Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(4), 226-243
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Research on Fault Detection and Classification of Industrial Equipment Based on Machine Learning

Gangying Cai

School of Mathematics and Statistics, Guilin University of Technology, Guilin, Guangxi, China
Email: cgy.paula@foxmail.com

How to cite this paper: Cai, G.Y. (2025) Research on Fault Detection and Classification of Industrial Equipment Based on Machine Learning. *Journal of Computer and Communications*, **13**, 226-243.
https://doi.org/10.4236/jcc.2025.134015

Received: April 2, 2025
Accepted: April 25, 2025
Published: April 28, 2025

Abstract

In the context of the rapid development of intelligent manufacturing, the stable operation of mechanical equipment is crucial for maintaining industrial production continuity and achieving economic benefits. Timely identification of potential fault risks based on equipment operation data can facilitate accurate maintenance and enhance production safety and efficiency. This study conducts fault detection and classification modeling using operational data from industrial equipment provided by a manufacturing enterprise. First, the raw data underwent data cleaning, outlier removal, and imputation of missing values. Key influencing factors, including plant temperature, equipment temperature, rotational speed, torque, hours of use, and equipment quality level, were identified through independence tests and ANOVA. For fault detection modeling, support vector machine (SVM) and decision tree (CART) algorithms were employed to determine whether the equipment experienced fault, with model performance evaluated using multiple evaluation metrics. For fault classification, a multi-classification model based on the random forest algorithm was developed to identify specific fault types. Furthermore, feature importance analysis was performed to quantify the impact of different features on various fault types, revealing the potential causes of faults. This study offers practical value for intelligent maintenance and predictive overhaul of manufacturing equipment, providing both data-driven insights and methodological reference for industrial applications.

Keywords

Industrial Equipment Fault Detection, Fault Diagnosis, Support Vector Machine, Decision Tree CART, Random Forest

1. Introduction

In the context of the global shift toward digital and intelligent manufacturing, the operational status of industrial equipment directly affects production efficiency and product quality. Prolonged operation in high-intensity, complex environments makes equipment prone to wear, aging, and faults. These faults can disrupt production, cause financial losses, and endanger personnel safety, thereby threatening the stability of production systems. Timely fault detection and accurate fault type identification are thus essential to improving production continuity and economic performance in manufacturing. Advances in cloud computing and the Industrial Internet of Things allow enterprises to monitor equipment data in real time, while machine learning is widely used in fault detection for its strong pattern recognition capabilities.

Qi *et al.* (2018) developed a fault diagnosis system for reciprocating compressors using big data and support vector machine (SVM)-based machine learning, achieving over 80% accuracy on real-world industrial data [1]. Carvalho *et al.* (2019) conducted a systematic review of machine learning methods in predictive maintenance, summarizing key techniques, performance, and challenges in industrial applications [2]. Cohen *et al.* (2022) proposed a hybrid fault diagnosis framework combining timed Petri nets and machine learning for event synchronization faults in discrete manufacturing systems, achieving 96% precision and recall [3]. With the increasing availability of data and computational power, deep learning has gained significant attention in fault diagnosis research. Gao *et al.* (2020) optimized the parameters of a Deep Belief Network using the Salp Swarm Algorithm to mitigate the impact of manual parameter tuning, thereby improving the accuracy of bearing fault diagnosis [4]. Cao *et al.* (2022) proposed an unsupervised domain-invariant Convolutional Neural Network (CNN) approach that extracts domain-invariant features under time-varying speeds to achieve fault diagnosis of mechanical equipment [5]. Ding *et al.* (2022) employed a Transformer model leveraging the self-attention mechanism to effectively extract fault features from vibration signals for comprehensive fault diagnosis [6]. Amin *et al.* (2023) proposed a fault detection method for wind turbines, combining CNN and spectral analysis to classify faults using two-dimensional feature images, which effectively reduced the cost of operation and maintenance [7].

Although deep learning methods have made significant progress in fault detection in recent years, traditional machine learning methods still have a strong advantage when dealing with small samples and high dimensional data. Zhang *et al.* (2022) and Okwuosa *et al.* (2022) employed an SVM-based binary classification model for fault diagnosis, validating the model's effectiveness using a data monitoring system [8] [9]. Purbowaskito *et al.* (2023) developed an integrated fault diagnosis framework that combines model-based diagnosis and machine learning classifiers, dynamically adapting the model to improve diagnostic accuracy [10]. Tao *et al.* (2024) proposed a machine learning-based method for photovoltaic fault diagnosis and localization using modulated photocurrent, achieving high ac-

curacy and low-cost deployment [11]. Vashishtha *et al.* (2024) comprehensively reviewed recent advancements in machine learning for industrial fault diagnosis, highlighting the transition from traditional ML to deep learning techniques [12]. Muzzammel (2025) developed a Gini-index-based ML algorithm for fault diagnosis in HVDC systems and highlighted the importance of preprocessing under resistance variability [13]. Wu (2025) proposed a motor fault diagnosis model integrating machine learning with fuzzy control, significantly improving diagnostic accuracy and stability [14].

Despite the significant progress in machine learning for fault diagnosis, noise, outliers, and class imbalance are often present in industrial data, posing challenges to the performance and stability of models. Duan *et al.* (2016) proposed a Support Vector Data Description-based method for machinery fault diagnosis, addressing unbalanced datasets by introducing a binary tree structure for multi-class classification [15]. Jablon *et al.* (2021) proposed a machine learning-based approach for diagnosing rotating machine unbalance using vibration orbital features, achieving robust performance even in noisy environments [16]. Huang *et al.* (2022) proposed a multi-scale fractional-order dimensionless metric combined with a random forest approach for fault diagnosis [17]. Jin *et al.* (2022) designed a residual preprocessing module and a multi-scale CNN to mitigate the influence of noise on diagnostic results [18]. Han *et al.* (2022) enhanced diagnostic capability in high-noise environments using the Transformer model to jointly extract global and local information [19]. Mian *et al.* (2023) developed a multi-sensor fault diagnosis system combining IRT and vibration data, using Deep Convolutional Neural Network and SVM to detect misalignment, unbalance, and rotor disk eccentricity faults [20]. Prawin (2025) proposed a hybrid 2DCNNLSTM algorithm for bearing fault diagnosis, combining CNN and LSTM to enhance diagnosis by capturing both spatial and temporal features, particularly under imbalanced data conditions [21].

