Front. Agr. Sci. Eng. 2025, 12(4): 837–856 https://doi.org/10.15302/J-FASE-2025634

RESEARCH ARTICLE

Fault diagnosis system for puffing machine with Bayesian-optimized convolutional neural network and multi-head attention based on multi-source fusion

Daolong HONG[1], Fuping ZHANG[2], Hua LI[1], Feiteng XIA[1], Yue SHEN[3], Xiche ZHANG[1], Xuebin FENG (✉)[1], Yongjian WANG (✉)[1]

1 College of Engineering, Nanjing Agricultural University, Nanjing 210031, China.
2 Changzhou Honghuan Machinery Co., Ltd., Liyang City, Liyang 213333, China.
3 College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210031, China.

KEYWORDS

Puffing machine, fault diagnosis, Bayesian optimization, convolutional neural network, multi-head attention

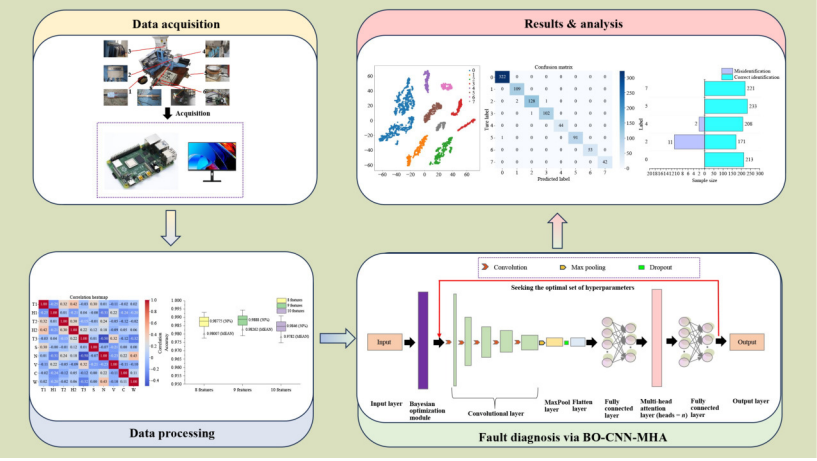
HIGHLIGHTS

- A BO-CNN-MHA-based fault diagnosis model is developed for puffing machines.
- Multi-source signals are effectively fused for fault feature analysis.
- The model achieves 98.8% accuracy and meets real-machine diagnosis requirements.

Received February 27, 2025;
Accepted May 7, 2025.

Correspondences: fxb9510@njau.edu.cn, yjwang@njau.edu.cn

GRAPHICAL ABSTRACT



ABSTRACT

Due to its high-temperature and high-pressure operating environment, food/feed puffing machines are prone to faults such as cavity blockage and cutter wear. This paper presents the design of a fault diagnosis system for puffing machines (food/feed processing equipment that expands or puffs agricultural products), based on a convolutional neural network and a multi-head attention mechanism model, which incorporates Bayesian optimization. The system combines multi-source information fusion, capturing patterns and characteristics associated with fault states by monitoring various sources of information, such as temperature, noise signals, main motor current and vibration signals from key components. Hyperparameters are optimized through Bayesian optimization to obtain the optimal parameter model. The integration of convolutional neural networks and multi-head attention mechanisms enables the simultaneous capture of both local and global information, thereby enhancing data comprehension. Experimental results demonstrate that the system successfully diagnoses puffing machine faults,

achieving an average recognition accuracy of 98.8% across various operating conditions, highlighting its high accuracy, generalization ability and robustness.

1 Introduction

The feed industry is the material basis for the development of modern animal husbandry and farming. Extrusion puffing technology, with its unmatched advantages in feed processing, has rapidly become one of the leading new feed processing technologies, both domestically and internationally. Currently, the intelligence level of puffing machinery on the market is relatively low. In the event of a failure, particularly a severe blockage, manual dismantling and cleaning of the expansion cavity are required. This process poses significant safety risks due to the high temperature of the expansion cavity, which can cause operator injuries. Therefore, there is an urgent need to develop a highly intelligent, stable and reliable fault diagnosis system for expansion machines to minimize the risk of manual troubleshooting.

Standard machinery fault diagnosis methods are mainly based on basic signal processing techniques, which often suffer from low accuracy due to harsh operating conditions. The signals are typically non-smooth, nonlinear and contain noise[1,2].

To enhance signal interpretability under such challenging conditions, advanced methods such as swarm decomposition combined with permutation entropy and the opposition-based slime mold algorithm have been used for accurate bearing fault identification[3]. In addition, to address the limitations of standard linear analysis, recent research has focused on leveraging nonlinear mechanisms, such as stochastic resonance, to enhance weak fault features in noisy environments[4]. For example, Qiao et al.[5] proposed a noise-boosted fault diagnosis method for machinery in which increasing the potential-well width enhances weak signal detection, highlighting its effectiveness in both overdamped and underdamped fractional-order systems.

With the rise of artificial intelligence, machine learning has become a research focus in the field of fault diagnosis. Researchers combine machine learning with signal processing, using various feature extraction methods to analyze monitoring signals using neural networks and other techniques to construct diagnosis models[6,7]. However, early machine learning methods face challenges in feature extraction and model parameterization, and perform poorly when dealing with large amounts of data[8,9]. In this context, recent research has introduced advanced machine learning techniques. Vashishtha & Kumar[10] introduced a gray wolf optimization algorithm with mutation strategy for Pelton wheel fault diagnosis, achieving 100% fault prediction accuracy by optimizing feature selection and model parameters.

In recent years, with the development of computers, sensors and communication technologies, the volume of monitoring data has been increasing, pushing fault diagnosis into the era of big data. The advantage of deep learning in constructing end-to-end diagnosis models by automatically learning features makes it widely used in the field of machinery fault diagnosis. Among deep learning models, convolutional neural networks (CNNs) are the most commonly applied models due to their ability to effectively extract spatial and temporal features from data[11]. Wang et al.[12] proposed a multiscale learning neural network containing one-dimensional and two-dimensional convolutional channels to learn the local correlation between neighboring and non-neighboring intervals in periodic signals for fault diagnosis of bearings. Xiao et al.[13] introduced fast machine fault diagnosis based on a marginalized denoising autoencoder for acoustic signals, using a variant of stacked denoising autoencoder to achieve rapid fault diagnosis of acoustic signals. Jang et al.[14] proposed cross-domain fault diagnosis of rotating machinery based on a discriminative feature attention network. This mechanism use spatial attention to extract focused information from both the feature generator and the discriminator, further enhancing task-specific features through the attention mechanism between the two extracted information types. Feng et al.[15] proposed the application of a residual multi-head self-attention network based on multimodal shallow feature fusion in electric motor fault diagnosis, where the fused features are introduced into the residual multi-head self-attention network to obtain more detailed information.

However, many fault samples suffer from issues such as small sample sizes and lack of fault labels, which pose challenges for deep learning methods. To address the issue of small fault sample sizes, Yue et al.[16] used a wavelet convolution model and a meta-learning training scheme to achieve diagnosis with limited samples across domains. Yu et al.[17] used a digital twin model to predict fault data from health data and used CNN to

achieve fault diagnosis with zero fault samples. For scenarios lacking fault labels, Qian et al.[18] developed a deep discriminative transfer learning network to achieve fault diagnosis using unlabeled samples from the target domain.

Despite the potential of combining deep learning with the operational characteristics of puffing machines, research and application of intelligent puffing machine fault diagnosis systems remain limited. Chu & Zhou[19] presented a fuzzy neural network-based fault diagnosis method for a single-screw extrusion puffing machine. This method includes an analysis of common faults, the application principle of fuzzy neural networks and experimental results, highlighting its advantages in improving fault diagnosis accuracy. IDAH Company, a feed machinery company in Taiwan, China, has developed a continuous lubrication system for extruders. This system not only minimizes downtime due to gearbox and bearing box failures but also enhances safety by directly connecting the continuous lubrication system to the main power supply. Swiss food and feed machinery company, Bühler, has developed error and downtime analysis, a digital tool that rapidly analyses production equipment performance. It examines the types and locations of machine failures, generates trend analyses illustrating common errors and assesses their overall impact on downtime.

Although earlier studies on fault diagnosis in puffing machinery have been effective, they often lack quantification of the impact of each feature on faults, relying heavily on experiential knowledge and limited data samples, and are based solely on simulation analysis[19]. Additionally, there is a challenge in sensor availability compared to other machinery fault diagnosis contexts[20,21]. To address these issues, this study integrated multiple types of sensors relevant to puffing machine faults, including temperature, humidity, noise, weight, current and feed speed, alongside vibration sensors, to comprehensively gather fault-related information from puffing machine. The study explored the most common types and levels of failures in puffing machines to design appropriate maintenance strategies. Also, correlation and importance analyses were conducted on these features to ensure the rationality of feature selection, and to identify key feature factors and optimal combinations influencing various types of faults. The integration of convolutional neural networks and multi-attention mechanisms facilitated the capture of both local and global information, thereby enhancing data comprehension and yielding high classification accuracy and robust generalization through the optimal parameter combination derived from Bayesian optimization. Bayesian-optimized CNN ensures efficient hyperparameter selection, accelerating convergence and improving model generalizability, while MHA enhanced the ability of the model to focus on critical features across multi-source data. The fused model analyses and trained the collected multi-source information and was validated on real-world machinery to demonstrate the efficacy of the method in puffing machine fault diagnosis.

2 Material and methods

2.1 Sensor selection and layout

Given the high-temperature and high-pressure conditions of puffing machine operation and its sensitivity to temperature, specific sensors are selected for monitoring various parameters. Temperature sensors were used to monitor temperature data in different compartments, waterproof feed temperature and humidity sensor measures the temperature and humidity of the feed material, and ambient temperature and humidity sensor captures ambient air temperature and humidity. Vibration of the puffing machine was monitored using an acceleration sensor and phase current of the main motor was measured with a Hall current sensor. Abnormal noise was detected using a noise sensor and a weighing sensor assesses feed product weights to judge production efficiency. Installation details of these components and measuring instruments in the puffing machine are given in Fig. 1. Table 1 provides details of the components and measuring instruments used.

2.2 BO-CNN-MHA fault diagnosis model

To enhance the accuracy of operational fault diagnosis for the puffing machine, a new model was developed that comprises a BO-CNN-MHA framework, combining Bayesian optimization (BO) with a CNN and a multi-head attention mechanism (MHA). This model extracts multi-source feature signals that reflect the operational state of the puffing machine components, serving as the foundation for fault assessment.

2.2.1 Bayesian optimization

Bayesian optimization is a robust global optimization strategy particularly suited for evaluating costly objective functions. In contrast to standard optimization methods that rely on gradient information or heuristic techniques, Bayesian optimization uses a probabilistic model to guide the search for the optimal solution. This approach is especially effective for black-box functions, where the form is unknown and the evaluation process can be time-consuming.

Fig. 1 Diagram of the installation of components and measuring instruments used in puffing machine. 1, Feed temperature and humidity sensor; 2, ambient temperature and humidity sensor; 3, vibration sensor; 4, temperature sensor; 5, weighing sensor; 6, Raspberry Pi data processor; 7, current sensor; and 8, noise sensor.

