RESEARCH ON FAULT DIAGNOSIS OF TOBACCO MACHINERY USING MULTI-DIMENSIONAL ATTENTION CONVOLUTIONAL NEURAL NETWORK

Zunwei Yang[1], Long Liao[1], Liming Jiang[1], Wenbin Feng[1], Quanyu Zhong[1], Ziyi Wang[1], Ruifang Gu[2], Wu Wen[2], Hao Jiang[*1], Rongya Zhang[*2]

The reliable operation of industrial equipment is essential for modern manufacturing systems. In cigarette production, as the core machinery in cigarette production, the operational status of tobacco processing machines directly affects production line efficiency and product quality. Traditional fault-diagnosis schemes built on CNNs (convolutional neural networks) and RNNs (recurrent neural networks) exhibit limited temporal modeling in the former and weak feature extraction in the latter. To address the above limitations, we present a new TDMA-CNN (Temporal-Dependency Multi-Dimensional Attention Convolutional Neural Network) architecture that leverages temporal dependencies and multi-dimensional attention for advanced fault detection in industrial machinery. The proposed architecture takes unprocessed vibration waveforms as input: a 1-D CNN first captures spatial patterns, after which a long short-term memory (LSTM) codes their temporal dynamics. Channel-wise and spatial attention modules are further introduced to dynamically emphasize salient features. Tested on the Case Western Reserve University benchmark, the method attains 98.53 % accuracy, surpassing competing techniques and offering a practical tool for the smart upkeep of machinery like tobacco processing lines.

Keywords: Convolutional Neural Network, Long Short-Term Memory, Attention Module, Fault Diagnosis

1. Introduction

With the advancement of science and technology, modern industrial equipment systems are evolving toward higher efficiency and greater complexity. As the core of the cigarette production line, the stability of tobacco machinery directly determines product quality and production throughput. Yet this growing system complexity has markedly increased both the difficulty and the criticality of fault diagnosis [1]. Without effective fault-diagnosis tools, latent anomalies may remain undetected, precipitating sudden equipment failures that reduce productivity, amplify economic losses, and even trigger safety incidents. Therefore, establishing a comprehensive and accurate fault-detection framework is not only

[1] Shifang Cigarette Factory of China Tobacco Sichuan Industrial Co., Ltd., Shifang, Sichuan Province 618400, China

[2] Technical Center of China Tobacco Sichuan Industrial Co., Ltd., , Chengdu, Sichuan Province 610100, China

pivotal for enhancing equipment reliability and extending service life, but also an indispensable prerequisite for safeguarding production continuity and operational safety [2].

Within the tobacco machinery system, bearings serve as critical moving components, the operating condition of which directly governs overall machine stability [3]. Bearing faults can immediately trigger abnormal vibration, increased noise, and functional failure, making bearing health a central indicator for assessing the condition of tobacco machinery. Consequently, research on bearing fault diagnosis in tobacco equipment holds significant practical engineering value.

Current fault-diagnosis techniques fall into three broad streams: signal processing, classical machine learning, and deep learning. Tang et al. [4] provided a comprehensive review of recently developed advanced signal-processing-based fault diagnosis methodologies. Nevertheless, these approaches still rely heavily on manual feature extraction, which inevitably introduces non-negligible human-induced errors. Machine-learning-based approaches—such as support-vector machines [5], multi-layer perceptrons [6], and K-nearest-neighbor algorithms—struggle to capture the nonlinear signal characteristics arising from the complex operating conditions of tobacco machinery, resulting in insufficient real-time performance and limited diagnostic accuracy. Deep learning has recently delivered a surge of breakthroughs in fault diagnosis. Its potent feature extraction, broad transferability, and precise recognition have drawn intense research interest [7].

Among deep-learning architectures, CNNs and RNNs are the most frequently employed. Wang et al. [8] proposed an upgraded CNN tailored for fault classification in noisy environments. This approach can accurately extract effective features from bearing vibration signals under noisy conditions, thereby achieving reliable fault identification. Lu Enhui et al. [9] devised a hybrid neural network that fuses CNN layers with attention blocks for underwater-robot fault detection. Although convolutional neural networks can efficiently extract key features from signals, relying solely on CNNs still exhibits evident limitations in capturing the temporal dependencies inherent in sequential data. Chen Yilong et al. [10] proposed an enhanced recurrent neural network that extracts temporal features from fault samples, enabling the diagnosis of compound faults in nuclear power plants. Although RNNs excel at modeling temporal relationships, they struggle to capture distinctive features. To tackle these limitations, we introduce a TDMA-CNN specifically crafted for tobacco-machinery fault diagnosis, uniting robust feature extraction with precise temporal-sequence modeling. The model takes raw vibration signals from bearing components as input. A 1-D CNN first extracts spatial features, while an LSTM layer models temporal dependencies. Channel- and spatial-attention modules are then introduced to enhance feature selection, enabling accurate fault classification. Besides providing an efficient diagnostic tool for

tobacco machinery, the method also establishes a reusable framework for the smart O&M (operation and maintenance) of comparable industrial assets.

