

Article

Bearing Fault Diagnosis Based on Golden Cosine Scheduler-1DCNN-MLP-Cross-Attention Mechanisms (GCOS-1DCNN-MLP-Cross-Attention)

Aimin Sun ¹, Kang He ², Meikui Dai ^{1,3}, Liyong Ma ^{1,3} , Hongli Yang ¹, Fang Dong ⁴, Chi Liu ⁵, Zhuo Fu ⁵ and Mingxing Song ^{1,3,*}

- ¹ School of Mechanical Engineering, Hebei University of Architecture, Zhangjiakou 075051, China; 2024225115@hebiace.edu.cn (A.S.); dmk1934@hebiace.edu.cn (M.D.); maliyong@buaa.edu.cn (L.M.); 15633135873@163.com (H.Y.)
- ² Fuzhou Education Development Research Center, Fuzhou 344000, China; 15179461728@163.com
- ³ Hebei Technology Innovation Center for Intelligent Production Line of Prefabricated Building Components, Zhangjiakou 075051, China
- ⁴ State Key Laboratory of Precision Manufacturing for Extreme Service Performance, Light Alloy Research Institute, Central South University, Changsha 410083, China; dongfang@csu.edu.com
- ⁵ School of Mechanical and Electrical Engineering, Changsha University, Changsha 410199, China; liu-chi@ccsu.edu.cn (C.L.); z20141061@ccsu.edu.cn (Z.F.)
- * Correspondence: smx1311@hebiace.edu.cn

Abstract

In contemporary industrial machinery, bearings are a vital component, so the ability to diagnose bearing faults is extremely important. Current methodologies face challenges in feature extraction and perform suboptimally in environments with high noise levels. This paper proposes an enhanced, multimodal, feature-fusion-bearing fault diagnosis model. Integrating a 1DCNN-dual MLP framework with an enhanced two-way cross-attention mechanism enables in-depth feature fusion. Firstly, the raw fault time-series data undergo fast Fourier transform (FFT). Then, the original time-series data are input into a multi-layer perceptron (MLP) and a one-dimensional convolutional neural network (1DCNN) model. The frequency-domain data are then entered into the other multi-layer perceptron (MLP) model to extract deep features in both the time and frequency domains. These features are then fed into a serial bidirectional cross-attention mechanism for feature fusion. At the same time, a GCOS learning rate scheduler has been developed to automatically adjust the learning rate. Following fifteen independent experiments on the Case Western Reserve University bearing dataset, the fusion model achieved an average accuracy rate of 99.83%. Even in a high-noise environment (0 dB), the model achieved an accuracy rate of 90.66%, indicating its ability to perform well under such conditions. Its accuracy remains at 86.73%, even under 0 dB noise and variable operating conditions, fully demonstrating its exceptional robustness.

Keywords: bearing fault diagnosis; 1DCNN-dual MLP framework; bidirectional cross-attention mechanism; feature extraction; high-noise environment robustness



Academic Editors: Xuefang Xu, Zijian Qiao, Mengdi Li and Peng Mei

Received: 23 July 2025

Revised: 1 September 2025

Accepted: 4 September 2025

Published: 6 September 2025

Citation: Sun, A.; He, K.; Dai, M.; Ma, L.; Yang, H.; Dong, F.; Liu, C.; Fu, Z.; Song, M. Bearing Fault Diagnosis Based on Golden Cosine Scheduler-1DCNN-MLP-Cross-Attention Mechanisms (GCOS-1DCNN-MLP-Cross-Attention). *Machines* **2025**, *13*, 819. <https://doi.org/10.3390/machines13090819>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the modern industrial system, rotating machinery occupies a pivotal position. As the rolling bearing is a core component, its operating condition plays a decisive role in the system's overall safety and reliability [1,2]. Condition monitoring and fault diagnosis of rolling bearings have therefore been the focus of research in recent decades [3,4].

Presently, the identification of faults in rolling bearings is predominantly predicated on signal processing methodologies, incorporating machine learning and deep learning algorithms. These methodologies entail the analysis of measured signals within the time domain, frequency domain, and time–frequency domain, thereby enhancing the efficacy of fault diagnosis [5]. As is well documented in the relevant literature, the following tools are commonly used: the fast Fourier transform (FFT) [6], the wavelet packet transform (WPT) [7], empirical mode decomposition (EMD) [8], and the Hilbert–Huang transform (HHT) [9]. However, the employment of signal processing-based methodologies frequently necessitates a sophisticated comprehension of signal analysis, thereby impeding the efficacy of fault diagnosis due to the inevitable presence of human interference.

Machine learning methods have been extensively utilized in the field of bearing fault diagnosis, with prominent examples including support vector machines (SVMs) [10], random forests (RFs) [11], and k-nearest neighbor algorithms (KNNs) [12]. However, it should be noted that conventional machine learning algorithms also impose limitations on the diagnostic capabilities of the model, primarily due to their relatively shallow structure.

In recent years, deep learning algorithms, including convolutional neural networks (CNNs) and multi-layer perceptrons (MLPs), have been extensively employed in the field of bearing fault diagnosis. The 1DCNN has been developed to simplify the structure of CNNs, with the objective of maximizing classification accuracy. The 1DCNN is combined with the self-correcting CEEMD algorithm, as proposed by Gao et al. [13]. This algorithm is an enhanced signal processing algorithm, derived from the CEEMD. This algorithm represents an enhanced signal processing technique, founded upon CEEMD, which incorporates the concepts of fuzzy entropy and kurtosis with a view to attenuating noise and identifying impulse signals, thereby enhancing the accuracy of classification. Guo et al. [14] employed motor speed signals and CNN models for the purpose of diagnosing a range of motor faults. Li et al. [15] combined the 1DCNN with gated recurrent units (GRUs). The 1DCNN has the capacity to automatically extract features from data, whilst GRUs can also perform this function. The GRU compensates for the deficiencies of CNN in processing time-series data, thereby ensuring the comprehensiveness of the extracted features and enhancing the accuracy of circuit fault diagnosis. He et al. [16] integrated 1DCNN with transfer learning (TL) to formulate a 1DCNN-TL model to predict the inlet gas volume fraction of a rotary vane pump, achieving a satisfactory fit. Mystkowski, Arkadiusz et al. [17] applied a multi-layer perceptron (MLP) with two hidden layers to detect faults in rotary weeders, achieving favorable results in both laboratory conditions and real-world tests. Zou, Fengqian et al. [18] proposed a depth-optimized 1DCNN that automatically extracts features from background noise while maintaining high accuracy even in noisy conditions. Xie, Suchao et al. [19] developed a multi-scale multi-layer perceptron (MSMLP) and used complementary integrated empirical modal decomposition (CEEMD) for fault diagnosis. However, the aforementioned studies are limited in scope as they begin from a single scale, such as the time or frequency domain, and the added noise is regular. In contrast, this study aims to address these limitations by expanding the scope to include multiple scales and by introducing a more diverse range of added noises.

In recent years, scholars have utilized fusion models for the purpose of fault diagnosis. For instance, Sinitsin et al. [20] combined CNN and MLP to form a CNN-MLP model for the diagnosis of bearings. Similarly, Song et al. [21] combined CNN and BiLSTM to construct the CNN-BiLSTM algorithm, which they then tuned using the improved PSO algorithm. In a recent study, Bharatheedasan, Kumaran et al. [22] explored the potential of combining a multi-layer perceptron (MLP) with long short-term memory (LSTM) to predict bearing life. Similarly, Wang, Hui et al. [23] investigated the efficacy of integrating multi-head attention with convolutional neural network (CNN) and optimizing the CNN

structure using the attention mechanism to enhance the recognition rate of fault diagnosis. Qing et al. [24] embedded Transformer in the dual-head integration algorithm (DHET) and used the dual-input time–frequency architecture to effectively capture the long-range dependencies in the data. Cui, Long et al. [25] introduced the self-attention mechanism into the bearing fault diagnosis. The study proposed a self-attention-based signal converter, which was able to learn the feature extraction capability from a large amount of unlabeled data by using a comparison learning method with only positive samples. In addition, Peng et al. [26] embedded the attention mechanism in a traditional residual network to evaluate the bearing state.

Notably, these studies were conducted in a laboratory setting with experimental conditions that excluded noise effects. The studies mainly focus on attention mechanisms within a single scale. In real industrial scenarios, the models may be disturbed by irregular random noise, which may significantly limit their effectiveness in applications. Therefore, the method proposed in this paper can maintain high accuracy under laboratory conditions and effectively deal with strong noise. Therefore, this method has the potential to achieve good performance in the face of a large amount of noise while maintaining high accuracy in the laboratory environment. The main contributions of this paper include the following:

1. The innovation of this work lies not in the individual components but in the integration strategy of dual MLP and 1DCNN and application scenarios under high noise. Our dual MLP extracts high- and low-frequency information from the data, solving the problem that the current Transformer-based models have a weak ability to capture local minute details and are vulnerable to noise interference at low signal-to-noise ratios.
2. The enhanced 1DCNN+dual MLP multimodal feature extraction structure has been designed to extract time- and frequency-domain features from the original data in depth.
3. A refined cross-attention mechanism is utilized to comprehensively capture signal characteristics from both the temporal and frequency domains. This approach demonstrates exceptional performance under diverse noise conditions.
4. The development of the GCOS learning rate scheduler constitutes a significant advancement in the field of automatic learning rate adjustment. This tool has been designed to address the challenge of manually selecting the learning rate by automatically adjusting it.
5. The experimental verification of the model's performance was conducted in two distinct settings: the laboratory environment and a high-noise environment.

2. Materials and Methods

2.1. 1DCNN-MLP Parallel Two-Stream Cross-Attention Fusion Mechanism

The proposed model architecture for the 1DCNN-MLP parallel dual-stream multi-scale feature fusion mechanism consists of a parallel 1DCNN, dual multi-layer perceptrons (MLPs), and serial cross-attention mechanisms. Initially, the timing data collected by the sensors are subjected to FFT. Thereafter, the timing data and the frequency-domain data after FFT are fed into the dual MLPs. Subsequently, the original timing data are fed into the 1DCNN for feature extraction. Then, the statistical features extracted by the dual MLP are aligned with the time-domain features extracted by 1DCNN and then sent to the cross-attention mechanism for deep feature fusion. Finally, forward and backward cross-attention are combined to enable fault diagnosis. The structure of the 1DCNN-MLP-Cross-Attention mechanism is illustrated in Figure 1 below.