Table 1 Details of components and measurement instrumentations used in this study

Sensor	Model	Operating voltage	Output	Measurement range
Temperature sensor	PT100	DC24V	Analog	–50 to 200 °C
Feed temperature and humidity sensor	SHT20	DC2.1-3.6V	Digital	–40 to 125 °C, 0%–100%
Ambient temperature and humidity sensor	SLS132R-25	DC9-30V	Analog	–20 to 80 °C, 0–100%
Vibration sensor	PR-3001-WZ1-V10-CX	DC24V	Analog	0–50 mm·s^{-1}
Current sensor	MIK-HRI	DC24V	Analog	0–50 A
Noise sensor	PR-300BK-ZS-N01	DC10-30V	Analog	30–120 dB
Weighing sensor	JLBU-1	DC24V	Analog	0–50 kg
Raspberry Pi	Raspberry Pi 4B	DC5V	/	/

The key components of Bayesian optimization include a surrogate model, which is typically a Gaussian process and an acquisition function. The surrogate model approximates the objective function based on prior evaluations, enabling the optimizer to make informed decisions when selecting the next sampling point. The acquisition function, such as expected improvement or upper confidence bound, determines the next evaluation point by balancing sampling in regions of high uncertainty with sampling in areas expected to yield high function values[22] (Fig. 2).

In Bayesian optimization, the parameter space of the black-box function was defined, and a surrogate model was created using a prior distribution. After each evaluation of the objective function, the model was updated with new data to improve its predictive accuracy. The acquisition function, which balances the predicted value and uncertainty, then directed the search toward areas of the parameter space with potential for better results. This cycle of inputting parameter combinations into the black-box function, updating the model, calculating acquisition function values, and selecting the next combination continued until stopping criteria were met.

The mathematical formulation of Bayesian optimization was:

$$\mu(x^*) = k(x^*, X) \cdot K(X, X)^{-1} \cdot y \quad (1)$$

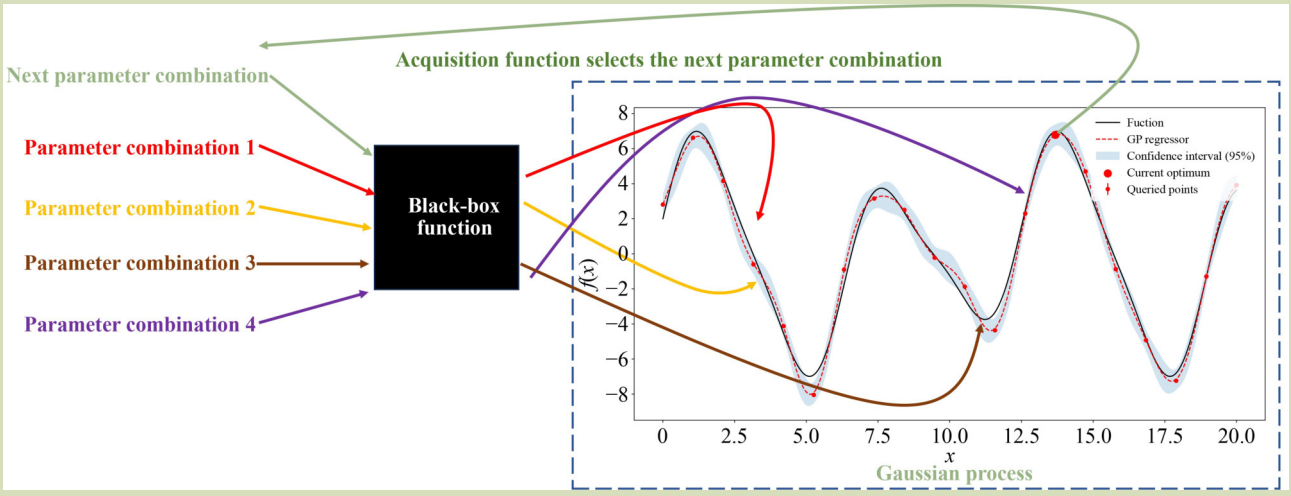


Fig. 2 Bayesian optimization flow chart.

$$\sigma^2(x^*) = k(x^*, x^*) - k(x^*, X) \cdot K(X, X)^{-1} \cdot k(X, x^*) \quad (2)$$

$$y^* = \underset{x}{argmax} E[f(x)] \quad (3)$$

$$x_{\text{next}} = \underset{x}{argmax}\, \alpha(x) \quad (4)$$

where, $\mu(x^*)$ is the predicted mean at the input point x^*, $\sigma^2(x^*)$ is the corresponding variance, X is set of known input points, $K(X, X)$ is the covariance matrix of the known points, y is the evaluation results of the objective function at the known points, $k(x^*, X)$ is the covariance between the point to be predicted x^* and the known points X, y^* is the optimal objective value, $f(x)$ is to the unknown objective function, $E[f(x)]$ is the expectation of the objective function f at a certain input point x, $a(x)$ is the acquisition function, and x_{next} is the next input point to be evaluated.

The iterative nature of Bayesian optimization enables efficient exploration of the input space and convergence to the optimal solution, highlighting its significant value in hyperparameter tuning within machine learning.

2.2.2 Convolutional neural network

CNN is a type of deep feedforward neural network characterized by local connectivity and weight sharing. It excels in representational learning, enabling effective local perception and feature extraction from data. CNNs have widespread applications in computer vision, natural language processing and speech recognition. CNNs typically comprise layers including input, convolutional, rectified linear unit (ReLU), pooling and fully connected layers (similar to ordinary neural networks). Figure 3 shows the architecture of a CNN.

Specifically, the convolutional layer performs convolution operations on the input data using a filter, which slides over the

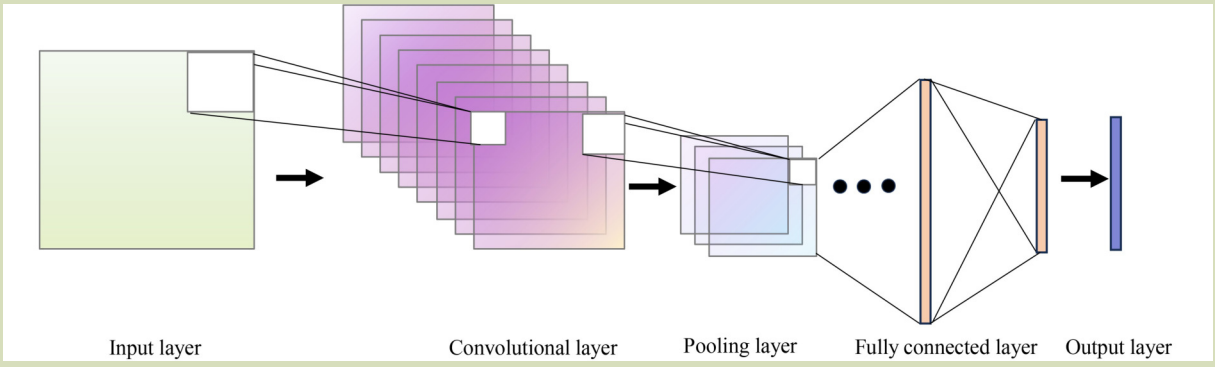


Fig. 3 Diagram of convolutional neural network architecture.

input, multiplying and summing elementwise to generate the feature map. The pooling layer subsamples the feature map, with max pooling taking the maximum value within a window and average pooling taking the mean value, thus reducing dimensionality to improve computational speed and robustness. The fully connected layer flattens the feature map and connects it to the neural network for model learning. In the convolutional and fully connected layers, parameters such as weights are optimized using gradient descent or other algorithms to minimize the loss function. This ensures that the predictions of the network on the training set closely match the labels, thereby enhancing classification accuracy and generalization for the training data. The formulas for the convolutional layer and ReLU, respectively, are:

$$h = f(x \otimes W + b) \tag{5}$$

$$ReLU(x) = max(0, x) \tag{6}$$

where, $\otimes$ is the convolution operation, x is the input data, W is the weight of the convolution kernel, b is the offset value, and $f(\cdot)$ is the activation function.

2.2.3 Multi-head attention

To better capture the global information and improve the fault diagnosis accuracy of puffing machine, this study introduced the attention mechanism to optimize the CNN model[23]. Attention is represented as the combination of query (Q) and key-value pairs $\{K_i, V_i \mid i = 1,2,\ldots,m\}$ mapping to the output, where the query, each key, and each value are vectors, and the output is the weighting of all values in V[24]. For the value q of the input query, the key dimensions are d_k, the value dimension is d_v, and the output attention matrix formula is:

$$\text{MultiHead}(Q,K,V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{7}$$

where, matrices Q, K and V are the query, keys and values matrices, respectively, where $Q \in R^{m\times d_k}$, $K \in R^{m\times d_k}$ and $V \in R^{m\times d_v}$. The dimension of the output matrix is $R^{m\times d_v}$.

A schematic diagram of scaled dot-product attention is given in Fig. 4.

To enhance the representational power of attention, multiple heads of attention were introduced through multiple parallel applications of scaled dot-product attention. Each attention head is a separate instance of scaled dot-product attention with different learning parameters. Initially, a linear mapping of Q, K and V is performed. The matrices Q, K and V, each with input dimensions of d_{model}, are mapped to $Q \in R^{m\times d_k}$, $K \in R^{m\times d_k}$, and $V \in R^{m\times d_v}$. Subsequently, the results computed from h

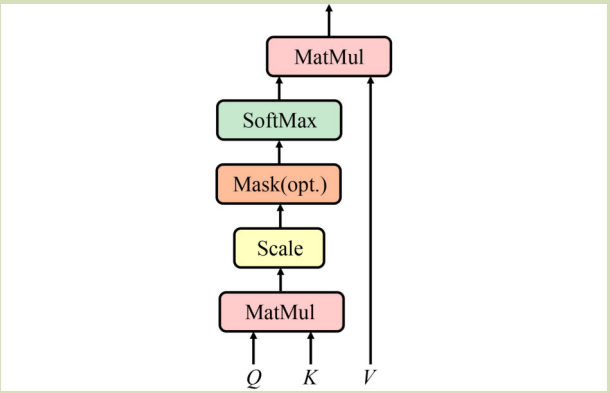


Fig. 4 Schematic of scaled dot-product attention.

instances of scaled dot-product attention are merged and then subjected to linear transformation. The output formula is:

$$\text{Attention}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \tag{8}$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{9}$$

where $W_i^Q \in R^{(d_{\text{model}}\times d_k)}$, $W_i^K \in R^{(d_{\text{model}}\times d_k)}$, $W_i^V \in R^{(d_{model}\times d_v)}$ and $W^O \in R^{(h\cdot d_v\times d_{\text{model}})}$ are the input and output dimensions at each step, h is the number of attention operations, and $d_k = d_v = d_{\text{model}}/h$ is the dimension after linear transformation and before the attention operation. The output matrix dimension after one attention operation is $R^{m\times d_v}$ and the final output after h operations is $R^{m\times(h\cdot d_v)}$. The dimensions of input and output matrices are identical.