2. Methodology

This research introduces a TDMA-CNN technique designed for fault identification in industrial machinery. The model takes raw signals as input, while a CNN extracts features and an RNN captures their temporal evolution. Channel- and spatial-attention mechanisms are integrated during feature extraction and temporal-dependency modeling to perform multi-dimensional feature refinement. A final dense layer then delivers the diagnostic label. Fig. 1 depicts the entire architecture.

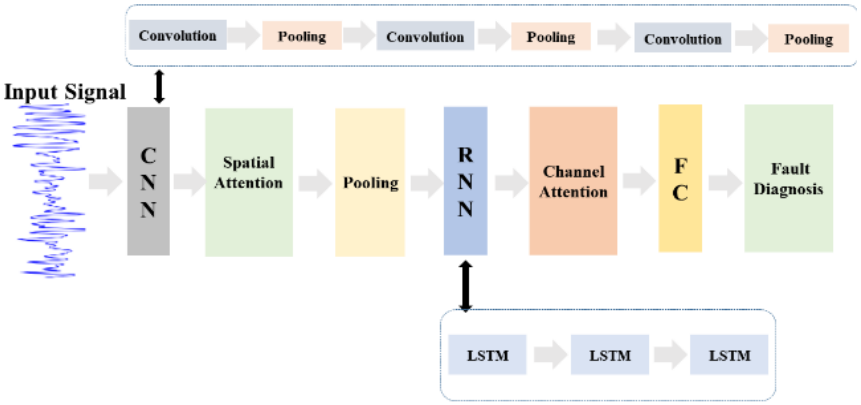


Fig. 1. Overall architecture of the proposed model.

2.1 Convolutional Neural Network

CNNs, among the most emblematic deep-learning architectures, are now staples in image recognition and computer vision [11]. A standard CNN stacks convolution, pooling and nonlinear activations. Convolution layers distill key patterns from the input with high efficiency, its mathematical formulation is given by Equation (1).

$$y_{i,j}^{k} = \sum_{m=1}^{M} \sum_{n=1}^{N} x_{i+m,j+n} \cdot w_{m,n}^{k} + b^{k} \tag{1}$$

Whereas $x_{i,j}$ denotes the input data at a given spatial (or temporal) location, $\omega_{m,n}^{k}$ represents the weight matrix of the k-th convolutional kernel, and b^{k} is the corresponding bias term.

The convolutional features are then nonlinearly mapped via an activation function, thereby enhancing the network's capacity to model complex patterns.

Commonly used activation functions include ReLU, Sigmoid, and Tanh. All CNN hyperparameters are specified in Table 1.

Table 1

Specific Parameters of Convolutional Neural Network

	Input Channels	Output Channels	Kernel Size	Stride	Padding
Conv 1	1	64	64	16	0
Pool 1	64	64	-	-	-
Conv 2	64	128	5	1	0
Pool 2	128	128	-	-	-
Conv 3	128	256	5	1	0
Pool 3	256	256	-	-	-

2.2 Recurrent Neural Network

RNNs can effectively capture the temporal dependencies in signals [12]. As a leading recurrent architecture, LSTM excels at modeling sequential data. LSTM mitigates gradient vanishing or explosion in lengthy sequences through its gating mechanism. LSTM incorporates three distinct gates to manage data flow. The input gate is responsible for updating the cell state. Meanwhile, the forget gate eliminates outdated information. Lastly, the output gate determines what data is released. The operation of LSTM unfolds in the following manner.

At each time step t, the LSTM receives several key inputs: the current input vector xt, the preceding hidden state ht−1, and the earlier cell state ct−1. The forget gate's operation is mathematically described in Equation (2).

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{2}$$

Here, σ represents the Sigmoid activation function. *ft* sets the fraction of memory to discard. *Wf* and *bf* represent the weight and bias, respectively.

The mathematical expressions for the input gate are shown in Equations (3) and (4).

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{3}$$

$$\tilde{c}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{4}$$

Here, tanh represents the activation function, i_t represents the proportion of input to be written, and $\tilde{c}_t$ is the candidate memory. The network will update the memory cell, with the mathematical expression shown in Equation (5).

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{5}$$

Here, $\odot$ represents element-wise multiplication. The output gate then yields the final result, with its formulation given in Eq. (6).

$$\begin{aligned} o_t &= \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \\ h_t &= o_t \odot \tanh\left(c_t\right) \end{aligned} \tag{6}$$

Table 2 summarizes the RNN parameters employed in our study.

