2024 6th International Conference on Intelligent Control, Measurement and Signal Processing (ICMSP) | 979-8-3315-1655-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMSP64464.2024.10866284

Attention-embedded multi-scale quadratic convolutional neural network for early bearing fault diagnosis

Yichen Li
China Tobacco Hubei Industrial Co. Ltd.
Hubei Xinye Reconstituted Tobacco Development Co. Ltd.
Wuhan, China
liyichen@hbtobacco.cn

Nina Song*
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology
Wuhan, China
*Corresponding author: m202373768@hust.edu.cn

Wei Zeng
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology
Wuhan, China
zengwei@hust.edu.cn

Lei Jiang
China Tobacco Hubei Industrial Co. Ltd.
Hubei Xinye Reconstituted Tobacco Development Co. Ltd.
Wuhan, China
jianglei@hbtobacco.cn

Abstract—**Bearings are vital components in various industries such as chemical, textile, food, and tobacco. Timely fault diagnosis is crucial for maintenance, performance, safety, and cost efficiency. Traditional methods rely on manual feature extraction and are time-consuming. This study proposes an attention-embedded multi-scale quadratic convolutional neural network (MQCNN) based on the attention-embedded quadratic network (QCNN). It utilizes an attention mechanism from quadratic neurons (qttention) to enhance noise resistance and automatically extracts frequency-specific signal features through varied convolution kernels and depths. This method improves diagnostic accuracy and offers significant economic advantages in engineering applications.**

Keywords: Bearing fault diagnosis, multi-scale, quadratic convolutional neural network (QCNN), quadratic attention (qttention)

I. Introduction

Bearing early fault diagnosis is essential in industries such as tobacco, textiles, and food production. It plays a crucial role in maintenance management, enhancing production efficiency, ensuring safety, minimizing downtime, and reducing costs. In the tobacco sector, where high-speed machinery is prevalent, bearing health directly impacts production stability and product quality. Effective monitoring often involves connecting instruments to rotating bearings to collect vibration signals for diagnosis. Past diagnostic approaches can be divided into two main types[1]: signal processing-based methods and data-driven methods.

Signal processing methods for bearing fault diagnosis, like Fast Fourier Transform and Hilbert Transform, analyze signals in the frequency or time-frequency domain. Recent advancements include the Fault Information-Guided Variational Mode Decomposition(FIVMD)[2], which effectively extracts weak bearing transients with improved stability and accuracy compared to traditional methods. However, these approaches face challenges with noise interference requiring complex denoising algorithms and issues with automation and scalability.

To address these limitations, data-driven methods, especially deep learning, have been widely adopted in bearing fault diagnosis over the past decade. The 1D-Convolutional Neural Network(1D-CNN)[3] can directly process raw signals without requiring preprocessing like feature extraction or denoising. The Wide First-Layer Kernel Deep Convolutional Neural Network(WDCNN)[4] uses wide kernels for effective signal extraction and has shown strong performance across various benchmarks, the 1D window-based multi-head self-attention (1D W-MSA)[5] and a Dual-Path Recurrent Neural Network with Wide First Kernel(RNN-WDCNN)[6] were introduced. Additionally, Chen *et al.*[7] proposed a multi-scale CNN-LSTM network for automatic feature learning with good experimental results. Liao *et al.*[8] enhanced classification accuracy and interpretability by integrating attention modules into their approach using quadratic neurons.

Despite advancements in fault detection models, accuracy drops with noise and interpretability is limited. Secondary networks excel in feature extraction, improving diagnostic performance for noisy bearing vibration signals. In summary, our contributions are twofold:

1. We propose MQCNN, as illustrated in Fig.1. This model integrates various kernel sizes to automatically extract signal features from raw data, enabling effective identification of fault characteristics.

2. The quadratic convolutional neural network pays attention to periodic high-amplitude vibrations, showing strong representation capability for early bearing fault diagnosis. It maintains high accuracy even in noisy tobacco production environments, reducing manual analysis time and costs while enhancing the timeliness and accuracy of diagnosis.