A schematic diagram of multi-head attention is given in Fig. 5.

Compared to single-head attention, multi-head attention can acquire and process input information more comprehensively

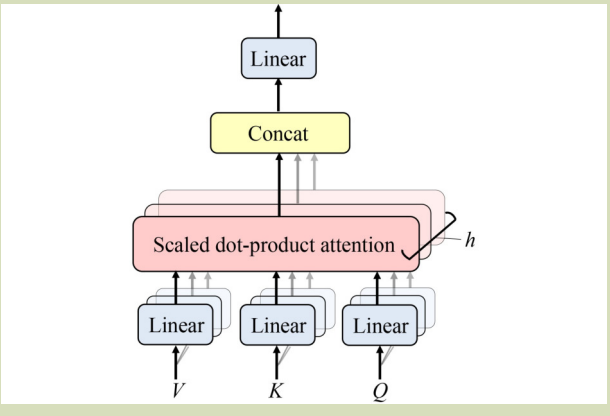


Fig. 5 Schematic of multi-head attention mechanism.

by concurrently processing and combining different attention heads. This enhances the generalization ability and learning efficiency of the model, while mitigating the risk of overfitting.

In the application of puffing machine fault diagnosis, the operation of the puffing machine involves data from various sensors (such as temperature, humidity and vibration). These data not only exist at different time scales but may also be influenced by different operating conditions. Standard single-feature analysis methods may not effectively capture the relationships among these diverse data. The multi-head attention mechanism, however, can process signals from different sensors in parallel through different attention heads, thereby extracting more useful features and achieving deep fusion of the data. This improves the accuracy and robustness of fault diagnosis. Therefore, MHA, when dealing with complex multi-source signals, can more effectively reveal the health status of equipment, especially in the case of agricultural-product puffing machinery, a nonlinear and high-dimensional industrial system, where its advantages are particularly prominent.

2.2.4 BO-CNN-MHA model

A BO-CNN-MHA model was developed in this study for the fault diagnosis for a representative puffing machine. This model integrates a CNN with a MHA, and its hyperparameters are optimized using BO to enhance performance and generalizability. In contrast to previous studies, which primarily focus on medical image classification and fault diagnosis using single-source signals, this work concentrated on fault diagnosis for agricultural-product puffing machinery, using multi-source, non-image time-series signals, such as temperature, humidity and vibration. These multi-source input data are more complex and challenging, requiring more sophisticated models for effective processing.

The CNN component extracts local spatial features, while the MHA module captures global contextual information by assigning adaptive attention weights in the feature space, enabling the model to process local and global features simultaneously[25,26].Additionally, the CNN model is optimized through BO, which tunes the hyperparameters and network architecture more effectively compared to standard optimization methods. This optimization process enhances the performance of the model by improving its generalization ability and reducing overfitting, ensuring better accuracy in fault diagnosis tasks.

The model is primarily structured with five convolutional layers, one multi-head attention layer and three fully connected layers. The convolutional layers form the initial segment of the model and consist of five layers designed to extract features from the input data. Each convolutional layer is followed by a ReLU activation function to enhance model nonlinearity. Subsequently, a maximum pooling layer reduces feature dimensionality while preserving essential information, followed by a dropout layer to mitigate overfitting. Following the convolutional layers, data flows through a fully connected layer before entering the multi-head attention layer, which is pivotal in assigning varying degrees of attention to different information segments. Post-attention, the model traverses another fully connected layer with subsequent ReLU activation. The final layer, also fully connected, generates the ultimate output of the model. Specifically, the model inputs 10 features pertaining to puffing machine operation, while the outputs encompass eight operational states, encompassing normal operation and seven fault conditions. The BO-CNN-MHA model network structure diagram is given in Fig. 6.

2.3 Fault diagnosis framework based on BO-CNN-MHA multi-source information fusion

To achieve intelligent fault diagnosis of puffing machine, this study used a novel method based on BO-CNN-MHA and multi-source information fusion. The process framework of this fault diagnosis method is shown in Fig. 7, and its specific steps are: (1) multi-source sensor signals are collected from a puffing machine using a Raspberry Pi 4B data processor, (2) the collected fault data samples are labeled, divided into training and test sets, and undergo batch processing and normalization, and feature correlation and importance analysis are also conducted, (3) the optimal hyperparameters and model architecture for BO-CNN-MHA are to be determined, and model training and testing are to be carried out, and (4) the resulting fault diagnosis outputs are compared with those of other classification models to assess the performance of the proposed method. Experimental validation was conducted to confirm the applicability of the model.

2.4 Experimental design and evaluation methodology

2.4.1 Experimental data collection method

This study used real data from expansion machines simulating various operating conditions. The data set records experimental data collected from December 2023 to January 2024, sampled at intervals of 1 s, comprising a total of 4760

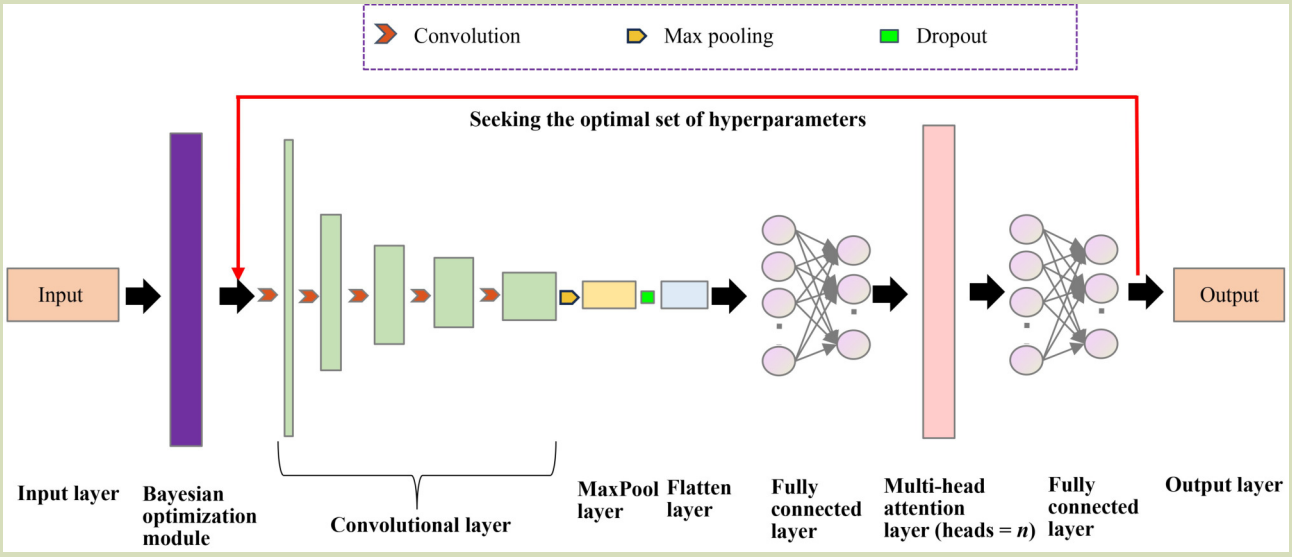


Fig. 6 BO-CNN-MHA model network structure.

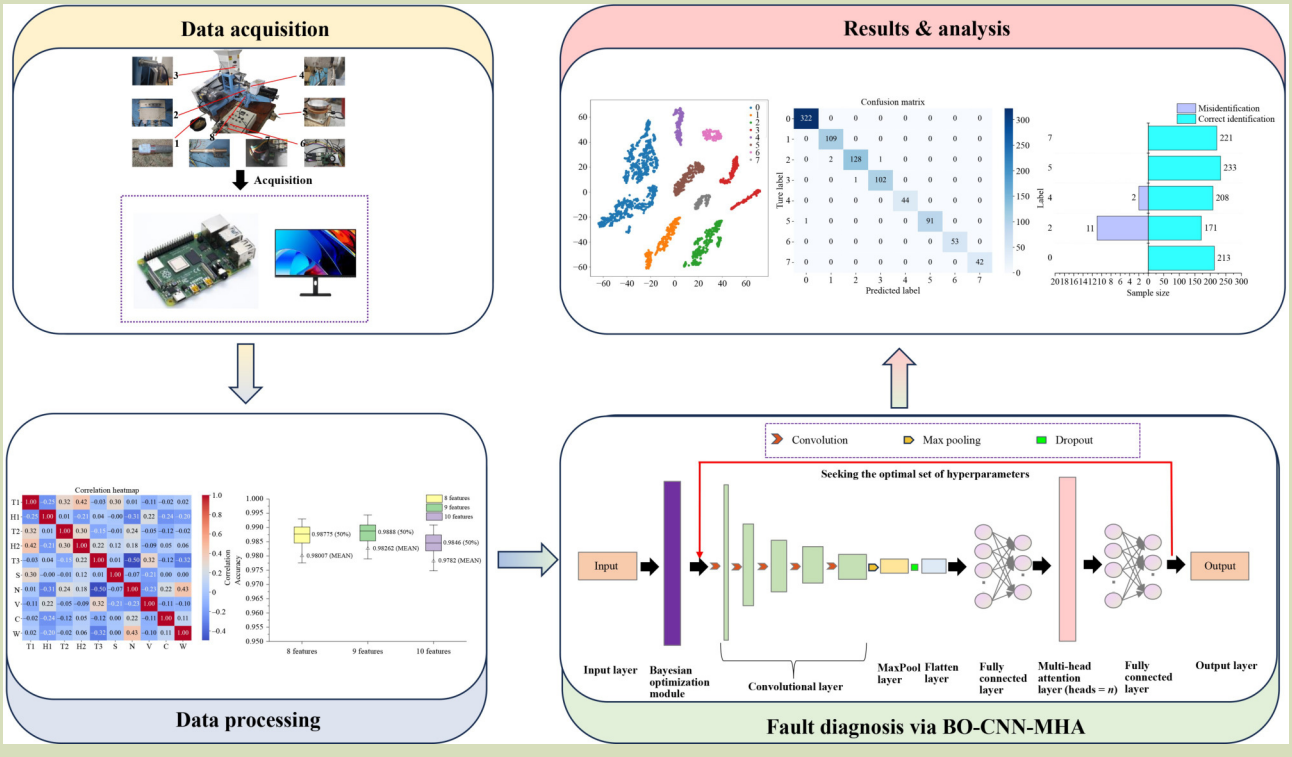


Fig. 7 Proposed fault diagnosis framework.

entries. Each record included component temperatures, humidity levels, speeds, noise levels, weights, vibrations and currents, with specific operating parameters as listed in Table 2. The puffing machine corresponds to state information, as illustrated in Table 3, primarily divided into normal operation and seven types of faults, totaling eight states. Normal operation refers to the puffing machine operating at temperatures of 85–105 °C, with feed material humidity at

Table 2 Operating parameters of puffing machine

Parameter	Position	Code
Temperature (°C)	Feed	T1
	Ambient	T2
	Internal cavity	T3
Humidity (%)	Feed	H1
	Ambient	H2
Speed ($r \cdot min^{-1}$)	Feed inlet	S
Noise (dB)	Cutter	N
Weight (kg)	Product outlet	W
Vibration ($mm \cdot s^{-1}$)	*Z*-axis	V
Current (A)	Main motor single phase	C

Table 3 State information on puffing machine

State	Sample size	Label
Normal	1818	0
Slight blockage	567	1
Moderate blockage	659	2
Severe blockage	488	3
Inlet clogged	290	4
Screw loosening	474	5
Moderate cutter wear	246	6
Severe cutter wear	218	7

65% ± 3%, a feed speed of 13 $r \cdot min^{-1}$ and other components functioning without damage. The seven types of faults were categorized into four main groups: cavity blockage, inlet clogged, screw loosening and cutter wear, as illustrated in Fig. 8. For cavity blockage faults, severity determines the appropriate countermeasures, classified as minor, moderate, and severe. Similarly, cutter wear severity was categorized into moderate wear and severe wear based on cutter conditions[27].

