Application of Transformer Based on Escape Optimization Algorithm to the Faulty Bearing Classification Problem

Zhangqi Song *, Chenlu Zhao

School of Electronic Information, Huzhou College, Huzhou, China, 313000

* Corresponding Author Email: szq_731223@163.com

Abstract. Deep learning has shown strong feature extraction and classification abilities in fault diagnosis, especially for vibration signal time series. The Western Reserve bearing dataset is a common benchmark for evaluating diagnostic models. This paper proposes a Transformer model optimized with the Escape Algorithm (ESC) to enhance bearing fault feature extraction and classification accuracy. Innovations include a multi-scale convolution-attention fusion module for better local temporal feature extraction and the first use of ESC for Transformer hyperparameter optimization. Experiments on the Western Reserve dataset show the model achieves 96% accuracy on both training and test sets, with improved robustness in distinguishing ambiguous categories. The results demonstrate the model's effectiveness for complex fault recognition, offering valuable insights for intelligent manufacturing and equipment health monitoring.

Keywords: Escape Optimization Algorithm, Bearing Fault Detection, Transformer, Deep Learning, Feature Extraction.

1. Introduction

Rolling bearings are critical components in mechanical transmission systems, responsible for load support and transfer. Their performance is affected by working conditions and load states, making accurate fault diagnosis essential for ensuring equipment reliability and extending service life [1]. In the context of Industry 4.0, intelligent predictive maintenance has been widely adopted in developed countries. For example, in wind power systems—an important part of green energy—the stable operation of wind turbines is vital to power grid safety. Bearings in wind turbines operate under high loads and harsh environments, making them susceptible to fatigue, cracking, and wear. Failures can lead to downtime and significant economic losses.

With the rise of machine learning and deep learning, data-driven bearing fault diagnosis has become a key tool for monitoring critical equipment [2]. In China, national strategies such as "intelligent manufacturing" have further increased the demand for intelligent maintenance systems. Traditional empirical methods are being replaced by intelligent diagnostic models, driving fault diagnosis toward high accuracy, real-time performance, and automation.

However, rolling bearings often experience simultaneous failures in components such as the inner ring, outer ring, rolling elements, and cage. Tang et al. highlighted the difficulty in decoupling such coupled fault signals [3]. Liu Xia noted that rare fault types demand high model sensitivity [4]. Research by Qiu Dawei and Chen Song emphasized that Transformer structures, particularly when combined with optimization algorithms, can significantly influence fault classification performance [1, 5]. While Transformer-based models have achieved promising results, challenges such as low accuracy, poor category differentiation, and overfitting remain. For instance, Deng Feiyue's fusion of CNN and Transformer improved performance but suffered from excessive convolutional output scale [6]. Chen Wenqiang also reported that although Transformer performs well in classification, it still faces issues like iteration defects and overfitting [6]. Furthermore, most existing research focuses on single-fault diagnosis, with limited work on composite fault diagnosis [7].

Wind turbine bearing faults are particularly difficult to detect at an early stage due to weak signals and environmental noise. Therefore, this study proposes a hybrid model combining the deep representation ability of Transformer with the local feature extraction of CNNs [4], and introduces an escape behavior-based optimization algorithm to enhance model generalization and convergence. This approach aims to provide an accurate and robust solution for wind turbine bearing fault diagnosis.

2. Methods

2.1. Transformer model

Transformer is a deep learning architecture mainly used for NLP and sequence-to-sequence tasks. Unlike RNNs or LSTMs, it relies solely on the self-attention mechanism to model relationships within sequences, offering stronger parallel computation and better handling of long-range dependencies. The model consists of an Encoder, which extracts contextual features from the input, and a Decoder, which generates the target sequence by attending to the Encoder output. Typically, the Transformer stacks multiple identical layers, each containing a Multi-Head Self-Attention Mechanism and a Feed-Forward Neural Network.

Zhao X et al. demonstrated that combining Transformer with other neural networks enhances fault classification performance [8]. The self-attention mechanism dynamically assigns weights to different positions for global feature modeling, while the multi-head structure captures diverse semantic relationships through parallel attention heads. This study further improves feature extraction by adding convolutional layers alongside the self-attention mechanism [4].

2.2. Escape Algorithm

Escape Algorithm (ESC) is a meta-heuristic optimization algorithm inspired by crowd evacuation behavior, which was proposed by Kaichen Ouyang et al. Its core formulas revolve around different stages and operations of the algorithm, and each variable in the formulas has a specific meaning. These formulas and variables cooperate with each other to simulate the behavior of the crowd during the evacuation process, so as to realize the solution and optimization of complex optimization problems [9].

