Bearing fault diagnosis method based on CNN-LightGBM

Xin Jia, Boping Xiao*, Zhanjuan Zhao, Lin Ma and Naichao Wang

1Beihang University, Beijing, China

E-mail: jiaxin_tec@buaa.edu.cn

Abstract. In recent years, deep neural networks have been widely used in bearing fault diagnosis. Aiming at the problem that the traditional convolutional neural network (CNN) has insufficient generalization ability and low accuracy in rolling bearing fault diagnosis, this paper presents a new fault diagnosis model based on CNN and light gradient boosting machine (LightGBM). By using the original vibration signal directly, the new model firstly uses the deep convolutional neural network with small kernel to extract features and introduces the batch normalization and the Adam algorithm to improve the rolling bearing fault diagnosis model convergence speed and generalization ability. Then, combined with the efficient and accurate characters of LightGBM in classification prediction, the extracted features are imported into LightGBM for training to complete bearing fault diagnosis of different fault types. This new model can realize the fault diagnosis of rolling bearing directly from end to end without extracting the fault feature of rolling bearing vibration signal by hand. The results of the comparative experiments on the Case Western Reserve University (CWRU) public bearing dataset show that the new model improves the accuracy compared to just using CNN. The experimental results also confirm that the proposed method has feature learning ability and has good ability for fault diagnosis.

1. Introduction

As one of the important parts of the mechanical rotation system, about 45% of the causes of mechanical equipment failure are caused by bearing failure[1], so it is very important to find the faults in time and effectively. At present, the methods of bearing fault diagnosis can be roughly divided into three categories: fault diagnosis based on analytical model, fault diagnosis based on signal processing and fault diagnosis based on artificial intelligence[2].

The earliest developed and most systematic fault diagnosis method is based on analytical model. Rafsanjani et al. (2009) established a nonlinear dynamic analysis model of rolling bearing system considering surface defects. According to the influence of nonlinear Hertz contact deformation and inner diameter clearance, considering the contact force of each rolling unit, the mathematical expressions of inner ring, outer ring and local defects of rolling piece are derived[3]. Kankar et al. (2009) uses the simplified method and the corresponding integral technology to predict the dynamic response of the rotor bearing system, which improves the numerical stability of the system[4].

The method based on signal processing usually uses statistical analysis to extract the time-domain features of vibration signals, wavelet transform, Fourier analysis and other methods to collect the frequency-domain features of vibration signals, then reduces the dimension of the features, and finally inputs the filtered features into the classifier to complete fault diagnosis. Misra et al. (2002) use the common principal component (PCA) analysis for feature dimensionality reduction[5]. Pandya et al.
(2013) use k-nearest neighbor (KNN)\cite{6} as classifier and Santos et al. (2015) use support vector machine (SVM)\cite{7}.

In recent years, the research of fault diagnosis method based on artificial intelligence is more and more thorough and extensive. Convolutional neural network has excellent performance in image recognition, which can take 2D spectrum image as input. Guo et al. (2018) use continuous wavelet transform to decompose vibration signal into continuous wavelet transform of different scale range according to rotation speed, obtain 2D image as input, and use convolutional neural network (CNN) to diagnose the fault of variable speed bearing\cite{8}. Convolutional neural network can also take 1D data as input directly. Zhang et al. (2017) uses 1-D original vibration signal as input and convolution neural network for bearing fault diagnosis, which realized end-to-end fault diagnosis\cite{9}. Convolutional neural network has the ability to extract deep-seated features, and some researchers use CNN and other methods to improve the diagnostic accuracy. Gu et al. (2017) extracts features based on empirical mode decomposition (EDM) and combine them with convolutional neural network for fault diagnosis\cite{10}.