2.4.2 Evaluation methodology

Accuracy, dynamically weighted focal loss and F1 score are used to evaluate the classification performance of the proposed model, while confusion matrix graphs are used for visualization to assess the performance of the classification model.

Accuracy (*ACC*) represents the proportion of correctly predicted samples out of the total number of samples:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where *FP* is the number of samples that are negative but predicted to be positive, *TN* is the number of negative samples that are predicted to be negative, *TP* is the number of positive samples that are predicted to be positive, and *FN* is the number of samples that are positive but predicted to be negative.

The dynamic weighted focal loss (*DWFL*) integrates focal loss with a dynamic weight adjustment mechanism to ensure that samples from minority classes receive higher weights for correction, thereby addressing the class imbalance problem:

$$num = \sum_{i=1}^{leny_{pred}} (y_{pred<0.5}) + \sum_{i=1}^{leny_{true}} (y_{true}) \tag{11}$$

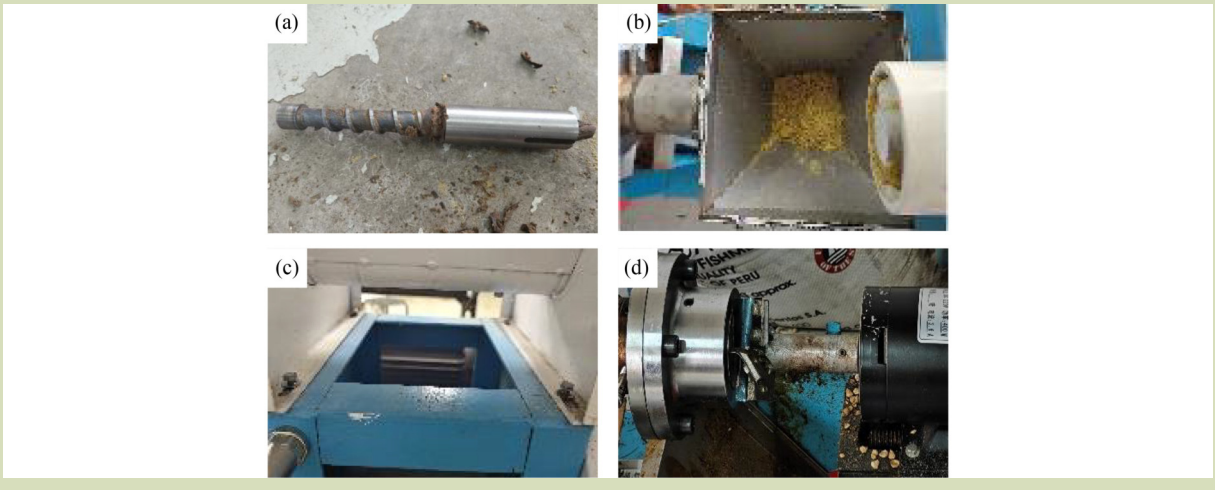


Fig. 8 The picture of puffing machine fault phenomenon. (a) Cavity blockage; (b) inlet clogged; (c) screw loosening; (d) cutter wear.

$$PosWeight = \frac{\sum_{i=1}^{leny_{true}} (y_{true})}{num + \varepsilon} \quad (12)$$

$$NegWeigh = \frac{\sum_{i=1}^{leny_{pred}} (y_{pred<0.5})}{num + \varepsilon} \quad (13)$$

$$Weighs = (1 - y_{true}) \cdot PosWeight + y_{true} \cdot NegWeight \quad (14)$$

$$DWFL = -\sum_{i=1}^{N} Weighs_i \cdot (1 - \hat{y}_i)^{\gamma} log(\hat{y}_i) \quad (15)$$

where, *num* is the sum of the predicted negative samples and the actual positive samples, *Weighs* is the dynamically calculated weight, N is the total number of samples, $\hat{y}_i$ is the predicted probability of the ith sample; and γ is the tuning factor that controls the weight of hard-to-classify samples.

The F_1 score combines information from both the precision rate and the recall rate to provide a more comprehensive assessment of the classification performance of the model, particularly when dealing with imbalanced data sets. The F_1 score better reflects the overall performance of the model in such cases. The formula is:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (16)$$

The formulas for precision and recall are:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (17)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (18)$$

Confusion matrix diagrams display the contents of the confusion matrix through heat maps. The horizontal axis of the confusion matrix diagram represents the predicted categories, the vertical axis represents the actual categories, and the color shade or numerical magnitude of each cell indicates the number of samples in the corresponding position. This helps to evaluate the performance of the model across different categories. The formula is:

$$\text{Confusion} \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \quad (19)$$

The t-distributed stochastic neighbor embedding (t-SNE) method was used to visually represent the features output from the model after training, allowing for the observation of the distribution of these features in a low-dimensional space to evaluate the classification effectiveness of the model for faulty data. t-SNE is a technique for visualizing high-dimensional feature outputs from the model in a reduced feature space, where similar samples are represented by nearby points and dissimilar samples by distant points.

Initially, t-SNE constructs a probability distribution for high-dimensional samples, where similar samples have a high probability of being selected and dissimilar points a very low probability. Subsequently, t-SNE defines a similar distribution for points in the low-dimensional embedding. Finally, t-SNE minimizes the Kullback-Leibler (*KL*) divergence between the two distributions concerning the positions of the embedded points and optimizes the *KL* divergence using a gradient descent method. This is formulated as:

$$p_{j|i} = \frac{\exp(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2})}{\sum_{k \neq i} \exp(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2})} \quad (20)$$

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \quad (21)$$

$$KL(P||Q) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad (22)$$

$$\frac{\partial C}{\partial_{y_i}} = 4 \sum_j (p_{j|i} - q_{j|i})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \quad (23)$$

where, σ_i is based on the distance of x_i from its nearest neighbor data point, adaptively selecting the Gaussian kernel bandwidth, y_i and y_j are the mappings of x_i and x_j in the lower-dimensional space, y_k and y_l are the low-dimensional representations of other data points, $p_{j|i}$ is the similarity between x_i and x_j, $q_{j|i}$ is the similarity between y_i and y_j, and $\partial C/\partial y_i$ represents the loss function of the *KL* divergence, which is adjusted through constant iterations of y_i values until the *KL* divergence is minimized.

Shapley additive explanations (SHAP) values provide a method for interpreting machine learning model predictions, rooted in the principles of Shapley values from game theory. Shapley values quantify the contribution of each player to the overall payoff of a game. In the context of SHAP values, each feature acts as a player in the game, and the prediction outcome of the model represents the payoff of the game. This approach calculates the contribution of each feature and determines its importance. The formula is:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (24)$$

where, ϕ_i is the SHAP value of feature i, N is the set of features, S is a subset that excludes feature i, $f(S)$ is the predicted output of the model given the subset S of features, and $|S|$ is the number of elements in set S. The formula indicates that ϕ_i, the SHAP value for feature i, is a weighted average of Shapley values. These values correspond to changes in the predicted output of the model when considering different combinations of feature subsets S.

3 Results

3.1 Characterization

3.1.1 Feature correlation analysis

The selected characteristic variables were assessed using the Pearson correlation coefficient, and their correlation heat map is given in Fig. 9. From this matrix, it is evident that the temperature of the feed material had a moderate positive correlation of 0.42 with ambient humidity. When ambient humidity was high, moisture in the air was more readily transferred to the surface of the feed material, leading to an increase in feed material temperature[28]. Additionally, there were weak correlations between feed material temperature and ambient temperature, feed rate and humidity. Feed material humidity had weak correlations with ambient humidity, noise, vibration speed, main motor current and product weight. Ambient temperature weakly correlated with ambient humidity and noise.

The cavity temperature had a moderate negative correlation of –0.50 with noise. An increase in cavity temperature caused expansion of its components, reducing internal gaps and minimizing vibration propagation, thereby reducing noise generation[29]. Cavity temperature was also weakly correlated with ambient humidity, product weight and main motor current. Feed rate weakly correlated with product weight. Noise and main motor current had a moderate positive correlation coefficient of 0.43. Increased motor load typically elevates current consumption; higher loads may increase friction within the motor, speeding up vibration and consequently increasing noise levels[30]. Noise had weak correlations with vibration speed and product weight.

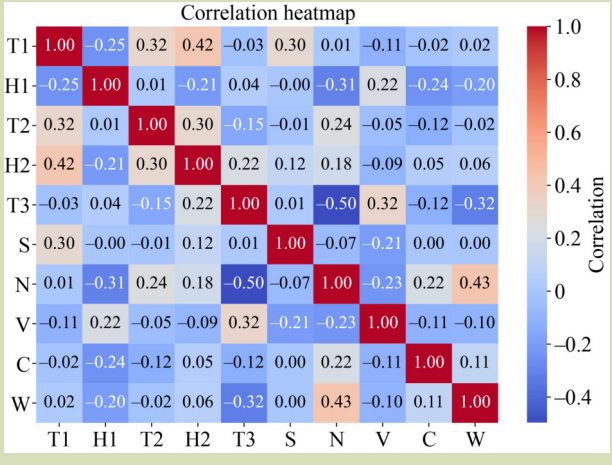


Fig. 9 Heat map of feature correlation coefficients.

Overall, the correlations between the 10 characteristic variables were generally weak, affirming their suitability as distinct failure characteristics of the puffing machine. This low inter-variable correlation indicates that each feature provides complementary information, thereby enhancing the effectiveness of multi-source data fusion. Integrating these heterogeneous features can significantly improve the fault diagnosis the ability of the model to capture diverse fault signatures.

3.1.2 Feature importance analysis

Using SHAP value theory, the SHAP value of each feature for its respective category was computed to determine the importance and contribution of each feature to the classification outcome[31]. Scatter plots depicting feature densities for the four types of faults are given in Fig. 10.