Despite significant advances in deep learning for fault detection, traditional machine learning methods (e.g., SVM, CART, and random forests) remain effective choices in many industrial environments, particularly in data-limited or high-dimensional scenarios. In contrast, traditional methods are less computationally demanding, making them suitable for resource-constrained environments, and they are easier to deploy. This study addresses the data characteristics and practical needs of industrial equipment fault diagnosis by leveraging real-world data to develop and evaluate machine learning models for fault detection and classification. In addition, it conducts a feature-based analysis of different fault types to investigate their root causes and underlying patterns.

The structure of the study is as follows: Section 1 provides an introduction to the research background and the significance of fault detection and classification for industrial machinery and equipment. It also reviews the current state of related studies and describes the organization of this study. Section 2 presents the research methodology, including the fault detection model based on Support Vector

Machine (SVM) and Decision Tree (CART), as well as the fault classification model based on Random Forest, providing a detailed explanation of the model construction and algorithmic principles. Section 3 analyzes and processes the experimental data, extracts key features through feature screening and correlation analysis, and performs dataset partitioning to support subsequent modeling. Section 4 presents the experimental results and analysis, including the performance evaluation of the fault detection and classification models, as well as an exploration of fault causes based on feature importance analysis, revealing the key influencing factors for different fault types. Finally, this study summarizes the main research findings and suggests directions for future research.

2. Fault Detection and Classification Modeling

2.1. Problem Description and Model Construction

In industrial production, the fault detection and classification of mechanical equipment are core tasks to ensure production safety and improve equipment management efficiency. Timely detection of equipment faults can help avoid production downtime, reduce maintenance costs, and ensure personnel safety. Furthermore, fault classification helps identify specific fault types and provide precise repair solutions for the maintenance personnel.

In this study, the fault detection task is formulated as a binary classification problem that determines whether a mechanical device has failed. Let $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^{n}$ represent the operational dataset of the equipment, where $x_i\in\mathbb{R}^d$ denotes the feature vector of the ith sample with a total of d features. The fault state label is represented by $y_i\in\{0,1\}$, with $y_i=0$ indicating normal operation and $y_i=1$ indicating a malfunction, where n denotes the total number of samples. For malfunctioning devices, the fault classification task is formulated as a multi-class classification problem, aiming to predict the specific fault category using the input data. Let the set of fault categories be denoted as $\mathcal{Y}=\{1,2,\cdots,C\}$, where C represents the total number of fault categories. The dataset can be expressed as $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^{n}, y_i\in\mathcal{Y}$. The objective of the fault detection problem is to construct a discriminant function $f:R^d\to\{0,1\}$ to predict the fault state of a device by training a classifier. The optimization goal is to minimize the classification error while ensuring robust generalization performance. The loss function $\mathcal{L}(f(x_i),y_i)$ is applied to evaluate the performance of the classifier, and a regularization term is incorporated to mitigate overfitting:

$$\min_f \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(f(x_i),y_i)+\lambda\Omega(f). \tag{1}$$

In Equation (1), $\Omega(f)$ represents the regularization term, and λ denotes the regularization parameter. Building on fault detection, fault classification further identifies the specific fault type when a device fails. The goal of fault classification is to construct a classifier $g:R^d\to\mathcal{Y}$. The optimization goal of the multiclass classification problem is likewise to minimize the classification error:

$$\min_{g} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(g\left(x_i\right), y_i\right) + \lambda \Omega\left(g\right). \tag{2}$$

By minimizing the loss function, the accuracy of the classifier can be effectively improved. To achieve this goal, this study employs the support vector machine (SVM), the CART decision tree algorithm, and the random forest model for fault detection and classification.

2.2. Support Vector Machine-Based Fault Detection Model

The Support Vector Machine (SVM) distinguishes between normal and faulty devices by constructing an optimal decision boundary. Compared to traditional classification methods, the SVM offers strong generalization ability and robustness, making it suitable for high-dimensional data and small sample scenarios. The core idea is to find the hyperplane that maximizes the interval in the feature space to ensure the generalization ability of the classifier.

Assume that the training dataset is $\mathcal{D} = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ denotes the feature vector of the ith sample with a total of d features, and $y_i \in \{-1, 1\}$ is the fault state label. SVM represents the classification hyperplane in the following form:

$$f\left(x\right) = \omega^{\mathrm{T}} x + b = 0, \tag{3}$$

In Equation (3), ω is the normal vector, which determines the direction of the hyperplane, and b is the bias term, which determines the position of the hyperplane. In the linearly separable case, the SVM finds the optimal hyperplane by maximizing the classification margin. Its optimization objective is:

$$\begin{aligned} &\min_{\omega, b} \frac{1}{2} \|\omega\|^2 \\ &\text{s.t. } y_i \left(\omega^{\mathrm{T}} x_i + b\right) \geq 1, i = 1, 2, \cdots, n. \end{aligned} \tag{4}$$

In practical applications, data are often non-linearly separable. In such cases, SVM introduces slack variables and employs a soft margin strategy. A penalty term is added to the optimization objective to balance the classification error and the maximization of the margin. The optimization objective is:

$$\begin{aligned} &\min_{\omega, b, \xi} \frac{1}{2} \|\omega\|^2 + \lambda \sum_{i=1}^{n} \xi_i \\ &\text{s.t. } y_i \left(\omega^{\mathrm{T}} x_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \end{aligned} \tag{5}$$

In Equation (5), λ is the regularization parameter used to balance the trade-off between the classification penalty and margin maximization, and ξ_i is the slack variable representing the penalty for misclassified samples. For linearly non-separable problems, SVM applies a kernel function to map the data into a higher-dimensional space, where it seeks the optimal hyperplane. After training, the classification decision function is:

$$f\left(x\right) = sign\left(\sum_{i=1}^{N} \alpha_i y_i \kappa\left(x_i, x_j\right) + b\right), \tag{6}$$

In Equation (6), α_i represents the Lagrange multiplier, $\kappa(x_i, x_i)$ denotes the kernel function, and $sign(\cdot)$ is the sign function. The optimization problem of the SVM is solved using its Lagrangian dual form. By introducing the Lagrange multiplier and applying the Karush-Kuhn-Tucker conditions, the support vectors and hyperplane parameters are determined. These parameters are then substituted into the classification decision function and the hyperplane equation to obtain the model training results, completing the fault detection task.

2.3. Decision Tree-CART-Based Fault Detection Model

A decision tree divides data using a tree structure, gradually categorizing samples into different classes. The Classification and Regression Tree (CART) is a classic algorithm commonly used for this purpose. The core idea of CART is to recursively split the dataset and construct a decision tree to distinguish between normal and faulty equipment. During the tree construction process, CART uses the Gini index as a criterion for feature selection. The Gini index measures the impurity of the dataset, representing the degree of uncertainty or misclassification within the categorization.

