Attention-Embedded Quadratic Network (Qttention) for Effective and Interpretable Bearing Fault Diagnosis

Jing-Xiao Liao, Graduate Student Member, IEEE, Hang-Cheng Dong, Zhi-Qi Sun, Jinwei Sun, Shiping Zhang, and Feng-Lei Fan, Member, IEEE

Abstract—Bearing fault diagnosis is of great importance to decrease the damage risk of rotating machines and further improve economic profits. Recently, machine learning, represented by deep learning, has made great progress in bearing fault diagnosis. However, applying deep learning to such a task still faces major challenges such as effectiveness and interpretability: 1) when bearing signals are highly corrupted by noise, the performance of deep learning models drops dramatically and 2) a deep network is notoriously a black box. It is difficult to know how a model classifies faulty signals from the normal and the physics principle behind the classification. To solve these issues, first, we prototype a convolutional network with recently invented quadratic neurons. This quadratic neuron-empowered network can qualify the noisy bearing data due to the strong feature representation ability of quadratic neurons. Moreover, we independently derive the attention mechanism from a quadratic neuron, referred to as qttention, by factorizing the learned quadratic function in analog to the attention, making the model made of quadratic neurons inherently interpretable. Experiments on the public and our datasets demonstrate that the proposed network can facilitate effective and interpretable bearing fault diagnosis. Our code is available at https://github.com/asdfghjk/QCNN_for_bearing_diagnosis.

Index Terms—Bearing fault diagnosis, deep learning, neural network, quadratic convolutional neural network (QCNN), quadratic neuron-induced attention (qttention).

I. INTRODUCTION

THE reliability of rotating machines such as wind turbines and aircraft engines is a critical issue and must be ensured at all times in the industrial field. Among major mechanical components of a rotating machine, bearings are the most popular source of faults. According to several studies [1], bearing faults are responsible for 40%–70% of electromagnetic drive system failures. Therefore, the diagnosis to bearing faults is of great criticality to improve the availability of rotating machines and further avoid economic loss. A common and viable method to detect bearing faults is analyzing vibration signals measured by attaching the measuring instrument to the rotating bearing [2]. Previous diagnosis works can be divided into two categories: signal processing-based methods and data-driven methods [3].

Signal processing-based methods usually utilize Fourier transform [4], short-time Fourier transform [5], Hilbert-Huang transform (HHT) [6], [7], and so on to transform a sequence of a signal into the frequency domain or time-frequency domain. Then, the spectral analysis is conducted to capture the features of failed bearings from the normal. Despite theoretical soundness, these methods suffer from the following weaknesses: 1) because bearings are always installed in a complex mechanical system, the measured signal is filled with interference from other mechanical components and background noise from the measurement device [1]. This requires practitioners to carefully design a sophisticated denoising model to extract various fault characteristics from the complex noise environment and 2) because different bearing structures and diameters cause different characteristic frequencies of the faults, the manually determined parameters are required in performing signal processing-based methods to maximize their diagnosis performance, which is neither intelligent nor scalable due to the complexity of industrial fields [3].

To overcome the above weaknesses, machine learning, particularly deep learning, is widely introduced for bearing fault diagnosis with the promise of big data-boosted end-to-end accurate feature discrimination. The past several years have seen a variety of deep learning models developed. Wen et al. [8] first folded 1-D time-domain signals into 2-D images, and then used LeNet-5 as a backbone to classify these images. Concurrently, more studies directly worked on 1-D time-domain signals with the 1-D-CNN models [8], [9]. For example, deep convolutional neural networks with wide first-layer kernels (WDCNN) adopted a wide convolution kernel in the first layer as the signal extractor for 1-D signals and achieved satisfactory results on several bearing fault benchmarks [10]. Due to the outperformance of the WDCNN, more advanced models have been devised based on the WDCNN [11], [12]. Moreover, the attention module was introduced in bearing fault diagnosis because it helps a network focus on fault features and further enhances feature extraction capabilities [11], [13], [14], [15], which not only improves the network’s classification accuracy but also makes the network interpretable.

Manuscript received 23 November 2022; revised 10 February 2023; accepted 6 March 2023. Date of publication 20 March 2023; date of current version 30 March 2023. The Associate Editor coordinating the review process was Dr. Md. Moinul Hossain. (Corresponding authors: Shiping Zhang; Feng-Lei Fan.)

Jing-Xiao Liao, Hang-Cheng Dong, Zhi-Qi Sun, Jinwei Sun, and Shiping Zhang are with the School of Instrumentation Science and Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: spzhang@hit.edu.cn).

Feng-Lei Fan is with Weill Cornell Medicine, Cornell University, Ithaca, NY 14850 USA (e-mail: hitfanfenglei@gmail.com).

Digital Object Identifier 10.1109/TIM.2023.3259031

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Despite that deep learning models have made a big stride in bearing fault detection, there are still challenges ahead that need to be addressed: 1) (effectiveness) currently, deep learning models can perform superbly, i.e., achieving over 90% accuracy, when the signals are clean or big data are curated. However, when signals are highly corrupted by noise, the performance of deep learning models drops dramatically. The above issue should be accommodated if one wants to implement a deep learning model in industrial fields and 2) (interpretability) deep learning, comprising a series of nonlinear or recursive operations, is notoriously a black box [16]. Due to the lack of interpretability, it is hard to know the physics principle used by a model to pick fault signals out and what changes can be made to further boost the model’s diagnosis performance.

