ISSN 2706-655X Vol.5, Issue 3: 26-31, DOI: 10.25236/IJFET.2023.050305

# Fault prediction of industrial machinery and equipment

**Mingmin Jiang, Wen Zhou*, Yan Gou**

*College of Information Engineering, Nanjing University of Finance & Economics, Nanjing, 210023, China*
**Corresponding author: 1632038454@qq.com*

***Abstract:*** *In this paper, we conduct a study on failure prediction of industrial machinery and equipment in order to improve the competitiveness of enterprises. We use Bootstrap to resample the dataset before the experiment. Then, we used four algorithms to build the prediction model, and used GridSearchCV to automatically adjust the parameters and train the equipment records extracted from the dataset. We selected the accuracy, F1 score, and ROC curve as the evaluation indexes of the model, and the evaluation results were compared, so we finally chose the fault prediction model built by LightGBMClassifier algorithm, and then used the confusion matrix to evaluate the model performance.*

***Keywords:*** *Fault Prediction; Integrated Learning; Decision Trees; Joint Analysis*

## 1. Introduction

Over the past decade, China's status as a major industrial manufacturing country has been firmly established as the first in the world, and it is currently in an important stage of transformation to a manufacturing power, and the manufacturing industry occupies a large proportion of the national economic system [1-2], so the efficiency requirements of industrial manufacturing in China appear to be particularly important. However, in the process of industrial production, due to a series of problems such as wear and tear, heat dissipation, electricity, and overload, machinery and equipment inevitably produce various types of failures, which also affect the quality and efficiency of industrial production.

Accurate and efficient advance prediction of failure risk can improve industrial productivity and economic efficiency, and also enhance the competitiveness of enterprises to a certain extent. In this paper, we use the data obtained in this paper to build a prediction model to determine whether the equipment will fail or not, to derive the training results, and to select suitable evaluation indexes to evaluate the prediction model. We use the processed data set to build a multi-classification model to determine which type of equipment failure belongs to TWF/HDF/PWF/OSF/RNF, derive the results, select suitable evaluation indexes to evaluate the prediction model, and quantitatively analyze each type of failure to find out the main relevant characteristic attributes of each type of failure as the main cause of the failure, and mine the potential association rules between the main cause and other characteristic attributes The potential association rules between the main cause and other characteristic attributes.

## 2. Model Establishments

### *2.1 The structure of Fault prediction model*

We analyzed the data of the "train data.xlsx" dataset with the attribute "whether a fault occurred", which is considered to be one of the most common binary classification problems. We selected several traditional and popular algorithms as well as several new algorithms, such as Support Vector Machine (SVM), Random Forest, XGBoost, Neural Network, Perceptual Neural Network, Stacking Integrated Learning, and Voting Integrated Learning.

We selected accuracy, F1 score, and subject operating characteristic curve (ROC) as a way to compare the predictive performance of several models built on different algorithms. Accuracy is the ratio of the number of correctly classified samples to the total number of samples, and is expressed by the formula:

$$accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (1)$$

ISSN 2706-655X Vol.5, Issue 3: 26-31, DOI: 10.25236/IJFET.2023.050305

The F1 score is the summed average of the precision rate P and the recall rate R, expressed by equation:

$$F1 = \frac{2PR}{P+R} \tag{2}$$

The exact rate is given by the formula:

$$P = \frac{TP}{TP+FP} \tag{3}$$

The formula for the recall rate is:

$$R = \frac{TP}{TP+FN} \tag{4}$$

The subject operating characteristic curve, or ROC curve, is a graph drawn using the true case rate TPR as the vertical axis and the false positive case rate FPR as the horizontal axis, describing the performance of the classifier as a function of the threshold, with a larger area representing the greater classification ability of the model. Among them, the true case rate TPR formula is:

$$TPR = \frac{TP}{TP+FN} \tag{5}$$

The false positive rate FPR equation is:

$$FPR = \frac{FP}{TN+FP} \tag{6}$$

### *2.2 The structure of Fault multi-classification model*

Since the "specific fault category" data of the source data is of object type, which is not easy for the model to operate on it, we need to quantize the specific fault type before building the model in order to make the classifier handle the attribute data better. Initially, we considered quantizing the data using unique thermal coding, which uses N-bit status registers to encode N states, each of which has a separate register bit and only one bit is valid at any time.

However, during the experimental process, we found that the prediction of the data transformed by the unique thermal coding was not very good on the model built by these five algorithms, as shown in Figure 1, which shows the variation of the accuracy with the training batch. " of the data, where the specific fault categories Normal, TWF, HDF, PWF, OSF, and RNF correspond to the numbers 0, 1, 2, 3, 4, and 5, respectively.