Assuming that the dataset at the current node is D, and the proportion of the kth class samples is p_k, the Gini index is defined as:

$$Gini(D) = \sum_{j=1}^{K} p_k (1 - p_k) = 1 - \sum_{j=1}^{K} p_k^2, \tag{7}$$

In Equation (7), $Gini(D)$ is the Gini index of the dataset D, K is the number of categories, and p_k represents the probability that a sample belongs to the kth category. When all samples within a node belong to the same category, the Gini index is 0; when the sample categories are evenly distributed, the Gini index approaches 1. In fault detection, CART traverses all features and partitioning points to select the partitioning method with the smallest Gini index, ensuring the maximum improvement in node purity. Although the generated decision tree can fit the training data well, an excessively large tree structure often results in overfitting. To address this issue, CART employs a pruning strategy to optimize model performance by reducing the complexity of the tree. The core of the pruning process involves introducing a regularization parameter to α control the tree's complexity. The loss function after pruning is defined as:

$$L(T) = \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda |T|, \tag{8}$$

In Equation (8), $L(T)$ is the loss function of the decision tree, $\ell(y_i, f(x_i))$ represents the classification error of the ith sample, $|T|$ denotes the number of leaf nodes in the tree, and λ is the regularization parameter used to balance the trade-off between tree complexity and classification accuracy. The optimal λ value is selected using the cross-validation method to determine the best-pruned subtree T^*. Pruning not only effectively prevents overfitting but also reduces the computational complexity of the model.

2.4. Random Forest-Based Fault Identification Model

Random Forest (RF) is an ensemble learning method based on decision trees. It enhances classification accuracy and generalization ability by aggregating the results of multiple decision trees. Compared with a single decision tree, RF demonstrates strong robustness when handling high-dimensional data, outliers, and noisy data. Additionally, it supports feature importance evaluation, providing insights for fault cause analysis. The core idea of RF involves using the Bagging method to resample the training data, generate multiple subsets, and train decision trees on these subsets. The final classification result is determined through a majority voting mechanism, effectively reducing model variance and improving stability. In this study, the RF model is employed for mechanical equipment fault identification.

Assume a training dataset $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, where $x_i \in \mathbb{R}^d$ represents the feature vector of the ith sample, with a total of d features, and $y_i \in \{0,1\}$ is the corresponding fault label. The random forest consists of M decision trees, and the training process is as follows:

1) Bootstrap sampling. Perform bootstrap sampling from the training data to generate M different training subsets D_m. The number of samples in each training subset is the same as the original dataset.

2) Construct a decision tree. For each training subset D_m, construct a CART. At each node split, k features ($k < d$) are randomly selected, and the best feature among them is chosen for node splitting. This process reduces the impact of feature correlation on the model.

3) Model voting. Input the test data into the trained random forest, with each tree independently predicting the result. The majority vote is used to aggregate the detection results from all decision trees, producing the final fault category for the test sample.

In this study, SVM, CART, and RF models are applied for fault detection and classification, in alignment with the characteristics of real-world industrial data and maintenance practices, offering insights into the practical implementation of fault diagnosis strategies.

3 Experimental Data Analysis and Model Evaluation Indicators

3.1 Dataset Description and Preprocessing

The dataset used in this study consists of 9000 records of operational and fault status data from industrial machinery of an enterprise. It is utilized for fault detection and fault classification analysis. The dataset includes key variables that reflect the equipment's operating environment, working status, and fault information. This dataset, sourced from a real-world industrial environment, provides a relevant basis for evaluating the performance of the proposed models in a practical context. Continuous variables include Factory temperature (K), equipment temperature (K), rotation speed (rpm), torque (Nm), and usage time (min), which describe the operating conditions of the equipment. Additionally, categorical variables such as machine number, unified specification code, and machine quality

level are provided. The machine quality level is categorized into three classes: high (H), medium (M), and low (L), indicating the equipment's performance level.

In terms of fault information, the data set contains two variables: "whether a fault has occurred" and "specific fault category". "whether a fault has occurred" takes a value of 0 or 1, 0 indicates that the equipment is operating normally, and 1 indicates that the equipment has failed. The "specific fault categories" include six types: NORMAL, TWF, HDF, PWF, OSF, and RNF. Among these, NORMAL indicates that the equipment is operating normally and corresponds to a value of 0 in the "fault occurrence" variable. TWF, HDF, PWF, OSF, and RNF represent wear faults, heat dissipation faults, power faults, overload faults, and other faults, respectively. In order to ensure the reliability and scientificity of the data, necessary judgments were made on the data.

For the fault category variable, the category label is digitized. Specifically, the label encoding method is used to assign "Normal" to 1, "PWF" to 2, "OSF" to 3, "RNF" to 4, "HDF" to 5, and "TWF" to 6. This method maintains the category information while facilitating the training and evaluation of subsequent models. For the discrete variable of machine quality level (L level, M level, H level), since it does not have a numerical relationship, the direct use of numerical encoding may cause the model to misjudge. Therefore, the paper adopts the one-hot encoding method to convert the quality level operation into binary data. Specifically, an independent binary feature column is created for each category, such as "machine quality level_H" and "machine quality level_M", whose values are 0 or 1, indicating whether the equipment belongs to the corresponding quality level. In this way, incorrect relationships caused by categorical data are avoided.

During the cleaning process, to ensure data consistency, feature markers that have no practical significance and do not contribute significantly to fault detection, such as machine numbers and unified specification codes, are deleted. In order to eliminate the impact of noise on model performance, abnormal data with fault markers marked as "1" but fault categories marked as "normal" are excluded. Some of the converted data are shown in Table 1.

Table 1. Example of feature transformation after data preprocessing.

Factory floor temperature	Equipment temperature	Rotation speed	Tor -que	usage time	Fault occurrence	Fault type	quality level_H	quality level_L	quality level_M
295.8	306.3	1235	76.2	89	1	2	0	0	1
295.7	306.2	2270	14.6	149	1	2	0	1	0
296.3	307.1	1534	33.8	151	0	1	0	0	1
296.3	307.1	1774	25.9	154	0	1	1	0	0
296.2	307	2119	18.3	159	0	1	0	0	1
296.2	307	1414	48.3	162	0	1	0	1	0
296.1	307	1523	42	164	0	1	0	0	1
296.1	307.1	1651	35.7	167	0	1	0	1	0
296.1	307.1	1485	36	169	0	1	0	0	1

Continued

296.2	307.2	1168	63.4	172	0	1	0	0	1
296.3	307.3	1566	35.8	175	0	1	0	1	0
296.3	307.2	1286	51.1	177	0	1	0	0	1

In addition, a visual analysis using the local outlier factor (LOF) algorithm was conducted on the factory temperature, equipment temperature, rotational speed, and torque to visualize the distribution characteristics and detect anomalies. The detection results are shown in Figure 1.