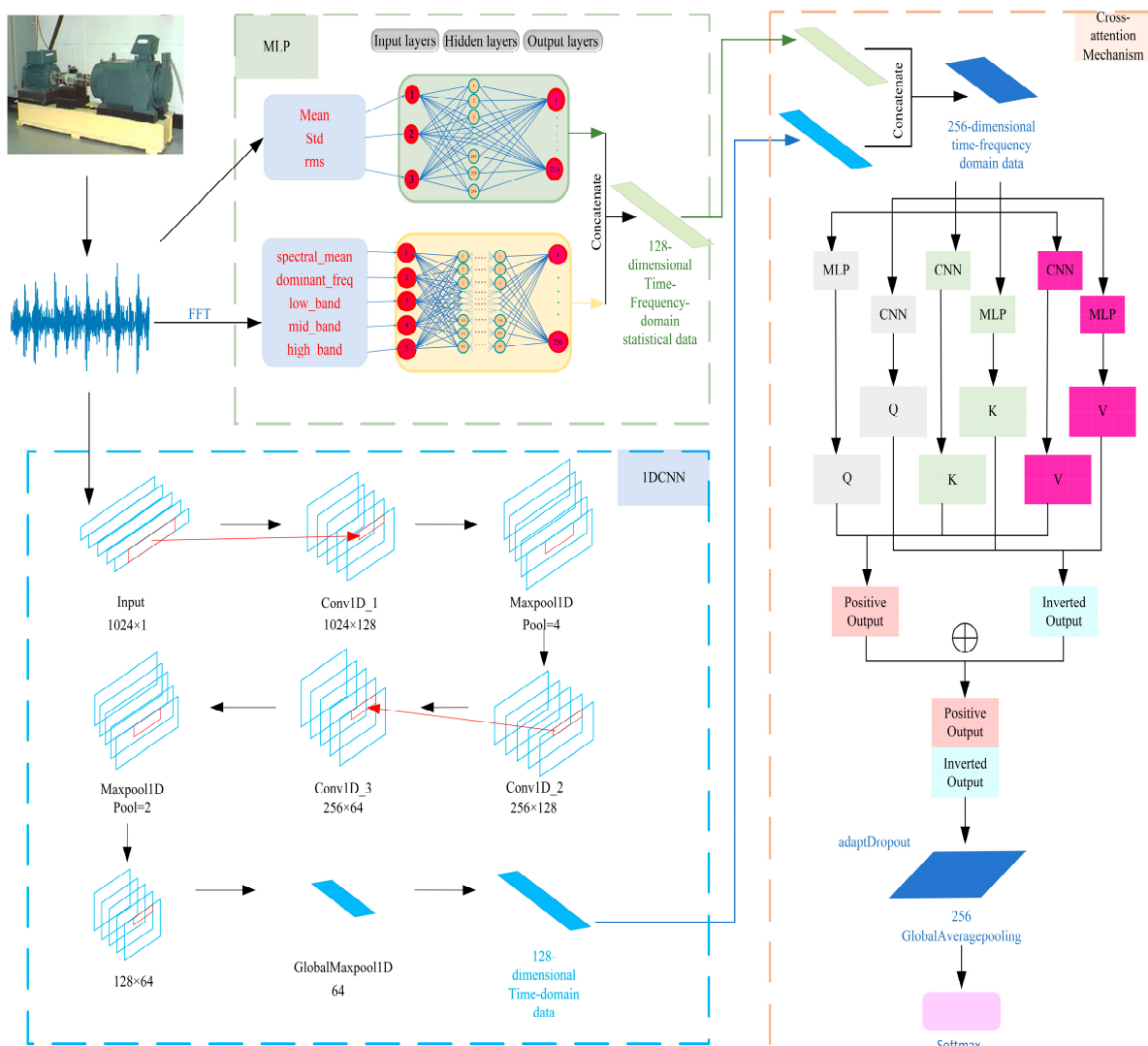


Figure 1. 1DCNN-MLP-cross-attention mechanism.

2.1.1. Two-Branch Multi-Scale Feature Fusion MLP

The present paper makes an innovative adoption of the dual MLP structure, whereby the dual-branch MLP structure is able to achieve independent inputs in the time domain and frequency domain. The inputs in the time domain are characterized by three elements: the mean, the standard deviation, and the root-mean-square (RMS) value. The mean represents the DC component of the signal, the average value of the vibration signal of a healthy bearing is usually close to zero, and the asymmetric shock caused by the fault will cause the mean to be offset; the standard deviation quantifies the fluctuation strength of the signal, the periodic shock when the bearing fails will significantly increase the standard deviation; and the root-mean-square value reflects the average energy of the signal. The frequency-domain inputs are the spectral mean, the dominant frequency position, and the superposition of the high-middle- and low-frequency energies extracted after the conversion of the time domain to the frequency domain by the FFT. The spectral mean is the standard of the frequency-domain energy concentration, the spectral energy distribution of healthy bearings is uniform, the energy is aggregated to the characteristic frequency during the fault, the main frequency position directly corresponds to the fault characteristic frequency, the high-frequency energy reflects the high-frequency resonance component of the early fault, the intermediate frequency energy covers the typical resonance band of the bearing fault, and the low-frequency energy captures the rotational fundamental frequency

of the bearing and the low-frequency harmonics of the fault. The detailed parameters of the time- and frequency-domain MLPs are given in Table 1 where the mean extraction equation is as follows:

$$u = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \quad (1)$$

Table 1. Detailed parameters of time- and frequency-domain MLPs.

	Layer Type	Neuron Number	Activation Function	Regularization	Dropout Rate
Time domain	Dense	256	Swish	L2 (0.01)	0.3
Frequency domain	Dense1	512	Swish	L2 (0.02)	0
	Dense2	256	Swish	L2 (0.01)	0.4

Standard deviation extraction equation:

$$\sigma = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - u)^2} \quad (2)$$

Root-mean-square value extraction equation:

$$x_{rms} = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x[n]^2} \quad (3)$$

The FFT converts the time domain to the frequency-domain equation:

$$X[k] = \sum_{n=0}^{N-1} x[n] \times e^{-j2\pi \frac{kn}{N}}, k = 0, 1, \dots, N-1 \quad (4)$$

Spectral amplitude calculation equation:

$$Amplitude[k] = |x[k]|, k = 0, 1, \dots, N/2 - 1 \quad (5)$$

Frequency-domain mean calculation equation:

$$u_{freq} = \frac{1}{n} \sum_{k=0}^{n-1} Amplitude[k], M = N/2 \quad (6)$$

Primary frequency position equation:

$$\operatorname{argmax}(magnitude) \times \Delta f \quad (7)$$

Initially, the three most significant statistics are extracted from the original time-domain data. The mean value reflects the average energy level of the signal, the standard deviation reflects the intensity of the fluctuation of the signal, and the root-mean-square (RMS) value reflects that the effective energy of the signal is directly related to the vibration energy. It is evident that these three statistics have the capacity to reflect the time-domain characteristics of the signal from a global point of view. The three time-domain features are entered into the time-domain multi-layer perceptron (MLP), while the fast Fourier transform (FFT) is used to convert the time-domain signal into a frequency-domain signal. This process enables the extraction of the mean value of the spectral components and the position of the dominant frequency, as well as the superposition of the high-middle- and

low-frequency energies. Subsequently, the five frequency-domain features are entered into the frequency-domain MLP, as illustrated in Figure 2. The time-domain plot is represented by Figure 2a, whilst the frequency-domain plot is represented by Figure 2b. Time-domain and frequency-domain diagrams are illustrated using a 12 k 0.014-inch rolling ball fault as an example.

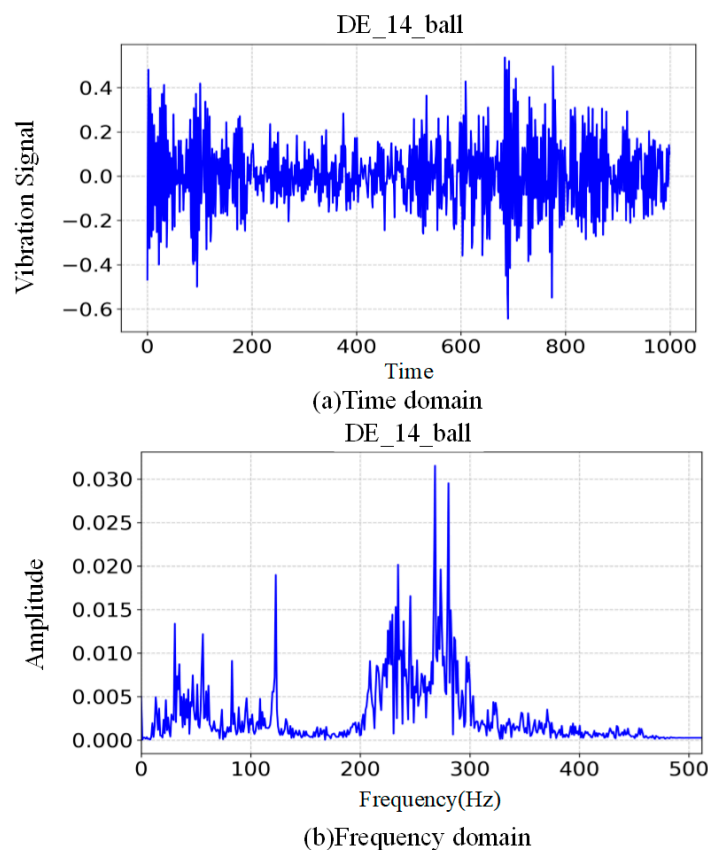


Figure 2. Time–frequency plot of 0.014-inch rolling ball faults at 12 k drive end.

As illustrated in the frequency-domain diagram, under standard operating conditions, the frequency domain of the bearing is predominantly concentrated within the 0–100 Hz range. However, under various types of fault conditions, the energy can be approximately divided into three distinct categories: low frequency (0–10 Hz), middle frequency (10–100 Hz), and high frequency (100–512 Hz). The high-, middle-, and low-frequency energy acquisition is based on this division. The following figure illustrates this. The third energy superposition diagram is representative of the high-, middle-, and low-frequency energy of the MLP. The superposition of the input diagram is illustrated by the example of a 12 k 0.014-inch rolling ball fault.

As demonstrated in Figure 3, the energy exhibited during the fault is predominantly concentrated within the high-frequency range of 100–512 Hz. It has been demonstrated that low-frequency energy (0–10 Hz) and mid-frequency energy (10–100 Hz) are significantly diminished in comparison to high-frequency energy.

The number of neurons in the first fully connected layer of the frequency-domain MLP is 512, and the process follows Equation (8):

$$h1 = Swish(X \cdot W1 + b1) \quad (8)$$

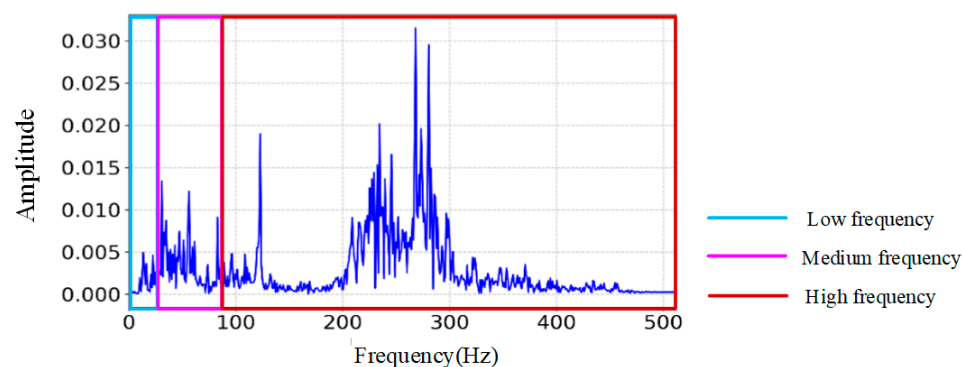


Figure 3. Energy superposition diagram.

The weight matrix **W1**, the bias vector **b1**, and the activation function Swish are defined as follows:

$$\text{Swish}(x) = x \cdot \sigma(\beta x) (\beta = 1.0) \quad (9)$$

In addition, this layer incorporates L2 regularization, a technique designed to limit the number of weight matrix paradigms (**W1**) and thereby mitigate the risk of overfitting.

The subsequent layer is the batch normalization layer, which contains 256 neurons. This layer normalizes each feature dimension independently using the following equation:

$$h_1 = \frac{h_1 - u}{\sqrt{\sigma^2 + \epsilon}} \quad (10)$$

In this context, “*u*” and “*ε*” are used to denote the mean and standard deviation, respectively, of the current batch.

The dropout technique is used to randomly mask a portion of the neurons during training. In this layer, 40% of the neurons are masked with the aim of forcing the network to learn redundant features and improve the generalization of the model

The multi-layer perceptron (MLP) in the time domain contains a fully connected layer consisting of 256 neurons with the dropout rate set to 0.3. The fully connected layer applies L2 regularization designed to prevent overfitting. After this layer, the subsequent output layer also produces 256-dimensional results

Finally, feature splicing combines the outputs of the time and frequency domains. This method offers good feature complementarity and preserves complete feature information. The feature dimensions are projected to 128 dimensions in order to maintain alignment with the 1DCNN.

A two-branch MLP can process features in the time and frequency domains separately, making the model robust in the face of noise.

2.1.2. Enhancement of 1DCNN

As the original vibration signal obtained from the bearing failure is one-dimensional time-series data, continuous downsampling enables 1DCNN to extract the time-series features from the points. The first layer of 1DCNN is Conv1D_1, with 128 filters and a convolution kernel size of 7. This sliding convolution uses a window of 7 time points, as illustrated in Figure 4 below.

