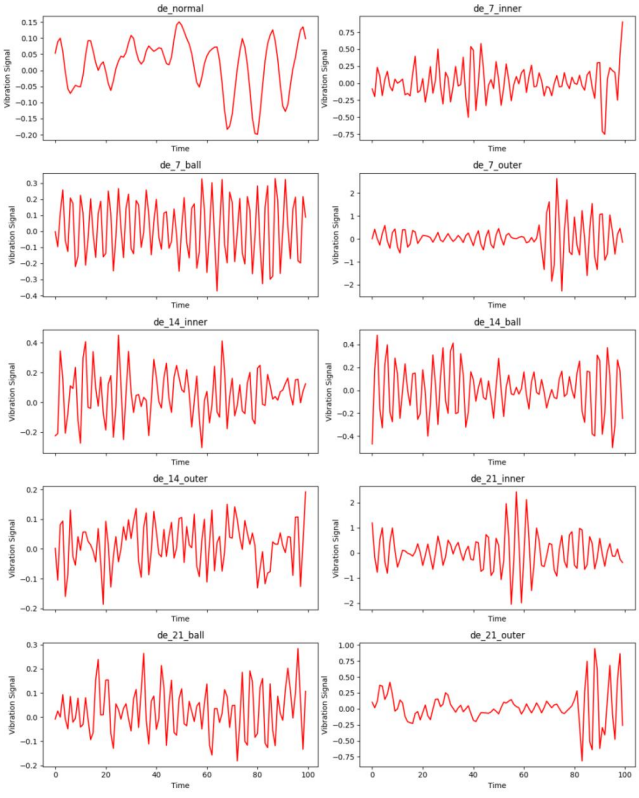


Figure 2 Part of Working Condition Data Visualization

Observe the curves of the four different types of vibration signals changing over time shown in the Figure 2. The amplitudes and frequencies of these signals both exhibit certain fluctuations, and at the same time, they also show a certain degree of periodicity and repeatability.

To optimize the processing and analysis of the original bearing vibration data, this paper divides every 1024 data points into a sample block for subsequent operations. This length approximately represents the time interval for the bearing to rotate three times. It can capture most of the key vibration information and effectively control the scale of data processing,as shown in Figure 3.

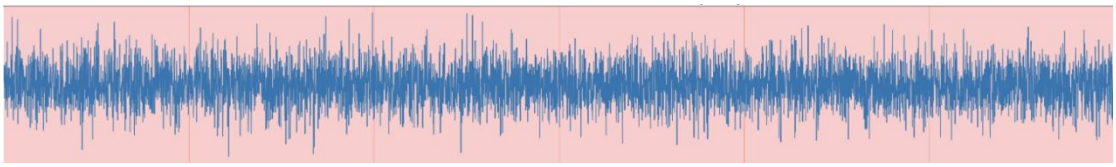


Figure 3 Schematic diagram of sliding window segmentation

Subsequently, to visually display the quantitative relationship among samples of different fault categories, this paper uses Figure 4 for visualization.

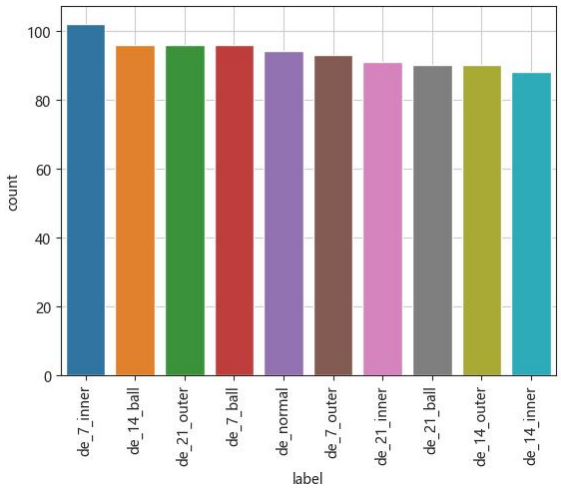


Figure 4 Number of Samples of Different Fault Categories

By observing Figure 4, it can be found that the quantity distribution of samples for each type of fault is relatively balanced. This indicates that the datasets is relatively evenly distributed among different classes, and there is no obvious imbalance.