II. Methodology

A. Quadratic Convolutional Neural Network

Quadratic Convolutional Neural Networks (QCNN) are a novel architecture of CNN designed to enhance the model's expressive power by introducing quadratic feature mappings. Compared to CNNs, QCNNs are capable of capturing more complex nonlinear relationships. A quadratic neuron[9] integrates two inner products and a power term of the input vector before applying a nonlinear activation function. Mathematically, the input vector is $(x_1, x_2, \cdots, x_n)$, the output of the quadratic neuron *f(x)* is:

$$\sigma\big(f(x)\big) = \sigma\left(\left(\sum_{i=1}^{n} w_i^r x_i + b^r\right)\left(\sum_{i=1}^{n} w_i^g x_i + b^g\right) + \sum_{i=1}^{n} w_i^b x_i^2 + c\right) \tag{1}$$

$$\sigma\big(f(x)\big) = \sigma\big((x^T w^r + b^r)(x^T w^g + b^g) + (x \odot x)^T w^b + c\big) \tag{2}$$

where $\sigma(\cdot)$ is a nonlinear activation function, $\odot$ represents the Hadamard product, w^r, w^g, $w^b \in R^n$ are weight vectors, and b^r, b^g, c are biases.

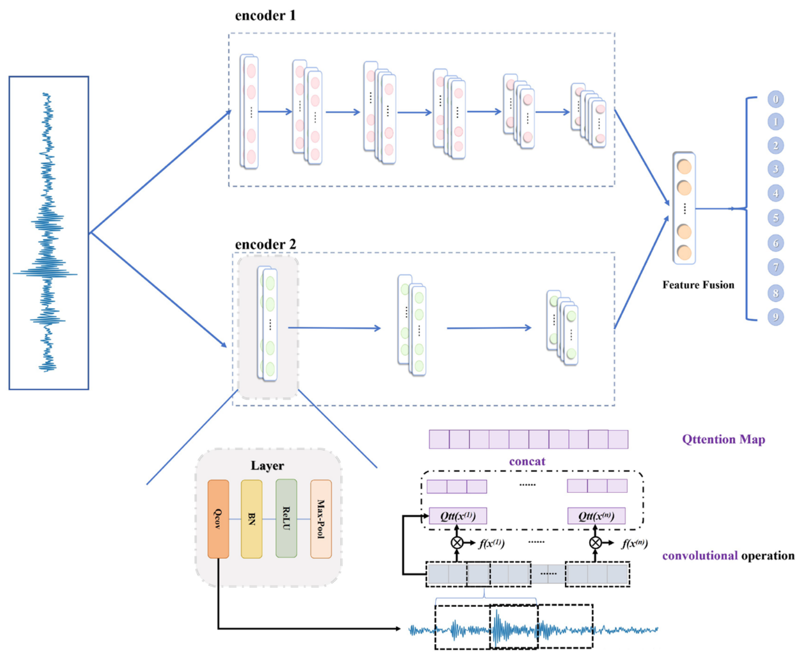

Figure 1 The proposed attention-embedded multi-scale quadratic convolutional neural network(MQCNN).

QCNN utilizes quadratic mappings in convolution operations, allowing outputs of the same layer to include interaction terms between features rather than just linear combinations. This enables the network to learn richer feature representations and extract higher-level features while enhancing feature relationships. Additionally, QCNN employs a parameter-sharing mechanism to reduce model complexity and improve training efficiency. The use of quadratic neurons aims to enhance diagnostic performance for early bearing faults

B. Efficient Neuron-induced Attention

In deep learning, attention mechanisms[10] help models focus on the most relevant parts of input data, improving performance and interpretability. While Liao *et al.*[8] found that a quadratic neuron integrates an attention mechanism called "qttention." By decomposing the quadratic neuron's functional formula, a mathematical representation akin to attention can be derived (omit the constant term for simplicity).

$$\sigma\left((x^T w^r + b^r)(x^T w^g + b^g) + (x \odot x)^T w^b + c\right) = \sigma\left(x^T (x \odot w^b + w^g x^T w^r + w^g b^r + w^r b^g)\right) \tag{3}$$

$$RawQtt(x) = x^T \odot w^b + w^g (x^T w^r) \tag{4}$$

The bias term can be neglected here, as it does not significantly affect the different inputs. Additionally, calculate the gradient of RawQtt (x), and take its absolute value to obtain the qttention map:

$$Qtt(x) = |Grad(RawQtt(x))| \tag{5}$$

This way, the gradient highlights the key parts of the attention weights, allowing for better visualization of changing trends and further mitigating common biases.