The activation function is Swish, and each filter produces one channel. Therefore, the input shape is (1024, 1), and the output shape after Conv1D is (1024, 128). The first layer

uses 128 channels for base feature extraction. The layer between the input and the output follows the following equation:

$$\mathbf{Z}^{(1)}[t, c] = \sum_{k=0}^6 \mathbf{X}[t + k - 3, :] \cdot \mathbf{W}^{(1)}[k, c] + \mathbf{b}^{(1)}[c] \quad (11)$$

$$\mathbf{A}^{(1)} = \text{swish}(\mathbf{Z}^{(1)})$$

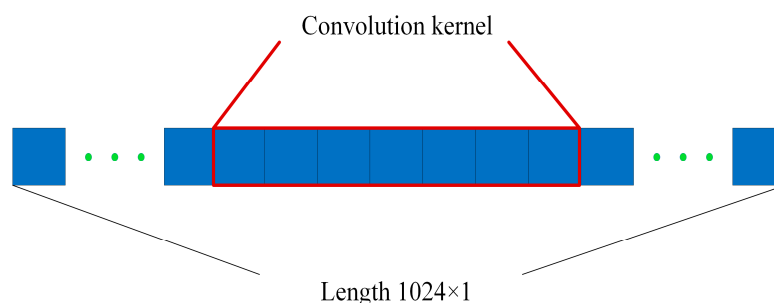


Figure 4. Convolution kernel.

The input $\mathbf{X} \in (\mathbf{R}^{1024 \times 1})$, convolution kernel $\mathbf{W}^{(1)} \in (\mathbf{R}^{7 \times 1 \times 128})$, bias $\mathbf{b}^{(1)} \in (\mathbf{R}^{128})$, and output $\mathbf{A}^{(1)} \in (\mathbf{R}^{1024 \times 128})$ are computed independently for each output channel, followed by weighted summation through a sliding window.

The subsequent layer is the first pooling layer, which is characterized by a pooling window of 4 and a step size of 4, as illustrated in Figure 5 below.

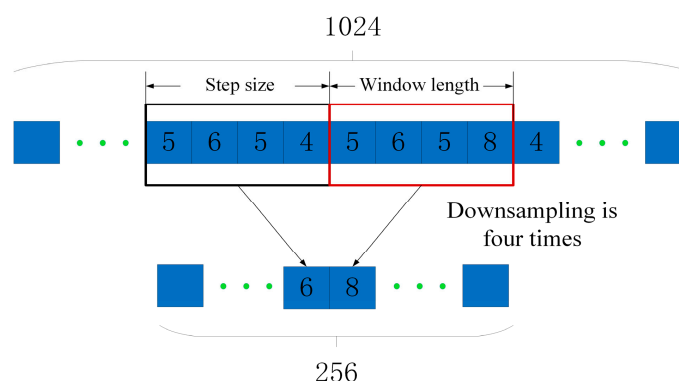


Figure 5. Pooling layer window and step size.

After the first pooling layer, a 4-fold downsampling operation was performed to reduce the temporal resolution to 256 points while preserving key features.

The second convolutional layer (Conv1D_2) has 128 channels and a convolution kernel size of 5. Sliding convolution uses a smaller 5-point window, which provides more detail than the 7-point window of the first layer. After this layer, the timing length remains at 256 and the number of channels remains at 128. The output shape is (256, 128). The purpose of the second convolutional layer is to obtain finer, more localized features. This is followed by the third convolutional layer (Conv1D_3), where the channels are compressed further to 64. The purpose of the 5-size convolution kernel is to prevent overfitting by balancing the ability to capture input information with the amount of computation. The output shape is (256, 64), followed by a second pooling layer. The pooling window and step size are changed to 2 to further compress the timing resolution to 128 points and downsample by a factor of two. The output shape of this layer is therefore (128, 64). Finally, the GlobalMaxPool1D global pooling layer takes the maximum over the full timestep for

each channel. This reduces (128, 64) to (64), i.e., the maximum for each channel forms a vector of length 64. This layer follows Equation (12):

$$\mathbf{P} = [t\max\mathbf{A}(2)[t, 1], t\max\mathbf{A}(2)[t, 2], \dots, t\max\mathbf{A}(2)[t, 64]] \quad (12)$$

The value of $\mathbf{p}$ is $\mathbf{R}^{64}$. Finally, there is the fully connected layer dense with output shape (128). The formula for this layer is the following:

$$\begin{aligned} \mathbf{Z}(3) &= \mathbf{P} \cdot \mathbf{W}(3) + \mathbf{b}(3) \\ \mathbf{A}(3) &= \text{swish}(\mathbf{Z}(3)) \end{aligned} \quad (13)$$

This layer maps 64 dimensions onto 128 dimensions, thereby improving the model's nonlinear representation. The detailed parameters of the enhanced 1DCNN are given in Table 2.

Table 2. Detailed parameters of enhanced 1DCNN.

Layer Type	Key Parameters	Exports
Input	1024 points	1024×1
Conv1D_1	Convolutional kernel size = 7	1024×128
Maxpool1D_1	Pooling window size = 4	256×128
Conv1D_2	Convolutional kernel size = 5	256×128
Conv1D_3	Convolutional kernel size = 5	256×64
Maxpool1D_2	Pooling window size = 2	128×64
GlobalMaxpool1D		64
Dense		128

2.1.3. The Integration of Two-Way Cross-Attention Mechanisms

In order to capture signal features more deeply, simply combining the two models, or using either the traditional or attention mechanisms, is not enough. Consequently, this paper introduces an enhanced two-way cross-attention fusion mechanism, as illustrated in Figure 6 below.

Initially, the attention mechanism was applied to the field of machine translation [27]. In contrast, the traditional attention mechanism only uses, for example, MLP features as the query and 1DCNN as the key and value. The features extracted by the two models are unidirectional, i.e., the multi-layer perceptron (MLP) is employed to calculate the attention weights, thereby integrating the features of the 1D convolutional neural network (CNN). For the deep extraction of features, this paper uses a bidirectional cross-attention mechanism instead of the traditional attention mechanism. In the context of the mechanism, the two-way cross-attention mechanism is subdivided into forward attention and reverse attention. In the forward attention configuration, an MLP is employed as the query, a 1DCNN is utilized as the key and value, and the MLP features are permitted to fuse the CNN features. Conversely, in the reverse attention configuration, a 1DCNN is used as the query, an MLP is employed as the key and value, and the 1DCNN features are permitted to fuse the MLP features. The forward and reverse attention layers each have Q, K, and V dimensions of 128. There are four attention heads, each with dimensions of 64 and a scaling factor of 8. Among them, positive attention is the following:

$$\text{Attn}_{\text{forward}} = \text{Softmax}\left(\frac{M \cdot C^T}{\sqrt{d_k}}\right) \cdot C \quad (14)$$

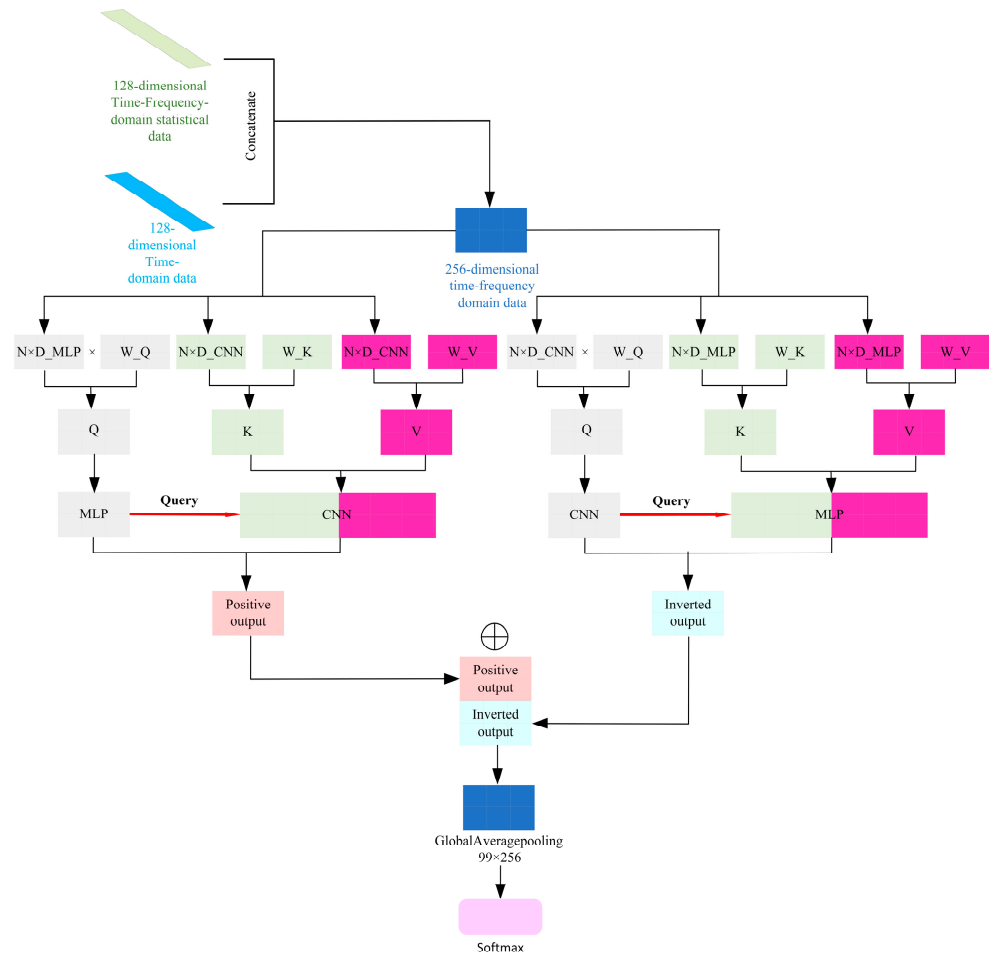


Figure 6. Cross-attention fusion mechanism.

Reverse attention is the following:

$$\text{Attn}_{\text{backward}} = \text{Softmax}\left(\frac{C \cdot M^T}{\sqrt{d_k}}\right) \cdot M \quad (15)$$

To illustrate this, we may consider forward attention as a case study. The initial step in this process is a bidirectional projection:

$$\begin{aligned} Q &= W_q \times x + b_q \\ K &= W_k \times Y + b_k \\ V &= W_v \times Y + b_v \end{aligned} \quad (16)$$

The second step is to scale the dot product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V \quad (17)$$

where the input QK^T is the attention score. $\sqrt{d_k}$ is the scaling factor, which aims to prevent the SoftMax gradient vanishing problem caused by too large dot product.

The third step focuses on designing the multi-head attention mechanism. This consists of four independent attention sub-modules that employ the scaled dot-product attention operation to integrate the four pieces of output information by splicing. This realizes the goal of feature fusion.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_4)W^O \quad (18)$$

The action mode of two-way attention weight includes the weight generation mechanism and the synergistic effect of weights, in which the weight generation mechanism is divided into forward weight and reverse weight, and the forward weight is as follows:

$$W_{\text{forward}} = \text{Softmax}(M \cdot C^T) \quad (19)$$

Positive weights reflect the attention of MLP statistics to the local fragments of the waveform.

The reverse weights are the following:

$$W_{\text{backward}} = \text{Softmax}(C \cdot M^T) \quad (20)$$

The sensitivity is of the reverse weight response waveform to a particular statistic.

The dominant feature of the positive synergy effect of the weight is the statistical feature, which is to extract the source of statistics from the waveform, and the dominant feature of the reverse synergy effect is the waveform feature, which is to screen the statistics that support the waveform mode. The bidirectional weights form closed-loop feedback, so that the statistical features and waveform features are mutually corrected.

Unlike the ordinary attention mechanism, which can only carry out unidirectional transmission of feature information, the bidirectional cross-attention mechanism breaks the unidirectionality of the ordinary attention mechanism through bidirectional information interaction. Through this mechanism, the MLP and the 1DCNN can pay attention to each other's features, enhancing the depth of feature extraction in the time–frequency domain.