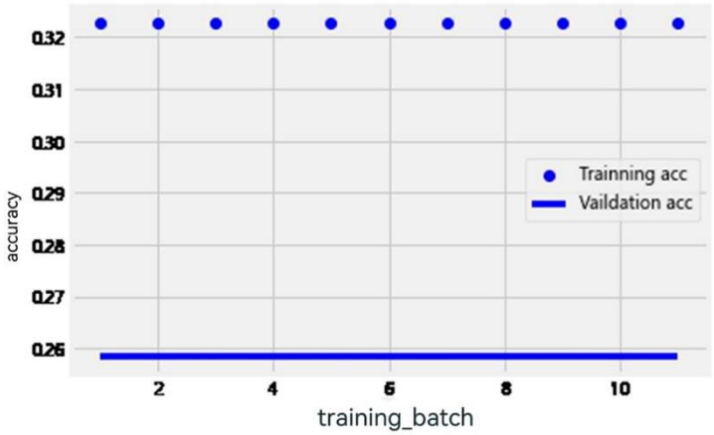


*Figure 1: Variation of accuracy with training batches*

## 3. Results

### *3.1 Comparison of different algorithms*

Based on the above analysis, we conducted qualitative analysis of the selected algorithms, and in order to derive the most optimal results and parameters, we used GridSearchCV to automatically adjust the parameters, build the corresponding prediction models and use the models to train on the "train data.xlsx" dataset to derive The respective evaluation results are shown in Table 1, and the combined ROC curves of

ISSN 2706-655X Vol.5, Issue 3: 26-31, DOI: 10.25236/IJFET.2023.050305

the various algorithms are shown in Figure 2.

*Table 1: Comparison of evaluation results of different algorithms*

| algorithm | accuracy | F1-score | ROC(area) |
|---|---|---|---|
| SVM | 96.50% | 0.00% | 80% |
| Random Forest | 98.44% | 75.00% | 97% |
| XGBoost | 98.39% | 72.90% | 97% |
| Neural Networks | 96.67% | | 50% |
| Perceptron | 96.89% | 0.00% | 92% |
| Stacking | 98.33% | 72.73% | 83% |
| Voting | 98.50% | 74.77% | 98% |

Through comprehensive comparison, we found that the evaluation scores and evaluation indexes of all aspects of Voting integrated learning are superior, and its accuracy is the highest, reaching 98.50%, F1 score is high at 74.77%, ROC curve area is the largest, reaching 98.00%, and the best prediction effect, so we finally choose the fault prediction model established by Voting integrated learning algorithm.

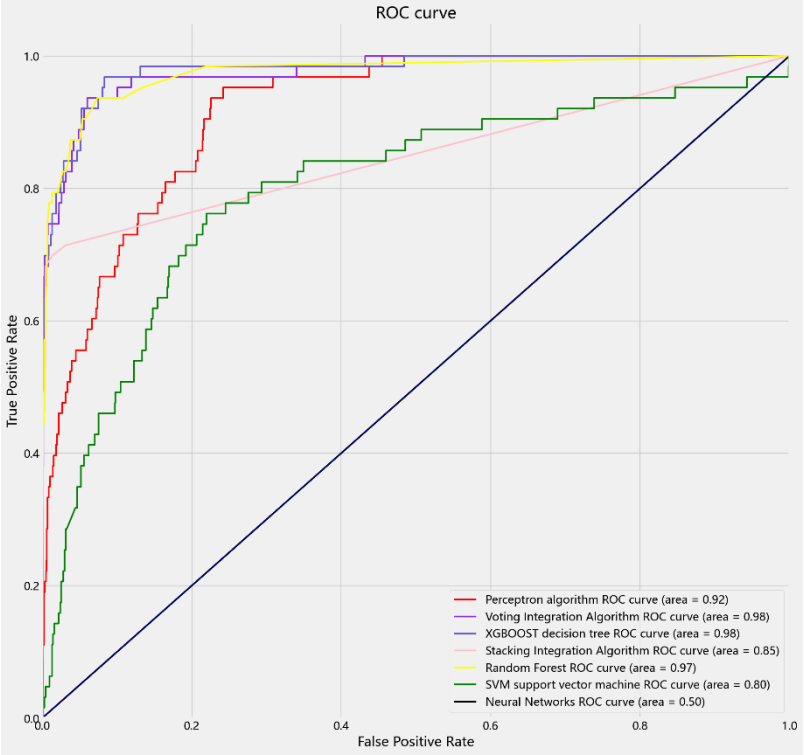


*Figure 2: Combined ROC curves of different algorithms*

### *3.2 Evaluation of Voting Integrated Learning Models*

Voting is an integrated learning model that follows the principle of minority rule. The method improves the robustness of the model by integrating multiple classification models and thus reducing the variance and reducing the error rate of the model.