From Fig. 10(a), it is evident that for cavity blockage faults, cavity temperature was the most critical feature, followed by feed humidity, suggesting that higher cavity temperatures or feed humidities increase the likelihood of blockage. In Fig. 10(b), for inlet clogged faults, feed speed emerged as the predominant feature, followed by feed temperature, indicating that higher feed speeds or temperatures elevate the risk of inlet clogged. Figure 10(c) shows that ambient temperature was crucial for screw loosening faults, where higher ambient temperatures correlate with increased screw loosening probabilities. Lastly, Fig. 10(d) shows that for cutter wear faults, cavity temperature was the primary feature, followed by ambient temperature and noise. Higher cavity temperatures reduced the likelihood of cutter wear, while elevated ambient temperatures or noise levels increased the risk.

In summary, cavity temperature, feed humidity, ambient temperature and feed rate were identified as significant factors, while vibration rate and main motor current contribute less to fault prediction. By integrating these diverse features into a unified model, multi-source data fusion enhances the accuracy and robustness of fault prediction, capturing a broader range of operational conditions and ensuring more reliable diagnosis across different fault types.

3.1.3 Feature selection

From the feature importance analysis, it can be seen that vibration speed and main motor current were the two features with the lowest importance for puffing machine faults. Therefore, data sets with eight features (excluding vibration speed and main motor current), nine features (excluding main motor current), and 10 features were selected for comparison.

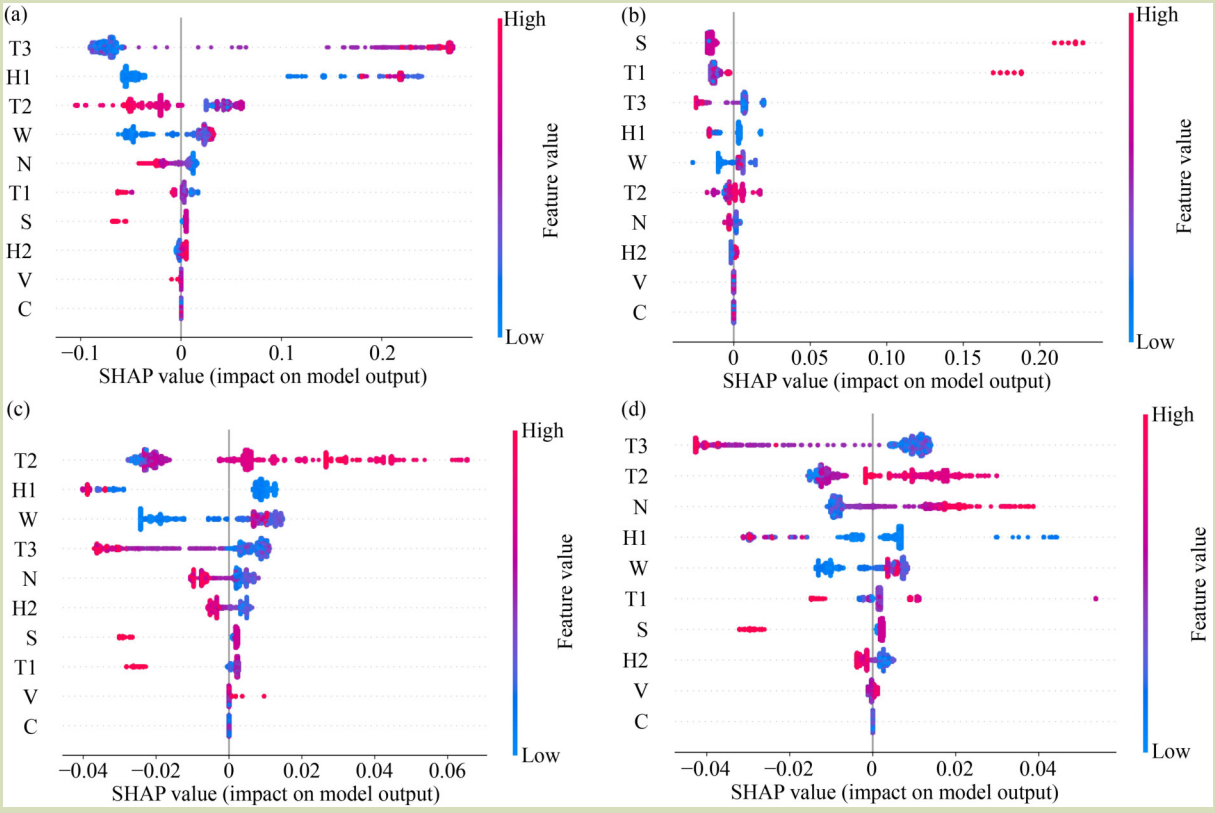


Fig. 10 Four types of fault scatter plots of feature density. (a) Beeswarm summary plot of cavity blockage; (b) beeswarm summary plot of inlet clogged; (c) beeswarm summary plot of screw loosening; (d) beeswarm summary plot of cutter wear.

the classification accuracies of their test sets after 100 training iterations are shown in the box plots in Fig. 11. The diagnosis performance was optimal with nine features, followed by eight features, and was worst when all 10 features were retained. This is primarily because low-importance features usually contain more noise and irrelevant information. Removing them

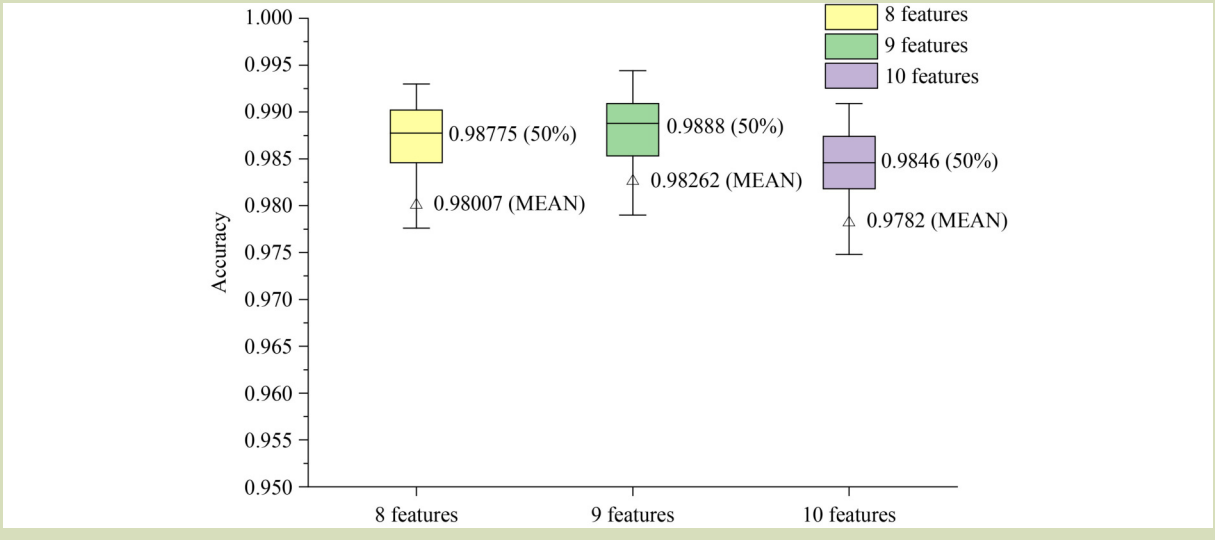


Fig. 11 Comparison of the diagnosis effect of different number of features.

allowed the model to focus more on the features that were truly meaningful for the classification task. Also, removing low-importance features simplified the model and improved its generalization ability, thereby enhancing classification accuracy[32]. Removing two features may result in more information being lost, and the model may not be able to capture the features of the data effectively. Thus, the diagnosis performance was slightly better when one feature was removed compared to when two features were removed.

3.2 Fault diagnosis performance analysis

3.2.1 Hyperparameter optimization of the CNN-MHA model

The data set, comprising 4760 samples representing eight distinct states, was partitioned into training, validation and test sets in a 6:2:2 ratio, followed by batch processing. The model developed for this study uses hyperparameter optimization via Bayesian optimization. The hyperparameter combinations for the baseline model and the optimal configurations obtained after 100, 200, 500 and 1000 iterations of Bayesian optimization are summarized in Table 4. The number of epochs was set to 500 and Adam optimizer was used.

Figure 12 shows the comparison curves of the F_1 scores for the validation set. Notably, the highest F_1 score of 0.995 was achieved with 1000 iterations, indicating that the overall performance of the model with optimal hyperparameters slightly exceeded that of the other configurations. This finding indicates that increasing the number of iterations during hyperparameter optimization enables a more thorough exploration of the parameter space, leading to the identification of superior model configurations[33].

To further investigate the optimal number of attention heads, we used the previously determined optimal hyperparameters to examine several configurations: 1, 2, 4, 8, 16, and 32 heads. The F_1 score curves on the validation set, after training the model for 500 epochs, are presented in Fig. 13. These results indicate that the different configurations of the multi-head attention mechanism significantly influenced the F_1 scores of the model. As the number of training epochs increased, the performance of the models with different numbers of attention heads tended to stabilize, ultimately approaching an F_1 score close to 1. This demonstrates that the model has an overall enhanced generalization capability.

Table 4 Optimal hyperparameter combinations for different Bayesian iteration counts

Bayesian iteration count	Number of filters	Kernel sizes (pixels)	Dropout rate	Learning rate	Batch size	Number of heads
Baseline	[64, 128, 256, 512, 102]	[3, 3, 3, 3, 3]	0.3	0.0001	128	8
100	[512, 512, 512, 380, 512]	[3, 4, 5, 5, 3]	0.1	0.000793	128	8
200	[223, 64, 512, 476, 393]	[3, 3, 3, 5, 5]	0.1	0.000479	128	8
500	[512, 64, 512, 64, 512]	[3, 3, 5, 3, 5]	0.5	0.000717	85	8
1000	[64, 64, 280, 512, 512]	[3, 3, 3, 5, 3]	0.1	0.00072	128	8

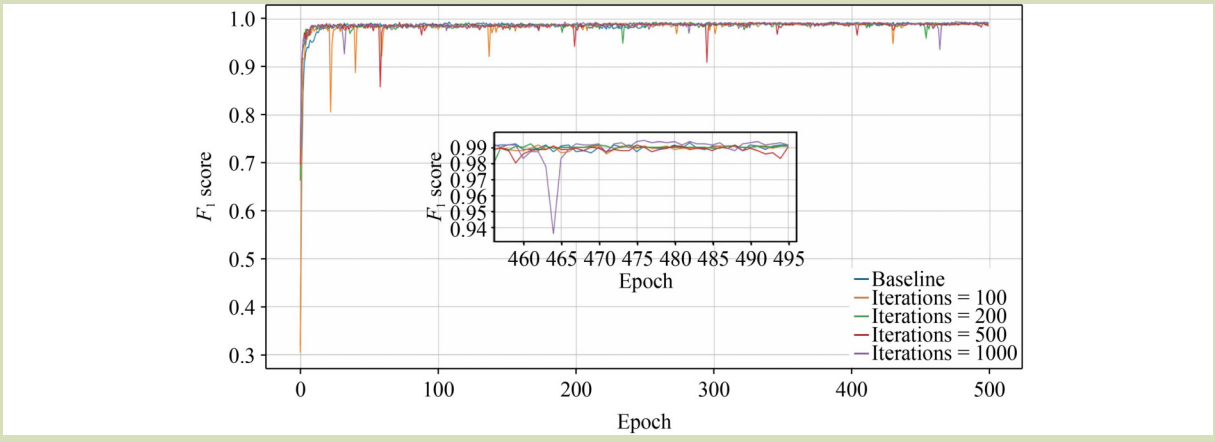


Fig. 12 Effect of different Bayesian optimization iterations on F_1 score.