Qttention is closely integrated with the convolution operation. As illustrated in Fig. 1, the process for generating the qttention map using convolution over the entire signal is described. Each convolutional kernel functions as a single neuron, moving with a constant stride and convolving with the input signal according to Equation (1). During this process, a qttention map is created, and the final qttention map for the entire signal is formed by concatenating all local receptive field qttention maps. For overlapping adjacent receptive fields, we average the qttention scores at the overlaps. Qttention boosts computational efficiency. . Its integration with convolution provides detailed granularity, making it ideal for diagnosing bearing faults characterized by periodic short-range high-amplitude vibrations.

C. MQCNN

The Multi-Scale Convolutional Neural Network (MSCNN)[11] is a specialized architecture that processes input across various scales to capture feature information at different sizes. This multi-scale feature fusion improves the network's ability to consider both long-range and short-range information, enhancing classification performance. MSCNN offers flexibility for complex scenes by allowing adjustments in the size and number of convolutional kernels based on task requirements. The MQCNN model is built upon this concept.

Prior to inputting the signals, we downsampled them to improve computational efficiency and performance. The powerful feature extraction capabilities of quadratic neurons are leveraged throughout the model, allowing for the derivation of qttention that focuses on periodic short-range high-amplitude vibrations, thereby enhancing performance in detecting early bearing faults. The first module features two one-dimensional QCNNs with different kernel sizes and depths as feature extractors. When the raw signal is input, these convolutional operations effectively extract features across various frequency domains. For one-dimensional vibration signals, a larger receptive field in encoder_1 uses wide convolutional kernels to capture low-frequency features, while encoder_2 focuses on high-frequency signals with a smaller receptive field. The feature vectors from both encoders are fused together. The second module employs fully connected layers with a softmax function for classification, using ReLU as the activation function for the hidden layer. The structural parameters of the model are in Table 1.

TABLE I. STRUCTURAL PARAMETERS

	Name	Kernel size/stride	Filters
Encoder1	Q1Conv1D_1	64/8	16
	Q1Conv1D_2-6	5/1	32/64
Encoder2	Q2Conv1D_1	16/8	16
	Q2Conv1D_2-3	16/8	32/64
QConv1D contains convolution operation, BN, Relu, Maxpool			

III. Experiments

The experimental data in this study is sourced from the Rolling Element Bearing Data Center at Case Western Reserve University (CWRU)[12]. The CWRU datasets is a globally recognized standard for bearing fault diagnosis. Therefore, this research employs the CWRU bearing datasets to validate the proposed MQCNN. We initially trained the network using the training datasets that includes nine different fault conditions and one normal condition, resulting in a total of ten classes. Experimental results indicate that the accuracy of MQCNN is 99.79%. Subsequently, we introduce noise into the raw signals to simulate a noisy tobacco manufacturing environment, then comparing the accuracy and robustness against noise of our model with that of other models under varying signal-to-noise ratios (SNRs).

A. Datasets and training set

The data acquisition system[12] at the CWRU Bearing Center involves two deep groove ball bearings installed at both the fan end (FE) and drive end (DE) of the motor. In this study, we utilize the vibration signals collected from the DE side at a sampling rate of 12 kHz for the classification tasks.In the experiments, each diagnostic instance consists of 2048 data points. To aid convolutional neural network training, we implement data augmentation through overlapping sampling, where each extracted training segment overlaps with the next, as shown in Fig. 2.

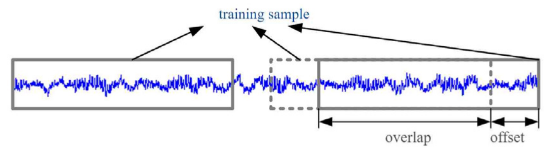

Figure 2 Overlapping sampling process diagram.