To deal with the challenges in effectiveness and interpretability, here we turn our eyes to the quadratic neurons and quadratic networks. Recently, motivated by introducing neuronal diversity in deep learning, a new type of neuron called quadratic neuron [17] was proposed by replacing the inner product in the conventional neuron with a simplified quadratic function. Why are quadratic neurons promising in solving the effectiveness and interpretability issues? For the former, it has been empirically and theoretically shown [17], [18], [19] that a quadratic network enjoys a high feature extraction ability. Due to such a character, quadratic models have the potential to qualify the noisy bearing diagnosis task. For the latter, as our independent methodological finding, the learned quadratic function in a quadratic neuron can induce an attention map through factorization, referred to as attention, making a quadratic network interpretable. Compared to the classical attention mechanism, which is global and computationally expensive, the attention spotlights local fine-grained importance maps and consumes fewer parameters, which is a valuable addition to the attention-based interpretability. In summary, our contributions are twofold.

1) We propose a simple and effective model that is made 120 of quadratic neurons for bearing fault diagnosis, as shown in Fig. 1. The quadratic neurons directly augment the model to outperform other state-of-the-arts in noisy data settings. Different from previous structural modifications, our innovation is at the neuronal level, applying different neurons to bearing fault diagnosis and upgrading CNNs-based bearing fault diagnosis models.

2) Methodologically, we derive the attention mechanism from quadratic neurons, which greatly facilitates the understanding to the model made of quadratic neurons. Our interpretation emphasizes the model’s spectral properties with the help of envelope spectrum analysis.

II. SPECTRAL CHARACTERIZATION OF BEARING FAULTS

As shown in Fig. 2, rolling element bearings consist of four clearly differentiated components: inner race, balls or rollers, cage, and outer race. The deterioration of each element will generate one or more characteristic frequencies in the frequency spectrum that allows a fast and easy identification. The four possible bearing failing frequencies are: 1) ball pass frequency outer (BPFO) or outer race failing frequency that corresponds to the number of balls passing through a given point of the outer race each time the shaft completes a turn; 2) ball pass frequency inner (BPFI) or inner race failing frequency that corresponds to the number of balls passing through a given point of the inner track each time the shaft completes a turn; 3) ball spin frequency (BSF) or rolling element failing frequency that corresponds to the number of turns a bearing ball make each time the shaft completes a turn; and 4) fundamental train frequency (FTF) or cage failing frequency that corresponds to the number of turns made by the bearing cage each time the shaft completes a turn.

The characteristic frequency of bearing defects is caused by balls passing through defect points. Given that \( n \), \( d \), and \( D \) are the number of balls, the ball diameter, and the bearing pitch diameter, \( f_r \) is shaft speed, and \( \phi \) is the angle of the load from the radial plane, the characteristic frequencies of bearing defects obey the following equations:

\[
\begin{align*}
    f_{BPFO} &= \frac{nf_r}{2}(1 - \frac{d}{D}\cos\phi), \\
    f_{BPFI} &= \frac{nf_r}{2}(1 + \frac{d}{D}\cos\phi), \\
    f_{FTF} &= \frac{f_r}{2}(1 - \frac{d}{D}\cos\phi), \\
    f_{BSF} &= \frac{Df_r}{2d}(1 - \left(\frac{d}{D}\cos\phi\right)^2),
\end{align*}
\]

where \( f_{BPFO} \), \( f_{BPFI} \) is the ball pass frequency of the outer race and inner race, \( f_{FTF} \) is the fundamental train frequency (cage speed), and \( f_{BSF} \) is the ball spin frequency [20], [21].

The most typical bearing defects can be identified in the envelope spectrum by utilizing the HHT. Fig. 3 shows the typical characteristics of bearing defects on the envelope spectrum. Because of the noise in the measurement, it is difficult to obtain a signal with all frequency components. In particular, the characteristic frequency of ball defects is hard to find because the ball is accompanied by spinning and rolling during motion.

III. METHODOLOGY

A. Quadratic Convolutional Neural Network

1) Quadratic Neuron: Encouraged by the neuronal diversity in a biological neural system, a new type of neuron called quadratic neuron [17] was proposed to promote neuronal diversity in deep learning. A quadratic neuron [17] integrates two inner products and one power term of the input vector before nonlinear activation. Mathematically, suppose that the input vector is \( \mathbf{x} = (x_1, x_2, \ldots, x_n)^T \), the output \( f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R} \) of a quadratic neuron is

\[
\sigma(f(\mathbf{x})) = \sigma \left( \sum_{i=1}^{n} w_i^T x_i + b^T \right) \left( \sum_{i=1}^{n} w_i^S x_i + b^S \right) + \sum_{i=1}^{n} w_i^b x_i^2 + c
\]

1https://power-mi.com/content/rolling-element-bearing-components-and-failing-frequencies
where $\sigma(\cdot)$ is a nonlinear activation function, $\odot$ denotes the Hadamard product, $\mathbf{w}', \mathbf{w}^e, \mathbf{w}^b \in \mathbb{R}^n$ are weight vectors, and $b', b^e, c \in \mathbb{R}$ are biases. Superscripts $r, g, b$ are just marks for convenience without special implications.

With the neuronal diversity, we argue that the network design consists of two parts: the neuronal design and the structural design. Fig. 1 shows the proposed quadratic convolutional neural network (QCNN). At the neuronal level, we adopt quadratic neurons in hope that they can facilitate bearing diagnosis via the strong feature representation power. Specifically, the QCNN substitutes the conventional convolution with the quadratic convolution operations in convolutional layers. Structurally, the QCNN inherits the structure of the WDCNN [10] as the structure of the proposed model because the WDCNN is a well-established model whose design has been considered by many follow-up studies. This structure stacks six CNN blocks and one fully connected layer.