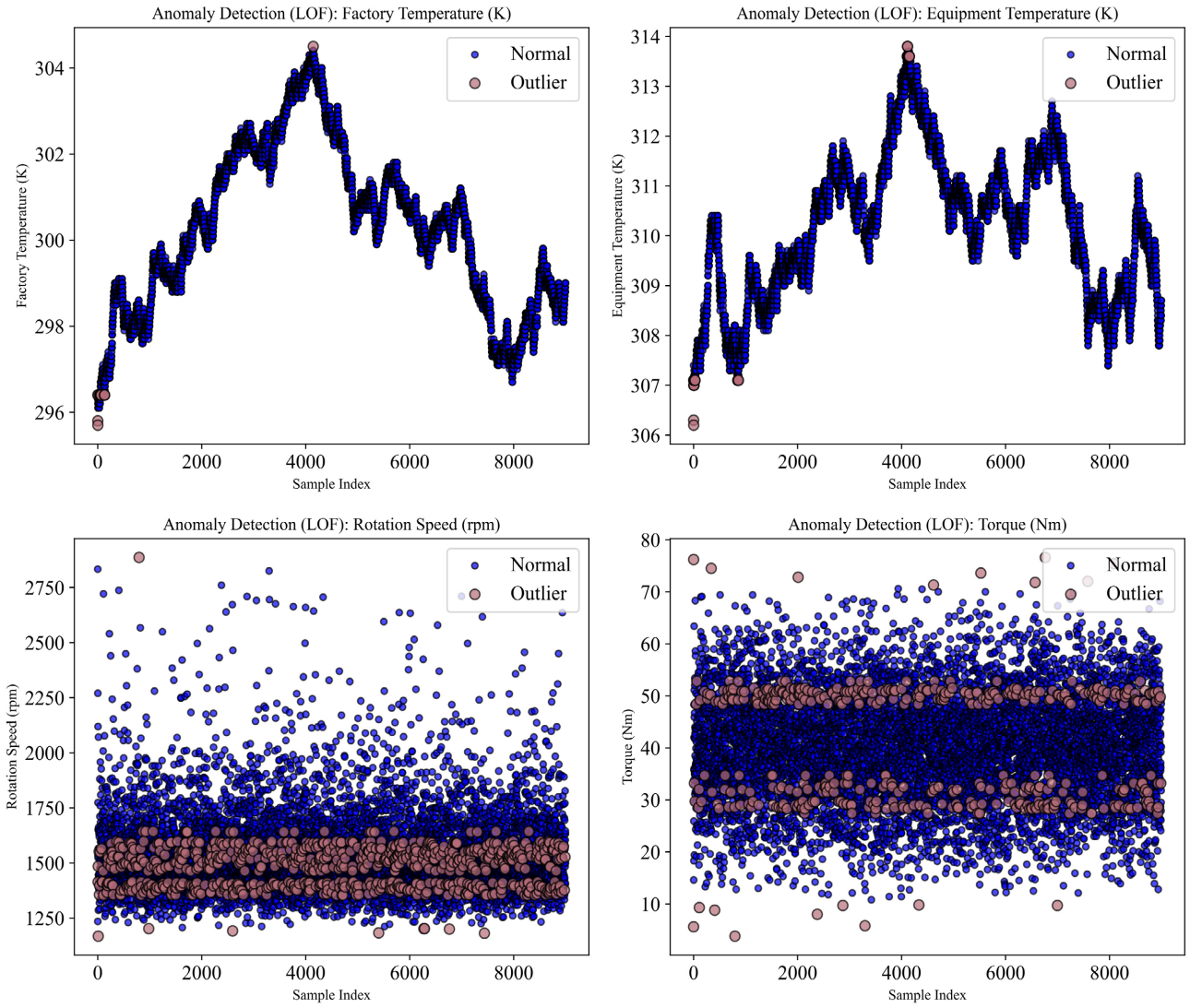

Figure 1. Data distribution and anomaly detection results.

The abnormal points are mainly concentrated in the rotational speed and torque features. Considering the differences in operating status and workload of different mechanical equipment, these anomalies may result from variations in equipment performance or actual working conditions. To ensure the integrity and

authenticity of the data, this study retains these anomalies to allow the model to more accurately reflect actual operating conditions.

3.2. Evaluation Metrics and Feature Analysis

To evaluate the performance of the fault detection and classification models, this study uses the following common classification evaluation metrics: accuracy, precision, recall, the F1 score, and AUC.

Accuracy measures the overall correctness of the model's detections and is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{9}$$

in Equation (9), TP denotes the number of samples correctly classified as positive, FN denotes the number of positive samples incorrectly classified as negative, FP denotes the number of negative samples incorrectly classified as positive, and TN denotes the number of samples correctly classified as negative.

Precision focuses on the model's false alarm rate, measuring the proportion of samples predicted as positive that are actually positive, as

$$P = \frac{TP}{TP + FP}, \tag{10}$$

recall measures the model's ability to identify actual positive samples and is defined as

$$R = \frac{TP}{TP + FN}, \tag{11}$$

when the dataset exhibits class imbalance, the F1 score provides a more balanced measure of model performance. It is calculated as the harmonic mean of precision and recall, as follows

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}, F_1 = \frac{2TP}{2TP + FP + FN}. \tag{12}$$

In addition, the AUC (Area Under the ROC Curve) represents the model's overall performance across different classification thresholds. The AUC value closer to 1 indicates better model performance. By comprehensively evaluating the above indicators, the reliability and effectiveness of the model in fault detection and fault identification tasks are demonstrated.

To identify relevant features, a statistical analysis was conducted. Both machine quality grade and fault occurrence are categorical variables; their association was examined using a chi-square test of independence. The results yielded a Pearson chi-square value of 13.018 with a p-value of 0.001, indicating a significant relationship.

For the continuous variables—plant room temperature, equipment temperature, rotation speed, torque, and operating time—one-way ANOVA was applied to assess their correlation with fault occurrence. As presented in **Table 2**, all variables showed p-values less than 0.05, suggesting significant group differences and

relevance to fault behavior.

Table 2. ANOVA table of each indicator and fault occurrence.