2.2. GCOS Golden Cosine Learning Rate Scheduler

In order to adapt the model to complex classification tasks, this paper uses a golden cosine learning rate scheduling strategy to dynamically adjust the learning rate. The golden cosine learning rate scheduler quickly explores the parameter space in the early stage of training to shorten the training cycle, gradually decays in the middle stage to avoid jumping out of the local optimum, and makes the parameters more accurate in the later stage. Specifically, the learning rate is calculated by the following formula:

$$\eta(t) = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left[1 + \cos\left(\frac{\pi t}{T} \cdot \phi\right) \right] \cdot (1 - e^{-\lambda t}) \quad (21)$$

Values $\eta_{\min}$ and $\eta_{\max}$ represent the minimum and maximum learning rates, respectively, i.e., the boundary conditions. “ T ” denotes the iteration count and “ Φ ” corresponds to the golden ratio of 1.618, which is used for dynamically adjusting the cycle phase. λ is the exponential decay factor that controls the long-term decay of the learning rate. t indicates the current training step count. GCOS strikes a balance between convergence speed and stability by employing golden ratio phase modulation and dynamic cycle restart. This makes it particularly well suited to large-scale deep learning tasks. Its mathematical design combines the smoothness of cosine annealing with the advantages of periodic restarting associated with SGDR, while optimizing long-term training performance through an exponential decay term.

3. Experimentation and Analysis

3.1. Experimental Data

The experimental data were obtained from the Case Western Reserve University (CWRU) rolling bearing dataset, which is a widely used dataset. The experimental platform, illustrated in Figure 7 below, comprises a 2 hp (1.5 kW) electric motor (positioned on the left), a torque transducer/translator (center), a power meter (right), and electronic controllers

(not depicted in the figure). The motor has four operating states: 0 HP, 1 HP, 2 HP, and 3 HP; 1 HP = 0.7457 kW. The function of the test bearing is to provide support for the motor shaft. The test bearings provide support for the motor shaft and are categorized into drive-side bearings and fan-side bearings. It has been determined that these bearings sustain damage at a specific point during the EDM machining process.

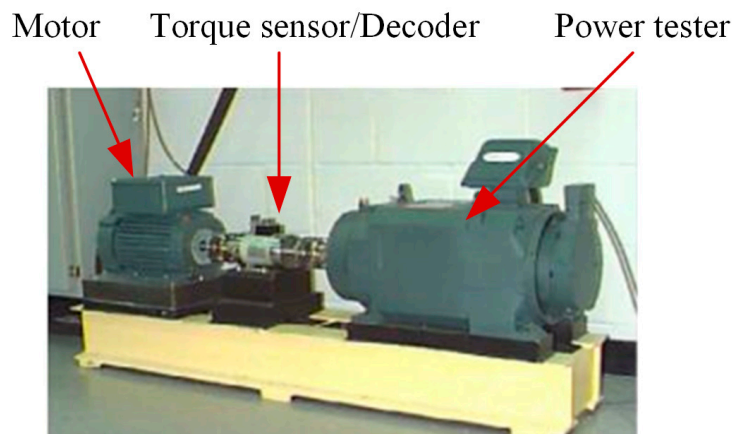


Figure 7. Experimental platform.

There are three types of bearing failure: inner ring (IR) failure, outer ring (OR) failure, and rolling element (ball) failure. There are four damage diameters, classified as follows: 0.007 inch (7 mils, 0.1778 mm), 0.014 inch (14 mils, 0.3556 mm), 0.021 inch (21 mils, 0.5334 mm), and 0.028 inch (28 mils, 1.016 mm) (1 inch = 25.4 mm). The typical characteristic points of damage to the bearing outer ring are located at 3, 6, and 12 o'clock on the dial. To monitor the failure state at the motor fan end and the drive end bearing seat, two acceleration sensors were installed in the relevant positions to collect vibration and acceleration data. A torque sensor/decoder was used to synchronize the measurement of power output and speed parameters, achieving the goal of comprehensive diagnostics. This study involved designing ten sets of experimental data to cover faulty inner rings, rolling elements, and outer rings at the 6 o'clock position (corresponding to working condition parameters of 0.007, 0.014, and 0.021). At the same time, a set of fault-free, normal operation data was introduced as a reference benchmark. The vibration signal was acquired by a 16-channel data logger, and the vibration signal sampling frequency was 12 kHz. A dataset is constructed using sliding window technology with the following parameters: window size: 1024 points; hop size: 1024 points; overlap: 0 points, with each bearing condition type containing 99 samples. Ultimately, the dataset was divided into three parts: a training set, a validation set, and a test set. The ratio of the training set, validation set, and test set is divided into 6:2:2, with the sample sizes for each group being 60, 19, and 20, respectively. The detailed fault classification information is shown in Table 3 below. To validate the model's performance in noisy environments, Gaussian noise is added to the original experimental data to simulate noise. This can be applied to each point of the original data over a wide range, which more closely resembles the noise generated by the bearing under actual operating conditions. The probability density function of the Gaussian noise is as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (22)$$

where “ μ ” is the mean, which determines the center of the noise, and “ σ ” is the standard deviation, which determines the spread of the noise.

Table 3. Experimental data.

Fault Category	Fault Diameter	Sample Size
Normalcy		99
Inner ring failure	7	99
Outer ring failure	7	99
Rolling body failure	7	99
Inner ring failure	14	99
Outer ring failure	14	99
Rolling body failure	14	99
Inner ring failure	21	99
Outer ring failure	21	99
Rolling body failure	21	99

3.2. Experimentation and Validation

3.2.1. Ablation Experiments on the Original Dataset

To make a more comprehensive comparison between each model, an ablation experiment is conducted for this model to compare the four evaluation indexes: accuracy, recall, precision, and F1 score. Fifteen independent experiments were conducted for each model, and the average result was calculated. Table 4 below shows the accuracy rates and standard deviations for the first five of fifteen experiments. The results of the ablation experiments for all indexes of the proposed model in this paper are shown in Figure 8. The detailed results are shown in Table 5. According to these results, it can be seen that the four indicators of the GCOS-1DCNN-MLP-Cross-Attention mechanism are optimal and that the accuracy rate has improved by 1.18% relative to the traditional 1DCNN-MLP fusion model, 1.764% relative to the traditional MLP, and 3.36% relative to 1DCNN. Comparing the accuracy rate of the 1DCNN-MLP-Cross-Attention mechanism model with that of the traditional 1DCNN-MLP model, it can be seen that the average accuracy of the model improves by 0.36%, rising from 98.65% to 99.01%. This indicates that the cross-attention mechanism module proposed in this paper improves the classification accuracy of the model. Based on the fair ablation results, replacing GCOS with cosine, StepLR, or the standard scheduler reduced the model accuracy by 0.46%, 0.64%, and 0.742%, respectively. These results demonstrate that the GCOS learning rate scheduler outperforms cosine, StepLR, and the standard scheduler. Observing Table 4 shows that the standard deviation of the model proposed in this paper is only 0.236 over five trials, indicating that the model achieves the highest accuracy while also showing high robustness. It is noteworthy that following the incorporation of the CNN-MLP with the learning rate scheduler proposed in this study, the accuracy rate is enhanced to 98.7%. The findings reveal that the golden learning rate scheduler proposed in this paper plays a significant role in enhancing the model's accuracy. Further investigation into the impact of the golden cosine learning rate scheduler on the model was conducted. To this end, a comparison is made between the model proposed in this paper and the 1DCNN-MLP-Cross-Attention mechanism model, with and without the golden cosine learning rate scheduler. The accuracy and loss curves of the 1DCNN-MLP-Cross-Attention mechanism model with and without the golden cosine learning rate scheduler are shown in Figure 9 below. It is evident from this figure that the accuracy of the GCOS-1DCNN-MLP-Cross-Attention mechanism fluctuates considerably from the 15th to the 35th round of training and then gradually converges to 1 and finally converges in the 60th round. The loss curves of the model also fluctuate in the 60th round. The loss curve of the model also reaches convergence at approximately round 60. In contrast, the accuracy of the 1DCNN-MLP-Cross-Attention mechanism model without the addition of the golden cosine learning rate scheduler fluctuates considerably between rounds 25 and 50 and finally converges at round 80. As demonstrated in Figure 10, the learning rate undergoes

a change following the implementation of the learning rate scheduler. In this paper, the initial learning rate is set to 10^{-3} , and the golden cosine learning rate scheduler establishes a minimum threshold for the learning rate, equivalent to 1% of the initial learning rate (illustrated as the red line in Figure 10). The golden learning rate scheduler causes the learning rate to vary in a cosine function-like manner, and in the process of automatically adjusting the learning rate, the model uses a dynamic learning rate that perfectly balances the disadvantages of overfitting and slow convergence.

Table 4. The average of 15 independent experiments and the accuracy rate of the first five trials.

Model	Accuracy					Average Accuracy	Standard Deviation
	1	2	3	4	5		
GCOS-1DCNN-MLP-Cross-Attention	99.49	100	99.49	100	99.89	99.83	0.236
1DCNN-MLP-Cross-Attention	99.32	98.89	98.45	98.98	99.49	99.01	0.38
GCOS-1DCNN-MLP	98.39	98.64	98.99	98.45	98.39	98.7	0.251
MLP-CNN	98.99	97.47	98.48	98.99	98.99	98.65	0.593
MLP	97.5	98.48	97.98	98.58	97.64	98.066	0.38
1DCNN	95.98	95.96	96.46	96.98	95.98	96.47	0.896

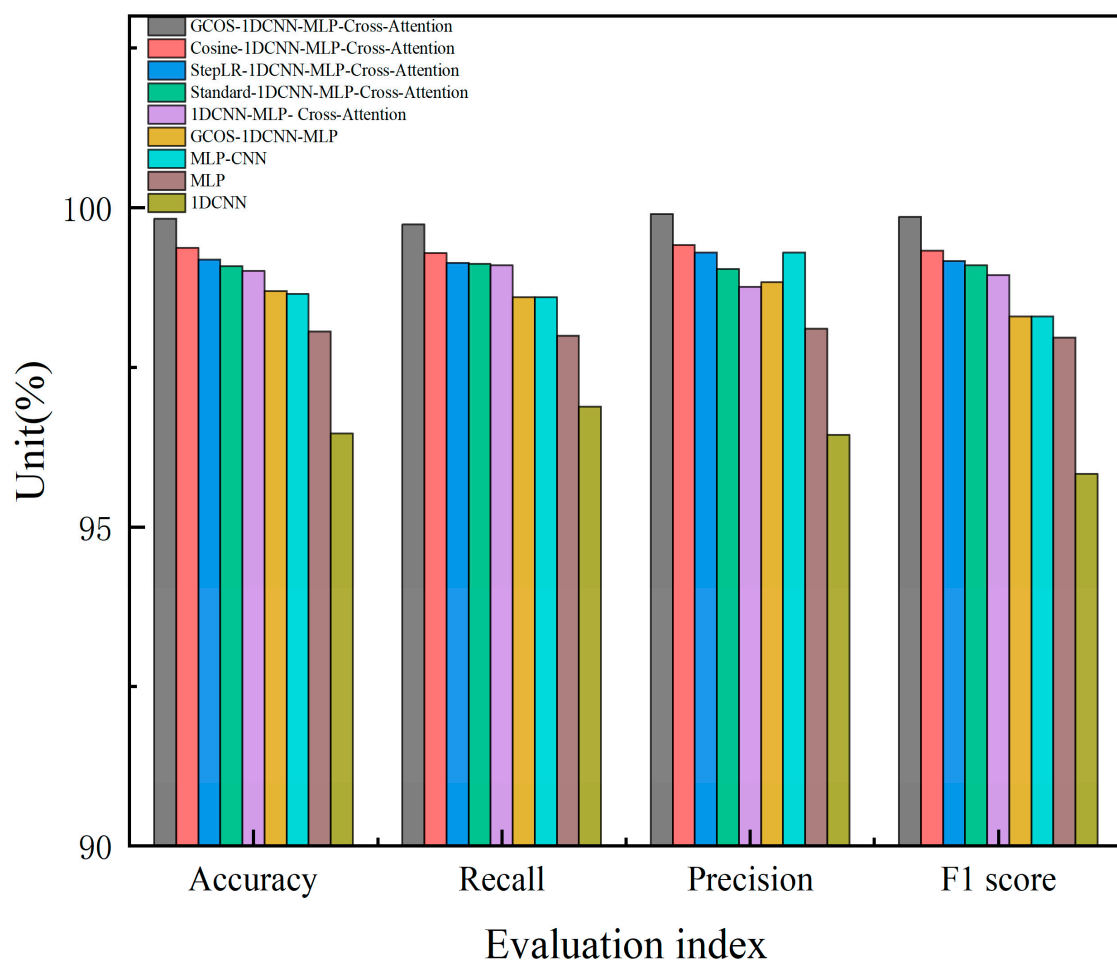


Figure 8. Results of ablation experiments.