Table 2

Specific Parameters of LSTM

	Input Dimension	Output Dimension	Batch_first
LSTM 1	256	128	True
LSTM 2	128	64	True
LSTM 3	64	64	True

2.3 Attention Module

By spotlighting salient features, attention mechanisms have become a staple in NLP, vision, and speech tasks. The attention mechanism can be interpreted through the relationships among Q, K, and V. Here, Q (query vector) represents the feature parameter values, K (key vector) represents the keys of the features, while V (value vector) denotes the associated data values.

Through the attention mechanism module, the input data is combined with predefined or trainable parameters to generate a weight matrix, thereby achieving weighted processing of features [13]. Our approach fuses channel-wise and spatial attention to boost multi-dimensional feature extraction.

2.3.1 Spatial Attention

Spatial attention can focus more on useful features when processing information. Fig. 2 illustrates the spatial attention block.

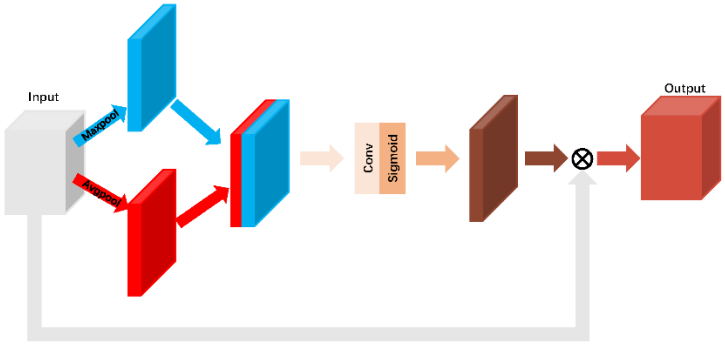


Fig. 2 Architecture Overview of the Spatial Attention Submodule

The features that are extracted by the convolutional layer are immediately channeled into the spatial attention module for further processing. The spatial attention mechanism can be expressed mathematically in the following way: Consider a four-dimensional input tensor [B, C, W, H], where B denotes the batch size, C represents the channel count, and W and H represent represent the spatial height and width, respectively. Spatial attention compresses the input tensor along the channel axis by computing its channel-wise mean and maximum, yielding two feature maps of shape [B, 1, W, H]; the corresponding operations are formalized in Equations (7) and (8).

$$Avgout(i,j) = \frac{1}{C}\sum_{c=1}^{C} x(i,j,c) \tag{7}$$

$$Max(i,j) = max_{c=1}^{c} x(i,j,c) \tag{8}$$

Here, x represents the input, $x(i,j)$ represents the feature value at position (i,j) of the input x, and C represents the channel. The mean feature map and the max feature map are combined along the channel dimension, resulting in a tensor with the shape [B, 2, W, H]. After concatenation, the features are processed through convolution and non-linear activation. The non-linear activation primarily utilizes the Sigmoid activation function. Sigmoid, akin to ReLU, serves as a nonlinear activation function. Unlike ReLU, Sigmoid confines its outputs strictly to (0, 1), making it a staple for classification tasks. In this attention module, the Sigmoid function is primarily used to assign weight coefficients. The Sigmoid function is mathematically represented in Equation (9):

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{9}$$

Here, x denotes the raw input value; after being passed through the Sigmoid activation, the transformed result is represented by σ(x). Finally, the initial input is multiplied by the tensor with the assigned weight coefficients to yield the ultimate feature representation, thereby finishing the extraction of the attended information. Table 3 details the exact parameter settings of the spatial-attention submodule.

Table 3

Parameters of the Spatial Attention Module

	Input Channels	Output Channels	Kernel Size	Stride	Padding
Conv	2	1	3	1	1
Max	256	1	-	-	-
Avg	256	1	-	-	-

2.3.2 Channel Attention

The channel-attention component is primarily implemented via a Squeeze-and-Excitation (SE) unit.By explicitly capturing cross-channel relationships, the SE block boosts the network's expressive capacity [14]. Fig. 3 depicts the SE block's operation.

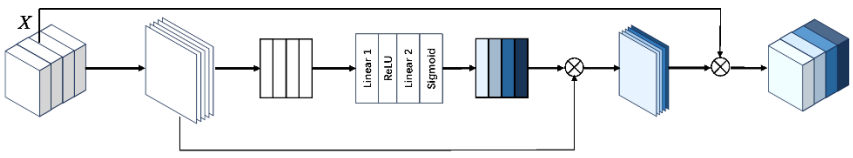


Fig. 3 Flowchart of the Channel Attention Module

The operation process of the SE module is as follows: The SE attention module takes the input X as the basis. Initially, channel-wise global average pooling compresses X, summarizing the overall context of every channel. Subsequently, the pooled features are processed through convolution and non-linear operations to compute the weight coefficients for different channels. In the process of generating weight coefficients, a hyperparameter called the reduction ratio is introduced to curb both computation overhead and parameter count. Finally, the original input X is multiplied element-wise by the tensor containing the weight coefficients to generate the output tensor. This output assigns specific weights to each channel, thus implementing a channel selection mechanism. Table 4 lists the detailed configuration of the channel-attention module.