Although many of the above works have achieved good results in the field of fault diagnosis, there is still a lot of room for improvement. It is difficult to obtain a system model, and because of the errors, disturbances and noises, the robustness problem gradually highlights in the meantime. The method based on signal processing needs complex signal processing and manual feature extraction, and the process is complex and the model robustness is weak. The accuracy of single CNN fault diagnosis needs to be improved. To solve the problems above, we proposed a fault diagnosis method named Convolution Neural Networks with Light Gradient Boosting Machine (CNN-LightGBM). The contributions of this paper are summarized below:

- The proposed method does not need to extract complex time-domain and frequency-domain signals, and uses the original signal as input to realize end-to-end fault diagnosis.
- This algorithm uses adaptive feature extraction, which can extract effective features, and uses an optical gradient intensifier as a classifier, which improves the accuracy and robustness of fault diagnosis.
- This algorithm can identify a variety of fault states and diagnose the fault location and damage diameter timely and effectively.

The remainder of this paper is organized as follows: The convolutional neural network (CNN) and light gradient boosting machine (LightGBM) are discussed in Section 2. The proposed CNN-LightGBM is specified in Section 3. After this, experiments and analysis on CWRU bearing faults dataset are presented in Section 4. We draw the conclusions and present the future work in Section 5.

2. Basic Theory

In this section, the architecture of CNN and LightGBM model are introduced.

2.1. A brief introduction of CNN
Convolution neural network is a kind of deep feedforward neural network with multiple hidden layers. Through the feature transformation layer by layer, the bottom features are transformed into the top features to realize the feature learning and expression. A typical CNN network consists of input layer, hidden layer, full connection layer and output layer. The hidden layer consists of a series of convolution layer and pooling layer. The output layer often uses softmax as classifier.

2.1.1. Convolutional layer
The role of convolution layer is to filter the features of input data, extract the feature space of input, and capture different visual patterns through multiple convolution kernels. It has the feature of weight sharing so as to reduce network parameters and memory consumption. Over fitting by too many parameters can be avoided. The convolution layer formula is described as follows:
\[ y^{l(i,j)} = K_i^j \ast x^{(r')} = \sum_{j=0}^{w-1} K_i^{j(f)} x^{(j+f)} \]  

(1)

where \( K_i^{j(f)} \) denotes \( j \)-th weight of the \( i \)-th convolution kernel in the \( l \)-th layer, \( x^{(r')} \) is the \( j \)-th convolved local area of the \( l \)-th layer, \( W \) is the width of the pooling region.

### 2.1.2. Pooling layer

The pooling layer, also known as the lower sampling layer, is to scale or reconstruct the sample size through the lower sampling of the image, to further highlight the extracted features while reducing the dimension of the data, so as to prepare for the next more refined features. The common pooling methods are max pooling and average pooling. We plan to use the max pooling as follows:

\[ p^{l(i,j)} = \max_{(j-1)W+t \in [jW]} \{ a^{l(i,j)} \} \]  

(2)

where \( a^{l(i,j)} \) denotes the value of \( t \)-th neuron in the \( i \)-th frame of layer \( l \), \( t \in [(j-1)W+1, jW] \). \( W \) is the width of the pooling region, and \( p^{l(i,j)} \) denotes the corresponding value of the neuron in layer \( l \) of the pooling operation.

The advantage of using the maximum pooling layer is that it can obtain position independent feature information, especially the vibration signal is periodic data.

### 2.1.3. Full connection layer

The full connection layer plays the role of classifier in the whole CNN. The features filtered by the previous convolution layer and pooling layer are flattened to form a one-dimensional feature vector, and then the input and output are fully connected. Eq. (3) represents the forward propagation formula of full connection layer:

\[ z^{l+1(j)} = \sum_{i=1}^{n} W_{ij} a^{l(i)} + b_j \]  

(3)

where \( W_{ij} \) represents the weight between the \( i \)-th neuron in the \( l \)-th layer and the \( j \)-th neuron in the \( l+1 \)-th layer, \( z^{l+1(j)} \) is the logits value of the \( i \)-th neuron in the \( l+1 \)-th layer, \( b_j \) is the bias term of all neurons in the \( l \)-th layer to the \( j \)-th neurons in \( l+1 \)-th layer.