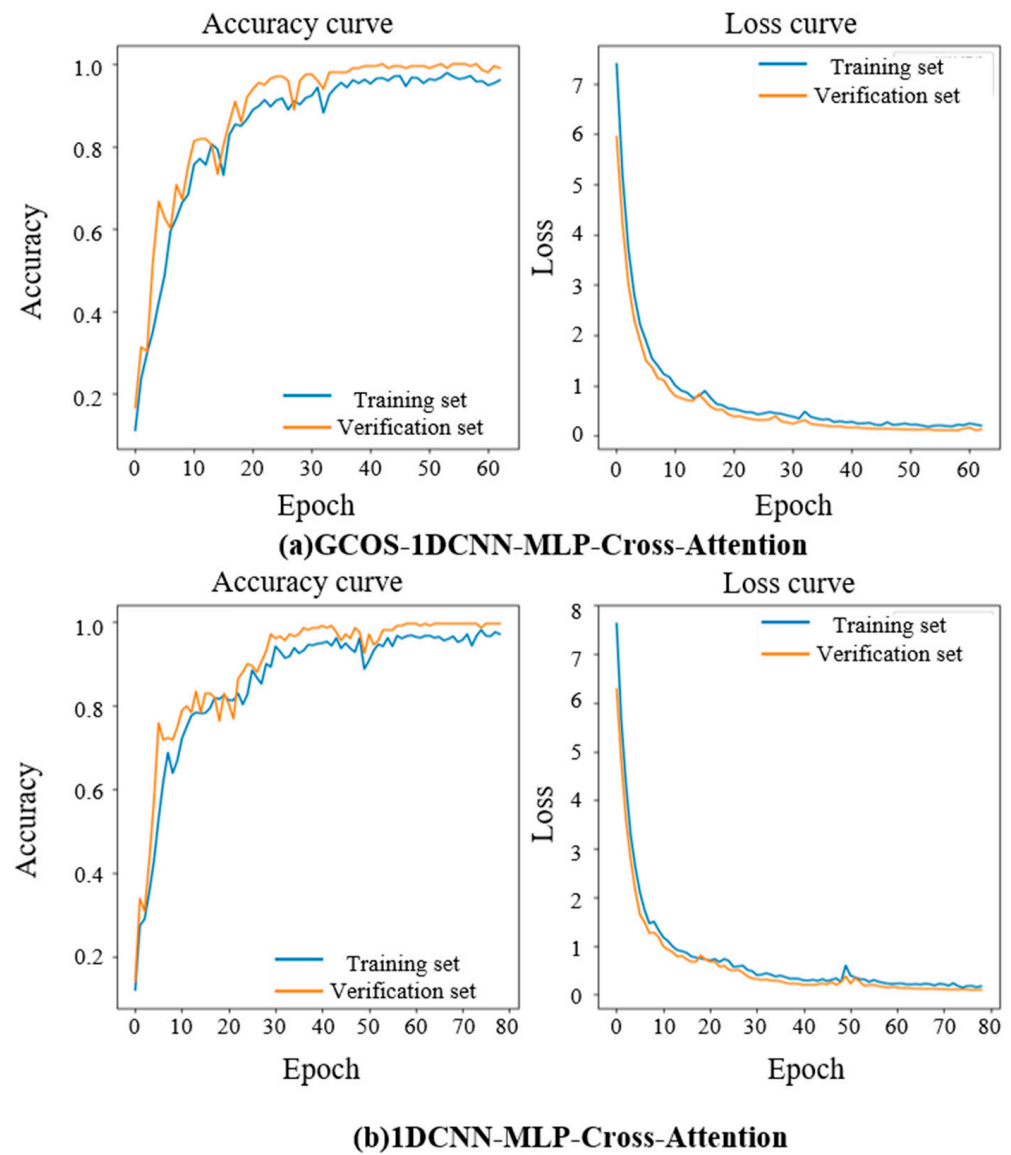


Figure 9. Effect of GCOS learning rate scheduler on accuracy loss rate.

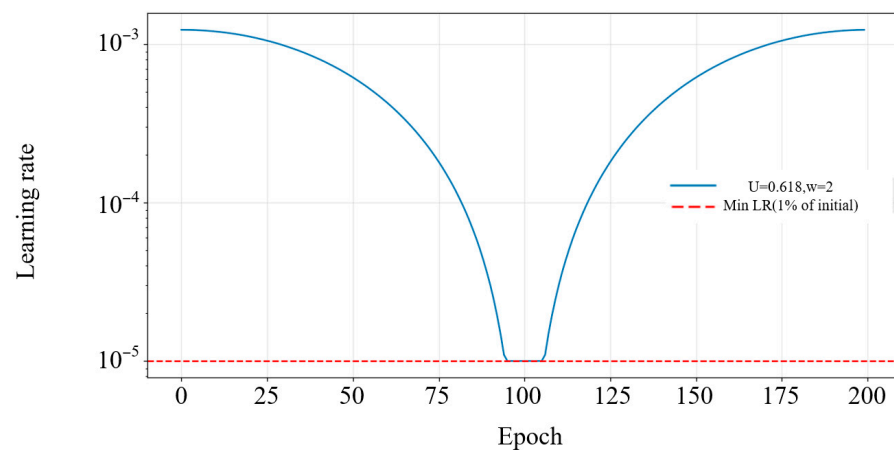


Table 5. Ablation experiments.

Model	1DCNN	MLP	GCOS Learning Rate Scheduler	Cross- Attention	Accuracy	Recall Rate	Precision	F1 Score
GCOS-1DCNN-MLP-Cross-Attention	✓	✓	✓	✓	99.83	99.74	99.9	99.86
Cosine-1DCNN-MLP-Cross-Attention	✓	✓		✓	99.37	99.29	99.42	99.33
StepLR-1DCNN-MLP-Cross-Attention	✓	✓		✓	99.19	99.14	99.3	99.17
Standard-1DCNN-MLP-Cross- Attention	✓	✓		✓	99.088	99.12	99.04	99.1
1DCNN-MLP-Cross-Attention	✓	✓		✓	99.01	99.1	98.76	98.95
GCOS-1DCNN-MLP	✓	✓	✓		98.7	98.6	98.84	98.3
MLP-CNN	✓	✓			98.65	98.63	99.31	98.34
MLP		✓			98.066	98.01	98.16	97.99
1DCNN	✓				96.47	96.86	96.46	95.88

3.2.2. Noise Environment Comparison Experiment

In industrial environments, data collected by sensors are often affected by various noise factors. To simulate these conditions, this study injects Gaussian noise of different intensities into the original vibration signals and evaluates the performance of the constructed model in high-noise environments. Noise is added uniformly to both the training and validation data during the loading phase. A fixed random seed ensures that the same noise sequence is generated each time. The noise intensity is adjusted using a noise factor, which remains constant and is applied to all samples. Signal-to-noise ratio (SNR) is a key metric for evaluating signal quality, defined as the ratio of useful signal power to noise power, typically expressed in decibels (dB). The formula for SNR is as follows:

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) \quad (23)$$

where P_{signal} denotes the effective power of the signal, and P_{noise} denotes the power of the added noise.

As illustrated in Figure 11 below, the fault signal of a 0.014-inch rolling ball fault at the 12 k drive end, with 20 dB and 0 dB Gaussian noise added, is shown. When the signal-to-noise ratio is 20 dB, the noise has little effect on the signal, and it is not difficult to extract the original vibration signal. When the signal-to-noise ratio is 0 dB, it indicates that the energy of the noise signal is the same as that of the original vibration signal, and it can be seen in the figure that the original vibration signal in blue has been completely destroyed when the signal-to-noise ratio is less than 0 dB, which indicates that the intensity of the noise signal has already exceeded the original vibration signal strength. At this time, it becomes extremely difficult to extract the original signal characteristics from the strong noise.

This paper presents an evaluation of signal-to-noise ratios across a range of noise intensities, from 20 dB to −5 dB. Figure 12 below provides a visual representation of the accuracy of various methods under different noise levels. Different methods employ the same regularization and stopping criteria. For a comprehensive analysis, please refer to Table 6, which presents detailed results. The results show that the proposed GCOS-1DCNN-MLP-Cross-Attention mechanism exhibits the highest accuracy at all signal-to-noise ratios, except for the model proposed in this paper, the accuracy of all the other models decays rapidly with the enhancement of the noise signal, and the accuracy of a single model decays to less than 80% at 10 dB; the model proposed in this paper is still at 97.47% accuracy, which is 14.64% higher than the MLP-CNN by 14.64% and CNN-Transformer by 13.63%. At 0 dB, the accuracy of the model proposed in this paper is 90.66%, while all the other models decay to below 80%, which is 40.66% higher than MLP-CNN and 20.46% higher than CNN-Transformer. Under the strong noise of −5 dB, the accuracy of the proposed model can still be maintained at 86.72%, while the accuracy of other models drops to less than 70%. Figure 13 presents the confusion matrix of the GCOS-1DCNN-MLP-Cross-Attention mechanism under 0 dB noise. It can be seen that the model in this paper only drops to 80% for the classification of the DE_14_outer class of faults under 0 dB of strong noise, and even six other classes remain above 90%. The components demonstrate the excellent performance of the GCOS-1DCNN-MLP-Cross-Attention mechanism in the face of strong noise. The experiment was conducted on a system comprising an Intel Core i9-13980HX CPU, an NVIDIA GeForce RTX 4060 GPU, and the Windows 11 operating system. The training parameters for the various methods are presented in Table 7.