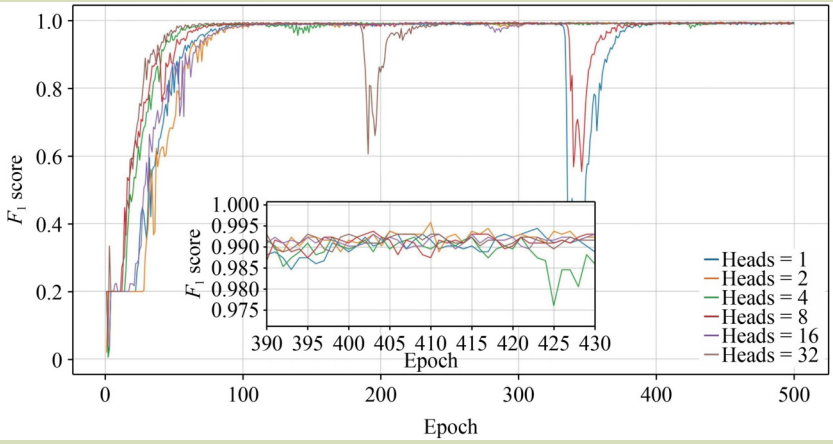


Fig. 13 Effect of different numbers of attention heads on F_1 score.

In particular, the configurations with 1, 8, and 32 attention heads converged rapidly during training and maintained relatively high stability across most epochs. However, these had abrupt drops in performance during certain training phases. The configurations with 4 and 16 heads also fluctuated during specific training intervals. In contrast, the configuration with 2 heads demonstrated a smoother performance trajectory and greater stability in the later epochs, achieving a peak F_1 score of 0.996. Therefore, we conclude that using 2 attention heads is the optimal choice, as it maintained a high F_1 score while having superior stability and reduced variability.

To determine the optimal number of epochs for model training, we selected various training iterations (200, 300, 500, and 800) alongside patience parameters values (50 and 100), implementing an early stopping mechanism to compare their effects on the convergence of validation loss[34]. The comparative curves are given in Fig. 14. Overall, model's validation loss showed a rapid decline during the initial stages, followed by a period of stabilization.

However, certain configurations, such as an early stopping patience of 100 combined with 300 training epochs (around the 198th epoch), and an early stopping patience of 100 combined with 500 training epochs (around the 164th epoch), exhibited significant fluctuations. This indicates that these settings may lead to model instability or overfitting. In contrast, the configuration with an early stopping patience of 100 and 800 training epochs achieved optimal performance at the 288th epoch, reaching a validation loss of 0.0180, which was maintained at a consistently low level thereafter.

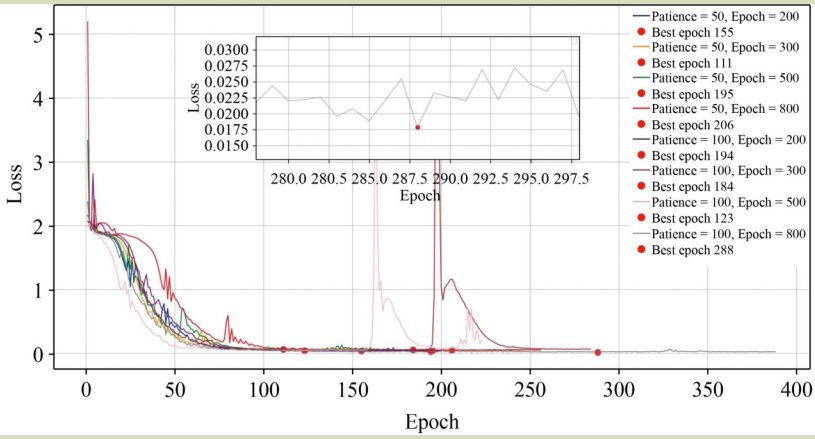


Fig. 14 Effect of different training epochs and patience on validation loss convergence.

Thus, the results indicate that simply increasing the number of training epochs does not necessarily lead to improved performance. The combination of a patience value of 100 and 800 training epochs demonstrated the highest stability and enabled the earlier identification of the optimal model.

Using the optimal hyperparameter combination obtained previously, the classification results of the model tested on the test set after training are given in Table 5. The overall accuracy for the eight categories of health states is 0.994. Labels 0 (normal), 1 (slight blockage), 4 (inlet clogged), 6 (moderate cutter wear) and 7 (severe cutter wear) achieved the highest accuracy of 100%. This high accuracy is attributed to the simplicity of the causes for these statuses. Conversely, the accuracy for status label 2 (moderate blockage) was 97.7%, label 3 (severe blockage) was 99.0% and label 5 (screw loosening) was 98.9%. The lower accuracy for these labels was primarily due to the complexity of the causes. For example, blockage is influenced by multiple factors, including cavity temperature and feed humidity, while the impact of screw loosening on the overall operation is minimal, thus reducing the classification accuracy for this fault type.

3.2.2 Comparison of results with other methods

The CNN-MHA model was validated against standard ANN, BP and CNN models using the same training and test data sets. To more effectively demonstrate the characteristics of different generation methods and classifiers, t-SNE used as a qualitative method to evaluate the extracted high-dimensional sample features. Figure 15 shows the t-SNE results of the features extracted by the four models, with each operational state indicated by a specific color. In Fig. 15(a,b), a significant overlap is evident, and Fig. 15(c) a some overlap is evident, indicating that these classifiers do not effectively distinguish samples with different health states. However, when the CNN-MHA model was used, there was minimal overlap, demonstrating the superiority of the proposed CNN-MHA. Compared with the standard CNN, the feature clusters are more compact with the addition of MHA. This indicates that incorporating self-attention enhanced the quality of the generated samples, enabling the classifier to learn deeper features more effectively. Additionally, Fig. 15(d) shows clear boundaries without obvious anomalies, indicating that the CNN-MHA algorithm was adequately robust for noisy and anomalous data. Also, the CNN-MHA model had a clear ability to differentiate between varying degrees of cavity blockage. Samples corresponding to slight blockage formed well-separated and tightly clustered groups. By comparison, samples representing moderate blockage had more dispersed clustering, while those of severe blockage tended to be separated into two distinct sub-clusters, reflecting the complex nature of the fault characteristics. These distribution patterns are largely attributed to the intricate interactions among factors such as temperature, humidity and operating conditions. Although the clustering was not flawless, it was markedly superior to that achieved by the other three models.

To further evaluate the performance of the classification model, a confusion matrix was used for analysis. The test set for the original fault data was evaluated using the trained ANN, BP, CNN and CNN-MHA models, with the resulting fault classification confusion matrix for eight classifiers is given in Fig. 16. The actual classification of each fault type can be visualized from the confusion matrix, showing that ANN had the poorest classification performance, BP performed slightly more effectively, CNN performed well and CNN-MHA gave the best overall classification performance. All four models did not effectively with classify label 1 (slight blockage) but perform well with label 4 (inlet clogged). BP performed quite poorly for label 6 (moderate cutter wear), while BP and CNN

Table 5 Test set validation results

Fault label	Sample size	Correct identification	Misidentification	Accuracy (%)
0	322	322	0	100
1	109	109	0	100
2	131	128	3	97.71
3	103	102	1	99.03
4	44	44	0	100
5	92	91	1	98.91
6	53	53	0	100
7	42	42	0	100

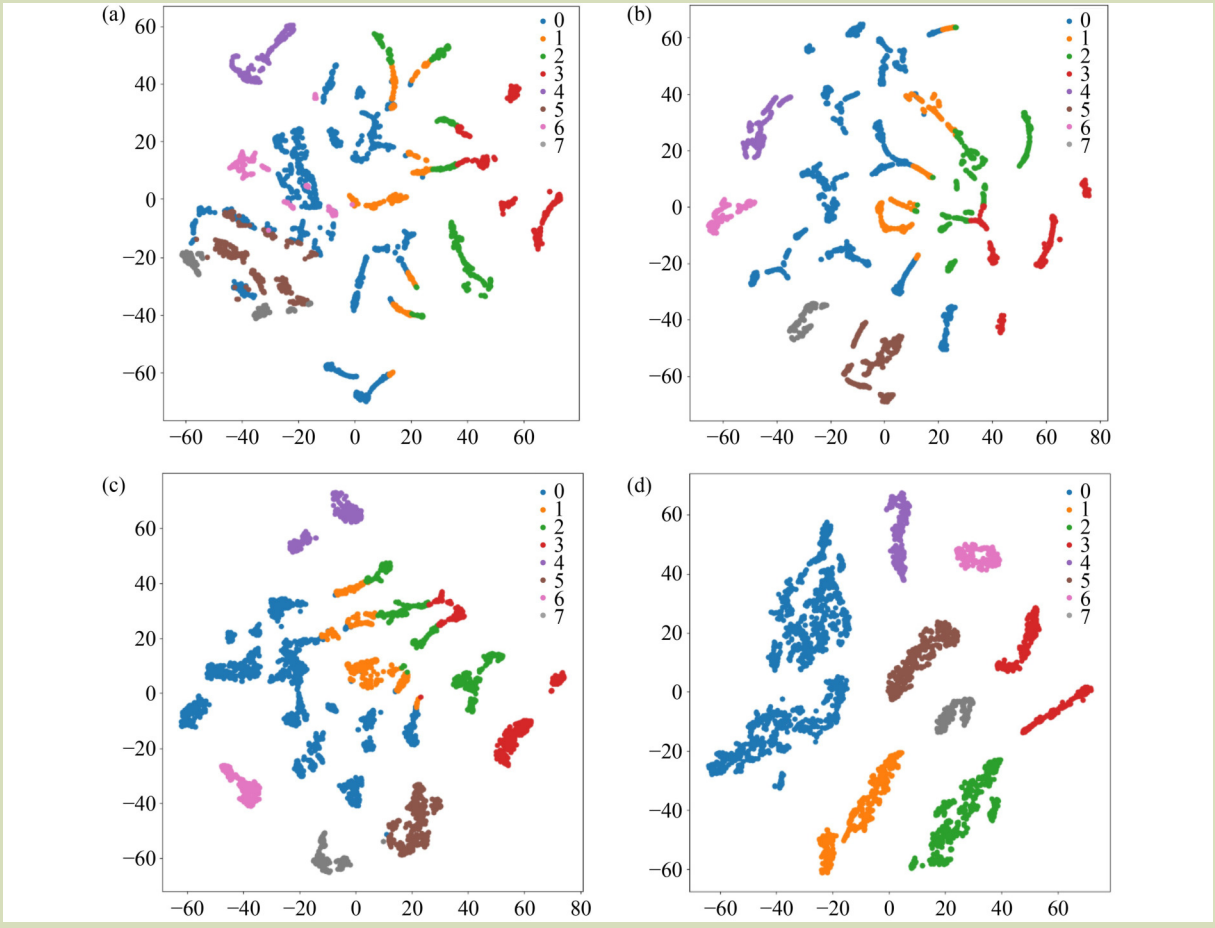


Fig. 15 Visualization of t-SNE with different classifiers. (a) ANN model; (b) BP model; (c) CNN model; (d) CNN-MHA model.

were slightly better than CNN-MHA for label 2 (moderate blockage). CNN also performed slightly better than BP and CNN-MHA for label 5 (screw loosening).