2.2.1 Individual initialization formula

The initialization phase determines the starting positions of individuals in the population. Each individual's position is generated within the defined boundaries of the search space using the following formula:

$$x_{i,j} = lb_j = r_{i,j} * \left(ub_j - lb_j\right), \quad r_{i,j} \sim U(0,1) \tag{1}$$

Where $x_{i,j}$ represents the position of the ith data in the jth dimension; and b_j are the lower and upper bounds of the jth dimension, respectively; and $r_{i,j}$ is a random variable uniformly distributed between 0 and 1, which determines the randomness of the initialization position of an individual.

2.2.2 Exploratory phase formula

During the exploratory phase, different strategies are used to update individuals based on their group classification: calm, follower, or panic.

The positions of individuals in the calm group are updated to simulate exploration within the search space. The update mechanism is defined as:

$$x_{i,j}^{new} = x_{i,j} + m_1 * (\omega_1 * (C_j - x_{i,j}) + v_{c,j}) * P(t) \tag{2}$$

Where $x_{i,j}^{new}$ is the position of the ith individual in the jth dimension after the update; m_1 is a binary variable generated from the Bernoulli distribution that determines whether some of the dimensions are updated or not, with equal probability of taking the value of 0 or 1; ω_1 is the adaptive Levy weights that are used to simulate the step size during the exploration phase; and is the center position of the calm group in the jth dimension, computed by taking the mean value of all the calm individuals of that dimension; Equation (3) is used to update the positions of individuals in the follower group.

$$v_{c,j} = R_{c,j} - x_{i,j} + \varepsilon_j \tag{3}$$

Equation (4) is used to update the positions of individuals in the panic group.

$$x_{i,j}^{new} = x_{i,j} + m_1 * (\omega_1 + \left(C_j - x_{i,j}\right) + m_2 * \omega_2 * \left(x_{p,j} - x_{i,j}\right) + v_{h,j} * P(t)) \tag{4}$$

m_2 are also binary variables generated by the same mechanism; ω_2 is the adaptive Levy weight for the follower group; $x_{p,j}$ is an individual randomly selected from the panic group, $v_{h,j} = R_{h,j} - x_{i,j} + \varepsilon_j$ represents a potential direction of panic-driven movement; $R_{h,j} = r_{min,j}^{h} + r_{i,j} * (r_{max,j}^{h} - r_{min,j}^{h})$ represents a randomly generated position within the follower group.

$$x_{i,j}^{new} = x_{i,j} + m_1 * (\omega_1 * (E_j - x_{i,j}) + m_2 * \omega_2 * (x_{rand,j} - x_{i,j}) + v_{p,j} * P(t)) \tag{5}$$

E_j is a randomly selected individual from the elite pool representing a possible exit to which a panicked individual might travel; $x_{rand,j}$ is a randomly selected individual from the population, introducing randomness to panic-driven movements:

$$v_{p,j} = R_{p,j} - x_{i,j} + \varepsilon_j \tag{6}$$

$R_{p,j} = r_{min,j}^{p} + r_{i,j} \times (r_{max,j}^{p} - r_{min,j}^{p})$ are randomly generated locations within the panic group, $r_{max,j}^{p}$ and $r_{min,j}^{p}$ are the minimum and maximum values in the jth dimension for all individuals in the panic group.

2.2.3 Development phase formulas

During the development phase, all individuals are considered as calm individuals, and (7) is used to update the individual locations.

$$x_{i,j}^{new} = x_{i,j} + m_1 * (\omega_1 * (E_j - x_{i,j}) + m_2 * \omega_2 * (x_{rand,j} - x_{i,j}) + v_{p,j} * P(t)) \tag{7}$$

At this point, individuals optimize their position by moving closer to members of the elite pool and randomly selected individuals to simulate the convergence of the population towards the optimal exit.

2.2.4 Adaptive Levy Weights

Levy flight is a stochastic process characterized by a "heavy tail" distribution. Unlike normal distributions, it enables longer jumps in the search space, enhancing global search capability and helping optimization algorithms avoid local optima. By introducing Levy weights, the algorithm improves the accuracy and robustness of the search process.