We use K-fold cross-validation to evaluate the combined predictive performance of models built on the Voting integrated learning algorithm. k-fold cross-validation splits the dataset into a training set and a test set for cross-training and validation to provide the most objective assessment of the model performance.

By running the code, we derived the integrated evaluation metrics for the prediction models built using Voting integrated learning as shown in Figure 3.

ISSN 2706-655X Vol.5, Issue 3: 26-31, DOI: 10.25236/IJFET.2023.050305

```
Voting test accuracy: 98.50%
Voting test F1 score: 74.77%
              precision    recall  f1-score   support

         yes       0.99      1.00      0.99      1737
          no       0.91      0.63      0.75        63

    accuracy                           0.98      1800
   macro avg       0.95      0.82      0.87      1800
weighted avg       0.98      0.98      0.98      1800
```

*Figure 3: Comprehensive evaluation metrics for Voting integrated learning models*

The prediction accuracy of the model built using voting integrated learning reached 0.99 and 0.91, and the recall reached 1.00 and 0.63, respectively, which shows that the model has a good prediction effect[3-4].

### *3.3 Evaluation of the LightGBMClassifier model*

LightGBMClassifier is a distributed gradient boosting framework based on decision tree algorithm, which supports efficient parallel training and has the advantages of faster training speed, lower memory consumption, higher accuracy, support for efficient parallelism, and can quickly handle large amounts of data.

We use the confusion matrix to evaluate the comprehensive prediction performance of the model built based on the LightGBMClassifier algorithm. The confusion matrix is an accuracy analysis table in machine learning that summarizes the prediction results of a classification model in the form of a matrix that judges the model performance as a whole by the records in the dataset according to the real categories and the categories predicted by the classification model.

By running the code, we derived the confusion matrix for the prediction model built using LightGBMClassifier as shown in Figure 4.

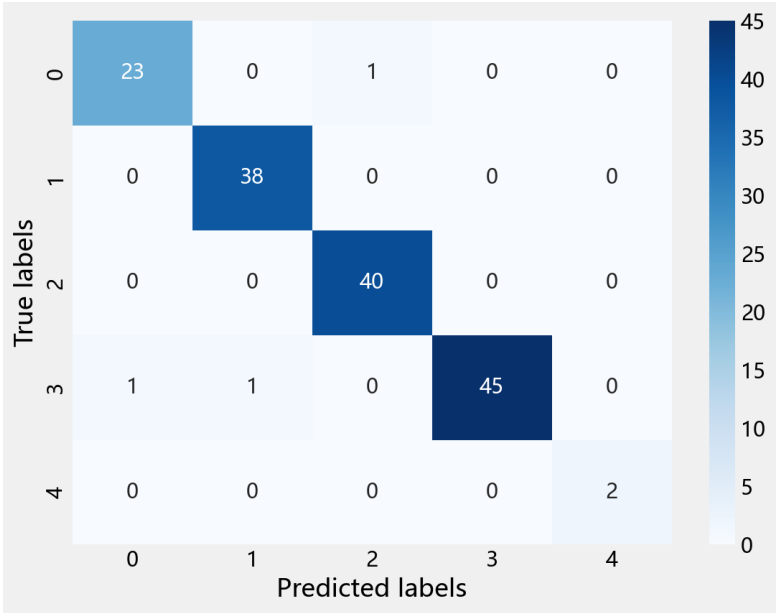


*Figure 4: Confusion matrix for LightGBMClassifier model*

As seen from the figure, the prediction model established by using LightGBMClassifier has a higher accuracy rate of classification and a lower error rate, which shows that the model has a better prediction effect.

ISSN 2706-655X Vol.5, Issue 3: 26-31, DOI: 10.25236/IJFET.2023.050305

## 4. Conclusions

### *4.1 The main causes of each type of failure*

We extracted the data for each class of faults individually and performed join analysis based on SPSSPRO [5] [6] to calculate the weights of each feature attribute and find out the main cause of each class of faults. Among them, for attributes with weights less than 1, we only keep the feature attribute with the largest weight as the main cause of each type of failure, while those with weights greater than 1, we consider them all to be the main cause of the failure.

### *4.2 Extracting association rules*

We use Pearson's correlation coefficient to explore the relationship mapping between the above main contributing factors and plot the relationship between two and two to explore the potential association rules.