### 2.2. A brief introduction of LightGBM

LightGBM is a gradient enhancement algorithm framework based on decision tree algorithms that Microsoft DMTK (Distributed Machine Learning Toolkit) team opened source on GitHub in 2016. It has the advantages as following: faster training efficiency, lower memory usage, higher accuracy, support for parallel and GPU computing. In addition, it can meet the needs of large-scale data processing.

#### 2.2.1. Gradient boosting decision tree (GBDT)

Ensemble Learning refers to the construction and combination of multiple weak classifiers to complete complex classification tasks to achieve better performance than a single classifier. There are two main kinds of ensemble learning methods. One is boosting algorithm and another is bagging algorithm. The idea of Gradient Boosting algorithm is described as follows.

The sample \( X \) is substituted into the basis function to obtain the sub-models \( f_1(X), f_2(X), \ldots, f_m(X) \) and the composite model \( F_m(X) \) is defined as:
\[ F_m(X) = \partial_0 f_0(X) + \partial_1 f_1(X) + \cdots + \partial_m f_m(X) \] (4)

Let the loss function be \( L(F_m(X), Y) \), after each new sub-model is added, let:

\[ L(F_m(X), Y) < L(F_{m-1}(X), Y) \] (5)

The gradient boosting decision tree (GBDT) is based on the decision tree (CART) as the gradient boosting algorithm, that is, each weak classifier \( f(X) \) in the above formula is a decision tree classifier.

2.2.2. Improvements of LightGBM

The LightGBM algorithm used in this paper is one of the ways to realize the idea of GBDT.

The basic idea of LightGBM is to combine many low accuracy tree models. Each iteration will generate a new tree based on all trees generated before, and then uses the method of gradient descent to make the loss function smaller and smaller and finally gets a better tree, which is used as the prediction model[11].

Based on the traditional gradient boosting decision tree, LightGBM is optimized by gradienct-based single-sided sampling, leaf growth strategy with depth limit and other algorithms[12].

- Gradient-based one-side sampling

Gradient-based one-side sampling (GOSS) is a sampling algorithm for training data. The purpose is to discard some examples that are not helpful for calculating information gain. Since data with a large gradient contributes a lot to the calculation of information gain, GOSS only keeps the data with a large gradient for data sampling. But if all gradient small data directly discarded certainly will affects the overall distribution of the data, so the GOSS first going to split the characteristics of all value according to the absolute size of descending sort, selection of absolute value of the largest \( a \times 100\% \) data, and then in the rest of the teach small gradient data randomly selected \( b \times 100\% \) data, and the data is multiplied by a constant \( ((1-a)/b)\% \), finally use it \( (a + b)\% \) data to calculate the information gain.

- Leaf growth strategy with depth limit

Leaf wise strategy with depth limitation. In the iterative process of extreme gradient boosting (XGBoost) algorithm, the level wise leaf growth strategy is adopted, that is to split the leaves of the same layer at the same time, as shown in Figure. 1.

But in fact, level wise does not distinguish the leaves in the same layer, and still searches and splits the leaves with low splitting gain, which brings a lot of unnecessary overhead. For this reason, LightGBM adopts a more efficient strategy. Every time it finds the leaf with the largest splitting gain from all the current leaves, and then splits, so it circulates, as shown in Figure. 2.

Therefore, compared with level wise, leaf wise can reduce more errors and get better accuracy under the same splitting times, but the disadvantage is that it may grow a deeper decision tree and produce over fitting. Therefore, LightGBM adds a maximum depth limit on leaf wise to ensure high efficiency and prevent over fitting.

In addition, LightGBM also optimizes and improves the optimal segmentation of category features and reduces the communication cost of parallel computing. Besides, it uses histogram optimization
algorithm to reduce memory consumption and significantly speed up the training speed of GBDT without compromising the accuracy.