Table 4

Specific Parameters of the Channel Attention Module

	Input Size	Output Size	Input Channels	Output channels
AdaptiveAvgPool	[1,1]	[1,1]	4	4
Linear1	-	-	4	4
Linear2	-	-	4	4

3 Experiments and Results Analysis

3.1 Experiment Setup

The workstation runs on an Intel (R) Core (TM) i7-14700HX. Throughout training, a batch size of 32 is selected to balance memory utilization and gradient stability, ensuring the entire experiment proceeds smoothly without abrupt resource spikes. For the implementation backbone we adopt PyTorch; optimization is handled by Adam, whose adaptive moment estimates and per-parameter learning rates provide robust convergence while continuously refining the network's weight space throughout training. Table 5 lists the model's training hyper-parameters.

Table 5

Specific Parameters for Model Training

Parameter	Value
Epoch	100
Batch Size	32
Learning Rate	0.0001

3.2 Dataset Introduction

The experiment adopts the standard bearing vibration dataset provided by Case Western Reserve University to assess the model's diagnostic accuracy. The CWRU bearing dataset, contains vibration signals from normal and faulty bearings under controlled shaft speeds and load conditions; its open availability and thorough labeling make it a widely accepted benchmark for validating fault-diagnosis and remaining-life-prediction methods. For systematic evaluation, the dataset is organized into four clear categories: normal condition, ball element defect, inner-

ring defect, and outer-ring defect. Each fault state is further subdivided into different fault magnitudes (7 inches, 14 inches, and 21 inches). Table 6 summarizes the key parameters, while Fig. 4 displays the corresponding time-domain waveforms.