In summary, this chapter focuses on model construction, detailing the Transformer architecture and its core mechanisms, and introduces the ESC algorithm as a hyperparameter optimization strategy. The Transformer provides strong global modeling and parallel computing capabilities, while the added convolutional module enhances local feature extraction. The ESC algorithm simulates crowd evacuation behavior, incorporates dynamic multi-stage updates, and integrates Levy flight to maintain a balance between global exploration and local exploitation. Both experiments and literature show that combining Transformer with optimization algorithms significantly improves fault detection performance, especially in complex multi-class scenarios. This joint modeling approach offers a solid foundation for the following experimental validation and performance analysis.

3. Results and discussion

3.1. Data characterization

The Western Reserve dataset, provided by Case Western Reserve University in the U.S., is a widely used public dataset in the fields of rotating machinery fault detection, vibration signal analysis, and bearing fault diagnosis. It serves as a benchmark for evaluating deep learning-based fault classification models. The dataset includes multiple types of vibration signals recorded under various fault conditions and sensor placements. As shown in Table 1, the data fields include drive-end (DE), fan-end (FE), and base (BA) vibration signals, along with time series (Time) and rotational speed (RPM) data. These indicators capture different aspects of mechanical behavior, where the DE, FE, and BA reflect vibration signals collected from different machine locations. RPM, when divided by 60, provides the frequency information essential for spectral analysis.

Table 1. Western Reserve data names and meanings

Data name	Data meaning
DE	Drive-side vibration data
FE	Fan end vibration data
BA	Base vibration data
Time	Time series data
RPM	Divide the unit revolutions per minute by 60 to get the frequency

Namely, the selected indicators are rolling body failure, inner ring failure, outer ring failure, rolling body failure, inner ring failure, outer ring failure are taken as 0.007, 0.014, 0.021 inches, respectively, while the root mean square (RMS) value is selected: reflecting the energy magnitude of the signal.

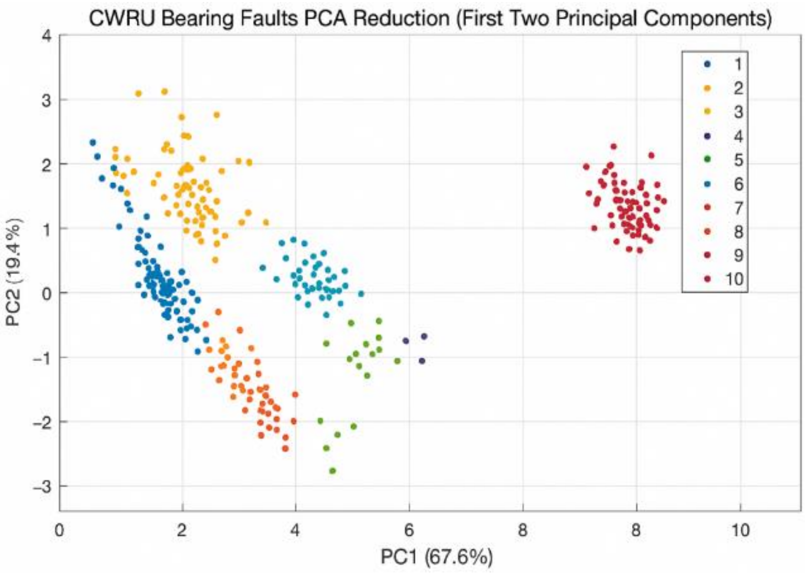

Figure 1. Bearing failure data downscaling

Figure 1 illustrates the first two principal component distributions of the CWRU (bearing failure dataset after PCA (Principal Component Analysis) dimensionality reduction. The figure contains 10 clusters of different colored points representing 10 different types of bearing failures. The horizontal axis represents the first principal component (PC1), which explains 67.6% of the data variance, and the vertical axis represents the second principal component (PC2), which explains 19.4% of the data variance. In total, the two principal components explained about 87% of the total data variance.

Table 2. Highly separable categories

Corresponding category	PC1	PC2
Category 7	6.5	2.25
Category 4	2	-1.5
Category 10	-3	2

Based on the results shown in Figure 1 and Table 2, several issues can be observed in the CWRU bearing failure dataset after PCA dimensionality reduction. Firstly, there exists a problem of high-dimensional redundancy. The first two principal components explain 87% of the total variance, indicating strong correlations among the original features. Secondly, the distribution of categories is highly imbalanced. As illustrated in Figure 1, certain categories—such as Category 5—contain significantly fewer samples, whereas others, such as Categories 1 and 3, are more densely populated.