3.2 Data Preprocessing

3.2.1 Time-domain and frequency-domain characteristics

To evaluate whether there are outliers in the time-domain and frequency-domain features (such as mean, standard deviation, and mean frequency) extracted from the original datasets, this paper uses box plots for intuitive detection. The box plots corresponding to some features are shown in Figure 5.

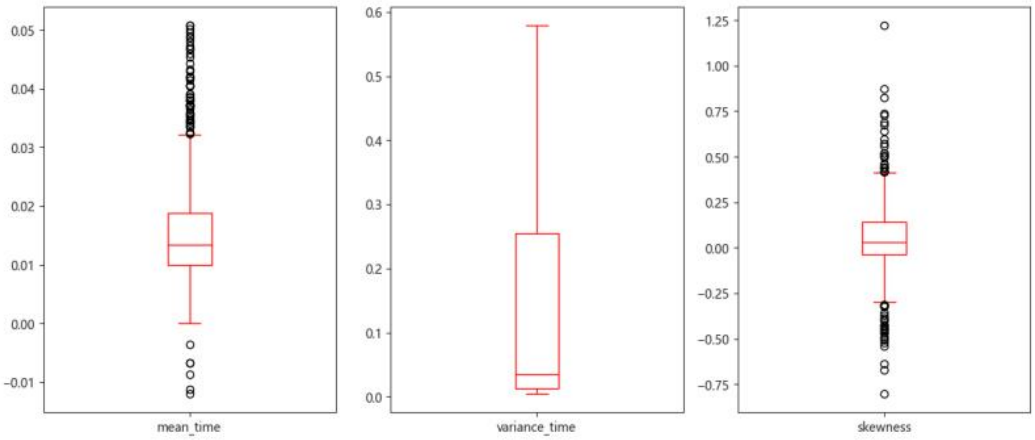


Figure 5 Box View of Some Features

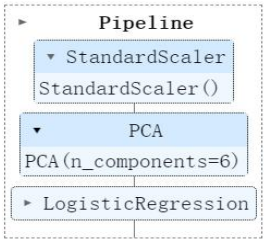
When conducting an in-depth analysis of the data distribution presented in Figure 5, it is noted that there are significant outliers in some variables. For example, in the features of "skewness" and "kurtosis", the number of outliers is particularly prominent, and their degree of deviation far exceeds the average level of other variables. The presence of such outliers may have an adverse impact on the subsequent data analysis and modeling processes, leading to inaccurate or biased results.

Therefore, to ensure the reliability and accuracy of the data analysis results, this paper uses the RobustScaler method to handle these outliers.

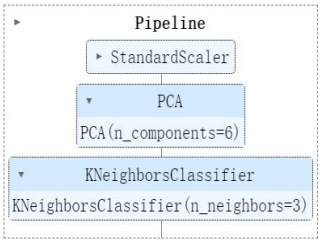
3.3 Prediction Results of the Aero-engine Bearing Fault Classification Model Based on Machine Learning

3.3.1 Model Structure

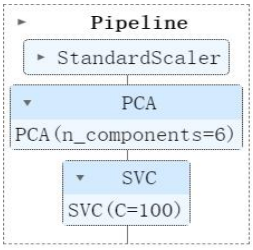
When using machine learning-related algorithms, the training set input consists of the extracted time-domain and frequency-domain features. The PCA method is adopted to reduce the dimensionality of the data. During model training, algorithms such as logistic regression, KNN, SVM, and decision tree are used to train the model respectively. The pipeline is employed to integrate the models, and the model structure obtained is shown in Figure 6.