In the confusion matrix, the CNN-MHA model performed strongly in distinguishing between different levels of cavity blockage. For slight blockage, the model accurately classified the samples, significantly outperforming the other models. However, there was some overlap between moderate blockage samples and slight blockage, with a few samples misclassified as either slight or severe blockage. For severe blockage, the model correctly identified the majority of samples, but due to the complexity of the fault characteristics, a small number of samples are misclassified as moderate blockage. These misclassifications were likely to have been influenced by factors such as temperature variations, humidity fluctuations, differences in operational conditions, and the inherent complexity of the fault features.

3.3 Experimental validation of fault diagnosis system for puffing machine

Human intervention is conducted under fault conditions to induce specific fault phenomena in the puffing machine, obtaining validation samples. To ensure the practical validity of the results, a balanced sampling method was used, maintaining a nearly 1:1 ratio of normal operation samples to various types of fault samples. After preprocessing the validation data, a total of 1645 data sets were obtained, comprising 800 sets of normal operation samples and 845 sets of failure data. The data collected include 175 samples of moderate blockage, 218 samples of inlet plugged, 190 samples of screw loosening, and 213 samples of severe cutter wear. The data were divided into five sets for validation, and the results are given in Fig. 17.

The results of the validation experiment indicate that when the ratio of normal operation samples to various types of fault samples in the puffing machine is close to 1:1, the system achieved an average recognition rate of 98.8%. Specifically, the

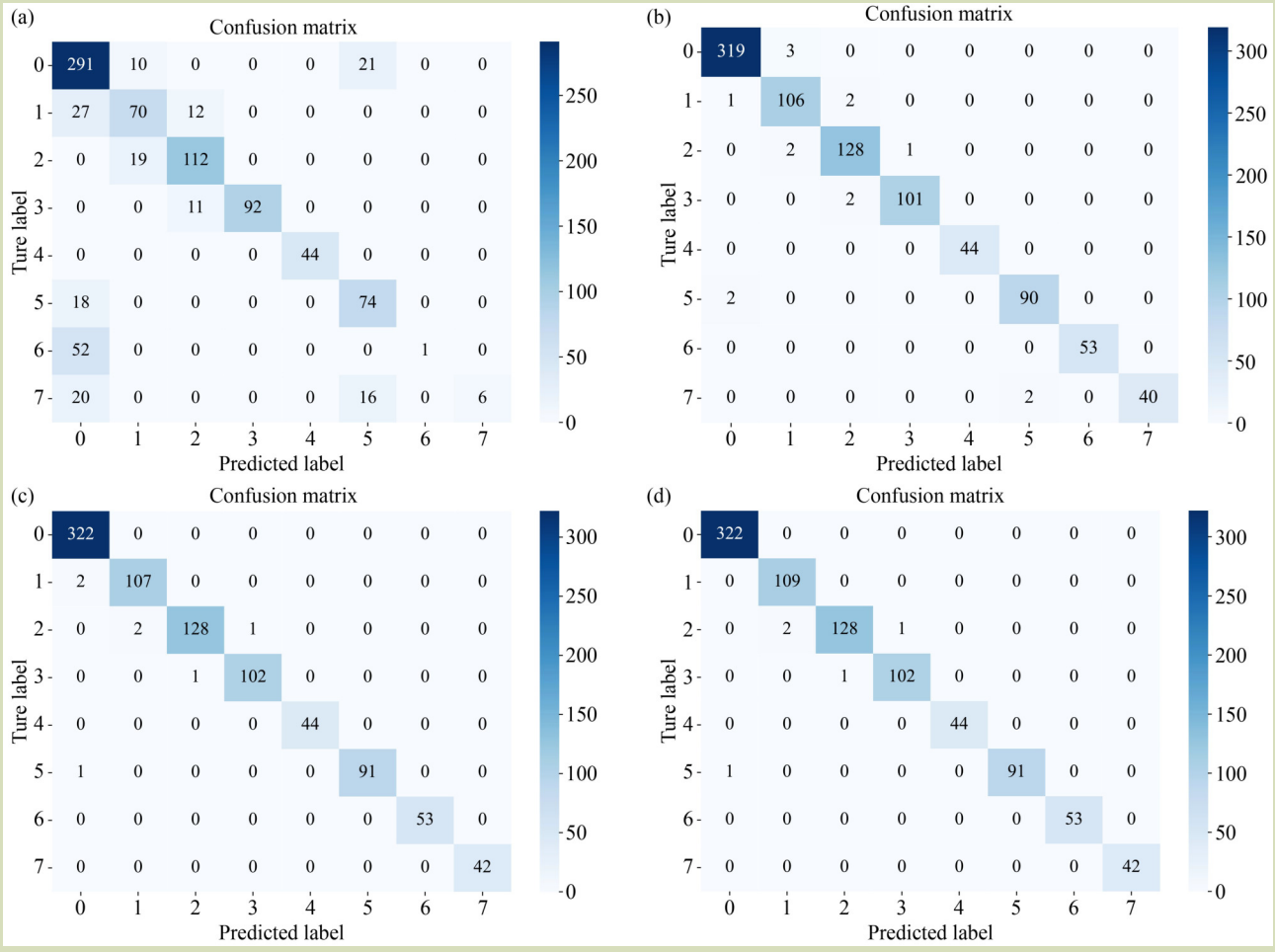


Fig. 16 Fault classification confusion matrix comparison figure. (a) ANN model; (b) BP model; (c) CNN model; (d) CNN-MHA model.

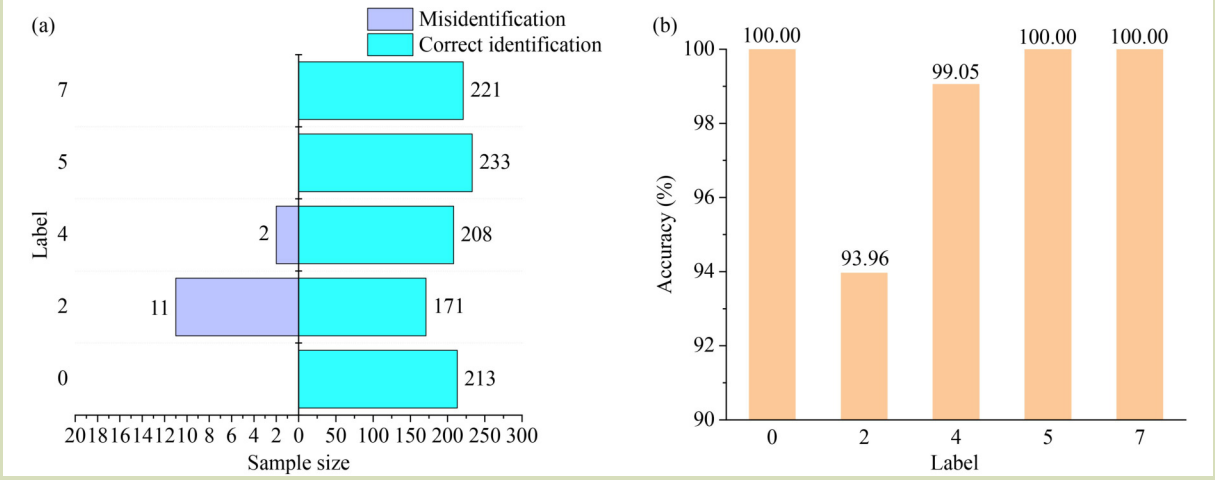


Fig. 17 Comparison of the diagnosis effect of different number of features. (a) Validation sample classification results; (b) validation sample classification accuracy.

recognition rate for moderate blockage was 94.0%, primarily due to overlapping data characteristics with severe blockage in puffing machine. The screw loosening area was closer to the vibration sensor, and cutter wear was closely related to noise. In addition, when faults occurred, the phenomena were obvious, so these two types of faults were recognized at a higher rate. For inlet clogged, the recognition accuracy reached 99.1%. Qiu et al.[27] achieved a diagnosis accuracy of 97.5% for a proposed adaptive fault diagnosis method for combine harvesters using multi-source information fusion. Xue et al.[35] proposed an improved Gaussian naive Bayes algorithm for diagnosing faults in the wet clutch control system of a hydrostatic power-split continuously variable transmission (CVT) for tractors, achieving an average accuracy of 98.2% using a time window and principal component analysis. Hou et al.[36] achieved 96.4% accuracy in fault diagnosis for the seeding metering drive motor of a plot seeder using support vector machine classification. Compared to similar agricultural machinery, the overall accuracy of fault validation for the puffing machine was better.

4 Discussion

During operation, operators commonly rely on experience to adjust parameters like feeding speed and temperature on puffing machine. The implementation of a fault alarm system for puffing machine will provide technical support but could be constrained by low precision and potential misjudgments due to simplistic fault criteria such as temperature thresholds and current abnormalities. These errors not only impact operator judgments but also shorten the lifespan of puffing machine by causing operational damage.

Chu & Zhou[19] introduced a fuzzy neural network-based fault diagnosis for single-screw extrusion puffing machine, marking the first use of machine learning in puffing machine fault diagnosis. However, there system relied on a single data source, lacked quantification of failure causes, and overlooked comprehensive production data indices beyond temperature, current and vibration abnormalities. Critical issues like blockage failures were not thoroughly analyzed, comparative experiments with other diagnosis methods were not conducted, and their study lacks objective assessments of the effectiveness and limitations the method. Also, it lacked real-machine experimental validation, relying only on simulation, thus raising doubts about its reliability and practicality.

Building on that research, the proposed fault diagnosis algorithm monitors key component temperatures, feed material temperature and humidity, noise signals, main motor current and vibration signals to extract fault characteristic features. By applying Bayesian optimization to fine-tune the hyperparameters, the CNN-MHA fault diagnosis model processes these signals more efficiently, integrating subjective operational factors to provide a comprehensive diagnosis of puffing machine failures. The model quantitatively analyses the primary causes of critical puffing machine blockages, classifies fault levels according to maintenance requirements, and achieves higher feature learning accuracy compared to previous deep learning algorithms.

Our approach not only improves puffing machine fault diagnosis accuracy to 98.8%, surpassing similar agricultural machinery, but also reduces false alarms and omissions. We demonstrate significant accuracy improvements when reducing sample type imbalance, showcasing model sensitivity and classification accuracy during real-machine testing. This underscores the applicability and effectiveness of this model for diagnosing common puffing machine faults.