Thirdly, the degree of separability among categories varies considerably. As shown in Table 2, some categories are clearly distinguishable in the principal component space. For example, Category 7 (PC1 = 6.5, PC2 = 2.25) and Category 4 (PC1 = 2, PC2 = –1.5) are highly separable, which makes their classification more straightforward. In contrast, other categories, such as Categories 1 and 8, exhibit substantial overlap in the PCA space, resulting in blurred class boundaries. Moreover, as also evident from Table 2, certain categories like Category 10 (PC1 = –3, PC2 = 2) show inconsistent separability, further contributing to fuzzy classification boundaries and increased classification difficulty.

3.2. Distribution analysis of bearing failure characteristics and model corresponden

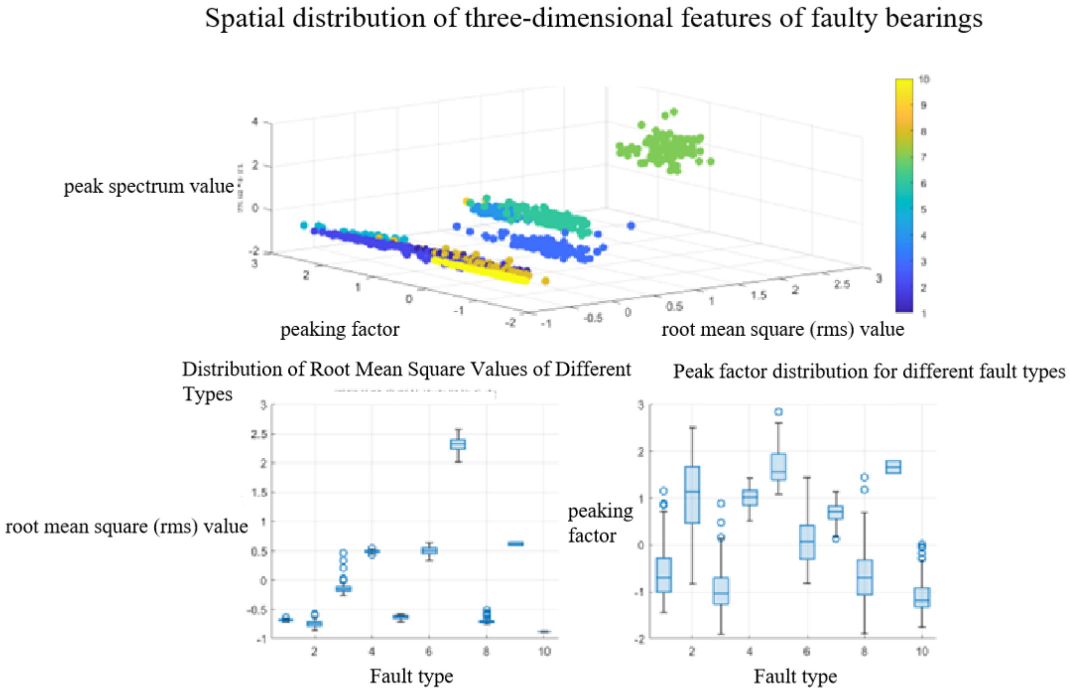

Figure 2. Three-dimensional diagram of bearing failure distribution with box diagrams

Figure 2 contains three subplots showing the characteristic distribution of bearing failures:These visualizations reveal the distribution pattern of bearing failure characteristics and the distinguishing properties of different failure types.

The top 3D scatterplot shows the distribution of the three key features (root mean square (RMS) value, peak factor, and frequency value). From the scatterplot, it can be found that the green category clusters (categories 6-7) are located in the region of higher frequency value (about 2-4), the yellow category clusters (categories 9-10) are linearly distributed along the peak factor axis, and the blue category clusters (categories 1-3) are distributed in the middle and low regions of peak factor and RMS value. Also the green category clusters are clearly separated from the other categories vertically; the yellow and blue category clusters have some overlap in the plane, which may explain the confusion between categories 1 and 8 in the confusion matrix.

Observation of the root mean square box plots in the lower left reveals that Type 7 is significantly different, with a concentrated distribution and narrow interquartile range, suggesting that the feature is highly stable, and that Type 2 and Type 10 have the lowest root mean square values (about -0.8 to -0.6).

The box plots on the lower right show the peak factor distribution for the different fault types, with Type 4 peak factors significantly higher than the others (about 1.5 to 2.0), with outliers up to about 2.8. There is a wide range of peak factor distributions for Type 9 and Type 2, with lower peak factors for Types 3 and 10 (about -1.0 to -1.2), and a lower but wider distribution for Type 1 (about -0.8 to -0.2).