2) Superiority of Representation: The quadratic neurons have the intrinsic enhancement in terms of representation ability, i.e., the enhancement is not due to the increased parameters but the involved nonlinear computation. First, the nonlinear mapping of a conventional neuron is only provided by the activation function, whereas a quadratic neuron gains an extra nonlinear mapping from the quadratic aggregation function. As such, a single quadratic neuron can realize XOR logic which is not doable by conventional neurons unless they are connected into a network. Second, a quadratic neuron is not equivalent to the combination or summation of three conventional neurons. Suppose that the activation function $\sigma(\cdot)$ is ReLU, the combination of conventional neurons can only be a piecewise linear function, but a quadratic neuron is a piecewise polynomial function [18]. As we know, a polynomial spline is better at approximating complicated functions than a linear spline.

3) Training Strategy: With the quadratic network as a more powerful model, a problem naturally arises: how can we ensure that the model can find a better solution? To address this problem, Fan et al. [18] proposed the so-called ReLinear algorithm,
where the parameters in a quadratic neuron are initialized as $w^e = 0, b^e = 1, w^b = 0, c = 0$, and $w^r, b^r$ follow the normal initialization. Consequently, during the initialization stage, every quadratic neuron in a network degenerates to a conventional neuron. Next, during the training stage, a normal learning rate $\gamma_r$ is cast for $(w^r, b^r)$, and a relatively small learning rate $\gamma_{r,b}$ for $(w^e, b^e, c)$.

### B. Efficient Quadratic Neuron-Induced Attention

The attention module [22] was originally derived for sequence tasks such as natural language processing. Mathematically, given the input signal $x \in \mathbb{R}^{n \times k}$, the commonly used scaled dot-product self-attention [23] is formulated as follows:

$$\text{softmax} \left( x W^Q (x W^K)^\top / \sqrt{d_k} \right) (x W^V)$$

where $W^Q, W^K, W^V \in \mathbb{R}^{k \times k'}$ are the projections for the query, key, and value, respectively, and $d_k$ is the scaling factor. The attention $\text{Att}()$ that reflects the importance of the input is taken as the following formula from (3):

$$\text{Att}(x) = \text{softmax} \left( x W^Q (x W^K)^\top / \sqrt{d_k} \right). \quad (4)$$

The above attention mechanism is widely adopted in the transformer models [24]. Concurrently, channel and spatial attention modules are developed to enhance the CNN [25]

$$F' = M_c(F) \odot F$$

$$F'' = M_s(F') \odot F'$$

where $F \in \mathbb{R}^{C \times H \times W}$ denotes intermediate feature maps produced by a convolutional layer, $M_c \in \mathbb{R}^{C \times 1 \times 1}$ denotes a 1-D channel attention map, and $M_s \in \mathbb{R}^{1 \times H \times W}$ denotes a 2-D spatial attention map. Furthermore,

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F)))$$

$$= \sigma \left( W_1 W_0 \left( F_c^\text{avg} \right) + W_1 W_0 \left( F_c^\text{max} \right) \right)$$

$$M_s(F') = \sigma(\text{Conv}([\text{AvgPool}(F'); \text{MaxPool}(F')])) \quad (6)$$

where $\sigma(\cdot)$ is the sigmoid function, AvgPool is the average pooling, MaxPool is the max pooling, multilayer perceptron (MLP) is the MLP layer, and Conv is the convolution filter. Regardless of different variants, the key to the attention mechanism is to induce an important map regarding the input that forces the network to discriminatively take advantage of the input information.

Based on such an observation, we find that a quadratic neuron also contains an attention mechanism, referred to as attention, by factorizing the learned quadratic function in analogy to attention. Mathematically, we factorize the expression of a quadratic neuron (2) as follows (we remove constant terms for conciseness):

$$\sigma \left( (x^\top w^r + b^r)(x^\top w^e + b^e) + (x \odot x)^\top w^b + c \right)$$

$$= \sigma \left( x^\top w^e (x^\top w^r + b^r) + b^r x^\top w^r + b^r b^r + (x \odot x)^\top w^b \right)$$

$$= \sigma \left( x^\top (w^e x^\top w^r + b^r) + x^\top (w^e b^r + (x \odot x)^\top w^b) \right)$$

$$= \sigma \left( x^\top (x \odot w^b + w^e (x^\top w^r + b^r) + w^e b^r) \right)$$

$$= \sigma (x^\top (x \odot w^b + w^e (x^\top w^r + b^r) + w^e b^r))$$

**Fig. 4.** Workflow of establishing the qttention map for the entire signal in the framework of the convolutional operation.

$$= \sigma \left( x^\top (x \odot w^b + w^e x^\top w^r + w^e b^r + \sigma (x^\top w^b)) \right) \quad (7)$$

where we let

$$\text{RawQtt}(x) = x^\top \odot w^b + w^e (x^\top w^r). \quad (8)$$

We exclude the bias terms $w^e b^r$, $w^e b^r$ because they keep intact for different $x$; therefore, they fall short of serving as importance scores. Furthermore, we calculate the gradient of $\text{RawQtt}(x)$ and take the absolute value to get the final qttention map

$$\text{Qtt}(x) = |\text{Grad}(\text{RawQtt}(x))| \quad (9)$$

The reason of doing so is that the gradient can better indicate the trend of the change and further eliminate the common bias.