Features Variables	Mean Squared Between (MS)	Mean Squared Error (MSE)	F	P-value
Factory Temperature	217.303	3.572	60.832	0
Equipment Temperature	20.913	2.01	10.407	0.001
Rotation Speed	747294.546	32193.695	23.212	0
Torque	34100.022	96.453	353.541	0
Usage Time	408749.27	4003.179	102.106	0

Based on these findings, six key features were selected for modeling: machine quality grade, plant temperature, equipment temperature, rotation speed, torque, and operating time.

4 Experimental Results Analysis

4.1 Model Performance Evaluation and Comparative Analysis

The initial feature selection was guided by both domain expertise and statistical significance, using chi-square tests for categorical variables and ANOVA for continuous variables. Before constructing the fault detection and classification models, we examined potential multicollinearity among the selected features. Pearson correlation coefficients were computed, and a heatmap was generated to visualize variable relationships. As shown in Figure 2, the results indicate a strong positive correlation between equipment temperature and factory temperature, with a correlation coefficient of 0.86. Additionally, a strong negative correlation is observed

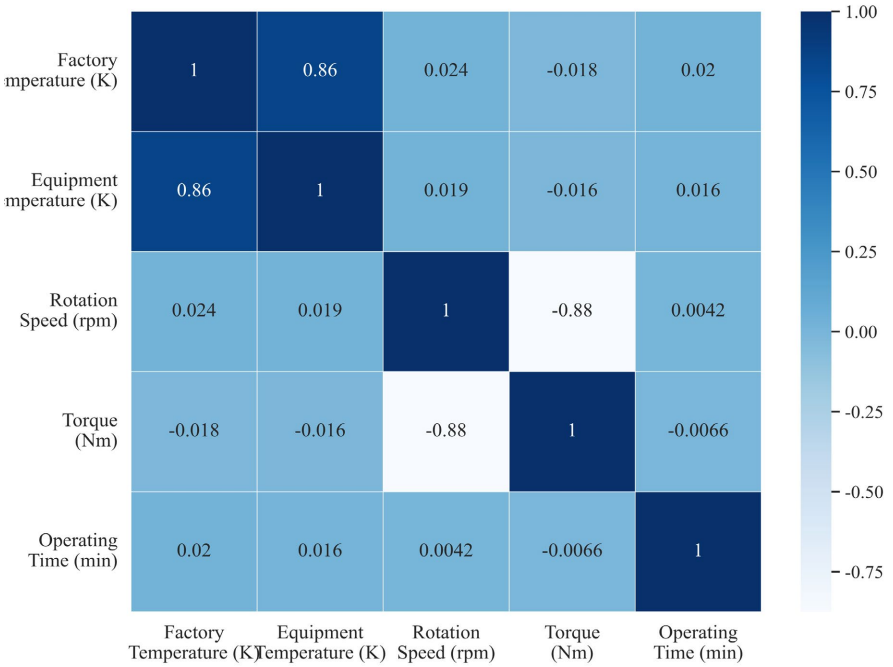

Figure 2. Heatmap of feature variable correlations.

between speed and torque, with a correlation coefficient of −0.88. To reduce the potential adverse effects of multicollinearity—such as instability in parameter estimation and reduced predictive performance—plant temperature and torque were excluded from model training. The remaining variables, equipment temperature and rotation speed, were retained along with the other non-redundant features to ensure model robustness.

Further analysis revealed a significant class imbalance in the sample data. In the binary variable representing "whether a fault occurred," normal operation data accounted for 96.65%, while fault data constituted only 3.35%. Regarding specific fault categories, "Normal" represented 96.65%, while the proportions of other fault categories were: TWF (0.46%), HDF (1.06%), PWF (0.82%), OSF (0.94%), and RNF (0.07%). Class imbalance can lead the model to favor the majority class during training, resulting in the neglect of minority class samples, which negatively impacts model robustness and classification performance. To address this issue, this study applied a class weight adjustment strategy when using the SVM and CART models, assigning higher weights to the minority class to mitigate the effects of data imbalance on model performance. During model training, the dataset was divided into a 70% training set and a 30% test set. Fault detection and classification models were established using the SVM and CART algorithms, respectively. Model performance was comprehensively evaluated based on indicators including F1 score, AUC, accuracy, recall, and precision. After training, the evaluation metrics were calculated using the test set, and the results are presented in Table 3.

Table 3. Model evaluation metrics.

Fault Classification Model	F1	AUC	Accuracy	Recall	Precision
SVM Model	0.86	0.9258	0.79	0.79	0.97
CART Model	0.90	0.9158	0.86	0.86	0.97

From the results in Table 3, it can be observed that the SVM model slightly outperforms the CART model in terms of AUC, indicating a marginally stronger ability to distinguish between positive and negative samples. However, the CART model demonstrates superior performance in terms of F1 score, accuracy, and recall, particularly excelling in the fault detection task by achieving higher accuracy and robustness. Specifically, the F1 score of the CART model reaches 0.90, representing a significant improvement compared to the SVM model's score of 0.86. Both the recall and accuracy of the CART model are also 0.86, further validating its effectiveness in fault detection. Additionally, although the precision of the SVM model is comparable to that of the CART model, its lower recall results in a reduced overall F1 score. In fault detection scenarios, timely and accurate identification of fault samples is critical. A higher recall rate plays a key role in reducing the false negative rate, making it particularly important. Considering this, the CART model was ultimately chosen as the primary model for fault classification

in this study to ensure more reliable fault detection in practical applications. Figure 3 further illustrates the ROC curves of the SVM and CART models, providing a visual comparison of their classification performance.

However, in the identification of specific fault categories, the CART model may suffer from model complexity and overfitting in multi-class classification scenarios. To address this, this study further employs the random forest model for detailed fault classification. As an ensemble learning method, the random forest model enhances the stability and generalization ability of the model by training multiple decision trees and applying a voting mechanism. Additionally, it offers robust feature importance analysis capabilities, which aid in identifying the key factors contributing to faults. Therefore, this study conducts detailed fault classification and feature importance analysis using the random forest model.