However, our method has some limitations. First, although the evaluated system had a comprehensive set of sensors, such configurations are relatively costly. In future work, it would be useful to examine low-cost, efficient sensor arrangements to develop a more economical version of the system, making it more suitable for deployment in small- and medium-sized enterprises. Second, while the proposed model delivered high diagnostic accuracy, its real-time performance still requires improvement. Subsequent research should focus on optimizing the model architecture and deployment strategies to enhance inference speed, enabling efficient and real-time fault monitoring and diagnosis on edge or embedded devices. Finally, complex operational conditions and overlapping data features occasionally lead to misclassifications. The performance of the new model in diagnosing multiple simultaneous faults remains unverified, and manual labeling may introduce inaccuracies at critical fault points[37,38]. Future research could enhance discrimination and fault prediction by incorporating additional sensors, such as triaxial vibration sensors, and using more adaptive fusion metric algorithms.

5 Conclusions and prospects

An intelligent monitoring and fault diagnosis system based on BO-CNN-MHA and multi-source information fusion for puffing machine was developed and tested. It collects information such as cavity temperature, feed humidity, main motor current and vibration signals of the puffing machine,

and integrates a CNN and multi-head attention mechanism to achieve intelligent fault diagnosis. The key success of this work and its prospects are summarized below.

The model combines a CNN, multi-head attention mechanism and Bayesian optimization to extract multi-source feature signals reflecting the operational state of key components in a puffing machine for fault diagnosis. By leveraging Bayesian optimization to obtain the optimal set of hyperparameters, the diagnostic accuracy and robustness of the model are further enhanced. Compared to standard methods, it reduces reliance on operator skills and experience. Additionally, compared to previous machine learning algorithms, it effectively enhances accuracy and reduces loss, making puffing machine fault diagnosis more intelligent and practical.

The model was applied to actual puffing machine operational data, demonstrating high accuracy in distinguishing various operational states and achieving high recognition rates for all fault types.

Given that data collection and experimental validation are conducted through simulation experiments in a laboratory environment, there are differences from real production environments. Therefore, future research should focus on puffing machine fault diagnosis and prediction methods.

In the future, there would be merit in focusing on enhancing system performance and adaptability of the new model, with an emphasis on optimizing its cost-effectiveness for broader deployment in small- and medium-sized enterprises. Additionally, efforts need to be directed toward improving real-time fault monitoring capabilities of the system to better support operational efficiency in dynamic industrial environments.

Acknowledgements

The work was sponsored by the Belt and Road Innovative Cooperation Project (BZ2022003).

Compliance with ethics guidelines

Daolong Hong, Fuping Zhang, Hua Li, Feiteng Xia, Yue Shen, Xiche Zhang, Xuebin Feng, and Yongjian Wang declare that they have no conflicts of interest or financial conflicts to disclose. This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Lu Y, Xie R, Liang S Y. Bearing fault diagnosis with nonlinear adaptive dictionary learning. *International Journal of Advanced Manufacturing Technology*, 2019, **102**(9–12): 4227–4239
2. Zhen D, Guo J, Xu Y, Zhang H, Gu F. A novel fault detection method for rolling bearings based on non-stationary vibration signature analysis. *Sensors)*, 2019, **19**(18): 3994
3. Vashishtha G, Chauhan S, Singh M, Kumar R. Bearing defect identification by swarm decomposition considering permutation entropy measure and opposition-based slime mould algorithm. *Measurement*, 2021, **178**: 109389
4. Qiao Z, Chen S, Lai Z, Zhou S, Sanjuán M A F. Harmonic-Gaussian double-well potential stochastic resonance with its application to enhance weak fault characteristics of machinery. *Nonlinear Dynamics*, 2023, **111**(8): 7293–7307
5. Qiao Z, He Y, Liao C, Zhu R. Noise-boosted weak signal detection in fractional nonlinear systems enhanced by increasing potential-well width and its application to mechanical fault diagnosis. *Chaos, Solitons, and Fractals*, 2023, **175**: 113960
6. Chen Z, Gryllias K, Li W. Mechanical fault diagnosis using convolutional neural networks and extreme learning machine. *Mechanical Systems and Signal Processing*, 2019, **133**: 106272
7. Xu X, Tao Z, Ming W, An Q, Chen M. Intelligent monitoring and diagnostics using a novel integrated model based on deep learning and multi-sensor feature fusion. *Measurement*, 2020, **165**: 108086
8. Li X, Jia X D, Zhang W, Ma H, Luo Z, Li X. Intelligent cross-machine fault diagnosis approach with deep auto-encoder and domain adaptation. *Neurocomputing*, 2020, **383**: 235–247
9. Jin Z, He D, Wei Z. Intelligent fault diagnosis of train axle box bearing based on parameter optimization VMD and improved DBN. *Engineering Applications of Artificial Intelligence*, 2022, **110**: 104713
10. Vashishtha G, Kumar R. An amended grey wolf optimization with mutation strategy to diagnose bucket defects in Pelton wheel. *Measurement*, 2022, **187**: 110272
11. Vashishtha G, Chauhan S, Sehri M, Hebda-Sobkowicz J, Zimroz R, Dumond P, Kumar R. Advancing machine fault

diagnosis: a detailed examination of convolutional neural networks. *Measurement Science & Technology*, 2025, **36**(2): 022001

12. Wang D, Guo Q, Song Y, Gao S, Li Y. Application of multiscale learning neural network based on CNN in bearing fault diagnosis. *Journal of Signal Processing Systems for Signal, Image, and Video Technology*, 2019, **91**(10): 1205–1217
13. Xiao D, Tao Z, Qin C, Yu H, Huang Y, Liu C. Fast machine fault diagnosis using marginalized denoising autoencoders based on acoustic signal. In: 2020 Prognostics and Health Management Conference (PHM-Besançon), Besancon, France. *IEEE*, 2020, 229–234
14. Jang G B, Kim J Y, Cho S B. Cross-domain fault diagnosis of rotating machinery using discriminative feature attention network. *IEEE Access: Practical Innovations, Open Solutions*, 2021, **9**: 99781–99793
15. Feng J, Su J, Feng X. A residual multihead self-attention network using multimodal shallow feature fusion for motor fault diagnosis. *IEEE Sensors Journal*, 2023, **23**(23): 29131–29142
16. Yue K, Li J, Chen J, Huang R, Li W. Multiscale wavelet prototypical network for cross-component few-shot intelligent fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 2023, **72**: 1–11
17. Yu X, Yang Y, Du M, He Q, Peng Z. Dynamic model-embedded intelligent machine fault diagnosis without fault data. *IEEE Transactions on Industrial Informatics*, 2023, **19**(12): 11466–11476
18. Qian Q, Qin Y, Luo J, Wang Y, Wu F. Deep discriminative transfer learning network for cross-machine fault diagnosis. *Mechanical Systems and Signal Processing*, 2023, **186**: 109884
19. Chu M, Zhou Z. Fault diagnosis study of single screw extruding machine based on fuzzy neutral network. *Journal of Chinese Agricultural Mechanization*, 2012, (5): 118–121 (in Chinese)
20. Xiang L, Wang P, Yang X, Hu A, Su H. Fault detection of wind turbine based on SCADA data analysis using CNN and LSTM with attention mechanism. *Measurement*, 2021, **175**: 109094
21. Wang Z, Wu Z, Li X, Shao H, Han T, Xie M. Attention-aware temporal–spatial graph neural network with multi-sensor information fusion for fault diagnosis. *Knowledge-Based Systems*, 2023, **278**: 110891
22. Frazier P I. A tutorial on Bayesian optimization. *arXiv Preprint*, 2018, arXiv:1807.02811
23. Wang H, Xu J, Yan R, Sun C, Chen X. Intelligent bearing fault diagnosis using multi-head attention-based CNN. *Procedia Manufacturing*, 2020, **49**: 112–118
24. Vaswani A, Shazeer N M, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017, **30**: arXiv.1706.03762
25. Zhao D, Wang J, Lin H, Wang X, Yang Z, Zhang Y. Biomedical cross-sentence relation extraction via multihead attention and graph convolutional networks. *Applied Soft Computing*, 2021, **104**: 107230
26. Wu F, He J, Cai L, Du M, Huang M. Accurate multi-objective prediction of CO_2 emission performance indexes and industrial structure optimization using multihead attention-based convolutional neural network. *Journal of Environmental Management*, 2023, **337**: 117759
27. Qiu Z, Shi G, Zhao B, Jin X, Zhou L. Combine harvester remote monitoring system based on multi-source information fusion. *Computers and Electronics in Agriculture*, 2022, **194**: 106771
28. Kong D, Fang P, Jin N, Duan E, Chen J, Wang H. Effect of temperature and particle size on thermophysical properties of different energy feedstuffs material. *Transactions of the Chinese Society of Agricultural Engineering*, 2019, **35**(6): 296–306 (in Chinese)
29. Gao P, Du Y, Ruan J, Yan P. Temperature-dependent noise tendency prediction of the disc braking system. *Mechanical Systems and Signal Processing*, 2021, **149**: 107189
30. Han J H, Park S U, Hong S K. A study on the effectiveness of current data in motor mechanical fault diagnosis using XAI. *Journal of Electrical Engineering & Technology*, 2022, **17**(6): 3329–3335
31. Wang H, Liang Q, Hancock J T, Khoshgoftaar T M. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, 2024, **11**(1): 44
32. Han H, Yang D. Correlation analysis based relevant variable selection for wind turbine condition monitoring and fault diagnosis. *Sustainable Energy Technologies and Assessments*, 2023, **60**: 103439
33. Dao F, Zeng Y, Qian J. Fault diagnosis of hydro-turbine via the incorporation of Bayesian algorithm optimized CNN-LSTM neural network. *Energy*, 2024, **290**: 130326
34. Bendebane L, Laboudi Z, Saighi A. AutoNLP for Optimal Number of Epochs in Multi-labeled Deep-Learning Models for Predicting Mental Disorders. In: Novel & Intelligent Digital Systems Conferences. Cham: *Springer Nature Switzerland*, 2024, 246–256
35. Xue L, Jiang H, Zhao Y, Wang J, Wang G, Xiao M. Fault diagnosis of wet clutch control system of tractor hydrostatic power split continuously variable transmission. *Computers and Electronics in Agriculture*, 2022, **194**: 106778
36. Hou Y, Wu Z, Cai X, Dong Z. Research on Fault Diagnosis and Prediction Method about Driving Motor for Seeding Metering of Plot Seeder Based on SVMs Classification and Regression. In: Proceedings of the IOP Conference Series: Earth and Environmental Science. *IOP Publishing*, 2020, **474**(3): 032012
37. Xiao Y, Zhou X, Zhou H, Wang J. Multi-label deep transfer learning method for coupling fault diagnosis. *Mechanical Systems and Signal Processing*, 2024, **212**: 111327
38. He Z, Chu P, Li C, Zhang K, Wei H, Hu Y. Compound fault diagnosis for photovoltaic arrays based on multi-label learning considering multiple faults coupling. *Energy Conversion and Management*, 2023, **279**: 116742