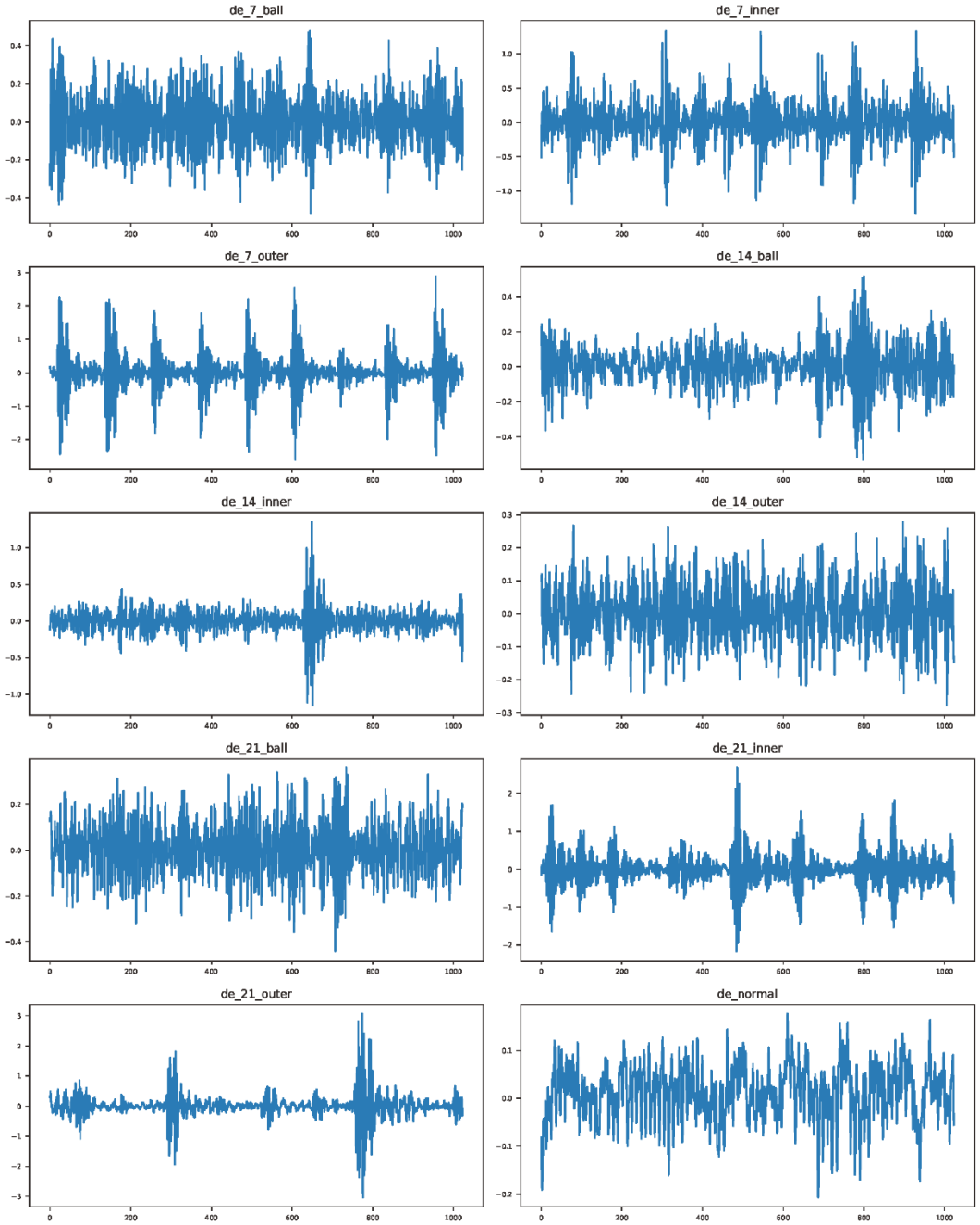


Fig. 4 Time-Domain Waveform of the Signal

Table 6

Parameters of the Case Western Reserve University Bearing Dataset

State	Label	Number of Samples	Signal Duration (s)
Healthy	De_normal	1024	0.085
Ball Fault_7	De_7_ball	1024	0.085
Ball Fault_14	De_14_ball	1024	0.085
Ball Fault_21	De_21_ball	1024	0.085
Inner Race Fault_7	De_7_inner	1024	0.085
Inner Race Fault_14	De_14_inner	1024	0.085
Inner Race Fault_21	De_21_inner	1024	0.085
Outer Race Fault_7	De_7_outer	1024	0.085
Outer Race Fault_14	De_14_outer	1024	0.085
Outer Race Fault_21	De_21_outer	1024	0.085

3.3 Experimental Results

In this experiment, the original vibration signals of the equipment are used as the input to the model, and the network is employed to extract features, ultimately achieving the fault detection task. Figs. 5 and 6 show the loss function curve and training accuracy curve of the model during training, as well as the final classification confusion matrix. Figs. 5 and 6 show that the proposed model converges smoothly and remains stable after training, ultimately attaining a fault-diagnosis accuracy of 98.53 %.