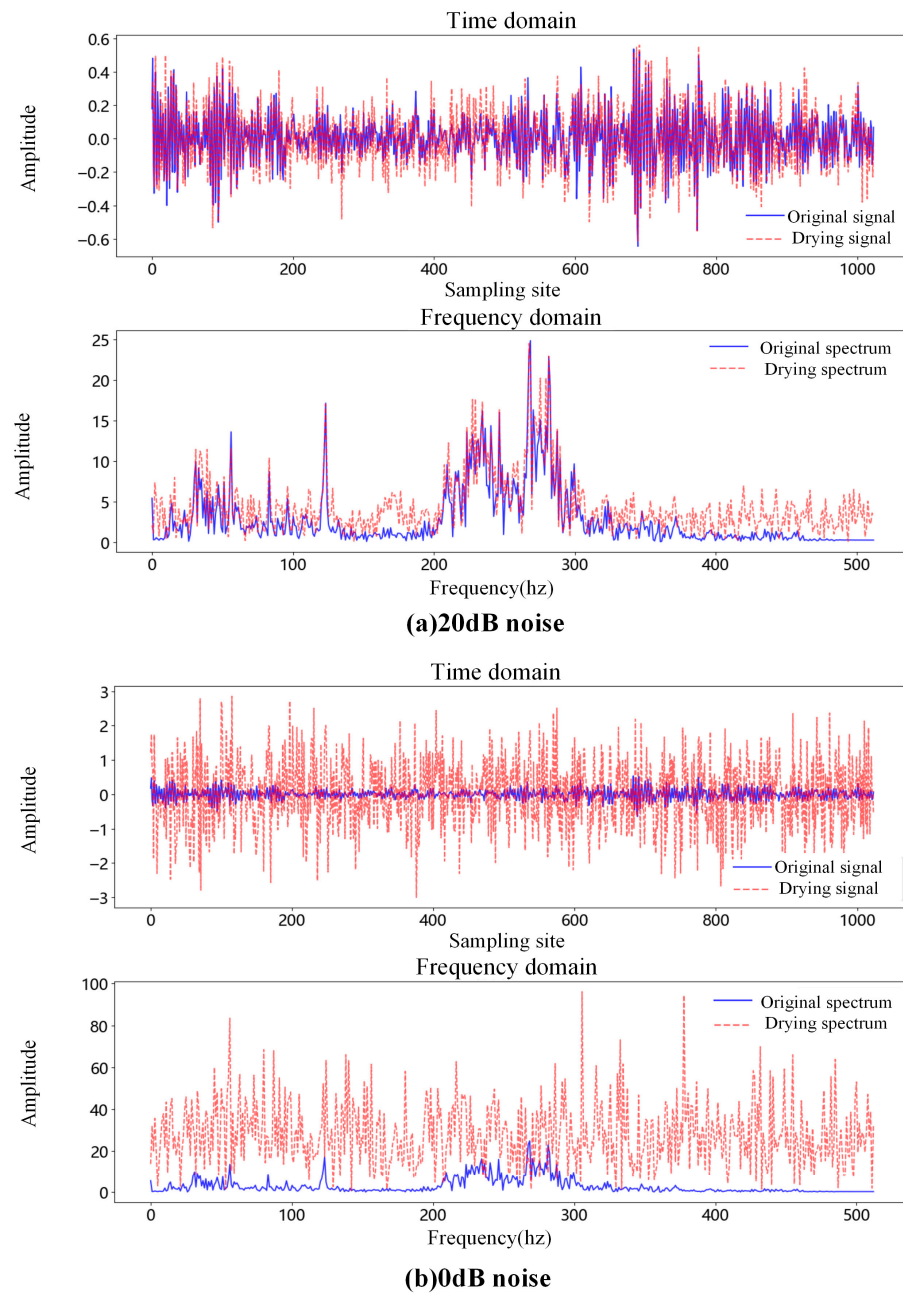


Figure 11. Comparison of 0.014-inch rolling seek fault noise at 12 k drive end.

Table 6. Accuracy of each model with different signal-to-noise ratios.

Model	SNR					
	−5	0	5	10	14	20
GCOS-1DCNN-MLP-Cross-Attention	86.72	90.66	94.7	97.47	98.23	98.74
CNN-Transformer	67.17	70.20	77.27	83.84	87.37	96.46
Traditional MLP-CNN	32.32	50	72.73	82.83	89.9	96
MLP	21.21	36.87	58.08	70.71	82.32	92.42
1DCNN	11.11	23.74	41.92	37.88	66.16	85.35
LSTM	36.11	48.44	58.8	69.77	70.53	71.99

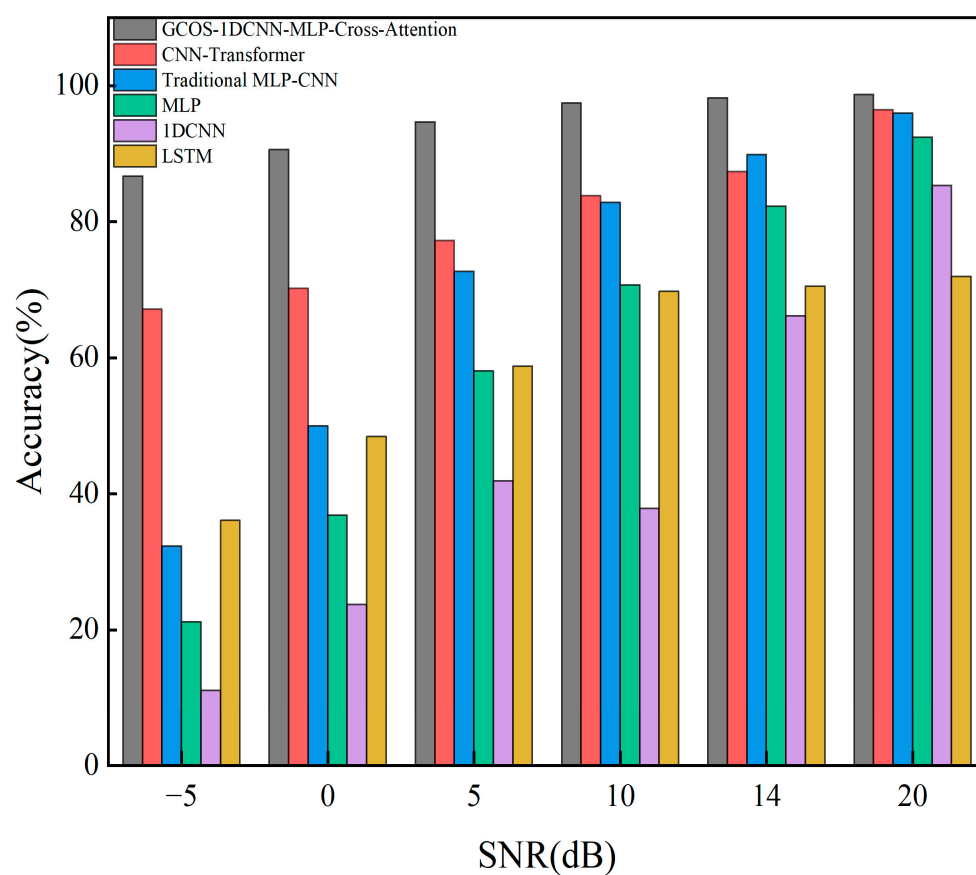


Figure 12. Accuracy of each model under different noise intensities.

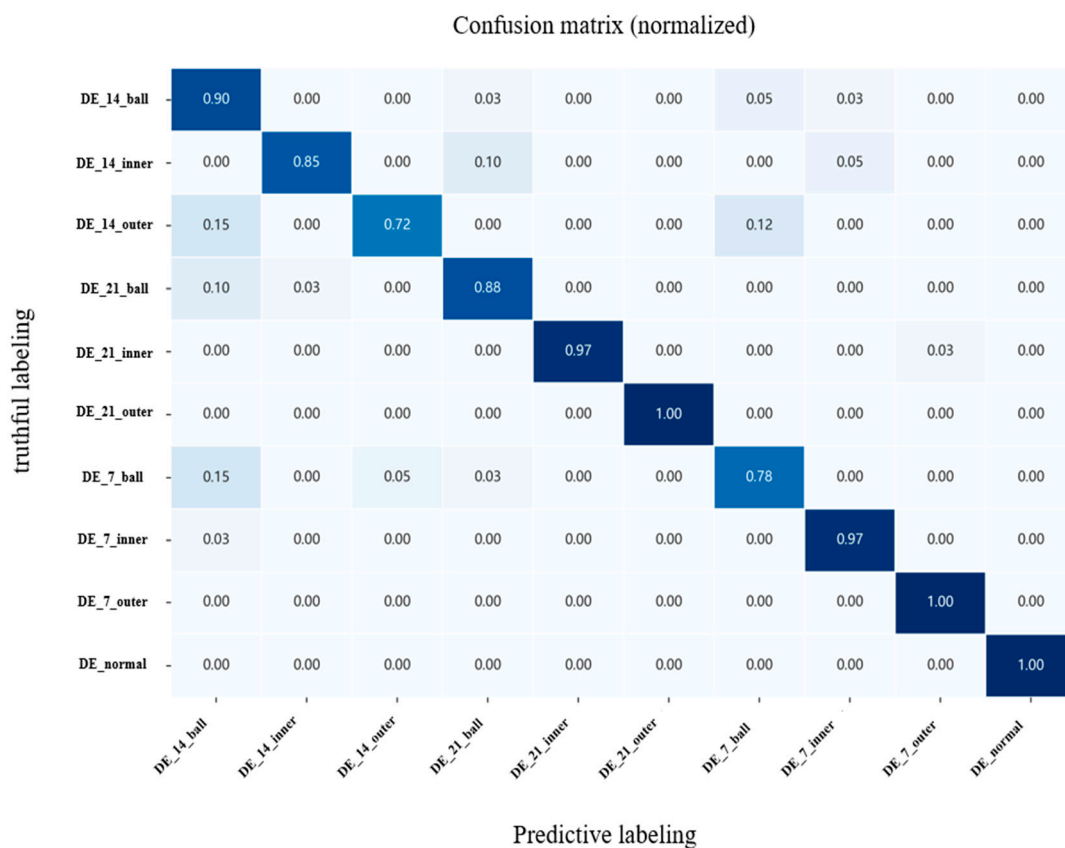


Figure 13. Confusion matrix for GCOS-1DCNN-MLP-Cross-Attention mechanism with 0 dB noise.

Table 7. Training parameters for different methods.

Model	Parameters/M	FLOPs/B	Inference Latency/ms	GPU Throughput/Samples/s	CPU Throughput/Samples/s	1 s Signal Delay at a 1 kHz Sampling Rate/ms	Training Time/min
GCOS-1DCNN-MLP-Cross-Attention	12.8	2.1	0.08	14,200	2100	81.9	2.6
CNN-Transformer	45.6	7.8	0.4	4800	650	400	8.2
Traditional MLP-CNN	8.3	1.5	0.6	7600	1200	614.4	3.5
MLP	0.9	0.3	0.3	28,500	3800	307.2	1.2
1DCNN	15.2	2.7	0.9	6200	1000	921.6	4.1
LSTM	24.7	4.2	2.1	1600	220	2150	15.3

3.2.3. Variable Load Noise Environment Comparison Experiment

To demonstrate the robustness of the proposed model and its performance under variable load noise conditions, data from the 12 K drive under 0 HP load conditions were used as the training set. Data from the 12 K drive end operating at a 1 HP load were used for both the test and validation sets (1 HP = 0.7457 kW). The noise addition method remained consistent with that used in the preceding section. Detailed data are presented in Table 8.

Table 8. Experimental data under variable load noise conditions.

Motor Load (KW)	Motor Speed (rpm)	Data Segmentation	Sample Size
0 KW	1797	Training set	99
0.7457 KW	1772	Validation set	33
0.7457 KW	1772	Test set	33

Figure 14 shows the experimental results, with the detailed outcomes presented in Table 9. The results indicate that, under variable load conditions, the performance of all models declined compared to single-load scenarios. However, the GCOS-1DCNN-MLP-Cross-Attention model demonstrated particularly robust performance, ranking first among the six models across various SNRs and maintaining an accuracy of 93.47% even in the presence of 5 dB of noise. At this point, the accuracy of the other models dropped below 80%. Under high noise conditions of −5 dB, the GCOS-1DCNN-MLP-Cross-Attention model achieved 25.3% higher accuracy than the CNN-Transformer model, while the other models nearly lost their classification capability. Figure 15 shows the confusion matrix for the GCOS-1DCNN-MLP-Cross-Attention model under variable loads ranging from 0 HP to 1 HP and 0 dB noise. These results demonstrate that the GCOS-1DCNN-MLP-Cross-Attention model exhibits robust performance under variable loads and noisy conditions.

Table 9. Variable load and accuracy rates of each model at different signal-to-noise ratios.