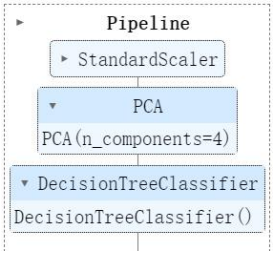
(a) Logistic Regression



(b) K-Nearest Neighbor



(c) Support Vector Machines

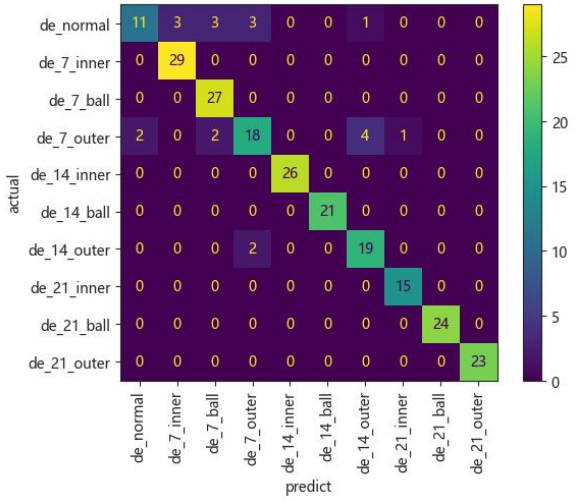


(d) Decision Tree

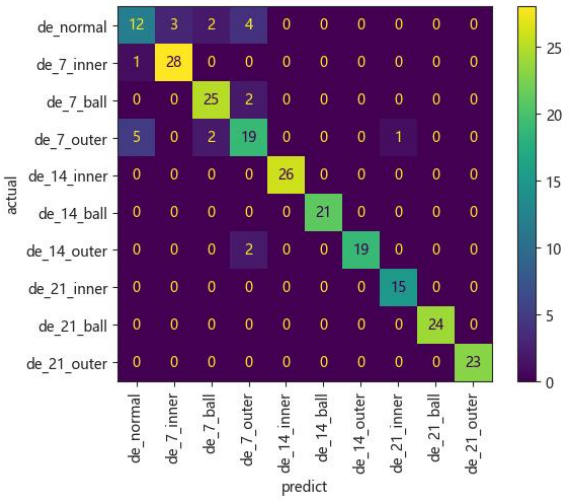
Figure 6 Structure of Each Machine Learning Models

3.3.2 Model prediction results

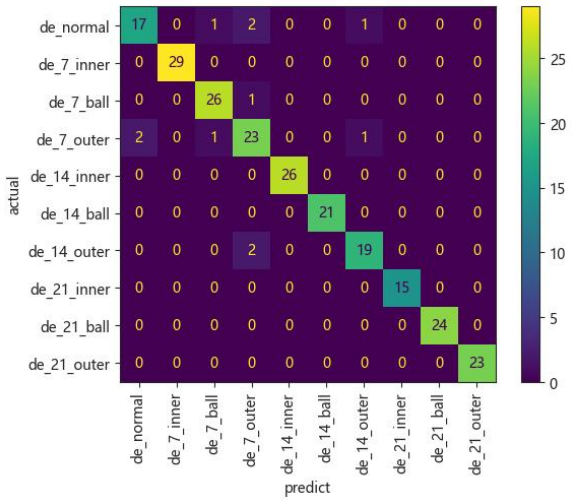
Bring the well-trained model into the test set, and the resulting confusion matrix is shown in Figure 7. In addition, the corresponding accuracy, precision, recall, and F1-score of the model are shown in Table 1.