The loss function used is cross-entropy, and the Stochastic Gradient Descent (SGD) optimizer is selected for network optimization. A grid search is performed to explore the hyperparameter space, resulting in a batch size set to 64. For MQCNN, the learning rates for and the quadratic neurons differ from the linear terms. We utilize accuracy scoring to validate the performance of the proposed method, which is defined as:

$$Acc = \frac{TP+TN}{TP+TN+FN+FP} \tag{6}$$

Where TP, TN, FP, and FN represent the counts of true positives, true negatives, false positives, and false negatives, respectively. Additionally, we introduce Gaussian noise into the original input signals to evaluate the model's performance in a noisy environment. The SNR is defined as *10log10(Ps/Pn)*, where P_s and P_n denote the average power of the signal and noise, respectively. All experiments are conducted in Windows 10 with an Intel i7 12700k CPU at 2.30 GHz and one NVIDIA RTX 3060Ti 8GB GPU. Our code is written in Python 3.8 with PyTorch, an open-source deep learning framework.

B. Experimental results

TABLE II. THE PERFORMANCE OF THE MODEL AT SNR = 0

metrics	TA	Val_acc	F1	FPR	PRE	Recall
score	99.94%	99.78%	0.99	0.00098	0.99	0.99

Table 2 shows that our experiments yield an accuracy near 100%. The False Positive Rate (FPR) and Precision (PRE) indicate high accuracy in predicting positive samples. Recall is 0.99, and the F1 Score is close to 1, demonstrating strong overall performance in detecting positive samples.

TABLE III. PERFORMANCE OF MQCNN AT SNR = -6 TO 4

SNR /dB	metrics					
	TA	Val_acc	F1	FPR	PRE	Recall
0	99.94%	99.79%	0.99	0.0010	0.991	0.99
2	99.18%	98.72%	0.98	0.0026	0.981	0.98
4	99.96%	100.0%	1.0	0.0	1.0	1.0
-2	99.24%	98.56%	0.95	0.0054	0.997	0.95
-4	99.08%	98.72%	0.98	0.0026	0.978	0.98
-6	99.20%	96.31%	0.87	0.0128	0.923	0.97

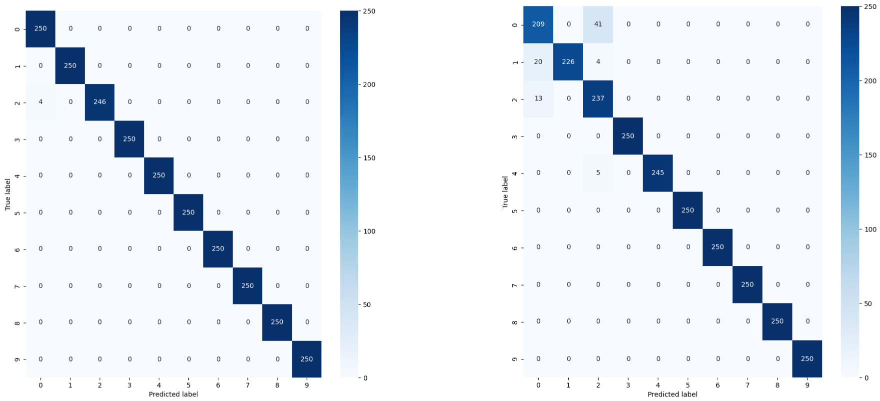

Figure 3 The confusion matrix of the experiment.

Figure 3 displays two confusion matrix for the training accuracy results. Multiple experiments indicate that the primary reason for the lower accuracy is the misclassification of the first three class labels. This may be attributed to the similarities among these fault signals, which complicate classification in the presence of noise. Therefore, it is crucial to evaluate the model's ability to accurately classify faults in complex operating environments.

To validate effectiveness of MQCNN, accuracy tests were conducted under different noise levels to assess the network's robustness in a noisy production environment. We used SNR as an indicator of the ratio between signal strength and background noise, with SNR set between 4 dB and -6 dB during processing. Through a series of experiments, the performance metrics of the MQCNN under varying noise conditions are presented in the Table 3.