In a convolutional layer, the qttention is tightly coupled to the convolution operation. Fig. 4 illustrates the workflow of establishing the qttention map for the entire signal. Each convolutional kernel is considered to be a single neuron, and a convolutional operation is shifting the kernel over the input with a constant stride. At each receptive field, the computation of the convolution kernel follows (2) and generates a qttention map. The final qttention map for the entire signal is derived by concatenating all qttention maps at local receptive fields. When the stride is smaller than the length of the receptive field, two neighboring receptive fields overlap. Then, we average the qttention scores at the overlapped locations.

**Remark** What makes the qttention special compared to the conventional attention module is that the conventional one usually is an independent plug-and-play module, but the qttention is neuron-induced. We summarize that the qttention has the following characteristics.

1. **Efficient:** Either the channel or the spatial attention computes an element-wise product for the entire input. Moreover, the channel attention module also includes an MLP network. Both the element-wise product for the entire input and the MLP network are computationally heavy. Assume a channel attention module (6) stacked after a convolutional layer $F \in \mathbb{R}^{C \times H \times W}$, its number of parameters is

$$\#(M_c(F)) = 2P \left( W_1 (c, c) \left( W_0 \left( F_c (c \times 1 \times 1) \right) \right) \right)$$

$$= 2 \times \frac{C^5}{r^2} \quad (10)$$

where $P$ is the projection size, $c$ is the channel number, and $r$ is the receptive field size.
where $C$ is an integer denoting the number of convolutional channels, and $r > 0$ is a scale factor. For an entire block that contains a convolutional layer with an attention module, the number of parameters is at least

$$\#(\text{Att}) = (C \times H \times W) + 2 \times \frac{C^5}{r^2}. \quad (11)$$

Moreover, with the addition of the spatial attention module, the number of parameters will be even larger.

In contrast, the qttention is generated by a convolution operation whose number of weights only scales with the kernel size, and parameters of qttention are the number of parameters of the quadratic convolutional layer

$$\#(\text{Qtt}) = 3 \times (C \times H \times W). \quad (12)$$

The number of parameters of a qttention is far fewer than that of a typical attention module. Therefore, even if we use qttention in every convolutional layer, the efficiency is much higher than adding the attention module.

2) Local: Compared to the typical attention module which is global, the qttention is local. The locality of the qttention is more suitable for bearing fault diagnosis. Usually, the attention module is applied after the signal is divided into several patches. Then, the attention score is assigned to each patch, and the attention map is for the entire signal but is coarse-grained. In contrast, the qttention is associated with the convolution, which emerges at each local receptive field. The locality of the qttention fits the diagnosis because the fault usually presents in the time domain as periodic short-range high-amplitude vibrations.

### IV. Experiments

In this section, we use two bearing fault datasets to investigate our proposed model. We first compare the QCNN with other SOTA methods under noisy conditions. The results show that the QCNN outperforms its competitors. Then, aided by the interpretability of qttention maps, we successfully decode the feature extraction process of the proposed model and the physics principle accounting for why the model can achieve good classification performance.

#### A. Datasets

1) CWRU Bearing Dataset: This widely-used dataset [26] is collected by Case Western Reserve University (CWRU) Bearing Data Center. Two deep groove ball bearings, 6205-2RS JEM SKF and 6203-2RS JEM SKF, are installed in the fan-end (FE) and drive-end (DE) of the electric motor. Single point defects with a diameter of 7, 14, and 21 mils are injected into the outer race, inner race, and ball of two bearings using electro-discharge machining. Therefore, this dataset has ten categories: nine types of faulty bearings and one healthy bearing. Specially, four levels (0, 1, 2, and 3 HP, where HP denotes Horsepower) of load are applied onto the shaft which slightly affects the motor speed (1797, 1772, 1750, 1730 r/min). Vibration data are collected at two sampling rates: 12 and 48 kHz for DE bearing faults. In this article, we use the vibration signal collected at DE side with the 12 kHz sampling rate.

2) HIT Angular Contact Ball Bearing Crack Dataset (HIT): We conduct a bearing fault test in MIIT Key Laboratory of Aerospace Bearing Technology and Equipment, Harbin Institute of Technology (HIT). Fig. 5 shows our bearing test rig. We utilize angular contact ball bearings HC7003 for the test. This bearing is for high-speed rotating machines compared to a deep groove ball bearing. The accelerometer is directly attached to a bearing to collect vibration signals produced by bearings. Consistent with the CWRU dataset, we inject faults at the outer race (OR), inner race (IR), and ball with three levels (minor, moderate, severe). Table I summarizes ten classes of bearings in our dataset. In the test, we choose the constant motor speed (1800 r/min) and use NI USB-6002 to acquire vibration signals with the 12 kHz sampling rate. We record 47 s of bearing vibration (561 152 points per category). Different from the CWRU dataset, bearing faults here are cracks of the same size yet different depths. Therefore, the vibration signals between different faults are more similar, making a diagnosis model more difficult to do classification accurately. Fig. 6 shows the raw signals in the time domain with respect to ten categories of our dataset.

#### B. Experiment Setup

1) Data Preprocessing: We add Gaussian noise into the raw input signal to verify model’s performance under noise environments. The signal-to-noise ratio (SNR) is $10 \log_{10}(P_i/P_n)$, where $P_i$ and $P_n$ are the average power of the signal and noise. Then, for both the original and the noised datasets, because the original signal is long, we randomly extract short sequences of 2048 numbers from each long sequence 1000 times. As such, we generate a total of 10,000 samples. Then, all samples are normalized into $[-1, 1]$. Both datasets are split into training, validation, and test sets with a ratio of 0.5:0.25:0.25.