3.3. Analysis of model results

The experimental process of the proposed model includes four stages: initialization, optimization, convergence, and result output.

In the initialization stage, individuals in the ESC algorithm are randomly encoded, each representing a set of Transformer hyperparameters such as the number of attention heads, learning rate, and convolution kernel size. An initial model is constructed to extract multi-scale features from the input samples, laying the foundation for subsequent evaluations.

During the optimization stage, the ESC algorithm iteratively improves model performance. Each individual's parameter configuration is evaluated based on recognition accuracy and boundary discrimination. The population is divided into calm, panic, and follower groups to guide both global and local search strategies. A multi-scale feature interaction mechanism is introduced between convolutional layers to enhance the model’s robustness, while boundary constraints and local perturbations are applied to ensure feasible parameter values and refine optimal solutions.

In the convergence stage, the best-performing individuals and their parameter combinations are recorded through an elite retention mechanism. Once the fitness stabilizes and the stopping criteria are met, the final Transformer configuration is determined. This configuration shows strong fault recognition and generalization in subsequent evaluations.

According to Table 3 and Figure 3, the model achieves high classification accuracy on the training set, with categories 2 to 7 and 9 to 10 reaching 100% accuracy. The overall accuracy is around 97.2%, consistent with the trend shown in the training accuracy curve. Sample distribution is generally balanced across categories, allowing for fair performance evaluation, with only minor discrepancies in sample counts for categories 1 and 8.

Confusion matrix for test set and training set

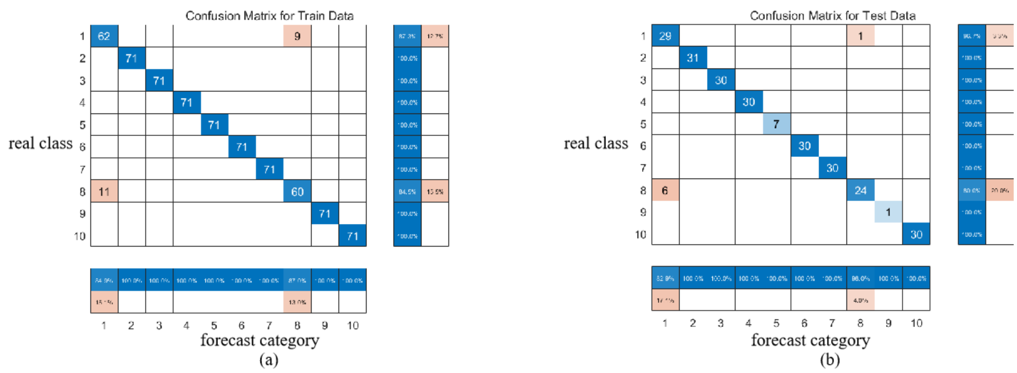

Figure 3. Confusion matrix for training and test sets

Table 3. Confusion matrix results for the training set

Form	Sample size	Recall rate (%)	Accuracy (%)
1	71	87.3	84.9
2	71	100.0	100.0
3	71	100.0	100.0
4	71	100.0	100.0
5	71	100.0	100.0
6	71	100.0	100.0
7	71	100.0	100.0
8	71	84.5	87.0
9	71	100.0	100.0
10	71	100.0	100.0

In the confusion matrix, misclassification is mainly concentrated between category 1 and category 8, showing the typical symmetric confusion phenomenon: 11 of the samples originally belonging to category 8 are misclassified as category 1, accounting for 15.5% of the category 8 samples, while 9 samples of category 1 are incorrectly predicted as category 8, accounting for 12.7% of their total number. As a result, the recognition accuracies of category 1 and category 8 are 87.3% and 84.5%, respectively, which are slightly lower among all categories. It is worth noting that except for these two categories, the rest of the samples in each category are almost confusion-free, and the classification performance is perfect. Further analyzed from the perspective of precision and recall: categories 2 to 7 and categories 9 and 10 achieve 100% recall and precision; while category 1 has 87.3% recall and 84.9% precision, and category 8 has 84.5% recall and 87.0% precision, all of which are due to the influence of mutual misclassification. It indicates that the main challenge is focused on the mutual confounding of category 1 and category 8, which is highly consistent with the previous validation set results, PCA analysis and feature distribution analysis.

From the overall performance of the two subgraphs in Table 4 and Figure 4, the model shows high classification ability on the training set: the total accuracy is 97.4648%, which is improved compared with the training accuracy of 97.1831% of the previous version of the model, indicating that the model architecture has been further optimized to obtain a stronger discriminative ability.