C. Compared with other methods

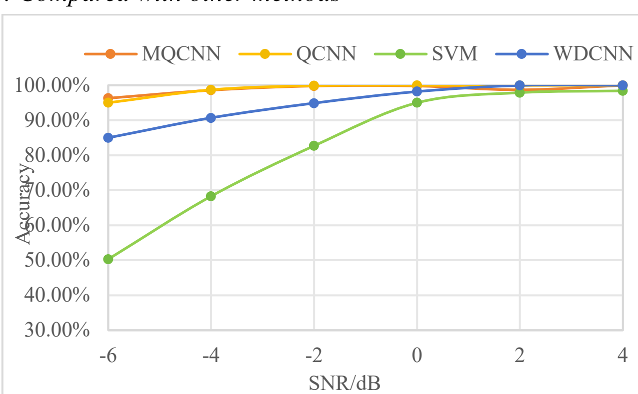

Figure 4 Accuracy of algorithms in different noise environments.

The comparative experimental results of the proposed model against SVM, QCNN, and WDCNN under varying SNR are illustrated in the Fig.4. While all models exhibit commendable performance in environments with high signal quality, it is evident that the model utilizing quadratic neurons demonstrates superior noise resistance. Additionally, the integration of a multi-scale convolution approach within the model configuration contributes to enhanced accuracy and robustness. While all models perform well in high-quality signal environments, the MQCNN, utilizing quadratic neurons, demonstrates superior noise resistance. The integration of a multi-scale convolution approach further enhances its accuracy and robustness, highlighting notable performance of MQCNN in both aspects.

IV. Conclusion

In this article, MQCNN is proposed based on QCNN and MSCNN, which incorporates quadratic neurons, blending traditional CNN characteristics with a unique attention mechanism. Our model automatically extracts low-frequency and high-frequency features through convolutions of varying sizes and depths. Testing on the CWRU datasets demonstrates that our model outperforms classical algorithms in interpretability and accuracy. Additionally, it maintains strong performance in noisy environments, like tobacco production workshops, showcasing its practical effectiveness and economic viability for engineering applications.

REFERENCES

[1] Y. Ding, J. Zhuang, P. Ding, and M. Jia, "Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings,"Reliability Engineering & System Safety, vol. 218, p. 108126, 2022.

[2] Ni Q, Ji J C, Feng K, et al. A fault information-guided variational mode decomposition (FIVMD) method for rolling element bearings diagnosis[J]. Mechanical Systems and Signal Processing, 2022, 164: 108216.

[3] Kiranyaz S, Ince T, Abdeljaber O, et al. 1-D convolutional neural networks for signal processing applications[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 8360-8364.

[4] Zhang W, Peng G, Li C, et al. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals[J]. Sensors, 2017, 17(2): 425.

[5] Wei Q, Tian X, Cui L, et al. WSAFormer-DFFN: A model for rotating machinery fault diagnosis using 1D window-based multi-head self-attention and deep feature fusion network[J]. Engineering Applications of Artificial Intelligence, 2023, 124: 106633.

[6] Shenfield A, Howarth M. A novel deep learning model for the detection and identification of rolling element-bearing faults[J]. Sensors, 2020, 20(18): 5112

[7] Chen X, Zhang B, Gao D. Bearing fault diagnosis base on multi-scale CNN and LSTM model[J]. Journal of Intelligent Manufacturing, 2021, 32(4): 971-987.

[8] Liao J X, Dong H C, Sun Z Q, et al. Attention-embedded quadratic network (qttention) for effective and interpretable bearing fault diagnosis[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-13.

[9] Fan F, Xiong J, Wang G. Universal approximation with quadratic deep networks[J]. Neural Networks, 2020, 124: 383-392.

[10] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation[J]. arXiv preprint arXiv:1508.04025, 2015.

[11] Cui Z, Chen W, Chen Y. Multi-scale convolutional neural networks for time series classification[J]. arXiv preprint arXiv:1603.06995, 2016.

[12] https://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website