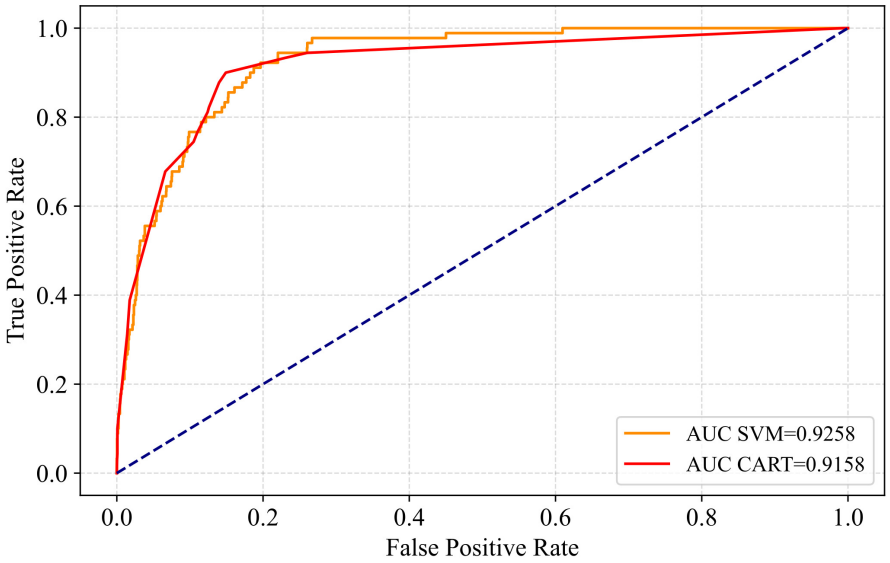

Figure 3. ROC curve Comparison of SVM and cart models.

4.2. Feature Importance Analysis and Fault Cause Exploration

Considering that the random forest model is insensitive to multicollinearity, all selected feature variables were input into the model for training to maximize the use of sample information. The model's performance was then evaluated using the test set. The results indicate that the random forest model performed well in fault detection and classification tasks. Specifically, the model achieved an F1 score of 0.88, an accuracy of 0.82, a precision of 0.97, and a recall of 0.82, demonstrating a well-balanced trade-off between precision and recall.

To further investigate the main causes of different fault types, this study analyzed the mean values of feature attributes corresponding to each fault category to identify variations in feature variables across different fault types. The main feature means for each fault category are presented in Table 4.

The following patterns can be observed from Table 4: (1) Correlation between speed and fault type: The average speed of power faults (PWF) is higher, whereas

Table 4. Mean values of key features for different fault types.

Features Variables	Normal	PWF	OSF	TWF	RNF	HDF
Factory Temperature	300.2	300.4	300.3	300.6	301.4	302.6
Equipment Temperature	310.2	310.1	310.3	310.4	310.7	310.8
Rotation Speed	1540.9	1802.8	1346.3	1573.1	1498.2	1339.2
Torque	39.6	46.6	58.4	38	42.2	52.2
Usage Time	106.6	89.1	206.9	216.7	107	101.7

the average speed of heat dissipation faults (HDF) is significantly lower. (2) Impact of torque: Overload faults (OSF) exhibit the highest average torque, reaching 58.4 Nm, while normal equipment has the lowest average torque at only 39.6 Nm. (3) Effect of usage time: Wear faults (TWF) and overload faults (OSF) are generally associated with longer usage times, further confirming the significant correlation between equipment aging or excessive use and fault occurrence.

Figure 4 illustrates the distribution of various types of faults across different machine quality levels. As shown in the figure, regardless of the fault type, equipment with quality level L exhibits a significantly higher probability of fault compared to equipment with quality levels M and H. Notably, in the case of TWF faults, although higher-quality equipment experiences fewer faults, the risk of fault increases with longer usage time. Even for equipment with higher quality levels, the likelihood of fault becomes substantial as operational time extends.

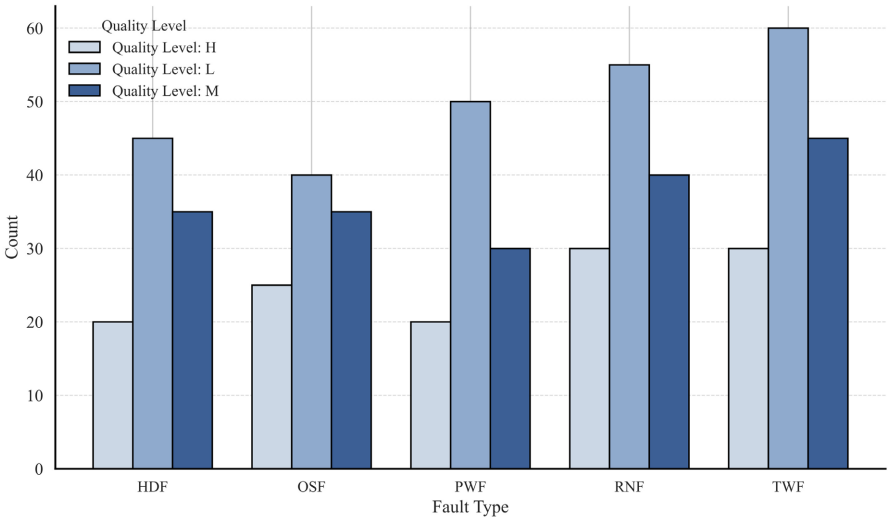

Figure 4. Distribution of fault types across different machine quality levels.

To further quantify the impact of each feature variable on fault occurrence, this study calculated feature importance using the random forest model. Figure 5. presents the feature importance rankings, revealing the following insights: 1) Usage time is the most important feature variable, with an impact on fault occurrence reaching 34.86%. This indicates that the likelihood of fault increases significantly as the equipment's usage time extends. 2) Torque and rotation speed have im-

portance scores of 30.23% and 14.06%, respectively, suggesting that these two variables are closely associated with the equipment's load and operating status. 3) Factory temperature and equipment temperature contribute to fault occurrence with importance scores of 8.84% and 7.45%, respectively, indicating that temperature has a moderate influence on equipment faults. 4) Machine quality grade has the lowest importance score, suggesting that in the same working environment, its impact on fault occurrence is relatively minor.

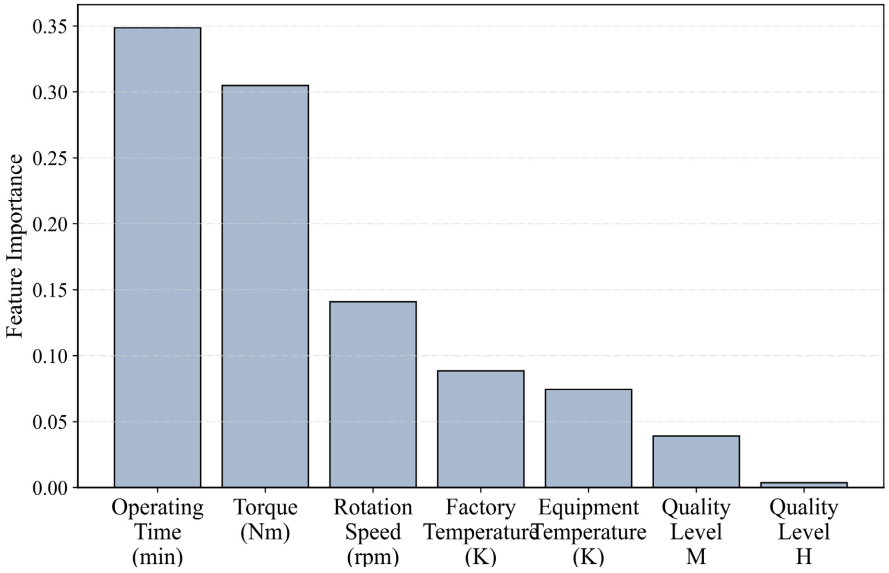

Figure 5. Feature importance ranking.