Model	SNR					
	−5	0	5	10	14	20
GCOS-1DCNN-MLP-Cross-Attention	85.72	86.73	93.47	94.8	95.23	96.54
CNN-Transformer	60.42	68.29	72.73	82.83	89.9	92
Traditional MLP-CNN	32.13	45.63	65.34	74.66	81.75	90.42
MLP	20.11	35.77	48.57	60.19	66.27	77.43
1DCNN	10.03	20.64	39.92	40.87	50.11	66.16
LSTM	32.16	40.11	51.3	56.22	60.52	63.55

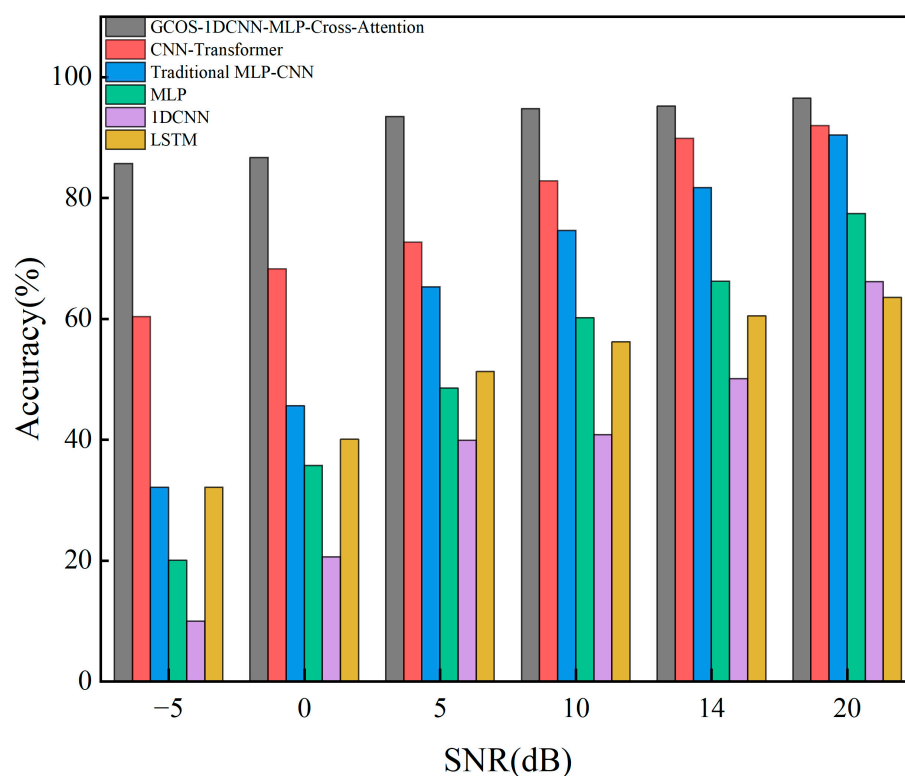


Figure 14. Accuracy of each model under variable loads and different signal-to-noise ratios.

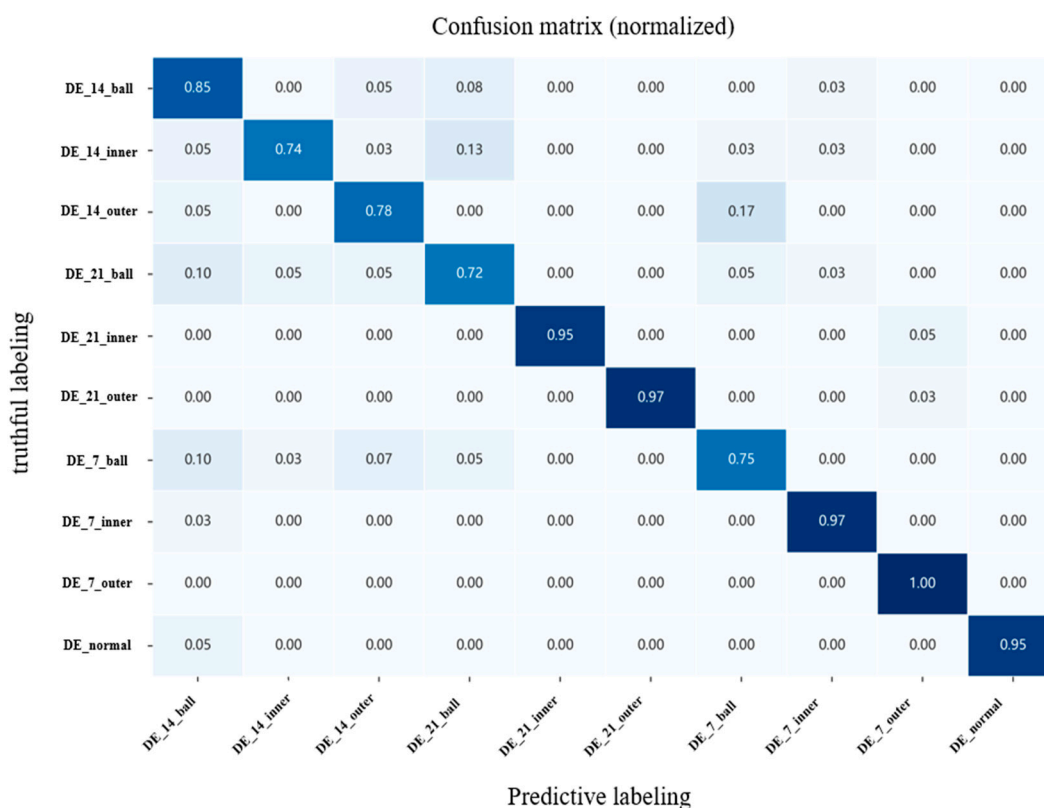


Figure 15. Confusion Matrix for GCOS-1DCNN-MLP-Cross-Attention mechanism under variable loading and 0 dB noise.

3.2.4. Feature Correlation Verification

To validate the rationale behind the feature selection in this paper, which includes three time-domain features and five frequency-domain features, a feature correlation analysis

was conducted. Figure 16 below shows the importance ranking, while Figure 17 depicts the Pearson correlation coefficient matrix for the eight features. Figure 16 shows that high-frequency energy is the most critical feature, demonstrating that energy distribution in the high-frequency band significantly impacts model performance. Feature importance values exceeding 0.14 were observed for time-domain standard deviation, frequency-domain mean, time-domain mean, and RMS, indicating that statistical characteristics in both time and frequency domains significantly influence model performance. Mid-frequency and low-frequency energy are ranked sixth and seventh in terms of importance, respectively, while the dominant frequency position is ranked as the least important. This suggests that the singular peak frequency location in the data provides weak discriminative power for classification or prediction. Determining feature correlation and redundancy requires an examination of the absolute values of the Pearson correlation coefficients between features. As Figure 17 shows, all the correlation coefficients in the current matrix have absolute values below 0.7, with the highest being 0.68. Therefore, there are no strong correlations among the eight features. Specifically, no correlation exists between the dominant frequency position and the other features. The correlation between low-frequency energy and the time-domain mean is 0.65, while the correlations between mid-frequency energy and the time-domain standard deviation, root mean square (RMS), and the frequency-domain mean are 0.68, 0.67, and 0.72. Feature subset ablation experiments were conducted. These defined the removal of mid-band energy, low-band energy, and the simultaneous removal of mid- and low-band energy as ablation groups 1, 2, and 3, respectively. The results of these experiments are shown in Table 10. Removing mid-band energy alone resulted in a 1.51% decrease in the model's average accuracy. Removing low-band energy alone resulted in an average accuracy decrease of 1.73%, while simultaneously removing mid- and low-band energy resulted in an average accuracy decrease of 2.6%. This indicates that mid-low-band energy significantly enhances the model's performance. To investigate the effect of mid-low-frequency energy on model performance in more detail, the feature distributions of mid-band and low-band energy were compared. The feature distribution comparison diagram clearly illustrates the differences in the distribution of mid-band and low-band energy across different categories. Figure 18 shows the feature distribution comparison diagram for mid-band and low-band energy.

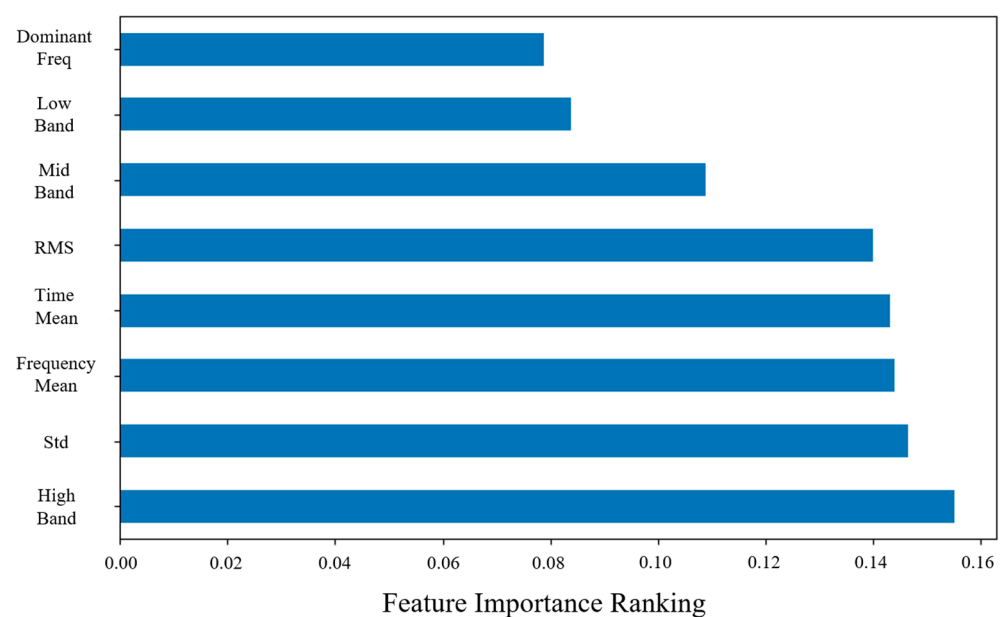


Figure 16. Feature importance ranking.

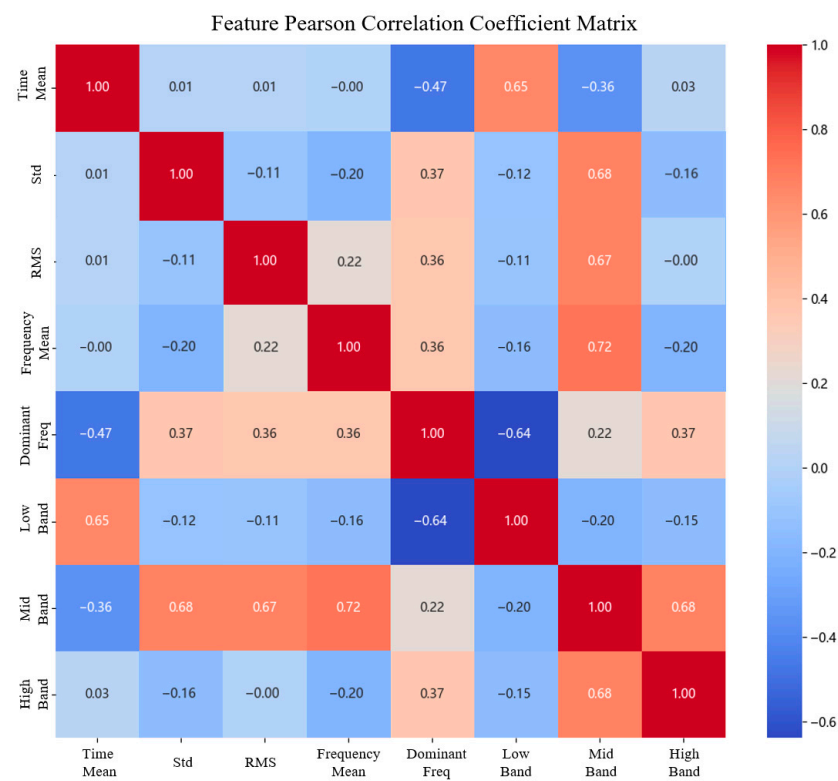


Figure 17. Feature Pearson correlation coefficient matrix.