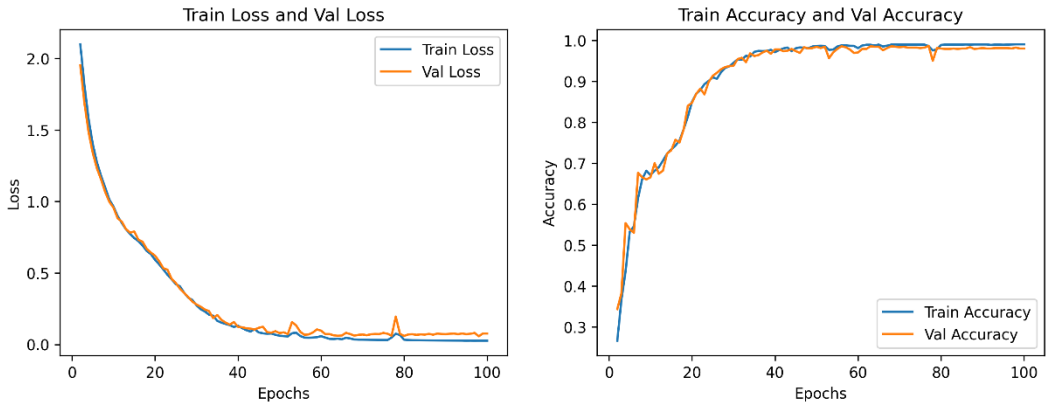


Fig. 5 Model Training Loss Function Curve and Accuracy Curve

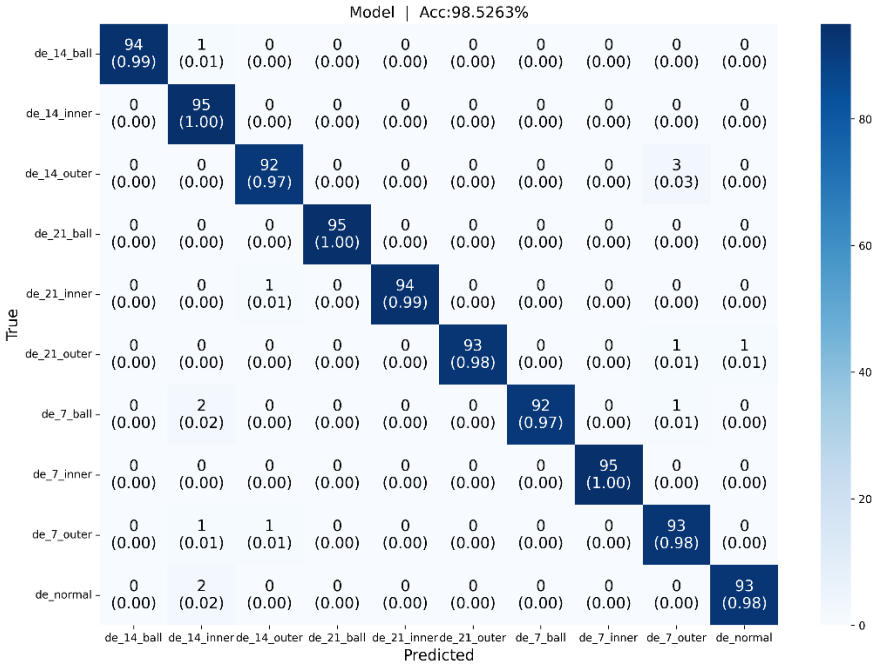


Fig. 6 Model Classification Confusion Matrix

To benchmark its performance, the proposed model is evaluated alongside several representative networks under identical conditions. Table 7 shows the accuracy of fault detection using different networks. Table 7 confirms that the proposed model delivers the strongest detection performance, achieving 98.53 % accuracy—outperforming all compared approaches.

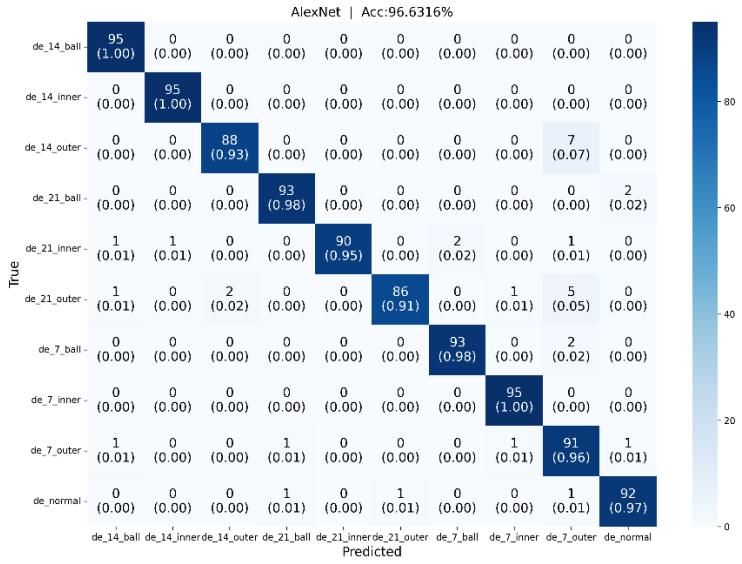
Table 7

Comparison of Experimental Results

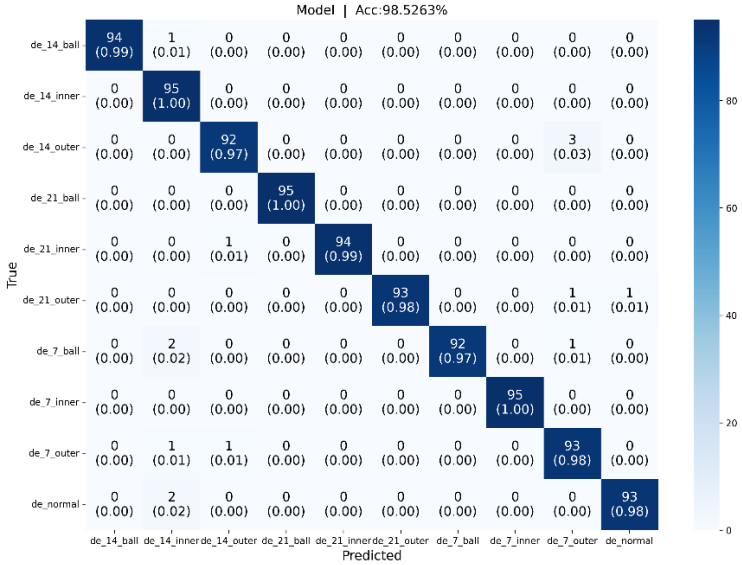
	Accuracy	Precision	Recall	F1 Score
Model	98.5263%	98.5766%	98.5363%	98.5354%
AlexNet	96.6316%	96.8142%	96.6316%	96.6549%
LSTM	47.8947%	48.5001%	47.8947%	46.1181%
GRU	44.2105%	46.6148%	44.2105%	42.5589%