2) Training Settings: The loss function is set to the cross-entropy loss, and we choose the stochastic gradient descent (SGD) [27] optimizer to optimize both networks. We train our
table and baseline methods using grid search with the same hyperparameter search space. After searching, we set batch size = 64 and training epochs = 50 for all methods. For the QCNN, the linear terms and quadratic terms are learned with different learning rates. Let $\gamma_c, b = \alpha \cdot \gamma_c$. We search $\gamma_c$ from $\{0.1, 0.3, 0.5, 0.8, 0.9, 0.05, 0.03, 0.008, 0.005, 0.003\}$ which is the same as all conventional methods and $\alpha$ from $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$.

We use the false positive rate (FPR), fault detection rate (FDR), precision (PRE), and F1 score to validate the performance of the proposed method, which is defined as follows:

$$FDR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$
$$PRE = \frac{TP}{TP + FP}, \quad F1 \text{ score} = \frac{2 \cdot \text{Precision} \cdot \text{FDR}}{\text{Precision} + \text{FDR}}$$

where TP, TN, FP, and FN denote the number of true positive, true negative, false positive, and false negative, respectively. We calculate the macro average of these metrics.

All experiments are conducted in Windows 10 with an Intel i9 10900 k CPU at 3.70 GHz and one NVIDIA RTX 3080 Ti 12 GB GPU. Our code is written in Python 3.8 with PyTorch, an open-source deep learning framework.

C. Classification Performance

The bearing signal is interfered with by other mechanical components and background noise from the measurement device, therefore, the noise-resistant capability of deep learning models is critical to the bearing fault diagnosis. We conduct experiments to verify the classification performance of the proposed model under noisy settings.

We compare our method with other SOTA methods, DCA-BiGRU [28], AResNet [29], RNN-WDCNN [12], MA1DCNN [11], and WDCNN [10]. Because codes of all these counterpart models are publicly available, we replicate them by their code and utilize the same data preprocessing. The optimizers, training strategies, and related hyperparameters are put into our Github repository for readers’ reference. Table II summarizes the sizes, #FLOPs, training, and inference time of all models. Compared to competitors, the QCNN has a relatively small model size, low computational complexity, and short inference time. Note that introducing quadratic neurons in convolutional layers only increases the number of parameters moderately, because the number of parameters in fully connected layers takes up a high percentage in the WDCNN.

We set three noise levels. The results are shown in Table III. All results are the average of ten runs. We draw the following highlights from Table III. First, the QCNN outperforms its competitors in terms of all metrics for three datasets. Although on the CWRU-2HP dataset, the QCNN is not the best, it admits a narrow gap to the first place. The QCNN shows the highest F1 score on the HIT dataset. In −6 dB noise, the QCNN is 16.82% higher than the second-best method. Therefore, we conclude that the QCNN is a versatile model that consistently delivers competitive performance at various noise levels.

Second, we choose −6 dB noisy data for the paired $t$-test. The $p$-values of F1 scores between QCNN and its competitors on the CWRU-0HP and HIT datasets are reported in Table IV. In a hypothesis test, the $p$-value is compared to the significance level to decide whether to reject the null hypothesis. Usually, the significance level is set to 0.05, which means the results have a 5% or lower chance of occurring under the null hypothesis to be considered statistically significant [30]. But we set the significance level to 0.01 for a harder test. The null hypothesis is that the two results are not from the same distribution. On the CWRU-0HP dataset, the DCA-BiGRU and MA1DCNN are significantly inferior to the QCNN. But the gaps between other competitors and QCNN are not significant. The reason is that CWRU is a relatively easy dataset, and so many methods can achieve nearly saturated high accuracy. On the HIT dataset, QCNN is significantly better than all other models. Because the noise interferes heavily with the signal, the HIT dataset is very challenging. Through significance analysis, we conclude that the improvement by QCNN is statistically meaningful.

Third, we calculate confusion matrices of the top 4 best methods to analyze the classification performance of each fault mode in detail. As shown in Fig. 7, although the noise severely interferes with the characteristics of the signal, all methods have the basic fault detection capability. Neither one misclassifies a faulty sample as a healthy sample (the last column of confusion matrices). However, the WDCNN misclassifies the health signal to the ball faulty 1. This might be because

![Fig. 6. Raw signals with respect to ten classes of our dataset.](image-url)
the amplitudes of these two types of signals are relatively close. Furthermore, the QCNN achieves the best results for classification among different faults. In the ball faulty 1, ball faulty 2, inner race faulty 3, and outer race faulty 2, the QCNN classifies all correctly, while other methods suffer serious misjudgment. We calculate the overall accuracy of defects in different locations. As shown in Table V, the QCNN outperforms others in all cases. All results suggest that QCNN has a better performance in severe noisy conditions.

At last, let us explain why under the same magnitude of noise, the HIT dataset is much more heavily corrupted than the CWRU dataset. We compare the time domain signals of the HIT and CWRU-2HP datasets. As shown in Fig. 8, the CWRU signal shows more significant impulsive characteristics. The peak value of the CWRU data is up to 3.8, while the peak value of the HIT data is only 0.18. This is because defects of the CWRU data are single-point fault injected using an electro-discharge, while defects of the HIT data are cracks generated by a laser engraving machine. Impulses produced by cracks have a lower amplitude than point defects, and therefore crack signals are more susceptible to noise. Moreover, as shown...
Fig. 8. Raw time domain signal on the CWRU and HIT datasets. (a) CWRU-2 HP. (b) HIT.

Fig. 9. Time and frequency domain waveforms of the outer race faulty signals on the CWRU-0HP and HIT datasets. The noise affects the HIT dataset more severely. (a) CWRU-0 HP. (b) HIT.