The relationship between machine temperature and plant temperature is shown in Figure 5.

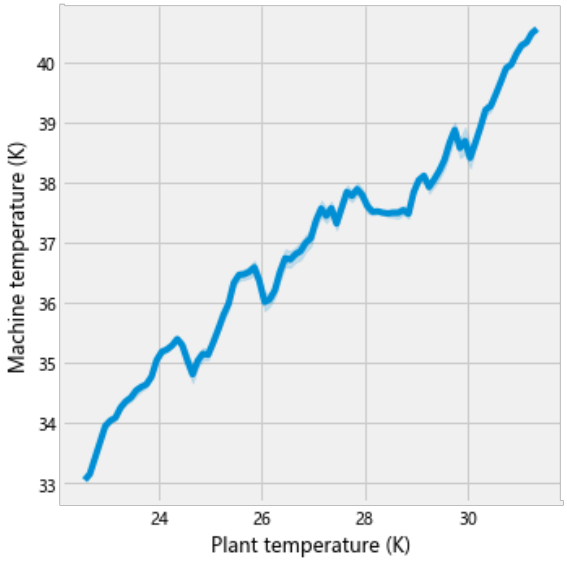


*Figure 5: Machine temperature and plant temperature relationship chart*

From the figure, it can be seen that there is a positive correlation between machine temperature and plant temperature.

The relationship between machine temperature and duration of use is shown in Figure 6.

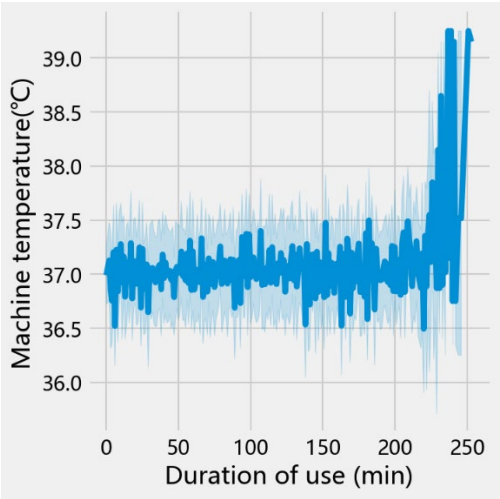


*Figure 6: The relationship between machine temperature and usage time graph*

ISSN 2706-655X Vol.5, Issue 3: 26-31, DOI: 10.25236/IJFET.2023.050305

As can be seen from the graph, when the use time is longer than 220 minutes, the temperature of the machine will rise sharply with the increase of use time.

The relationship between speed and torque is shown in Figure 7.

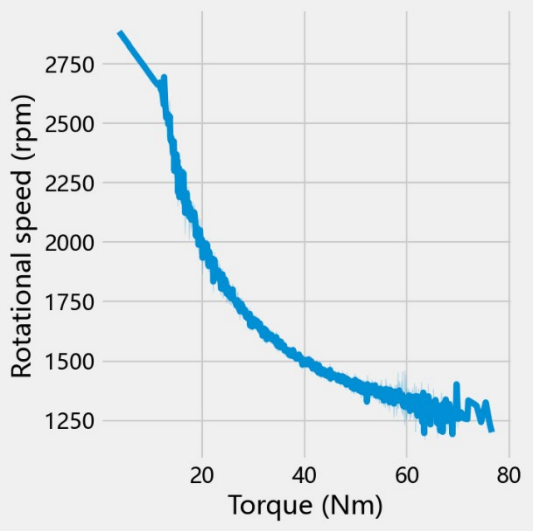


*Figure 7: Rotational speed vs. torque graph*

As can be seen from the graph, there is a negative correlation between speed and torque.

Viewing the attribute weights of each indicator, we found that the main causes of mechanical equipment failure are machine temperature, speed, torque, plant temperature, and length of use.

## References

*[1] Hu Yujian. Application of mechanical automation technology in machinery manufacturing industry [J]. Modern Industrial Economics and Informatization, 2021, 11(08):140-146.*
*[2] Cao Yingming. Portable rotational speed measuring instrument[J]. Measurement and Testing Technology, 2017, 44(09):27-29.*
*[3] Wang Lin. Common methods of mechanical equipment fault diagnosis and monitoring and its development trend [J]. Journal of Wuhan University of Technology, 2000(03):62-64.*
*[4] Fu Xitao. Research and prospect of reciprocating compressor fault diagnosis [J]. Technology and Market, 2014, 21(07):119-120.*
*[5] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.*
*[6] Jia JP, He XQ, Jin Y. Statistics (4th ed.) [M]. People's University of China Press, 2009.*