Table 7 shows that our time-aware multidimensional attention CNN surpasses AlexNet, lifting diagnostic accuracy by a clear margin. Its advantage stems from jointly extracting rich spatial patterns and explicitly modeling the temporal evolution of those features, thereby capturing the sequential relationships within the vibration signals. Complementing this, spatial and channel attention mechanisms selectively amplify task-relevant features while attenuating redundant ones, sharpening the network's focus on the most informative components. Although LSTM and GRU have strong capabilities for modeling temporal dependencies, there is some noise in the signal that LSTM and GRU cannot effectively suppress. As a result, they fail to achieve high fault diagnosis accuracy.

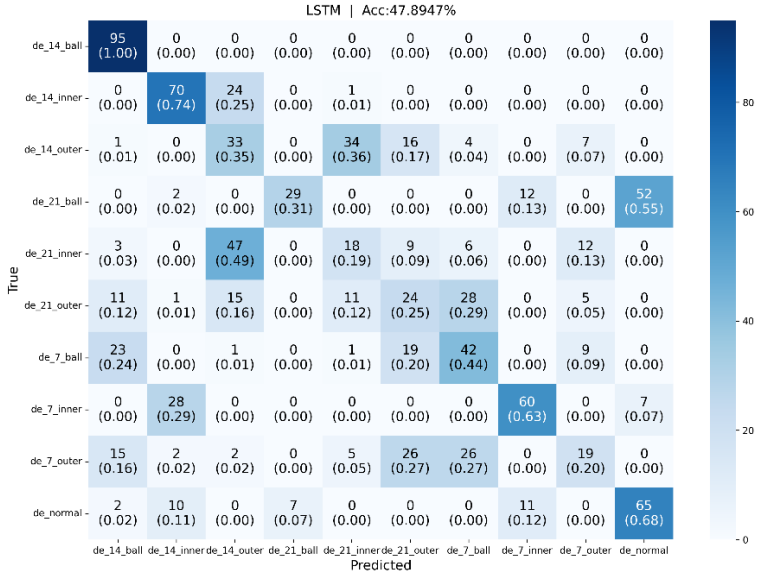
Fig. 7 presents the confusion matrices for all compared models, offering a side-by-side view of their classification performance.



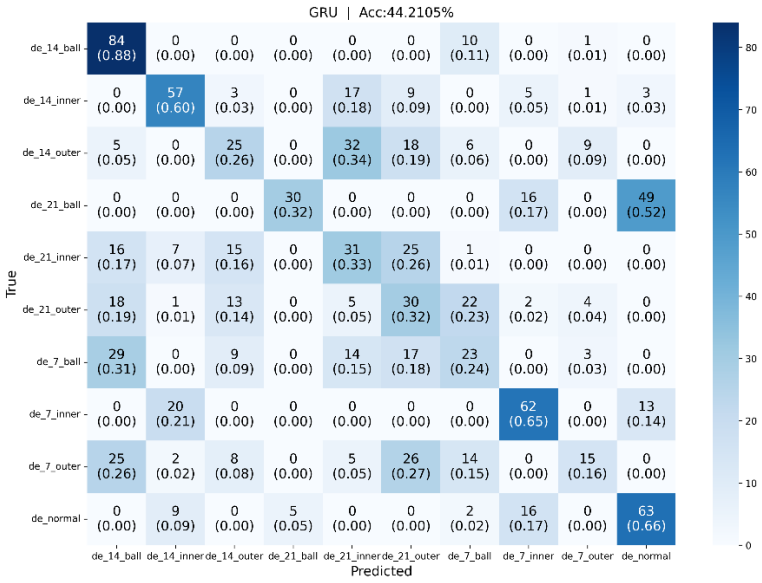
(a) AlexNet



(b)Model



(c) LSTM



(d) GRU

Fig. 7 Confusion Matrices of Different Models

We conducted an ablation study to isolate and quantify how much the spatial-attention and channel-attention units each boost overall performance. Table 8 shows the accuracy of fault detection using the ECA attention module, CBAM attention module, and without using any attention module [15]. Table 8 shows that

integrating both channel and spatial attention yields higher detection scores than using either module alone.