In most cases, QCNN outperforms WDCNN. Particularly, the QCNN leads by a large margin on severely noisy signals, e.g., in SNR $= -6$ dB CWRU-0HP dataset, the average F1 score of the QCNN is 10.1% higher than the WDCNN. Since the QCNN and WDCNN share the same structure, it confirms that a quadratic neuron is a simple and effective way of augmenting a network’s performance. To further compare what the QCNN and WDCNN learn from data, we use t-distributed stochastic neighbor embedding (t-SNE) [31] to visualize output features of the last convolutional layers of the QCNN and WDCNN. As shown in Fig. 10, different colors denote different categories of bearings. On the CWRU-2HP dataset, although both models exhibit promising feature extraction capabilities, evidenced by discernible clusters of ten categories of data, the QCNN is more desirable. For example, the yellow cluster in the result of the WDCNN is contaminated by purple points, but the purple and red clusters are well separated in the result of the QCNN. On the HIT dataset, the difference is quite significant, the WDCNN fails to gain discernible clusters, but the QCNN clearly manifests ten distinguishable clusters. It suggests that quadratic neurons have a powerful feature representation capability, enhancing the neural network to extract signal features even under heavily noisy conditions.

D. Feature Extraction Capability of QCNN

We compare the QCNN and WDCNN to illustrate that quadratic neurons have a stronger feature extraction capability.

E. Interpretability of Qttention

1) Qttention versus Grad-CAM++: Visualization of the intrinsic weights of a deep learning model is considered as an understanding of the model learning process, offering an intuitive evaluation of the model’s performance [32]. In the bearing fault diagnosis, we can obtain the interpretability of the QCNN through some visualization techniques, i.e., class activation mapping (CAM) [33], Grad-CAM [34], and Grad-CAM++ [35]. Here, we use Grad-CAM++ to generate a salience map for QCNN. Grad-CAM++ is an advanced version of the basic CAM, which has been applied to the interpretable visualization of the intelligent diagnosis models of bearing faults [28], [32]. As shown in Fig. 11, Grad-CAM++ produces a salience map by interpolating and smoothing the
convolutional weights. It only exhibits the approximate area of the model’s attention. In contrast, the attention indicates the point-by-point attention of the convolutional layers, which is more precise than Grad-CAM++. Therefore, the attention is better at revealing the key information about the decision process of the model and more helpful in understanding the relationship between learned features and physical properties of the signal.

2) Comparisons Between Attention and Convolution: Previous works have shown that the attention mechanism has the property of orienting the model to focus on key features of the data [11], [13], [14], [36]. As we deduced before, attention is a quadratic neuron-induced attention mechanism showing what a convolutional layer focuses on from input signals. The values of the attention represent the importance levels of the input \( x \) so that a quadratic network admits a self-explanatory feature extraction capability. This is a unique advantage of using quadratic neurons. In contrast, the conventional neuron does not enjoy such a kind of explainability because the linear coefficients are independent of the input, and cannot serve as the attention map. Here, we compare the attention map and the convolution layer. We compute the output using the following equation:

\[
\text{Out}(x) = |\text{Grad}(x^T w)|. \quad (13)
\]

The results are shown in Fig. 12. The attention map of the first layer more accurately conforms to the raw signal. This means that the quadratic network notices where the vibration intensity of the signal is higher. In contrast, the conventional network has larger weight values around high amplitude signals, but is not aligned with the raw signal and contains interference noise. Therefore, it lacks interpretability of the features extracted from the network.

Furthermore, we conduct a classification experiment to verify whether a attention map indeed captures important features. We use a attention map and the output of the first convolutional layer as the input of an SVM classifier. We feed the original signals directly into an SVM classifier as a baseline. As Table VI shows, the highlight is that QCNN+ SVM outperforms others by a large margin, suggesting that the attention map is more effective. Compared to the SVM, QCNN+SVM shows a significant performance improvement. In particular, on the HIT dataset, the QCNN+ SVM is 18.21% higher than SVM. What’s more, note that WDCNN+SVM has a surprisingly low classification F1 score. This phenomenon indicates that conventional neural networks may not identify as important bearing fault features as attention does. This also agrees with the visualization in Fig. 12.

3) Visualization in Time Domain: Attention informs what local information a quadratic convolutional layer highlights and the flow of important information across layers. Here, we compute attention maps from the earlier three quadratic convolutional layers for healthy, faulty, and noisy faulty bearing signals. Because each convolutional layer is followed by a max-pooling layer, the lengths of signals, as well as the corresponding attention maps, become short. So at the second and third layers, we need to interpolate the attention map to the same dimension of the input. Fig. 13 showcases the attention maps of three signals, where a higher value of attention means the healthy area might be contacted to generate a severe vibration, are captured. As the layer goes deep, the attention to key faulty sequences retains. At last, in regard to the noisy faulty bearing signal, the attention map can still attend to the faulty positions even when the noise almost dominates the raw signal. Moreover, the attention map of the third convolutional layer keeps similar features to that of the noiseless faulty bearing. This phenomenon accounts for why the QCNN can
maintain good performance for noisy signals; the QCNN can consistently acquire useful features even if the signal is noisy.

4) Visualization in Envelope Spectrum: The bearing vibration signal is mixed with multiple frequency components. The sign of a bearing defect is a significant impulse in the frequency domain. While in the low-frequency domain, there exist the shaft frequency and its harmonics. Although we show that the attention map can extract fault features to enhance the downstream classifier, we are curious to the questions: 1) does the QCNN learn features consistent with the physical interpretation? 2) what is the inherent mechanism of the QCNN for learning fault features? To resolve these questions, we validate the envelope spectrum of the attention map of the first convolutional layer by considering the physics principle of bearing defect vibration signals.