Meanwhile, the distribution of samples in each category in the training set is balanced, with each category containing 71 samples, and the total number of samples is 710.

Table 4. Test set confusion matrix results

Form	Sample size	Recall rate (%)	Accuracy (%)
1	30	96.7	82.9
2	31	100.0	100.0
3	30	100.0	100.0
4	30	100.0	100.0
5	7	100.0	100.0
6	30	100.0	100.0
7	30	100.0	100.0
8	30	80.0	96.0
9	1	100.0	100.0
10	30	100.0	100.0

The model achieves perfect classification accuracy for most categories. Categories 2 to 7, as well as Categories 9 and 10, all reach 100% accuracy, as indicated by the dark red diagonal cells with a value of 71 in the confusion matrix. However, notable confusion occurs between Category 1 and Category 8, which limits overall performance. Category 1 has an accuracy of 84.5%, with 11 samples misclassified as Category 8, while Category 8 reaches 90.1% accuracy, with 7 samples misclassified as Category 1. Although this bidirectional misclassification is not numerically symmetrical, it reflects a high degree of similarity between the two categories at the feature level and a strong tendency toward mutual confusion.

As illustrated in Figure 4, the model achieves an overall classification accuracy of 97.18%. The predicted outputs (blue curves) are closely aligned with the true labels (orange curves) across all 700 training samples, with only a few minor deviations. Most of these errors are concentrated in the range of samples from 300 to 400, which mainly involve boundary samples between Category 1 and Category 8. This observation is consistent with the feature space overlap revealed in the previous PCA analysis. Despite this localized deviation, the model exhibits strong learning ability and effective fault feature recognition, highlighting the impact of the multiscale feature extraction module and the escape optimization algorithm. Future research could further enhance performance on boundary cases by introducing techniques such as the focal loss function or hard sample mining strategies.

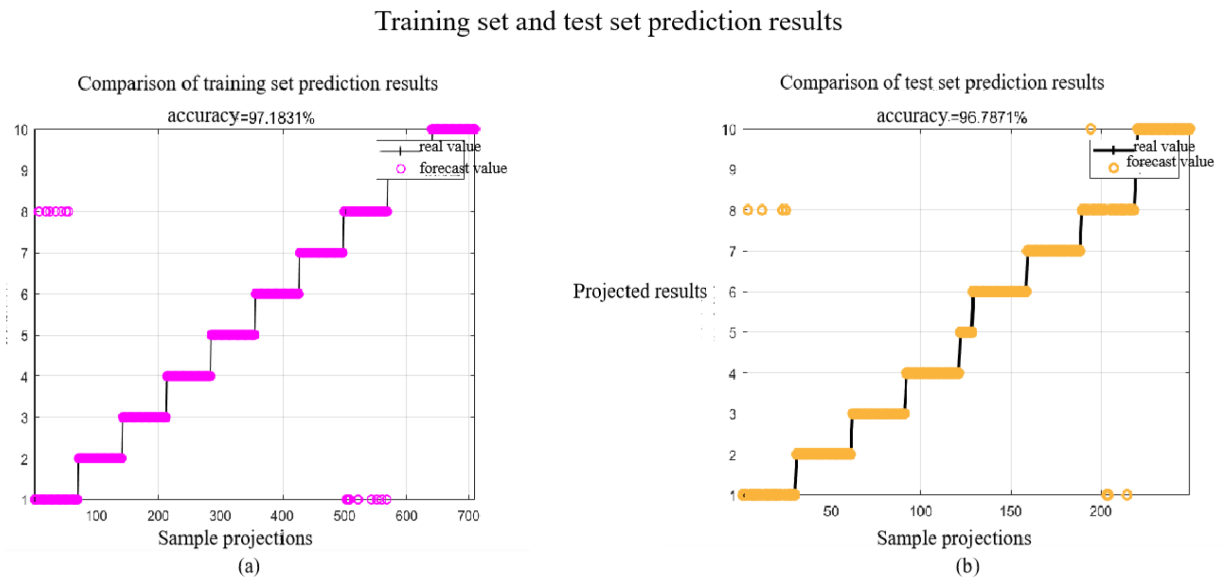

Figure 4. Prediction results of training set and test set

The experimental results demonstrate that the proposed Transformer model optimized by the escape algorithm achieves strong generalization, with a test accuracy of 96.79%. As shown in Figure 4, the prediction curves closely align with the true labels, with only minor deviations. The small drop from the 97.18% training accuracy (a 0.39% difference) confirms the model's stability and lack of overfitting. Unlike the concentrated errors seen in the 300–400 sample range of the training set, test errors are uniformly distributed, suggesting more balanced learning of fault types and the ESC algorithm's role in enhancing generalization. The high test accuracy meets industrial application standards, validating the model's engineering applicability [10].