Table 8

Comparison of Ablation Study Accuracy

	Accuracy	Precision	Recall	F1 Score
Model	98.5263%	98.5766%	98.5363%	98.5354%
ECA Module	97.7895%	97.8518%	97.7895%	97.8075%
CBAM Module	97.4211%	97.4580%	97.4211%	97.4216%
No Attention Module	93.6462%	92.9975%	93.6462%	93.6213%

Table 8 indicates that the fault diagnosis accuracy achieved with the ECA module and CBAM attention module falls short compared to the accuracy of the model introduced in this paper. The joint use of spatial and channel attention enables the network to refine signal features along complementary dimensions, boosting its ability to isolate critical patterns and suppress noise. In contrast, the ECA and CBAM modules enhance the signal's feature representation from only one perspective. The accuracy is the lowest when no attention module is used, as the absence of an attention mechanism fails to suppress redundant feature information, thereby interfering with the final fault diagnosis.

4. Conclusion

This study introduces a convolutional neural network that couples temporal modeling with multidimensional attention, tailored for reliable fault diagnosis in industrial equipment. Initially, a 1-D CNN captures local spatial patterns in the raw vibration traces, after which stacked LSTM layers learn their extended temporal dynamics, jointly yielding rich spatio-temporal representations for fault discrimination. Channel- and spatial-attention blocks are further inserted to spotlight salient features across both dimensions, with a final fully-connected layer mapping the refined representation to fault classes. Relative to competing architectures, the proposed approach retains strong feature-extraction power while explicitly capturing the sequential correlations inherent in vibration signals. Complementing this, channel- and spatial-attention modules selectively enhance informative signal components while suppressing noise, sharpening the model’s focus on fault-related patterns. Evaluated on the CWRU bearing dataset, the network delivered a fault-classification accuracy of 98.53 %. This result surpasses all compared baselines, offering clear evidence of the model’s superior diagnostic capability. Although the model performs excellently under typical working conditions, there is still room for improvement in terms of diagnostic robustness in scenarios involving multiple coupled faults in electrical equipment, feature stability in extremely noisy electromagnetic environments, and the processing efficiency of high-dimensional real-time signals. In response to the intelligent diagnosis needs of electrical faults, the model's spatiotemporal feature modeling method can be

directly applied to the vibration and current signal analysis of transformers, motors, and other equipment. By solidifying the feature processing logic of temporal dependence and attention mechanisms, it provides technical support for early warning of motor bearing faults and the assessment of the sub-health status of electrical systems.

REFERENCES

[1] Qiu S, Cui X, Ping Z, et al. Deep learning techniques in intelligent fault diagnosis and prognosis for industrial systems: A review. Sensors, 2023, 23(3): 1305.

[2] Yu J, Zhang Y. Challenges and opportunities of deep learning-based process fault detection and diagnosis: a review. Neural Computing and Applications, 2023, 35(1): 211-252.

[3] Souza R M, Nascimento E G S, Miranda U A, et al. Deep learning for diagnosis and classification of faults in industrial rotating machinery. Computers & Industrial Engineering, 2021, 153: 107060.

[4] Tang L, Tian H, Huang H, et al. A survey of mechanical fault diagnosis based on audio signal analysis. Measurement, 2023, 220: 113294.

[5] Li Y, Xu M, Wei Y, et al. A new rolling bearing fault diagnosis method based on multiscale permutation entropy and improved support vector machine based binary tree. Measurement, 2016, 77: 80-94.

[6] Vieira R G, Dhimish M, de Araújo F M U, et al. Comparing multilayer perceptron and probabilistic neural network for PV systems fault detection. Expert systems with applications, 2022, 201: 117248.

[7] LeCun Y, Bengio Y, Hinton G. Deep learning. nature, 2015, 521(7553): 436-444.

[8] Wang H, Liu Z, Peng D, et al. Interpretable convolutional neural network with multilayer wavelet for Noise-Robust Machinery fault diagnosis. Mechanical Systems and Signal Processing, 2023, 195: 110314.

[9] Lu Enhao, Ji Daxiong, Wang Xia. The research on fault diagnosis of Autonomous Underwater Vehiclebased on Convolutional Neural Networks and attention mechanisms. Marine Information Technology and Applications, 2025, 40(02): 65-73.

[10] Chen Yilong, Lin Meng, Zhou Shiqi. Compound fault diagnosis method of nuclear powerplant based on recurrent neural network. Journal of Naval University of Engineering, 2025, 37(01): 36-42.

[11] Zhou Yu, Wang Yan. Review of Bearing Fault Diagnosis Based on Convolutional Neural Network. Journal of Beijing Institute of Graphic Communication, 2024, 32(08): 6-12.

[12] Salehinejad H, Sankar S, Barfett J, et al. Recent advances in recurrent neural networks. arXiv preprint arXiv:1801.01078, 2017.

[13] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing, 2021, 452: 48-62.

[14] Smith W A, Randall R B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. Mechanical systems and signal processing, 2015, 64: 100-131.

[15] Chen L, Yao H, Fu J, et al. The classification and localization of crack using lightweight convolutional neural network with CBAM. Engineering Structures, 2023, 275: 115291.