HHT is a powerful technique to analyze nonstationary signals [20]. By extracting the envelope of the signal through HHT and calculating its frequency spectrum, the characteristic frequencies of the nonstationary signal can be clearly observed [6]. In [37], the characteristic frequencies of faults in the CWRU dataset are computed. Here we adopt those computations. Fig. 14 shows the envelope spectra of the raw signals and attention maps. We use $f_c$ to uniformly represent $f_{BPFO}$, $f_{BPFI}$, $f_{BSF}$ for simplicity. First, both the envelope spectra of the outer race defect and inner race defect signals exhibit significant shaft frequency and characteristic frequencies, but the characteristic frequency of ball defects is hard to find because the ball is accompanied by spinning and rolling during motion. Second, by comparing the envelope spectra of the raw signal and the attention map, we find that the QCNN favors bearing fault characteristics over the shaft. For the outer raw signal and the attention map, we find that the QCNN performs others. In particular, QCNN-base shows significant performance. Moreover, a quadratic network prefers a larger normal learning rate $\gamma_r > 0.08$ for faster convergence and a smaller scaling factor $0.001 < \alpha < 0.1$ for better performance.

1) Discussion of the Quadratic Neuron: Here, we conduct an ablation study to investigate the effects of different terms in a quadratic network. Classification experiments with no noise and an SNR $= 0$ dB are performed on our dataset. We modify the equation of a quadratic neuron and examine the effect on the classification results. Here, QCNN-ng denotes that a quadratic neuron does not have $x^T w^g + b^g$ term

$$\sigma(f_{ng}(x)) = \sigma((x^T w^f + b^f) + (x \odot x)^T w^b + c).$$

QCNN-np denotes that a quadratic neuron does not have power term

$$\sigma(f_{np}(x)) = \sigma((x^T w^f + b^f)(x^T w^b + b^b)).$$

QCNN-base denotes the original design in (2).

Table VII shows the results. First, the proposed method outperforms others. In particular, QCNN-base shows significant advantages in noise environments, with the accuracy $4.32\%$ higher than that of a WDCNN. Second, the performance of QCNN-np and QCNN-ng is similar in no noisy conditions, but a QCNN-ng shows superior results in noisy conditions. The lack of the first-order item does not cause significant performance loss, while the power term of a quadratic neuron provides an important role in the network.

2) Hyperparameter Sensitivity: The important hyperparameters of a quadratic network are a normal learning rate $\gamma_r$ and a quadratic terms learning rate $\gamma_{f,b}$ [18]. As we introduce before, we use a scale factor $0 < \alpha < 1$ to regulate the relationship between learning rates ($\gamma_r = \alpha \cdot \gamma_{f,b}$). Here, we conduct experiments to investigate the influence of $\alpha$ and

![Fig. 13. Three convolutional layers’ attention maps on CWRU 1HP. From shallow to deep, the network gradually intensifies its focus on the strongest vibration features. (a) Healthy bearing. (b) Fault bearing. (c) Fault bearing on −6 dB noise.](image-url)
In real-world applications, only a limited number of faulty samples are available, which often causes deep learning models to overfit [38]. Furthermore, the operating conditions for bearings in the industrial field can vary significantly, requiring the model to possess robust cross-domain capabilities. Recently, contrastive learning [44], [45] and transfer learning [41], [42] have been adopted to address these challenges. Despite their effectiveness, there remains room for improvement. The QCNN offers a promising solution, as the enhancement of quadratic networks at the neuron level allows for a better combination of these techniques.

In this article, we have proposed a convolutional network made of recently developed quadratic neurons for the end-to-end time-domain bearing fault diagnosis. Moreover, not only the proposed quadratic convolutional network can have competitive classification performance on two bearing faults datasets, but also we have independently derived that a quadratic neuron inherently contains an attention mechanism, referred to as qttention. Furthermore, we have discussed differences between qttention and CNNs, and have conducted a classification test to verify that the bearing fault features attended by qttention are indeed important. We have utilized the qttention map to interpret the features learned by the proposed quadratic model in both time and frequency domains, and elucidate the physics principle behind. In the future, more efforts are required to refine the qttention mechanism and find real-world applications.

**ACKNOWLEDGMENT**

The experiments in this work were carried out with the support of Prof. Xiaoli Zhao and Hongyuan Zhang from the MIIT Key Laboratory of Aerospace Bearing Technology and Equipment at HIT. We express our heartfelt gratitude for their contributions and support.

**REFERENCES**

[1] A. H. Bonnett and C. Yung, "Increased efficiency versus increased reliability," *IEEE Ind. Appl. Mag.*, vol. 14, no. 1, pp. 29–36, Jan./Feb. 2008.
[2] P. D. McFadden and J. D. Smith, "Model for the vibration produced by a single point defect in a rolling element bearing," *J. Sound Vib.*, vol. 96, no. 1, pp. 69–82, Sep. 1984.
[3] Y. Ding, J. Zhuang, P. Ding, and M. Jia, “Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings,” *Rel. Eng. Syst. Saf.*, vol. 218, Feb. 2022, Art. no. 108126.
X. Zhang, C. He, Y. Lu, B. Chen, L. Zhu, and L. Zhang, “Fault diagnosis for small samples based on attention mechanism,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

Jing-Xiao Liao (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in instrumentation engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree with the School of Instrumentation Science and Engineering.