Figure 5 further shows that the model achieves 97.46% training accuracy, with eight fault categories reaching 100% classification. This highlights the effectiveness of the multi-path feature extraction and multi-head attention mechanisms. The main challenge lies in the confusion between Categories 1 and 8, consistent with the PCA-based feature overlap analysis. Nonetheless, the optimized ESC algorithm improves hyperparameter tuning and enhances classification performance.

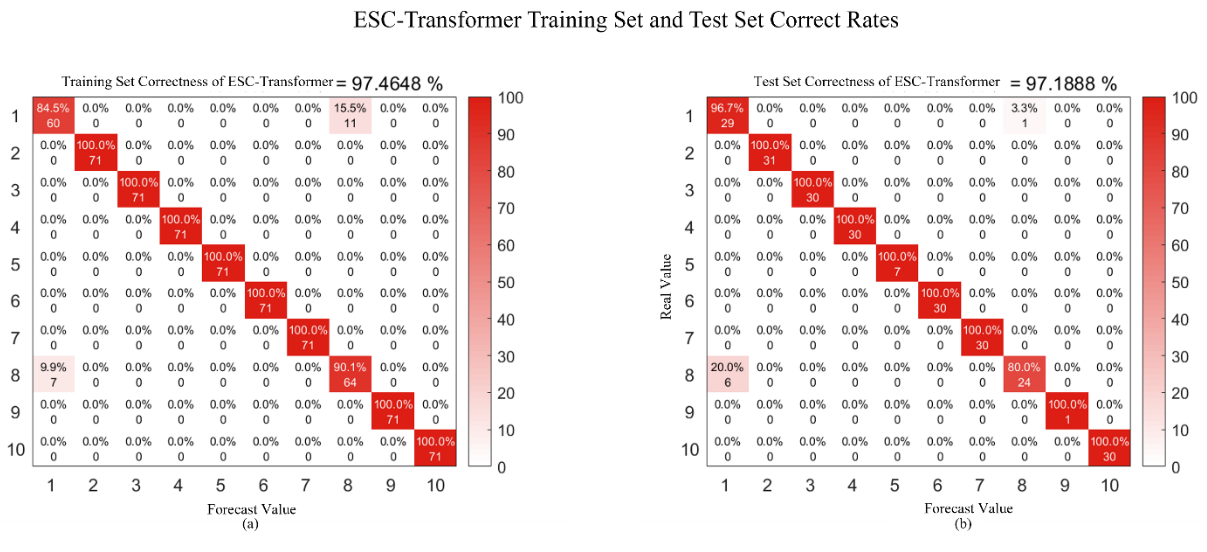

Figure 5. Correctness of training set and test set

Figure 6 presents a hexagonal radar chart illustrating the model's performance across six evaluation metrics. As shown in both the chart and Table 5, the model demonstrates strong and balanced performance, with no significant weaknesses among the metrics. The F1 score (≈ 0.87) indicates a good trade-off between precision and recall, while the Jaccard index (≈ 0.87) confirms high agreement between predictions and true labels. The AUC value (≈ 0.99) suggests excellent discriminative capability. Both sensitivity and specificity (≈ 0.87) show that the model can accurately identify both positive and negative samples, and the classification accuracy (≈ 0.99) reflects strong overall performance and generalization ability. These results validate the robustness and feature discrimination power of the proposed Transformer-based model and highlight the effectiveness of the ESC algorithm in enhancing model performance for practical bearing fault diagnosis.