To further identify differentiated features, the feature distributions of different fault categories were analyzed. **Figure 6**. presents the box plots of factory temperature, equipment temperature, rotation speed, torque, and usage time.

The following insights can be drawn from **Figure 6**: 1) HDF: The temperature distribution of heat dissipation faults is relatively concentrated, with the minimum temperature significantly higher than the median of other fault types. Notably, the probability of heat dissipation fault increases sharply when the equipment temperature exceeds 310K. 2) PWF: PWF also exhibit a concentrated temperature distribution and typically occur in high-temperature environments. Special attention should be paid to the risk of power faults when the equipment temperature is elevated. 3) Relationship Between Rotation Speed and Torque: HDF and OSF are more likely to occur when the speed falls below 1500 rpm or the torque exceeds 40 Nm. In contrast, PWF are more common at extremely high speeds or exceptionally low torque levels. 4) Impact of Usage Time: OSF and TWF predominantly occur during extended usage periods. Therefore, regular maintenance and monitoring of long-running equipment can effectively reduce fault rates. This chapter analyzed the model's performance and fault causes, identifying key factors influencing faults and providing support for equipment maintenance and detection.

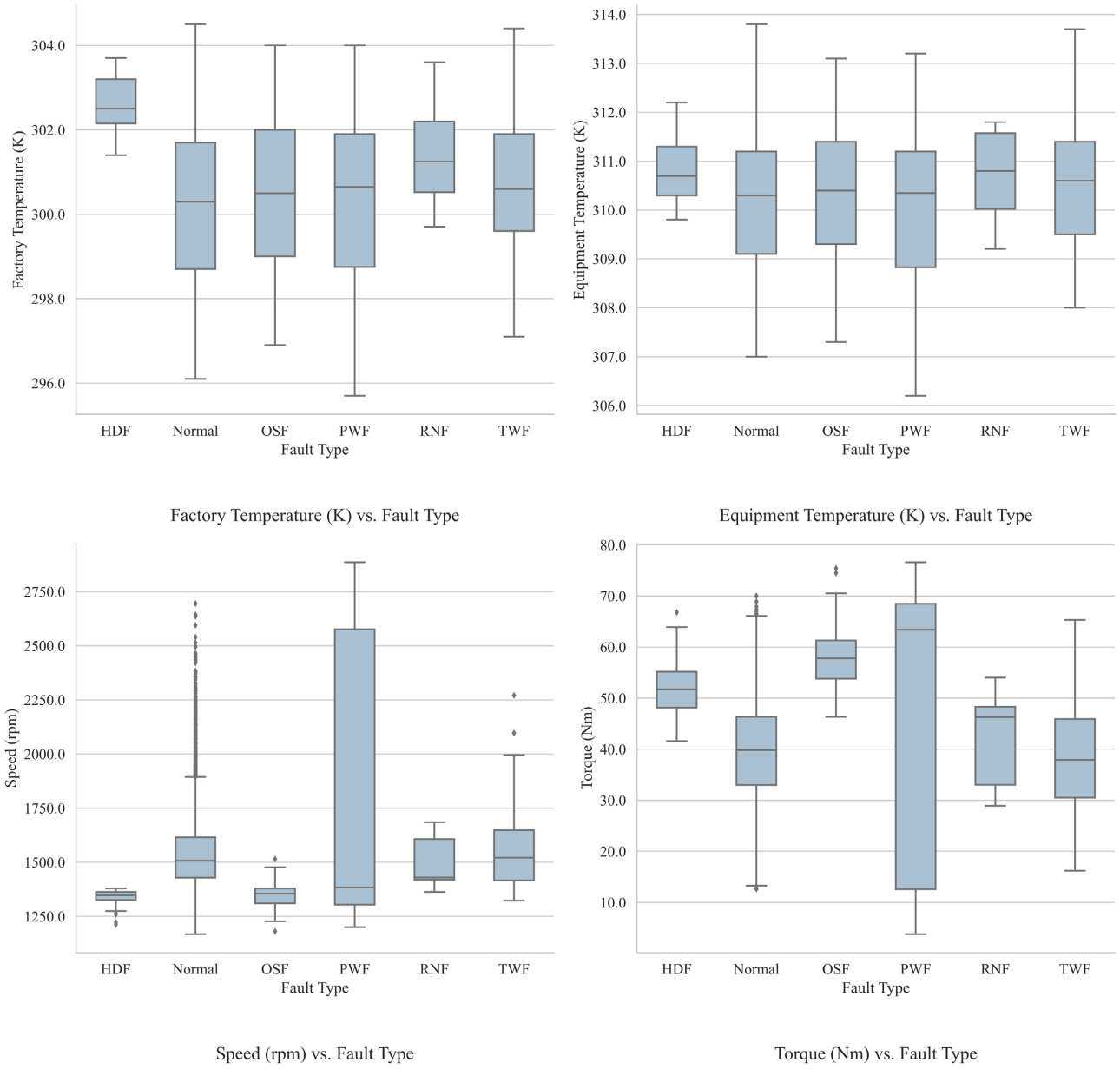

Figure 6. Box plot of feature distributions for different fault types.