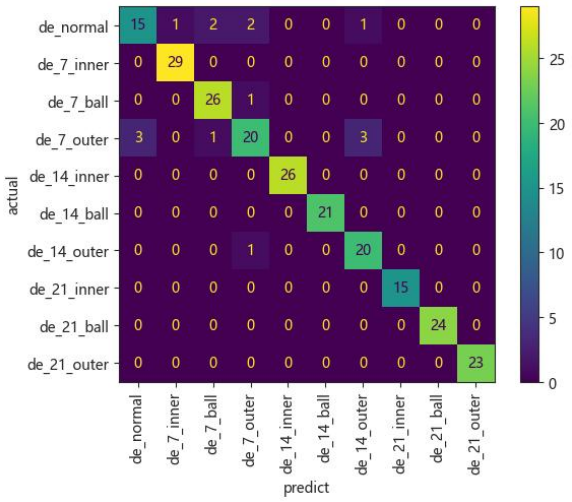
(a) Logistic Regression



(b) K-Nearest Neighbor



(c) Support Vector Machines



(d) Decision Tree

Figure 7 Confusion Matrix for Machine Learning Models

Table 1 Performance Comparison of Classification Algorithms

Model	Accuracy	Precision	Recall	F1-Score
LR	0.95	0.95	0.95	0.95
KNN	0.96	0.96	0.96	0.96
SVM	0.99	0.99	0.99	0.99
DT	0.96	0.96	0.96	0.96

After comparison, it's not difficult to find that in the current situation, the Support Vector Machine(SVM) algorithm has the best overall performance.

4 CONCLUSIONS

This study deeply analyzes the bearing vibration data and reveals the effectiveness and universality of machine learning algorithms such as logistic regression, KNN,and SVM for fault diagnosis.The PCA method is adopted for dimensionality reduction, which not only retains the main variation information of the data but also improves the computational efficiency and prediction performance of the model. The hyperparameters are optimized through grid search and 5-fold cross-validation, which proves the effectiveness of this method. Moreover, the optimal hyperparameter combination can significantly improve the classification accuracy.

In practical applications, the research results of this study have wide applicability in the fields of industrial equipment fault diagnosis and predictive maintenance. They can reduce costs and downtime while enhancing the reliability and safety of equipment operation.

Meanwhile, traditional machine learning methods have problems such as high computational costs and difficulty in capturing non-linear relationships when dealing with large-scale high-dimensional data. It is crucial to introduce deep-learning methods in the future. Models like CNN, RNN, and LSTM have strong automatic feature extraction capabilities. Autoencoders, combined with various techniques, can improve the model performance to achieve more efficient and intelligent industrial production.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Li Dongjun, Li Ya, Li Dongwen, et al. Remaining useful life prediction of aero-engines based on MIC feature extraction and BO-CatBoost. Journal of Air Force Engineering University, 2024, 25(01): 31-38.

[2] Sha Yundong, Zhao Junhao, Luan Xiaochi, et al. A fault identification method for main bearings based on multi - feature parameter fusion and dimensionality reduction. Journal of Shenyang Aerospace University, 2024, 41(05): 15-25.

[3] Yin Kang. Research on Fault Diagnosis Algorithms for Aero - engine Bearings and Design of Fault Diagnosis Software. Supervisor: Zhang Weitao; Sun Xiaochao. Xidian University, 2021: 17-34.

[4] Liu Junli. Research on Fault Diagnosis and Remaining Useful Life Prediction Methods for Rolling Bearings Based on Machine Learning. Southwest Jiaotong University, 2023: 10-23.

[5] Zhang Jian, Qian Haiting. Classification and Performance Comparative Analysis of Bearing Vibration Data Based on Machine Learning. China Tire Resources Comprehensive Utilization, 2025(01): 163-165.
[6] Cheng Xiang. Research on Fault Diagnosis of Rolling Bearings Based on Machine Learning. Anhui University of Science and Technology, 2024: 10-17.
[7] Cai Zhengyin. Research on Fault Diagnosis of Rolling Bearings Based on Vibration Signal Processing and Machine Learning. Heilongjiang University, 2024: 10-16.
[8] Zhang Weitong. Research on fault diagnosis of aero engine bearing based on unbalance dataset. Harbin Engineering University,2023: 10-19.