3. Proposed Method

3.1. the CNN-LightGBM model structure

In view of the analysis of convolution neural network and LightGBM, it is found that convolution neural network has the ability of self-adaptive extraction of deep-seated features, while LightGBM has the advantages of high classification accuracy and fast training speed. We propose a bearing fault diagnosis method based on CNN-LightGBM, and the model structure is shown in Figure 3.

![Figure 3. Architecture of CNN-LightGBM.](image)

As shown in Figure 3, this proposed model mainly consists of two parts: one is the CNN part and another one is the LightGBM part. The CNN part used to extract features is composed of several convolution layers and pooling layers. The original vibration signal can be input into the model directly without any other processing. The BN layer is added after the convolutional layer. After extracting and filtering features of the convolution pooling layer and the pooling layer, the results are input into the full connection layer. Finally, the results of the full connection layer are extracted, and the LightGBM is used as a classifier for fault diagnosis of various fault types. In addition, we add batch normalization after convolution pooling layer to improve the convergence speed and generalization ability of the model. This model uses CNN to extract features, and then uses LightGBM as classifier to achieve complex feature cross, which can improve the accuracy of fault diagnosis.

3.2. the CNN-LightGBM model building

The construction of CNN-LightGBM can be divided into three steps as shown in Figure 4: data set processing, model training and model testing.
The vibration signal samples are obtained from the original vibration signals. The training set and test set are divided according to a certain proportion. We use training set data to train the model and use test set data to verify the model.

3.3. Model optimization method

Batch normalization[13] and Adam algorithm[14] are introduced to optimize the model.

3.3.1. Batch normalization

The batch normalization (BN) layer first subtracts the mean value of the mini batch from the characteristic graph before the input activation function, and then divides it by its standard deviation. However, the input data is limited to the range with the mean value of 0 and the variance of 1, so two learnable parameters are introduced to scale and translate the data. The introduction of BN layer in the model can reduce the transfer of internal covariates, effectively prevent gradient dispersion and accelerate the training speed. The specific calculation process of BN layer is shown in Eq. (6)-(9).

\[
\mu_B = \frac{1}{m} \sum_{i=1}^{m} x_i
\]

\[
\sigma_B^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2
\]

\[
\hat{x}_i = \frac{x_i - \mu_B}{\left(\sigma_B^2 - \epsilon\right)^{1/2}}
\]

\[
y_i = \gamma \hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i)
\]
where \( B = \{x_1, \ldots, x_m\} \) is the enter of mini-batch, \( \mu_B \) and \( \sigma_B^2 \) are the mean and variance of batches respectively. \( \hat{x}_i \) is the normalized vector, \( y_i \) is the output of BN layer, \( \gamma \) and \( \beta \) are the scaling factor of the learnable parameter, \( \epsilon \) is constant term.

### 3.3.2. Adam algorithm

Adam algorithm is a first-order optimization algorithm which can replace the traditional stochastic gradient descent (SGD) process. It can update the weights of neural networks iteratively based on the training data. Adam not only calculates the learning rate of adaptive parameters based on the first-order moment mean as root mean square prop (RMSProp) algorithm, but also makes full use of the second-order moment mean of gradient (i.e. biased variance).

# 4. Experiments

In this section, experiments data, model parameter setting description, result analysis are described.

### 4.1. Experiments data

In order to evaluate the proposed method, we used the rolling bearing dataset of Case Western Reserve University (CWRU), which is commonly used in bearing fault diagnosis literature\(^\text{[15]}\). Figure. 5 presents the testbed of CWRU.

![Case Western Reserve University testbed](image)

Figure 5. Case Western Reserve University testbed.

As shown in Figure. 5, the test bench consists of four parts, one motor (left) with fan end bearing and drive end bearing, one torque sensor/encoder (middle), one power meter (right) and control electronic equipment (not shown). Rolling bearings were seeded with faults using electro-discharge machining and vibration data is collected by accelerometer, which collects digital data at the rate of 12000 or 48000 samples per second.