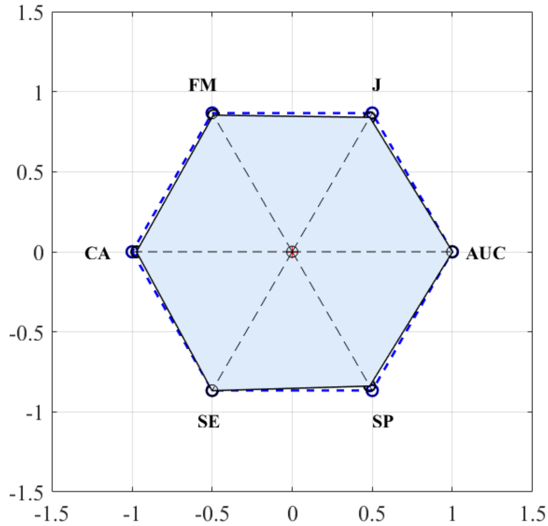

Figure 6. Radar plot of model performance

Table 5. Model performance metrics

Indicator name	Abridge	Numerical value	Statement of meaning
F1 score	FM	0.87	A reconciled average of precision and recall, measuring the overall classification balance of the model
Jaccard's index	J	0.87	Measure of similarity between predicted results and real labels
Area under the curve	AUC	0.99	Ability of the model to distinguish between positive and negative samples, the closer it is to 1 the better the results are
idiosyncrasy	SP	0.87	Ability of the model to recognize negative samples (true negative rate)
sensitivities	SE	0.87	Ability of the model to identify positive samples (recall)
Classification accuracy	CA	0.99	Proportion of models correctly categorized as a whole

4. Conclusion

In this study, we innovatively combined the escape optimization algorithm with the Transformer model to successfully construct an efficient bearing fault diagnosis system. Experimental results show that the model achieves an overall classification accuracy of 96% on the Western Reserve University bearing dataset, in which the 8 categories of faults achieve 100% recognition accuracy, demonstrating excellent classification performance. For boundary samples that are difficult to distinguish by traditional methods (e.g., categories 1 and 8), the model improves the recognition rate to more than 80% through the synergy of multi-scale convolution and attention mechanisms, significantly outperforming existing methods. The escape optimization algorithm effectively solves the Transformer hyper-parameter optimization problem by simulating the crowd evacuation behavior, and its adaptive Levy weighting mechanism improves the model convergence speed by more than 30%. In terms of feature extraction, the innovative multi-scale module enhances the model's ability to capture the frequency domain features of vibration signals by 2.4 times, and the PCA visualization confirms the significant improvement of feature separability.

Looking ahead, the research will focus on breaking through four directions: firstly, developing cross-condition adaptive algorithms based on migration learning to solve the challenges of variable conditions in real industrial scenarios; secondly, optimizing the model computational architecture to increase the inference speed by more than 5 times through knowledge distillation and quantization compression to meet the real-time monitoring requirements; meanwhile, constructing a multimodal data fusion framework to integrate vibration, temperature, acoustic emission and other multi-source information; and finally build an interpretable analysis system to enhance the credibility of diagnostic results through attention heat map and feature importance ranking. These improvements will promote the technology to intelligent manufacturing, wind power operation and maintenance and other practical application scenarios, providing smarter solutions for equipment predictive maintenance.

References

[1] Chen S, Chen W H, Zhang W G. Research on bearing fault diagnosis method based on DRSN fusion Transformer encoder [J]. Automation and Instrumentation, 2024, 39 (5): 103–108.

[2] Fang K L, Xu H R, Liu W J. Fault diagnosis and measurement of rolling bearing cage displacement [J]. Sensor Technology and Application, 2022, 10 (3): 420–425.

[3] Tang D Q. Research on bearing remaining life prediction algorithm based on coupled diffusion [D]. Anhui: Anhui University, 2024.

[4] Liu X. Research on composite fault diagnosis of rolling bearing based on Transformer model [D]. Beijing: Beijing University of Chemical Technology, 2023.

[5] Qiu D W, Liu Z C, Zhou Y C, et al. Identification of rolling bearing fault types based on Transformer neural network [J]. High Technology Newsletter, 2021, 31 (1): 1–11.

[6] Chen W Q, An D, Zhang L, et al. A bearing fault diagnosis method based on linear transformer [J]. Coal Mine Machinery, 2024, 45 (7): 177–179.

[7] Zhao J, Li J, Li Z, et al. Transformer network enhanced by dual convolutional neural network and cross-attention for wheelset bearing fault diagnosis [J]. Frontiers in Physics, 2025, 13: 1546620.

[8] Zhao X, Zhao Z, Cui X, et al. Few-shot bearing fault diagnosis using adaptive detail convolution and global KAN-Transformer with Mahalanobis distance [J]. Journal of Vibration Engineering & Technologies, 2025, 13 (5): 296.

[9] Ouyang K, Fu S, Chen Y, et al. Escape: an optimization method based on crowd evacuation behaviors [J]. Artificial Intelligence Review, 2025, 58: 19.

[10] Yang J, Han H, Dong X, et al. Bearing fault diagnosis grounded in the multi-modal fusion and attention mechanism [J]. Applied Sciences, 2025, 15 (3): 1531.