His research interests include deep learning, polynomial neural networks, and intelligent fault diagnosis.

[4] S. E. Pandarakone, M. Masuko, Y. Mizuno, and H. Nakamura, “Deep neural network based bearing fault diagnosis of induction motor using fast Fourier transform analysis,” in *Proc. IEEE Energy Convers. Congr. Expo. (ECCE)*, Sep. 2018, pp. 3214–3221.

[5] H. Gao, L. Liang, X. Chen, and G. Xu, “Feature extraction and recognition for rolling element bearing fault utilizing short-time Fourier transform and non-negative matrix factorization,” *Chin. J. Mech. Eng.*, vol. 26, pp. 106–105, 2014.

[6] V. K. Rai and A. R. Mohanty, “Bearing fault diagnosis using FFT of intrinsic mode functions in Hilbert–Huang transform,” *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2607–2615, 2007.

[7] S. Osman and W. Wang, “A morphological Hilbert–Huang transform technique for bearing fault detection,” *IEEE Trans. Instrum. Meas.*, vol. 65, no. 11, pp. 2654–2656, Nov. 2016.

[8] W. L. Wang, X. Li, M. Gao, and Y. Zhang, “A new convolutional neural network-based data-driven fault diagnosis method,” *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2017.

[9] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, “Deep residual shrinkage networks for fault diagnosis,” *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020.

[10] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, “A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals,” *Sensors*, vol. 17, no. 2, p. 425, Jan. 2017.

[11] H. Wang, Z. Liu, D. Peng, and Y. Qin, “Understanding and learning discriminant features based on multiattention IDCNN for wheelset bearing fault diagnosis,” *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 5735–5745, Sep. 2020.

[12] A. Shenfield and M. Howarth, “A novel deep learning model for the detection and identification of rolling element-bearing faults,” *Sensors*, vol. 20, no. 18, p. 5112, Sep. 2020.

[13] X. Li, W. Zhang, and Q. Ding, “Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism,” *Signal Process.*, vol. 161, pp. 136–154, Aug. 2019.

[14] Z.-B. Yang, J.-P. Zhang, Z.-B. Zhao, Z. Zhai, and X.-F. Chen, “Interpreting network knowledge with attention mechanism for bearing fault diagnosis,” *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106829.

[15] H. Wang, J. Xu, R. Yan, C. Sun, and X. Chen, “Intelligent bearing fault diagnosis using multi-head attention-based CNN,” *Proc. Manuf.*, vol. 49, pp. 112–118, Jan. 2020.

[16] B. Zhou, X. Dong, H. Zhou, and G. Tang, “Deep dynamic adaptive transfer network for rolling bearing fault diagnosis with considering cross-machine instance,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2021.

[17] G. Xu et al., “AFARN: Domain adaptation for intelligent cross-domain bearing fault diagnosis in nuclear circulating water pump,” *IEEE Trans. Ind. Informat.*, vol. 19, no. 3, pp. 3229–3239, Mar. 2023.

[18] Y. Zhou, Y. Dong, H. Zhou, and G. Tang, “Deep dynamic adaptive transfer network for rolling bearing fault diagnosis with considering cross-machine instance,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[20] P. Bhandari. (2022). *An Easy Introduction to Statistical Significance*. Cham, Switzerland: Springer.

[21] X. Zhang, C. Hu, Y. Liu, B. Chen, L. Zhu, and L. Zhang, “Fault diagnosis for small samples based on attention mechanism,” *Measurement*, vol. 187, Jan. 2022, Art. no. 110242.
Hang-Cheng Dong received the B.S. and M.S. degrees in instrumentation engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree with the School of Instrumentation Science and Engineering.

His research interests include explainable AI and its applications, piecewise linear models, and defect detection.

Zhi-Qi Sun received the B.S. and M.S. degrees in instrumentation engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2019 and 2022, respectively.

She is currently a Single Board Hardware Engineer at ZTE Corporation, Shenzhen, China. Her research interests include deep learning, neural network, and bearing fault diagnosis.

Jinwei Sun received the B.S. and M.S. degrees in electromagnetic measurement technology and instrumentation from the Harbin Institute of Technology (HIT), Harbin, China, in 1987 and 1990, respectively, and the Ph.D. degree from Saga University, Saga, Japan, in 2000.

He engaged as a Post-Doctoral Researcher with the University of Nancy, Nancy, France, from 2003 to 2004. He is currently a Professor with the School of Instrumentation Science and Engineering, HIT. His research interests include brain–computer interface systems, active noise control, sensor technology, testing instruments, and signal processing.

Shiping Zhang received the B.S., M.S., and Ph.D. degrees in instrumentation science and technology from the Harbin Institute of Technology (HIT), Harbin, China, in 1997, 1999, and 2003, respectively.

He is currently an Associate Professor with the School of Instrumentation Science and Engineering, HIT. His research interests include signal processing, deep learning, and sensors fusion.

Feng-Lei Fan (Member, IEEE) received the B.S. degree in instrumentation engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2017, and the Ph.D. degree from Rensselaer Polytechnic Institute (RPI), Troy, NY, USA, in 2021.

He is currently a Research Assistant Professor with the Department of Mathematics, The Chinese University of Hong Kong, Hong Kong. He has authored 26 papers in flagship AI and medical imaging journals. His primary research interests lie in deep learning theory and methodology, neuroscience, and medical image processing.

Dr. Fan served as a PC Member in many conferences such as the International Joint Conference on Artificial Intelligence and the Association for the Advancement of Artificial Intelligence. He was a recipient of the 2021 International Neural Network Society Doctoral Dissertation Award.