We select 12kHz drive end fault data covering ten bearing states and four different working states, including normal bearing data and 0.007 inch, 0.014 inch and 0.021 inch damage diameter in inner ring, outer ring and rolling element. The experiments data are also collected under different working conditions having four different industries and mines, i.e. zero to three horsepower (motor speed from 1720 RPM to 1797 RPM). We take samples from the original vibration signal to obtain signal data with the length of 1024. Details of the experiments data are shown in Table 1.

### Table 1. Description of bearing experiments data

| Labels | Fault Place | Fault Diameter (inch) | Sample Size |
|--------|-------------|----------------------|-------------|
| 0      | None        | 0                    | 1657        |
| 1      | Inner Race  | 0.007                | 476         |
| 2      | Inner Race  | 0.014                | 472         |
| 3      | Inner Race  | 0.021                | 474         |
| 4      | Ball        | 0.007                | 473         |
| 5      | Ball        | 0.014                | 475         |
8

4.2. Model parameter setting description
For the original vibration signal with 1024 dimensions, the CNN-LightGBM model is designed as shown in Figure. 5, and the three-layer convolutional pooling layer is designed to extract deep features. The reduced convolutional kernel layer by layer can effectively compress the parameters of the model.

The parameters of the convolutional layer, pooling layer and full connection layer of the CNN part in the proposed model are shown in Table 2.

| Layer Type   | Kernel | Stride | Output Size (Width × Depth) |
|--------------|--------|--------|----------------------------|
| Convolution1 | 64 × 1  | 16     | 64 × 16                    |
| Pooling1     | 2 × 1  | 2      | 32 × 16                    |
| Convolution2 | 3 × 1  | 1      | 32 × 32                    |
| Pooling2     | 2 × 1  | 2      | 16 × 32                    |
| Convolution3 | 3 × 1  | 1      | 16 × 64                    |
| Pooling3     | 2 × 1  | 2      | 8 × 64                     |
| Convolution4 | 3 × 1  | 1      | 8 × 64                     |
| Pooling4     | 2 × 1  | 2      | 4 × 64                     |
| Fully-connected | 100 / |        | 100 × 1                    |
| Softmax      | 10     | /      | 10                         |

The experiment samples in table 1 are divided into 4:1 ratio to get training set and test set.

Experiment 1: We use the parameters in table 2 to build the CNN model, use the cross entropy loss function and softmax classifier. In addition to the parameters shown in table 2, the batch normalization layer is added between convolution layer and activation layer to improve the training efficiency and generalization ability of the network. Besides, we use the Adam algorithm to accelerate model learning time.

Experiment 2: Based on experiment 1, the model shown in Figure. 3 used and LightGBM is introduced as a classifier to further optimize the model. In addition, we use Bayesian Optimization as super parameter optimization because of its sample efficiency and five main parameters are adjusted: leaf node number, tree depth, tree maximum tree, bagging fraction and learning rate.

4.3. Analysis of experimental results
The deep neural network framework used in the experiment is a high-level neural network API Keras with Tensorflow as the back end proposed by Google. Firstly, this paper compares the accuracy of fault diagnosis before and after using LightGBM model and then verifies the effect of model feature learning.

4.3.1. Model accuracy
We use train set to train the model, then use the trained model to classify and predict, and compare the predicted results with the actual results.
The classification accuracy of training set and test set in Experiments 1 and 2 are shown in Figure 6. The results show that the accuracy of the proposed CNN-LightGBM model increases with the number of experiments, and compared with single CNN model in Experiments 1, the accuracy of the proposed model is improved especially when epoch is small.

The confusion matrix is to count the real and predicted categories of each sample fault type. The number on the diagonal of the matrix represents the number of samples correctly classified and predicted to this category. From the distribution of the confusion matrix in Figure 7, it can be seen that most bearing samples can be correctly classified into corresponding categories. The accuracy of fault diagnosis is up to 98.92%.