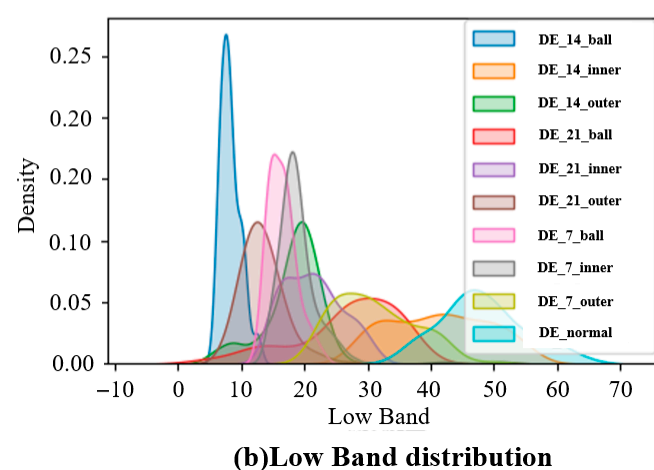
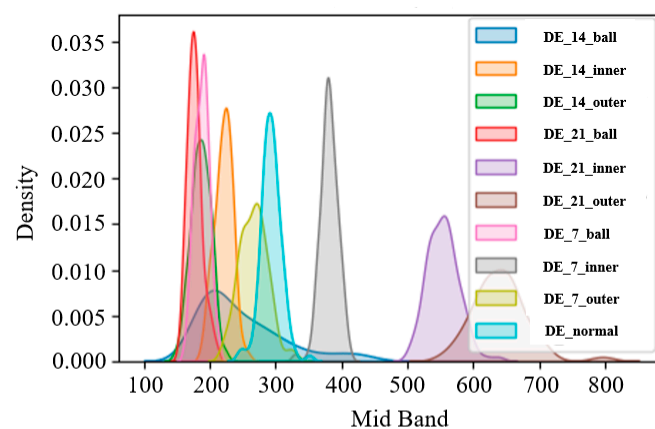


Figure 18. Distribution of mid-band and low-band energy characteristics.

Table 10. Feature subset ablation experiment.

Group	Accuracy	Recall Rate	Precision	F1 Score
Primitive	99.83	99.74	99.9	99.86
Group1	98.32	98.44	98.37	98.41
Group2	98.1	98.07	98.13	98.11
Group3	97.23	97.12	97.33	97.4

Figure 18 shows that the mid-frequency energy distribution exhibits a multi-peak characteristic. The DE_14_ball category has the highest mid-frequency energy peak, forming a distinct single peak between 300 and 400 Hz, which differs from the peaks observed in the other categories. Meanwhile, DE_21_ball and DE_7_ball exhibit a bimodal distribution between 200 and 400, and the low-frequency energy distribution shows distinct zones of energy concentration. DE_14_outer demonstrates an extremely high peak in low-frequency energy, forming a single peak near zero. This indicates that its signal components are predominantly of a low frequency. This suggests that, although mid-frequency and low-frequency signals are somewhat correlated with other features, they significantly contribute to specific faults. Thus, joint analysis confirms that the eight selected features demonstrate excellent performance in terms of both data distribution and model contribution, making their selection both reasonable and necessary.

4. Conclusions and Future Work

In order to cope with the problems of insufficient feature extraction and poor model performance in high-noise environments in current bearing fault diagnosis methods, this paper designs a GCOS-1DCNN-dual MLP-bidirectional cross-attention mechanism model, which gives full play to the advantages of 1DCNN for extracting temporal data and bidirectional cross-attention mechanism for feature fusion. The model constructed in this paper provides a new way and idea of extracting fault diagnosis for bearings, including the ability of dual MLP to deeply extract time-domain and frequency-domain features while using bidirectional cross-attention mechanism to deeply fuse multimodal features, and also designed a learning rate optimizer, which can automatically adjust the learning rate during the model learning process to cope with the shortcomings of slow learning speed and overfitting of the model. In fifteen independent validation experiments using the Case Western Reserve University bearing dataset, the model demonstrated an average accuracy of 99.83%, exhibiting exceptional robustness. Meanwhile, noise is added to the original dataset, noise experiments are conducted, and the results show that the model proposed in this paper still maintains a high accuracy of 90.66% and 86.72% under the strong noise levels of 0 dB and −5 dB. This shows that the model proposed in this paper has the advantages of high recognition accuracy, good robustness, and strong noise resistance.

Future work should evaluate the model's adaptability to noise scenarios beyond Gaussian noise—such as pink noise and impulse/impact noise—to better address diverse noise conditions encountered in industrial applications. Additionally, research into scheduler sensitivity is warranted.

Author Contributions: Conceptualization, A.S.; methodology, A.S. and L.M.; software, A.S.; validation, A.S.; formal analysis, M.D. and H.Y.; investigation, A.S., M.D. and M.S.; resources, L.M.; data curation, K.H. and F.D.; writing—original draft preparation, A.S.; writing—review and editing, A.S.; visualization, L.M.; supervision, K.H., Z.F., C.L. and M.S.; project administration, L.M.; funding acquisition, L.M. and F.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the outstanding youth project of the Education Department of Hunan Province, China (24B0792, 23B0799), and the Changsha Municipal Natural Science Foundation (kq2402042), the Science and Technology Research and Development Command Plan Project of Zhangjiakou, China (2311005A), Hebei Province graduate student innovation ability training funding project (CXZZSS2025127), Graduate Student Innovation Fund Grants, Established Projects of Hebei University of Architecture (XY2025033), Graduate Student Innovation Fund Grants Established Projects of Hebei University of Architecture (XY2024080), and the 2024 Annual Project Topics of the “14th Five-Year Plan” for Educational Science Research in Hebei Province (2403199).

Data Availability Statement: The original contributions presented in this study are included in this article. Further inquiries can be directed to the corresponding author(s).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Song, Q.; Jiang, X.; Du, G.; Liu, J.; Zhu, Z. Smart multichannel mode extraction for enhanced bearing fault diagnosis. *Mech. Syst. Signal Process.* **2023**, *189*, 110107. [\[CrossRef\]](#)
2. Zhao, W.; Wang, Z.; Cai, W.; Zhang, Q.; Wang, J.; Du, W.; Yang, N.; He, X. Multiscale inverted residual convolutional neural network for intelligent diagnosis of bearings under variable load condition. *Measurement* **2022**, *188*, 110511. [\[CrossRef\]](#)
3. Hoang, D.; Kang, H. Rolling element bearing fault diagnosis using convolutional neural network and vibration image. *Cogn. Syst. Res.* **2019**, *53*, 42–50. [\[CrossRef\]](#)
4. Meng, Z.; Zhan, X.; Li, J.; Pan, Z. An enhancement denoising autoencoder for rolling bearing fault diagnosis. *Measurement* **2018**, *130*, 448–454. [\[CrossRef\]](#)
5. Jin, Y.; Song, X.; Yang, Y.; Hei, X.; Feng, N.; Yang, X. An improved multi-channel and multi-scale domain adversarial neural network for fault diagnosis of the rolling bearing. *Control. Eng. Pr.* **2025**, *154*, 106120. [\[CrossRef\]](#)
6. Rai, V.K.; Mohanty, A.R. Bearing fault diagnosis using FFT of intrinsic mode functions in Hilbert–Huang transform. *Mech. Syst. Signal Process.* **2007**, *21*, 2607–2615. [\[CrossRef\]](#)
7. Rajeswari, C.; Sathiyabhama, B.; Devendiran, S.; Manivannan, K. Bearing Fault Diagnosis using Wavelet Packet Transform, Hybrid PSO and Support Vector Machine. *Procedia Eng.* **2014**, *97*, 1772–1783. [\[CrossRef\]](#)
8. Saidi, L.; Ben Ali, J.; Fnaiech, F. Bi-spectrum based-EMD applied to the non-stationary vibration signals for bearing faults diagnosis. *ISA Trans.* **2014**, *53*, 1650–1660. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Liu, Z.; Peng, D.; Zuo, M.J.; Xia, J.; Qin, Y. Improved Hilbert–Huang transform with soft sifting stopping criterion and its application to fault diagnosis of wheelset bearings. *ISA Trans.* **2022**, *125*, 426–444. [\[CrossRef\]](#)
10. Wang, B.; Qiu, W.; Hu, X.; Wang, W. A rolling bearing fault diagnosis technique based on recurrence quantification analysis and Bayesian optimization SVM. *Appl. Soft Comput.* **2024**, *156*, 111506. [\[CrossRef\]](#)
11. Mansouri, M.; Fezai, R.; Trabelsi, M.; Mansour, H.; Nounou, H.; Nounou, M. Fault Diagnosis of Wind Energy Conversion Systems Using Gaussian Process Regression-based Multi-Class Random Forest. *IFAC-PapersOnLine* **2022**, *55*, 127–132. [\[CrossRef\]](#)
12. Pandya, D.; Upadhyay, S.; Harsha, S. Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN. *Expert Syst. Appl.* **2013**, *40*, 4137–4145. [\[CrossRef\]](#)
13. Gao, S.; Li, T.; Zhang, Y.; Pei, Z. Fault diagnosis method of rolling bearings based on adaptive modified CEEMD and 1DCNN model. *ISA Trans.* **2023**, *140*, 309–330. [\[CrossRef\]](#)
14. Guo, Z.; Yang, M.; Huang, X. Bearing fault diagnosis based on speed signal and CNN model. *Energy Rep.* **2022**, *8*, 904–913. [\[CrossRef\]](#)
15. Li, Z.-B.; Feng, X.-Y.; Wang, L.; Xie, Y.-C. DC–DC circuit fault diagnosis based on GWO optimization of 1DCNN-GRU network hyperparameters. *Energy Rep.* **2023**, *9*, 536–548. [\[CrossRef\]](#)
16. He, D.; Ye, K.; Yuan, J.; Li, S.; Xu, S. Prediction of inlet gas volume fraction of rotary vane pump under variable operational conditions with 1DCNN-TL model utilizing vibration signals. *Int. J. Heat Fluid Flow* **2024**, *110*, 109588. [\[CrossRef\]](#)
17. Mystkowski, A.; Wolniakowski, A.; Idzkowski, A.; Cieżkowski, M.; Ostaszewski, M.; Kociszewski, R.; Kotowski, A.; Kulesza, Z.; Dobrzański, S.; Miastkowski, K. Measurement and diagnostic system for detecting and classifying faults in the rotary hay tedder using multilayer perceptron neural networks. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108513. [\[CrossRef\]](#)
18. Zou, F.; Zhang, H.; Sang, S.; Li, X.; He, W.; Liu, X.; Chen, Y. An anti-noise one-dimension convolutional neural network learning model applying on bearing fault diagnosis. *Measurement* **2021**, *186*, 110236. [\[CrossRef\]](#)
19. Xie, S.; Li, Y.; Tan, H.; Liu, R.; Zhang, F. Multi-scale and multi-layer perceptron hybrid method for bearings fault diagnosis. *Int. J. Mech. Sci.* **2022**, *235*, 107708. [\[CrossRef\]](#)

20. Sinitsin, V.; Ibryaeva, O.; Sakovskaya, V.; Ereemeeva, V. Intelligent bearing fault diagnosis method combining mixed input and hybrid CNN-MLP model. *Mech. Syst. Signal Process.* **2022**, *180*, 109454. [[CrossRef](#)]
21. Song, B.; Liu, Y.; Fang, J.; Liu, W.; Zhong, M.; Liu, X. An optimized CNN-BiLSTM network for bearing fault diagnosis under multiple working conditions with limited training samples. *Neurocomputing* **2024**, *574*, 127284. [[CrossRef](#)]
22. Bharatheedasan, K.; Maity, T.; Kumaraswamidhas, L.; Durairaj, M. Enhanced fault diagnosis and remaining useful life prediction of rolling bearings using a hybrid multilayer perceptron and LSTM network model. *Alex. Eng. J.* **2025**, *115*, 355–369. [[CrossRef](#)]
23. Wang, H.; Xu, J.; Yan, R.; Sun, C.; Chen, X. Intelligent Bearing Fault Diagnosis Using Multi-Head Attention-Based CNN. *Procedia Manuf.* **2020**, *49*, 112–118. [[CrossRef](#)]
24. Snyder, Q.; Jiang, Q.; Tripp, E. Integrating self-attention mechanisms in deep learning: A novel dual-head ensemble transformer with its application to bearing fault diagnosis. *Signal Process.* **2025**, *227*, 109683. [[CrossRef](#)]
25. Cui, L.; Tian, X.; Wei, Q.; Liu, Y. A self-attention based contrastive learning method for bearing fault diagnosis. *Expert Syst. Appl.* **2024**, *238*, 121645. [[CrossRef](#)]
26. Jiang, P.; Wang, Y.; Yang, C.; Zhang, L.; Duan, B. Fault diagnosis of wind power pitch bearings based on spatiotemporal clustering and a deep attention subdomain adaptive residual network. *Measurement* **2025**, *242*, 116187. [[CrossRef](#)]
27. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.