5. Conclusion

This study investigates the fault detection and classification of industrial mechanical equipment using machine learning methods. Fault detection and multi-class classification models were developed based on SVM, CART, and Random Forest, and their performance was evaluated through comprehensive experiments. The results indicate that the models exhibit reasonable accuracy and robustness, and offer practical reference value for fault diagnosis and maintenance decisions in industrial scenarios. Feature importance analysis further reveals the contribution of key variables such as equipment temperature, rotation speed, and torque to various fault types. To further enhance the adaptability and scalability of the pro-

posed approach, subsequent research will consider evaluation across more diverse and heterogeneous datasets, enabling broader applicability in varied industrial settings. This may also provide a foundation for comparing traditional machine learning methods with deep learning models, offering deeper insights into their relative strengths under different data conditions.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Qi, G., Zhu, Z., Erqinhu, K., Chen, Y., Chai, Y. and Sun, J. (2018) Fault-Diagnosis for Reciprocating Compressors Using Big Data and Machine Learning. *Simulation Modelling Practice and Theory*, **80**, 104-127. https://doi.org/10.1016/j.simpat.2017.10.005

[2] Carvalho, T.P., Soares, F.A.A.M.N., Vita, R., Francisco, R.D.P., Basto, J.P. and Alcalá, S.G.S. (2019) A Systematic Literature Review of Machine Learning Methods Applied to Predictive Maintenance. *Computers & Industrial Engineering*, **137**, Article ID: 106024. https://doi.org/10.1016/j.cie.2019.106024

[3] Cohen, J., Jiang, B. and Ni, J. (2021) Machine Learning for Diagnosis of Event Synchronization Faults in Discrete Manufacturing Systems. *Journal of Manufacturing Science and Engineering*, **144**, Article ID: 071006. https://doi.org/10.1115/1.4052762

[4] Gao, S., Xu, L., Zhang, Y. and Pei, Z. (2020) Rolling Bearing Fault Diagnosis Based on Intelligent Optimized Self-Adaptive Deep Belief Network. *Measurement Science and Technology*, **31**, Article ID: 055009. https://doi.org/10.1088/1361-6501/ab50f0

[5] Cao, H., Shao, H., Zhong, X., Deng, Q., Yang, X. and Xuan, J. (2022) Unsupervised Domain-Share CNN for Machine Fault Transfer Diagnosis from Steady Speeds to Time-Varying Speeds. *Journal of Manufacturing Systems*, **62**, 186-198. https://doi.org/10.1016/j.jmsy.2021.11.016

[6] Ding, Y., Jia, M., Miao, Q. and Cao, Y. (2022) A Novel Time-Frequency Transformer Based on Self-Attention Mechanism and Its Application in Fault Diagnosis of Rolling Bearings. *Mechanical Systems and Signal Processing*, **168**, Article ID: 108616. https://doi.org/10.1016/j.ymssp.2021.108616

[7] Amin, A., Bibo, A., Panyam, M. and Tallapragada, P. (2022) Vibration Based Fault Diagnostics in a Wind Turbine Planetary Gearbox Using Machine Learning. *Wind Engineering*, **47**, 175-189. https://doi.org/10.1177/0309524x221123968

[8] Zhang, H., Pan, C., Wang, Y., Xu, M., Zhou, F., Yang, X., *et al.* (2022) Fault Diagnosis of Coal Mill Based on Kernel Extreme Learning Machine with Variational Model Feature Extraction. *Energies*, **15**, Article 5385. https://doi.org/10.3390/en15155385

[9] Okwuosa, C.N. and Hur, J. (2022) A Filter-Based Feature-Engineering-Assisted SVC Fault Classification for SCIM at Minor-Load Conditions. *Energies*, **15**, Article 7597. https://doi.org/10.3390/en15207597

[10] Purbowaskito, W., Lan, C. and Fuh, K. (2024) The Potentiality of Integrating Model-Based Residuals and Machine-Learning Classifiers: An Induction Motor Fault Diagnosis Case. *IEEE Transactions on Industrial Informatics*, **20**, 2822-2832. https://doi.org/10.1109/tii.2023.3299111

[11] Tao, Y., Yu, T. and Yang, J. (2024) Photovoltaic Array Fault Diagnosis and Localization Method Based on Modulated Photocurrent and Machine Learning. *Sensors*, **25**, Article 136. https://doi.org/10.3390/s25010136

[12] Vashishtha, G., Chauhan, S., Sehri, M., Zimroz, R., Dumond, P., Kumar, R., *et al.* (2025) A Roadmap to Fault Diagnosis of Industrial Machines via Machine Learning: A Brief Review. *Measurement*, **242**, Article ID: 116216. https://doi.org/10.1016/j.measurement.2024.116216

[13] Muzzammel, R. (2025) Comprehensive Exploration of Limitations of Simplified Machine Learning Algorithm for Fault Diagnosis under Fault and Ground Resistances of Multiterminal High-Voltage Direct Current System. *Journal of Sensor and Actuator Networks*, **14**, Article 29. https://doi.org/10.3390/jsan14020029

[14] Wu, W. (2025) Automotive Motor Fault Diagnosis Model Integrating Machine Learning Algorithm and Fuzzy Control Theory. *International Journal of Fuzzy Systems*.

[15] Duan, L., Xie, M., Bai, T. and Wang, J. (2016) A New Support Vector Data Description Method for Machinery Fault Diagnosis with Unbalanced Datasets. *Expert Systems with Applications*, **64**, 239-246. https://doi.org/10.1016/j.eswa.2016.07.039

[16] Jablon, L.S., Avila, S.L., Borba, B., Mourão, G.L., Freitas, F.L. and Penz, C.A. (2020) Diagnosis of Rotating Machine Unbalance Using Machine Learning Algorithms on Vibration Orbital Features. *Journal of Vibration and Control*, **27**, 468-476. https://doi.org/10.1177/1077546320929830

[17] Huang, Y., Xu, Z., Cao, L., Hu, H. and Tang, G. (2022) Fractional Dimensionless Indicator with Random Forest for Bearing Fault Diagnosis under Variable Speed Conditions. *Shock and Vibration*, **2022**, Article ID: 1781340. https://doi.org/10.1155/2022/1781340

[18] Jin, Y., Qin, C., Zhang, Z., Tao, J. and Liu, C. (2022) A Multi-Scale Convolutional Neural Network for Bearing Compound Fault Diagnosis under Various Noise Conditions. *Science China Technological Sciences*, **65**, 2551-2563. https://doi.org/10.1007/s11431-022-2109-4

[19] Han, S., Shao, H., Cheng, J., Yang, X. and Cai, B. (2023) Convformer-NSE: A Novel End-To-End Gearbox Fault Diagnosis Framework under Heavy Noise Using Joint Global and Local Information. *IEEE/ASME Transactions on Mechatronics*, **28**, 340-349. https://doi.org/10.1109/tmech.2022.3199985

[20] Mian, T., Choudhary, A. and Fatima, S. (2023) Multi-Sensor Fault Diagnosis for Misalignment and Unbalance Detection Using Machine Learning. *IEEE Transactions on Industry Applications*, **59**, 5749-5759. https://doi.org/10.1109/tia.2023.3286833

[21] Prawin, J. (2024) Deep Learning Neural Networks with Input Processing for Vibration-Based Bearing Fault Diagnosis under Imbalanced Data Conditions. *Structural Health Monitoring*, **24**, 883-908. https://doi.org/10.1177/14759217241246508