4.3.2. Verification of feature learning effect
In order to verify the adaptive feature learning ability of the model for the original input signal, t-distributed stochastic neighbor embedding (t-SNE) method is used to visualize the features learned in the full connection layer of the proposed model.
5. Conclusions

In this paper, we present a bearing fault diagnosis method based on convolution neural network and light gradient boosting machine. Compared with traditional single convolution neural network, which is used to both extract features and classify, light gradient boosting machine is used as classifier to realize the fine recognition of different faults. Besides, the batch normalization layers are added to improve the convergence speed and classification accuracy of the model. We verify the model on CWRU bearing fault dataset and the experimental results show that this method is an effective method for bearing fault diagnosis. In addition, this method can extract features adaptively without any denoising signal pre-processing and feature extraction. This method is also effective in fault diagnosis of different load and speed.

In this paper, we mainly focus on the bearing fault diagnosis under the condition of constant speed and pressure. However, in the actual project, the speed and pressure are usually changing. In future work, we plan to extend the research results of this paper to the fault diagnosis under complex conditions such as variable speed and pressure.

References

[1] Nandi S, Toliyat H A, and Li X 2005 Condition Monitoring and Fault Diagnosis of Electrical Motors - A Review J. IEEE Transactions on Energy Conversion 20(4) 719-29
[2] Guo Yifan and Tang Jiayin 2019 A Review of The Development of Life Prediction and Fault Diagnosis Technology Based on Machine Learning Algorithm. Computer Measurement & Control 27(03) 13-9
[3] Rafsanjani, A., Abbasion, S., Farshidianfar, A., and Moeenfard, H. 2019 Nonlinear dynamic modeling of surface defects in rolling element bearing systems. Journal of Sound and Vibration 319(3-5) 1150-74
[4] Kankar, P.K., Harsha, S.P., Kumar, P., and Sharma, S.C. 2009 Fault diagnosis of a rotor bearing system using response surface method. European Journal of Mechanics, A/Solids 28(4) 841-57
[5] Misra, M., Yue, H.H., Qin, S.J., and Ling, C. 2002 Multivariate process monitoring and fault diagnosis by multi-scale PCA. Computers & Chemical Engineering 26(9) 1281-93
[6] Pandya, D.H., Upadhyay, S.H., and Harsha, S.P. 2013 Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN. Expert Systems with Applications 40(10) 4137-45
[7] Santos, P., Villa, L., Reñones, Aníbal, Bustillo, A., and Maudes, Jesús 2015 An SVM-based solution for fault detection in wind turbines. Sensors 15(3) 5627-48
[8] Sheng, Guo, Tao, Yang, Wei, Gao, and et al 2018 An intelligent fault diagnosis method for bearings with variable rotating speed based on pythagorean spatial pyramid pooling CNN. Sensors 18(11) 3857
[9] Wei, Z., Gaoliang, P., Chuanhao, L., Yuanhang, C., and Zhujun, Z. 2017 A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors 17(2) 425-
[10] Nianlong Gu, Hao Pan, and Peng He 2017 Bearing Fault Diagnosis Method Based on EMD-CNNs. 3rd International Conference on Computer Science and Mechanical Automation (CSMA 2017)
[11] Yasser, K., and Hemayed, E. 2017 Novelty Detection for Location Prediction Problems Using Boosting Trees. *International Conference on Computational Science & Its Applications*. Springer, Cham

[12] Ke, G., Meng, Q., Finley, T. and et al 2017 LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 3146-54

[13] Ioffe, S., and Szegedy, C. 2015 *Batch normalization: accelerating deep network training by reducing internal covariate shift*

[14] Kingma, D. P. and Ba, J. *Adam: a method for stochastic optimization*. *Computer Science*

[15] Smith, W. A. and Randall, R. B. 2015 Rolling element bearing diagnostics using the case western reserve university data: a benchmark study. *Mechanical systems and signal processing*, 64-65(dec.) 100-31